UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Departamento de Lingua Española

TESE DE DOUTORAMENTO

EXTRACÇÃO DE RELAÇÕES SEMÂNTICAS. RECURSOS, FERRAMENTAS E ESTRATÉGIAS

Autor:

Marcos Garcia González

SANTIAGO DE COMPOSTELA

2014

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Departamento de Lingua Española

TESE DE DOUTORAMENTO

EXTRACÇÃO DE RELAÇÕES SEMÂNTICAS. RECURSOS, FERRAMENTAS E ESTRATÉGIAS

Autor:

Marcos Garcia González

Orientador:

Pablo Gamallo Otero

SANTIAGO DE COMPOSTELA

2014

D. Pablo Gamallo Otero, Professor Contratado Doutor da Área de Língua Espanhola da Universidade de Santiago de Compostela

FAZ CONSTAR:

Que a memória intitulada **EXTRACÇÃO DE RELAÇÕES SEMÂNTICAS. RECURSOS, FERRAMENTAS E ESTRATÉGIAS**, realizada por **D. Marcos Garcia González** sob a minha direcção no Departamento de Língua Espanhola da Universidade de Santiago de Compostela, reúne os requisitos exigidos no artigo 34 do regulamento de Estudos de Doutoramento, e constitui a Tese que defende para optar ao grau de Doutor.

2014

Orientador	**Doutorando**
Pablo Gamallo Otero	Marcos Garcia González

Agradecimentos

Devo começar por agradecer ao orientador desta tese pela dedicação, atenção e formação que me ofereceu ao longo dos últimos anos. Também aos membros das diferentes equipas de trabalho em que me inseri durante a realização da tese. Continuo agradecendo às pessoas que partilharam comigo locais de trabalho e de lazer, tanto na Faculdade de Filologia como no CiTIUS (e noutros lugares, digamos, menos institucionais).

Fora do âmbito académico, agradeço a todas as pessoas que —de maneira consciente ou inconsciente— me ajudaram a finalizar este projecto, nomeadamente àquelas que me permitiram desfrutar mais da vida e me ensinaram a manter um equilíbrio entre as obrigações pessoais e profissionais.

Por último, tenho que agradecer o apoio da Universidade de Santiago de Compostela, através de um Contrato Predoutoral (2010), e aos financiadores dos seguintes projectos: Ontopedia, do Ministerio de Educación y Ciencia (referência FFI2010-14986); projectos do Governo Galego: referência 2008/101 e, HPCPLN (referência EM2013/041), e projecto Feder-Interconnecta: Celtic (referência 2012-CE138).

Santiago de Compostela, 2014

Abstract

Relation extraction is a subtask of information extraction that aims at obtaining instances of semantic relations present in texts. This information can be arranged into machine-readable formats, useful for several applications that need structured semantic knowledge.

This thesis explores different strategies to automate the extraction of semantic relations from texts in Portuguese, Spanish and Galician. Both machine-learning (distant-supervised and supervised) and rule-based techniques are investigated, and the impact of the different levels of linguistic knowledge is analyzed for the various approaches. Regarding domains, the experiments are focused on the extraction of encyclopedic knowledge, by means of the development of biographical relations classifiers (in a closed domain) and the evaluation of open information extraction systems.

In order to implement the extraction systems, several natural language processing tools have been built for the three research languages: from sentence splitting and tokenization modules to part-of-speech taggers, named entity recognizers and coreference resolution systems. Furthermore, several lexica and corpora have been compiled and enriched with different levels of linguistic annotation, which are useful for both training and testing probabilistic and rule-based models. As a result of the work carried out in this thesis, new resources and tools are available for automated processing of texts in Portuguese, Spanish and Galician.

Keywords: information extraction, natural language processing, named entity recognition, part-of-speech tagging, coreference resolution

Resumen

La extracción de relaciones, encuadrada dentro de las tareas de extracción de información, pretende obtener ejemplos de relaciones semánticas presentes en textos. Información

que puede ser posteriormente organizada en formatos legibles por ordenadores, siendo útil para diversas aplicaciones que necesiten conocimiento semántico estructurado.

La presente tesis evalúa diferentes estrategias para la extracción automática de relaciones semánticas de textos en portugués, español y gallego. Con ese fin, son utilizadas tanto técnicas de aprendizaje automático (con supervisión-distante y supervisadas) como sistemas basados en reglas, siendo analizado el impacto de diferentes niveles de conocimiento lingüístico en las varias aproximaciones evaluadas. En relación al dominio, las extracciones se centran en la obtención de conocimiento de carácter enciclopédico, mediante la creación de clasificadores de relaciones biográficas (en dominio cerrado) y la evaluación de sistemas de extracción de información abierta.

Con el objetivo de implementar los sistemas de extracción, también han sido construidas diversas herramientas para el procesamiento del lenguaje natural en los tres idiomas referidos: desde módulos de segmentación de oraciones y de tokenización, a sistemas de desambiguación morfosintáctica, de reconocimiento de entidades con nombre y de resolución de correferencia. Además, han sido compilados y adaptados lexicones y corpus con anotación lingüística de diferentes niveles, útiles para entrenar y evaluar modelos estadísticos y basados en reglas. Como resultado del trabajo realizado en esta tesis, se disponibilizan nuevas herramientas y recursos para el procesamiento automático de textos en portugués, español y gallego.

Palabras clave: extracción de información, procesamiento del lenguaje natural, reconocimiento de entidades con nombre, anotación morfosintáctica, resolución de correferencia

Resumo

A extracção de relações, enquadrada dentro das tarefas de extracção de informação, visa obter automaticamente exemplos de relações semânticas presentes em textos. Esta informação pode ser posteriormente organizada em formatos legíveis por computadores, sendo útil para diversas aplicações que necessitem conhecimento semântico estruturado.

A presente tese avalia diferentes estratégias para a extracção automática de relações semânticas de textos em português, espanhol e galego. Com esse fim, são utilizadas tanto técnicas de aprendizagem automática (com supervisão-distante e supervisionadas) como sistemas baseados em regras, sendo analisado o impacto de diferentes níveis de conhecimento lin-

guístico nas várias abordagens avaliadas. Em relação ao domínio, as extracções lidam com conhecimento de carácter enciclopédico, mediante a criação de classificadores de relações biográficas (em domínio fechado) e a avaliação de sistemas de extracção de informação aberta.

Com o objectivo de implementar os sistemas de extracção, foram também construídas diversas ferramentas para o processamento da linguagem natural nos três idiomas referidos: desde módulos de segmentação de orações e de tokenização, a sistemas de desambiguação morfossintáctica, de reconhecimento de entidades mencionadas e de resolução de correferência. Além disso, foram compilados e adaptados léxicos e corpora com anotação linguística de diferentes níveis, úteis para o treino e avaliação de modelos probabilísticos e baseados em regras. Como resultado do trabalho realizado nesta tese, disponibilizam-se novas ferramentas e recursos para o processamento automático de textos em português, espanhol e galego.

Palavras-chave: extracção de informação, processamento da linguagem natural, reconhecimento de entidades mencionadas, anotação morfossintáctica, resolução de correferência

Conteúdo

Capítulo 1

Introdução

A emergência da sociedade da informação provocou um aumento exponencial na produção e difusão de dados nas últimas décadas (Castells, 1996). Embora quantificar a sua dimensão não seja uma tarefa fácil calcula-se que actualmente se gera a mesma quantidade de informação em poucos dias do que a criada pelo ser humano até 2002 (Moore, 2011). Além disso, uma parte importante desses dados contém informação não estruturada, como por exemplo o texto livre.[1]

Por um lado, o enorme tamanho dos dados existentes impede que as pessoas acedam a toda essa quantidade de informação através da leitura. Pelo outro, o facto de parte de essa informação se encontrar em formatos não estruturados impossibilita que os computadores a possam *compreender* e que muitas aplicações tirem proveito dela.

Com o fim de lidar com estes problemas, disciplinas como o Processamento da Linguagem Natural (PLN), em que se inclui a presente tese, desenvolvem ferramentas que permitem o tratamento de texto por parte dos computadores, facilitando assim o processamento de dados de origem linguística.

1.1. Processamento da linguagem natural

O PLN estuda e implementa mecanismos de interacção em língua natural entre seres humanos e máquinas, como a compreensão das próprias línguas naturais ou a geração de discurso. Entre outras, o PLN engloba áreas tão diversas como a tradução automática, o reco-

[1] A sua quantificação varia entre $\approx 30\%$ e $\approx 80\%$ (Swoyer, 2007).

nhecimento de fala ou a extracção de informação, tema principal deste trabalho (Jurafsky e Martin, 2009).

Diferentes campos do conhecimento, tais como a inteligência artificial, as ciências da computação ou a linguística têm abordado várias tarefas de PLN utilizando diversas estratégias em função dos objectivos de cada aplicação. Assim, ao lado de modelos construídos com conhecimento linguístico profundo, como as gramáticas HPSG (Pollard e Sag, 1994) ou algumas propostas de fonologia computacional (Bird e Ellison, 1994), têm-se implementado sistemas estatísticos que obtêm, com informação linguística superficial, melhores resultados em diversas tarefas do que abordagens que utilizam informação mais complexa (Ratnaparkhi, 1996).

1.2. Extracção de informação e extracção de relações

A Extracção de Informação (EI) é uma área do PLN cujo objectivo é a obtenção automática de informação estruturada a partir, normalmente, de texto, e cujos sistemas vêm sendo avaliados desde a década de 90 em várias conferências como as MUC[2] (Message Understanding Conference), as CoNLL[3] (Conference on Computational Natural Language Learning) ou as ACE[4] (Automatic Content Extraction).

Um tipo de tarefa da EI, conhecida como Extracção de Relações (ER), consiste na identificação de relações semânticas entre entidades ou conceitos. Como exemplo, veja-se o seguinte texto:

> *"John A. Garcia (nascido em 1949 na Galiza) é um dos pioneiros da indústria moderna americana de videojogos e o atual presidente da NovaLogic".*

Um sistema de ER poderia obter desta oração a seguinte informação estruturada, onde cada extracção se compõe de uma relação semântica e dous argumentos:

- `DatadeNascimento`, *John A. Garcia – 1949*
- `LocaldeNascimento`, *John A. Garcia – Galiza*
- `éPresidenteDe`, *John A. Garcia – NovaLogic*

[2] `http://www-nlpir.nist.gov/related_projects/muc/`
[3] `http://ifarm.nl/signll/conll/`
[4] `http://www.itl.nist.gov/iad/mig/tests/ace/`

A obtenção deste tipo de informação de fontes não estruturadas permite a sua transformação em conhecimento organizado, que pode ser processado por computadores, e ser utilizado em diversas aplicações tais como sistemas de resposta a perguntas (Mann, 2002) ou de recuperação de informação (Wan *et al.*, 2005), entre outros.

Em função do tipo de extracção que realizem, os sistemas de ER podem ser divididos em dous grandes grupos (apresentados pormenorizadamente no Capítulo 4):

– Domínio fechado: este tipo de aproximação tem como objectivo extrair exemplos de relações previamente definidas (normalmente, um conjunto pequeno), tais como as referidas `DatadeNascimento` ou `éPresidenteDe`.

– Domínio aberto e extracção de informação aberta: paradigmas mais recentes da EI utilizam grandes repositórios de informação para adaptar sistemas capazes de extrair milhares de relações semânticas, bem como para treinar modelos de Extracção de Informação Aberta (OIE, do inglês *Open Information Extraction*), que obtêm automaticamente todo o tipo de relações de base verbal (Banko *et al.*, 2007). Do exemplo anterior, um sistema de OIE poderia obter os seguintes triplos (compostos de dous argumentos ligados por uma relação não definida previamente):

  - *John A. Garcia* `é` *um dos pioneiros da indústria moderna americana de videojogos*
  - *John A. Garcia* `é` *o atual presidente da Novalogic*
  - *John A. Garcia* `é_o_atual_presidente_da` *Novalogic*

Como se verá ao longo do trabalho, a presente tese analisa principalmente estratégias de extracção de relações em domínio fechado, embora diversas avaliações utilizem também um sistema de OIE.

1.3. Objectivos

Na sua formulação inicial, o objectivo da presente tese consistia na avaliação de diferentes estratégias para a extracção, em domínio fechado, de relações enciclopédicas de textos em português (pt), espanhol (es) e galego (gl). Contudo, no início da realização deste trabalho, algumas das ferramentas de PLN necessárias para construir sistemas de extracção de relações

nas três línguas alvo não existiam, não se distribuíam livremente ou tinham sido realizadas para fins específicos diferentes dos desta tese.

Portanto, o objectivo inicial do projecto foi ampliado, ao ser necessário o desenvolvimento ou a adaptação de várias ferramentas de PLN orientadas ao desenho de sistemas de ER em português, espanhol e galego. Assim sendo, a presente tese tem os seguintes objectivos:

Principal: O objectivo principal consiste na avaliação de diferentes estratégias para a extracção em domínio fechado de relações de carácter enciclopédico —especificamente biográfico— em português, espanhol e galego.

Paralelos: Para a consecução do objectivo principal, vários objectivos paralelos foram definidos, que se podem englobar em um único: o desenvolvimento ou adaptação das ferramentas de PLN necessárias para a ER em português, espanhol e galego.

Línguas

Tanto as extracções realizadas como as diversas ferramentas e recursos apresentados neste trabalho foram feitas com o fim de processar textos em português, espanhol e galego.[5]

A escolha destas línguas deveu-se, por um lado, ao próprio carácter geográfico e cultural em que se insere a tese, já que espanhol e galego são idiomas oficiais na Galiza. De modo similar, a utilização do português revelou-se natural por ser uma variedade linguística próxima do galego —consideradas a mesma língua por diferentes autores (Cunha e Cintra, 1984, por exemplo)— e com maior quantidade de dados a serem analisados do que este. Por outro lado, os estudos sobre a ER eram escassos em quaisquer das três línguas, pelo que se considerou oportuno tratá-los em todas elas.

Em termos gerais, antes da realização deste trabalho existia um maior número de ferramentas de PLN disponíveis para espanhol, pelo que grande parte dos sistemas adaptados e/ou desenvolvidos são para português e galego.

Em relação às diferentes variedades nacionais do português, tentou-se utilizar tanto o Português Europeu (PE) como o Brasileiro (PB) —e outras variedades africanas—, embora o português europeu foi a variedade prioritária naqueles casos em que uma delas tinha de ser escolhida.

[5] Galego e português, nesta tese, são diferenciados pela utilização de diferentes sistemas ortográficos: é considerada galega a língua que utiliza as normas ortográficas apresentadas em Real Academia Galega e Instituto da Lingua Galega (2004), enquanto é português a que segue as diferentes ortografias da Academia Brasileira de Letras e da Academia das Ciências de Lisboa.

Além disso, aquelas estratégias que requereram grandes quantidades de informação para serem aplicadas satisfatoriamente, só foram avaliadas em português e espanhol, devido à escassez de dados em galego existentes na Web.

1.4. Ferramentas e metodologia utilizadas

Do ponto de vista metodológico, este trabalho baseia-se principalmente na utilização de conhecimento linguístico para a realização de processamento da linguagem natural, mas combina este conhecimento com abordagens próprias de outras disciplinas (como a aprendizagem automática) com o fim de atingir os seus objectivos de modo eficaz. Assim, na criação e adaptação das diferentes ferramentas aproveitam-se as formulações propostas em trabalhos de carácter teórico, mas prioriza-se a qualidade dos resultados sobre a consistência formal. Trata-se, portanto, de uma tese fundamentalmente prática.

Para além das diferentes ferramentas desenvolvidas durante a realização do trabalho (apresentadas ao longo da tese), foram escolhidas duas *suites* de PLN multilíngue para levar a cabo vários dos objectivos propostos:

FreeLing

FreeLing[6] é um conjunto de livrarias de análise linguística que contém diversos módulos de processamento tais como segmentadores de orações, anotadores morfossintácticos ou reconhecedores de entidades mencionadas, entre outros (Padró e Stanilovsky, 2012). As razões para a escolha deste *software* foram as seguintes:

– Desempenho: FreeLing contém diversos módulos de PLN com desempenhos ao nível do estado-da-arte.

– Arquitectura: FreeLing adapta-se a outras ferramentas utilizadas no processo de extracção de relações semânticas.

– Licença: FreeLing disponibiliza-se sob licença livre GPL.

[6] `http://nlp.lsi.upc.edu/freeling/`

DepPattern

DepPattern[7] é uma *suite* de análise sintáctica que inclui gramáticas de dependências (explicadas a seguir) para diversas línguas, um compilador de gramáticas e analisadores sintácticos automáticos (*parsers*) (Gamallo e González López, 2011).

As gramáticas de DepPattern são baseadas em regras, que são escritas num formalismo próprio que facilita tanto a modificação como a adição ou remoção de regras de sintácticas.

DepPattern é utilizado nesta tese como analisador sintáctico para as três línguas alvo pelas seguintes razões:

– Desempenho: os *parsers* de DepPattern são rápidos e robustos.

– Arquitectura: inclui módulos de compatibilidade com FreeLing.

– Formalismo: o formalismo de DepPattern permite criar ou adaptar regras para fins específicos, tais como a extracção de relações.

Apesar de que as análises de DepPattern não são sempre completas (algumas dependências podem não ser cobertas pelas regras incluídas nas gramáticas), os *parsers* disponíveis tiveram melhor desempenho do que o Maltparser[8] treinado com os corpora Bosque 8.0[9] e AnCora[10] para português e espanhol, respectivamente. Diversas avaliações mostraram que os analisadores de DepPattern atingiram valores de 88%/79%/83% (pt) e de 85%/74%/79% (es) em precisão, *recall* e medida F, respectivamente (Gamallo, 2012). Para além disso, esta *suite* também é disponibilizada sob licença GPL.

Uma vez que as análises sintácticas realizadas em diferentes capítulos do presente trabalho utilizam representações de dependências, este tipo de gramáticas são agora apresentadas sucintamente. Além disso, ao longo da tese são aplicados repetidamente sistemas de aprendizagem automática, pelo que esta metodologia é também definida a seguir:

[7] `http://gramatica.usc.es/pln/tools/deppattern.html`
[8] `http://www.maltparser.org`
[9] `http://www.linguateca.pt/floresta/corpus.html#bosque`
[10] `http://clic.ub.edu/corpus/ancora`

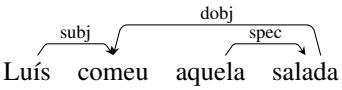


Figura 1.1: Exemplo de uma análise de dependências.

Gramática de dependências

A gramática de dependências é um conjunto de teorias linguísticas que considera que a informação sintáctica pode ser codificada —principalmente— através de relações binárias entre dous elementos de uma oração.

De modo geral, assume-se que os trabalhos em gramática de dependências moderna começaram com as publicações de Lucien Tesnière (Tesnière, 1959), sendo popularizados recentemente em diferentes tarefas de PLN (Kübler *et al.*, 2009).

As gramáticas de dependências consideram que cada palavra está relacionada com outra palavra da mesma oração, mas —à diferença das gramáticas de constituintes— não agrupam conjuntos de palavras em unidades maiores (p. ex., frases nominais). Cada dependência estabelece uma relação binária entre dous elementos (um núcleo e um dependente), atribuindo-lhe uma função sintáctica (sujeito, modificador, etc.). Na Figura 1.1 mostra-se um exemplo de análise de dependências: em cada dependência, a seta sai do dependente e chega ao núcleo (incluindo a função sintáctica acima).[11] O elemento que não é dependente (só núcleo) é a raiz da oração (neste caso, a forma verbal *comeu*).

Note-se que a saída de um analisador de dependências pode ser convertida em representação de constituintes mediante ferramentas específicas, pelo que ao longo da tese haverá referências a constituintes sintácticos (como frases nominais ou preposicionais, por exemplo) quando a sua utilização se considerar vantajosa.

Aprendizagem automática

A aprendizagem automática (normalmente conhecida como *machine learning*, em inglês) é um campo da inteligência artificial orientado ao desenvolvimento de algoritmos que aprendam, através de um conjunto de dados, a realizar uma determinada tarefa (Mitchell, 1997).

Entre os diferentes tipos de aprendizagem automática, esta tese utiliza algoritmos de classificação supervisionados. Estes métodos consistem na aplicação de algoritmos de aprendi-

[11] Onde *subj* é sujeito; *spec*, especificador e *dobj* significa objecto directo.

zagem em exemplos previamente classificados (conjunto de treino ou de aprendizagem) dos quais o computador generaliza uma função, podendo depois classificar novos exemplos desconhecidos. Assim, dado um conjunto de treino que inclua exemplos de informações clínicas de pessoas e classificações de cada uma dessas pessoas em relação a uma doença, o sistema poderá predizer (com maior ou menor precisão) se um indivíduo não analisado previamente tem ou não a doença, em função das características que contenha a sua informação clínica.

Para levar a cabo o processo de aprendizagem, os dados de treino devem ser previamente processados para extrair deles um conjunto de elementos considerados relevantes. No exemplo anterior, é preciso escolher do historial clínico qual é a informação que possa ser necessária para saber se a pessoa tem ou não a doença. Cada um dos elementos utilizados pelo algoritmo durante o processo de aprendizagem é denominado atributo (referidos habitualmente pelo termo em inglês *feature*).

1.5. Estrutura

Para além deste capítulo introdutório e das conclusões, o presente trabalho estrutura-se em sete capítulos organizados em três partes, em função do seu tema principal.

Os diferentes capítulos têm como base fundamental um conjunto de trabalhos publicados durante a realização da tese, e referidos ao longo da mesma. Assim, o conteúdo dos artigos foi seleccionado, actualizado e ampliado e/ou corrigido quando foi considerado oportuno. Depois, o resultado foi traduzido e adaptado à estrutura da tese, mostrada a seguir e resumida na Figura 1.2:

Parte I: Processamento Prévio à Extracção de Relações

A primeira parte da tese centra-se na apresentação das ferramentas e recursos de PLN necessários para aplicar estratégias de extracção de relações, constando de dous capítulos:

Capítulo 2: Este capítulo descreve a adaptação e avaliação de módulos de tokenização, segmentação de orações, lematização e anotação morfossintáctica para português e galego, bem como uma estratégia de correcção desta última ferramenta e diversos recursos como corpora anotados e dicionários morfossintácticos.

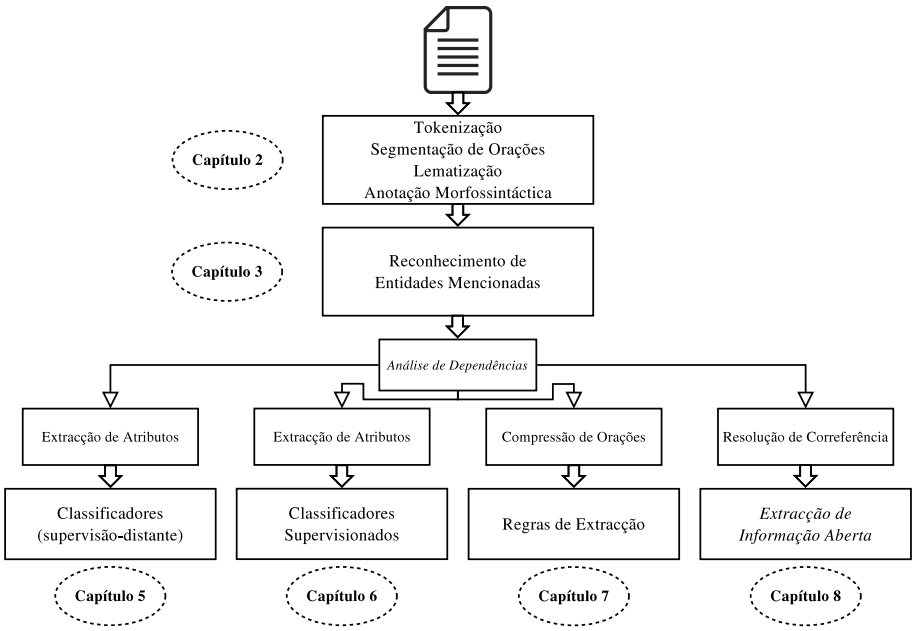


Figura 1.2: Diagrama dos processos realizados em cada capítulo. A entrada (acima) é texto plano, sendo avaliadas quatro estratégias de extracção de informação (abaixo) nos Capítulos 5 a 8. Os elementos em itálico não foram realizados especificamente nesta tese.

Capítulo 3: O Capítulo 3 apresenta a adaptação e criação de ferramentas para o reconhecimento e classificação de vários tipos de entidades mencionadas, como pessoas, localizações, organizações, datas, quantidades ou moedas, entre outras.

Parte II: Estratégias para a Extracção de Relações

A Parte II contém quatro capítulos que tratam sobre a extracção de relações semânticas em domínio fechado:

Capítulo 4: Este capítulo faz uma revisão de várias estratégias para a extracção de relações que têm sido aplicadas em diferentes contextos (pelo que não aparece na Figura 1.2). A revisão é feita tendo em conta os objectivos de cada sistema, bem como as línguas para as quais foram desenhados.

Capítulo 5: Neste capítulo é apresentado um conjunto de testes para a extracção de relações mediante a estratégia de supervisão-distante, que consiste na obtenção semiautomática de conjuntos de treino utilizados por algoritmos de aprendizagem automática.

Capítulo 6: Este capítulo analisa a efectividade de diferentes atributos de base linguística na construção de classificadores supervisionados para a extracção de relações. Além disso, apresenta dous corpora com anotação de relações biográficas corrigida manualmente, em português e espanhol.

Capítulo 7: O último capítulo da Parte II apresenta uma estratégia de extracção de relações baseada em regras de alta precisão obtidas semiautomaticamente, que utiliza métodos de compressão de texto para aumentar o *recall*.

Parte III: Resolução de Correferência e Extracção de Informação Aberta

A última parte da tese contém um único capítulo, que avalia o impacto da resolução de correferência no processo de extracção de informação:

Capítulo 8: Este capítulo apresenta um sistema de resolução de correferência de entidades pessoa, três corpora com anotação correferencial anotada manualmente, e um conjunto de avaliações da combinação da resolução de correferência com a extracção de informação aberta.

Finalmente, o Capítulo 9 mostra as conclusões da presente tese, bem como as principais contribuições do trabalho realizado e caminhos para o trabalho futuro.

Parte I

Processamento Prévio à Extracção de Relações

CAPÍTULO 2

PROCESSAMENTO INICIAL

2.1. Introdução

As diferentes tarefas iniciais do processamento da linguagem natural constituem uma etapa com enorme importância em processos posteriores, tais como a extracção de informação.[1]

Entre elas, o presente capítulo trata das seguintes tarefas: tokenização, segmentação de orações, análise morfológica com lematização e anotação morfossintáctica (também conhecida como *PoS-tagging*, do inglês *Part-of-Speech tagging*). Uma vez que no início da realização deste trabalho já existiam ferramentas que fizessem este tipo de processamento para espanhol, este capítulo centra-se unicamente no desenvolvimento de ferramentas para português e galego.

Estas variedades linguísticas têm uma flexão verbal complexa, para além de formas homógrafas e de contracções de tokens ambíguas, pelo que precisam ser tratadas com ferramentas desenvolvidas especificamente para elas (Graña *et al.*, 2002; Branco e Silva, 2004). Assim, tanto a criação como a adaptação de recursos para estas línguas devem ter em conta os problemas concretos que apresentam, com o fim de evitar erros de análise em etapas posteriores do processamento.

Tanto o português (pt) como o galego (gl) dispõem de ferramentas de análise morfossintáctica de alta precisão (Bick, 2000; Marques e Lopes, 2001; Ribeiro *et al.*, 2003; Branco e

[1] Aqui, *iniciais* refere-se às tarefas que são aplicadas sequencialmente no começo de muitos sistemas de PLN, e que transformam um texto em unidades com informação linguística acessíveis para um computador, mas ainda sem conteúdo semântico.

Silva, 2004, para português, e Domínguez Noya *et al.* (2009) para galego), mas durante a realização do presente trabalho desconhecíamos *software* com licenças livres para este fim (salvo os anteriores ficheiros de treino para galego de FreeLing, desenvolvidos na Universidade de Vigo (Carreras *et al.*, 2004)).

Este capítulo descreve os procedimentos de adaptação de diferentes módulos de FreeLing (apresentado na Secção 1.4) para português e galego. São mostrados os principais casos problemáticos e indicadas as soluções adoptadas em cada um deles. Para além disso, o capítulo inclui um conjunto de avaliações do módulo de etiquetação morfossintáctica para diferentes variedades do português, bem como uma estratégia de correcção de erros produzidos por esse sistema de anotação.

As adaptações realizaram-se fundamentalmente com recursos linguísticos de livre distribuição acessíveis *on-line*, e são disponibilizadas sob licenças GPL no *software* FreeLing desde a versão 2.1 (Padró *et al.*, 2010).

A seguir a esta introdução, a Secção 2.2 revisa o trabalho relacionado. Os recursos utilizados são apresentados na Secção 2.3. A Secção 2.4 explica a estratégia de tokenização utilizada, enquanto a Secção 2.5 descreve o processo de segmentação de orações. A seguir, o módulo de análise morfológica com lematização mostra-se na Secção 2.6. Depois, a Secção 2.7 foca-se nos diferentes aspectos da etiquetação morfossintáctica, sendo as principais conclusões deste capítulo apresentadas em 2.8.

O conteúdo deste capítulo tem como base os seguintes trabalhos já publicados: Garcia e Gamallo (2010a,b,c) e Garcia *et al.* (2014).

2.2. Trabalho relacionado

Esta secção apresenta alguns dos trabalhos relacionados mais relevantes sobre a anotação morfossintáctica de português e galego, a principal ferramenta das descritas neste capítulo. As referências relacionadas com outros módulos são realizadas à medida que são apresentados.

Português

Para português, vários etiquetadores foram desenvolvidos tanto para a variedade brasileira como para a portuguesa. A maior parte deles utilizam algoritmos probabilísticos, embora exista algum método baseado em regras.

Português Europeu (PE): O sistema PALAVRAS utiliza um vasto conjunto de regras e um léxico de perto de 50.000 lemas para realizar *PoS-tagging* e análise sintáctica (Bick, 2000).

Empregando modelos estatísticos, em Marques e Lopes (2001) é apresentado um método de redes neuronais para a etiquetação morfossintáctica, que obtém resultados de alta precisão ($\approx 96\%$) utilizando corpora de treino reduzidos.

Em Ribeiro *et al.* (2003) são comparados modelos de Markov e *PoS-taggers* de base transformacional (baseados em Brill (1995)) —orientados ao pré-processamento de um sistema de síntese de fala— obtendo resultados de precisão de $\approx 92\%$ a $\approx 95\%$.

Branco e Silva (2004) compararam diferentes algoritmos —de base transformacional, de máxima entropia (Ratnaparkhi, 1996), modelos ocultos de Markov (HMM, do inglês *Hidden Markov Models*) (Tufis e Mason, 1998) e modelos de Markov de segunda ordem (Brants, 2000)— para o processamento do português europeu. Os melhores resultados são obtidos com o modelo transformacional, com valores de precisão de 97.09%.

Português do Brasil (PB): Para a variedade brasileira, Aires (2000) também comparou vários algoritmos supervisionados para a etiquetação morfossintáctica, tendo os melhores resultados uma precisão de 90.25%, utilizando o sistema MXPoST (Ratnaparkhi, 1996). Desenvolvimentos posteriores (com um conjunto de etiquetas simplificado) melhoraram o desempenho até 97%.[2]

O modelo de máxima entropia MXPoST também obteve os melhores resultados em Aluísio *et al.* (2003), com uma precisão de 95.92%.

Galego

Para galego, Graña *et al.* (2002) apresentaram um método de tokenização orientado à anotação morfossintáctica, com análise de contracções, formas compostas de verbo e pronome clítico, numerais, ou locuções, entre outras. A implementação destes métodos no etiquetador XIADA resultou num sistema *PoS-tagger* com valores de precisão superiores a 95% (Domínguez Noya *et al.*, 2009; Domínguez Noya, 2013).

Para além deste etiquetador, o Seminario de Lingüística Informática da Universidade de Vigo treinou, antes da realização da presente tese, o módulo *PoS-tagger* de FreeLing para

[2] `http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html`

galego. Como se mostrará ao longo do capítulo, este módulo foi melhorado com diferentes recursos, tanto de acesso livre como realizados *ad-hoc* para esta língua.

2.3. Recursos utilizados

Esta secção apresenta sucintamente os recursos utilizados para a adaptação dos diferentes módulos de FreeLing descritos neste capítulo. Além disso, descreve os processos de conversão feitos durante o desenvolvimento.

Corpora

Para treinar o módulo de anotação morfossintáctica, foram utilizados os seguintes corpora:

Para português europeu o corpus foi criado a partir do Bosque 8.0, na altura o único disponível livremente com informação morfossintáctica detalhada.[3] Este recurso contém aproximadamente os 1.000 primeiros extractos dos corpora CETEMPúblico e do CETEMFolha (este último, não empregado, de português do Brasil), o que faz um total de mais de 138.000 tokens. O Bosque foi anotado automaticamente e, posteriormente, revisto de forma manual por linguistas. Sendo um corpus com informação sintáctica, esta foi eliminada na conversão para o formato requerido por FreeLing. Para as diferentes avaliações, este corpus foi dividido em conjuntos aleatórios de treino e teste.

O corpus utilizado para treinar o módulo *PoS-tagger* de galego foi criado no projecto Gari-Coter (Barcala *et al.*, 2007), e contém mais de 237.000 tokens; o corpus, gerado a partir de notícias jornalísticas, é especializado em economia. Uma vez que a anotação morfossintáctica deste recurso seguiu os *standards* do Grupo EAGLES (Leach e Wilson, 1996), as únicas adaptações precisas para o treino foram relativas à homogeneização de alguns elementos das etiquetas (*tags*): Modificou-se, por exemplo, o caso de alguns pronomes (de nominativo para oblíquo), ou o género dos determinantes indefinidos (de comum para neutro), de acordo com o conjunto de etiquetas (*tagset*) definido e utilizado no dicionário. Uma vez que o corpus galego se compõe de notícias económicas, um pequeno corpus jornalístico de 6.200 tokens foi criado e revisto manualmente para as avaliações.

[3] `http://www.linguateca.pt/Floresta/corpus.html#bosque`

Dicionários

Para além do corpus, o treino do *PoS-tagger* de FreeLing requer também um dicionário de formas flexionadas (que contenha os lemas e *tags* possíveis para cada token).

Para português, utilizou-se o léxico de formas simples LABEL-LEX (SW) (Eleutério *et al.*, 2003), que contém mais de 1.257.000 formas, geradas a partir de perto de 120.000 lemas.

Para galego, o trabalho partiu do dicionário criado pelo Seminario de Lingüística Informática da Universidade de Vigo, que fazia parte de anteriores versões de FreeLing. O dicionário foi ampliado com entradas verbais e nominais extraídas de diferentes corpora e flexionadas automaticamente com ajuda de sistemas de flexão nominal e de conjugadores verbais (Gamallo *et al.*, 2013). Actualmente, o dicionário contém mais de 428.000 entradas, o que se corresponde com mais de 577.000 formas se tivermos em conta aquelas entradas com mais de uma análise.

Adaptação

Os dous recursos referidos para português têm características diferentes em relação à anotação morfossintáctica, pelo que foi preciso fazer uma conversão de cada um deles para o formato aqui utilizado. Neste processo surgiram algumas incoerências que implicaram tomadas de decisão do ponto de vista linguístico. Assim, *tags* como "pron-indp: pronome independente" (utilizado no Bosque), não tinham correspondente directo nas etiquetas do léxico, pelo que não foi possível uma transferência automática entre os formatos. A conversão destes casos teve de ser incluída individualmente no processo de transformação, e decidir em cada ocorrência dos tokens no corpus qual era a etiqueta que lhe correspondia de acordo com o dicionário.

Para além das inconsistências no nível morfossintáctico, a conversão do corpus e do dicionário apresentou problemas em termos de lematização nominal. Assim, enquanto o Bosque lematiza os adjectivos superlativos como elementos não derivados (*altíssimo* é o lema de *altíssimo/a/(s)*), o LABEL-LEX (SW) opta por decisões mais coerentes do ponto de vista teórico: *altíssimo/a/(s)* $>$ *alto*. De modo similar, outras diferenças notórias entre as lematizações do corpus e do dicionário foram as relacionadas com a derivação semântica: O LABEL-LEX (SW) considera que formas como *mulher* (nome) e *melhor* (adjectivo) derivam de *homem* e *bom*, respectivamente, enquanto o Bosque atribui *mulher* e *melhor* como lemas dos mesmos tokens.

Nestes casos, a solução adoptada foi de modo geral aquela que tivesse como base processos morfológicos e não semânticos. Assim, no primeiro dos casos, optou-se por considerar os adjectivos superlativos como derivados do adjectivo simples; no segundo exemplo, a decisão tomada foi consistente com a lematização utilizada no Bosque, que diferencia as formas que não apresentam uma relação morfológica directa.

No tratamento das locuções e dos nomes próprios compostos por mais de um elemento, o Bosque apresenta algumas inconsistências que, para os objectivos deste trabalho, não permitiram avaliar o desempenho do reconhecedor de expressões multipalavra com precisão. Assim, enquanto *Conselho de Administração da PEC-Alimentação* é dividido em quatro tokens ("Conselho de Administração", "de", "a" e "PEC-Alimentação"), uma expressão como *director-clínico do Hospital Prisional S. João de Deus* é anotada no corpus como um único token/lema. A solução adoptada nestes casos foi a seguinte: os elementos marcados como locuções no Bosque foram extraídos automaticamente, e adicionados à lista de expressões multipalavra de FreeLing, depois de serem revistos manualmente. Nos corpora de treino e avaliação, porém, estas formas foram divididas em tokens individuais, pelo que o treino e a avaliação foram realizadas sem locuções.

Em galego, a etiquetação entre os recursos escolhidos tinha sido mais consistente, pelo que processo de adaptação foi menos complexo.

2.4. Tokenização

A primeira ferramenta adaptada foi o tokenizador. Este módulo converte, através de regras, um texto plano num vector de palavras. É uma tarefa relativamente simples, que consiste em identificar as fronteiras de palavras (e outros tokens tais como signos de pontuação) através dos espaços e da própria pontuação, pelo que a maior dificuldade encontra-se na identificação de algumas contracções.

Formas ambíguas como *desse* (em português, ou *dese* em galego) podem ser um verbo (*dar*) ou uma contracção de preposição e demonstrativo (*de+esse/ese*). Este tipo de ambiguidades provoca uma circularidade entre o etiquetador morfossintáctico e o tokenizador. Este último não poderá decidir se separar *desse* em *de+esse* sem conhecer a sua categoria morfossintáctica, mas o *PoS-tagger* não pode ser aplicado sob um texto não tokenizado. As soluções que FreeLing permite adoptar nestes casos encontram-se na análise morfológica e morfossintáctica (dicionário, afixos e *PoS-tagger*), pelo que neste primeiro processo as contracções não

serão separadas. A saída do tokenizador, portanto, manterá ainda a ambiguidade neste tipo de formas.

Por este motivo, é importante ter em conta a ordem de aplicação entre o tokenizador e o *PoS-tagger*, a qual influencia o modo como as contracções ambíguas são tratadas (Graña *et al.*, 2002; Branco e Silva, 2003).

Outro aspecto a considerar é a interacção entre o tokenizador e o segmentador de orações. Na ordem de aplicação proposta (tokenizador $>$ segmentador), o primeiro dos módulos deve reconhecer as abreviaturas (identificando o ponto como parte da abreviatura: "*Sr.*" e não "*Sr*" "."), por exemplo) para evitar os casos de ambiguidade mais comuns na entrada do segmentador de orações. A diferença entre as configurações do tokenizador de português e galego está, portanto, na lista de abreviaturas.

Esta estratégia de tokenização já tinha sido utilizada com êxito para o português europeu, obtendo valores de precisão superiores a 99% (Silva, 2007).

2.5. Segmentação de orações

O segmentador de orações recebe a saída do tokenizador e devolve uma nova oração cada vez que detecta uma fronteira. As línguas românicas não apresentam muitas diferenças nos marcadores ortográficos, pelo que a adaptação deste módulo para português e galego não teve grandes dificuldades. Uma vez que as ambiguidades mais frequentes entre os pontos finais e os pontos de abreviação já foram resolvidas pelo tokenizador, o segmentador não precisa tratar especificamente estes casos.

Entre as duas variedades analisadas, as diferenças de segmentação não são significativas, dizendo respeito a especificidades ortográficas, como a utilização dos pontos de interrogação e exclamação no início de orações (não utilizados em português, mas facultativos em galego).

Este tipo de estratégias também tinham sido avaliadas em português europeu, com resultados de mais de 99% de precisão (Silva, 2007).

2.6. Análise morfológica

O sistema de análise morfológica de FreeLing é um conjunto de módulos que realizam tarefas como a identificação de numerais e de datas, o reconhecimento de entidades mencionadas e de expressões multipalavra, bem como a pesquisa no dicionário (que inclui lematização) e o tratamento dos afixos.

Esta secção descreve a adaptação do módulo de pesquisa em dicionário, para a qual foi precisa a transformação do formato dos léxicos disponíveis e a criação de regras de lematização de afixos verbais e nominais.

Este módulo compõe-se de dous submódulos que actuam em paralelo: um deles procura no dicionário todas as possibilidades de análise de cada um dos tokens encontrados na entrada, enquanto o outro aplica as regras de lematização de afixos, que permitem que alguns tokens que não se encontram no dicionário sejam analisados pelo sistema.

O dicionário de português europeu contém mais 1.257.000 formas, enquanto o de galego supera as 577.000 formas (Secção 2.3). Note-se que FreeLing não possui um lematizador próprio, senão que o lema de cada token é procurado no léxico. Isto implica a necessidade de léxicos amplos, com o fim de atingir níveis altos de precisão nesta tarefa.

O submódulo de tratamento de afixos permite criar regras de lematização de formas com prefixos e sufixos. Assim, não é preciso incluir no dicionário todas as possibilidades de combinação de formas verbais com clíticos, nem diminutivos, aumentativos, advérbios terminados em *mente*, ou formas prefixadas.

Em relação às formas compostas por verbo e pronome clítico, é preciso referir que a ortografia do português separa o verbo e o pronome com um hífen, e mantém independente a acentuação da forma verbal (e.g., *conhecem-me*). Isto faz com que este processo seja uma tarefa trivial. Porém, em galego a forma composta é escrita como uma única palavra (*coñécenme*), pelo que foi preciso criar um conjunto de regras de análise de sufixos que tenham em conta tanto a identificação dos tokens (verbo + pronome), como a adição ou supressão de acentos gráficos (*coñécen* > *coñecen*).[4]

A pesquisa em dicionário e o tratamento de afixos permitem que na execução do etiquetador morfossintáctico sejam tratadas as contracções não divididas pelo tokenizador. O funcionamento é o seguinte: as contracções não ambíguas (por exemplo *do*: preposição *de* + artigo *o*), estão presentes no dicionário com o formato "do de+o SPS00+DA",[5] pelo que estas formas são divididas em dous tokens na saída final.[6]

[4]Uma versão posterior do módulo de tratamento de afixos em galego foi apresentada em Solla Portela (2010).

[5]Formato de três colunas separadas por um espaço, em que a primeira é o token, a segunda o lema e a terceira a etiqueta morfossintáctica (veja-se o *tagset* na Tabela A.1, página 158).

[6]Pode entender-se que a análise de *do* contém ambiguidade relativa à categoria de *o*, que além de artigo, poderia ser pronome nos casos em que o núcleo da frase nominal não está preenchido: "O homem do qual ele falou", pelo que a entrada do dicionário incluiria SPS00+DA/PD. No caso que nos ocupa, unicamente nos referimos à ambiguidade em que uma única forma pode ser analisada como contraída ou não: *deste* como contracção de preposição+demonstrativo ou como verbo.

Porém, os casos de ambiguidade (*desse/dese, destes, pelo/polo*, etc.) podem ser tratados —fundamentalmente— de duas maneiras: incluindo-as no dicionário, ou acrescentando regras de lematização destas formas ao submódulo de tratamento de afixos. As duas soluções referidas permitem evitar a circularidade referida na Secção 2.4, uma vez que todas as alternativas existentes no módulo de análise morfológica são avaliadas pelo desambiguador morfossintáctico e pelo *PoS-tagger* que, se for preciso, realizará uma retokenização.

Qualquer das duas estratégias é similar às adoptadas em Graña *et al.* (2002) ou em Branco e Silva (2003), uma vez que deixam a decisão de separar (ou não) os casos de ambiguidade aos módulos de análise morfossintáctica, e não ao tokenizador.

Avaliação

A lematização foi avaliada dividindo o número de lemas correctamente atribuídos pelo número total de lemas do corpus de teste. Em português europeu, os resultados foram obtidos num extracto do corpus Bosque de 50.000 tokens, com uma precisão de $98,58\%$. Em galego, o corpus de teste foi de 6.200 tokens, sendo a precisão de $99,41\%$.

Dentro do conjunto de módulos de análise morfológica, a seguinte ferramenta de FreeLing (o desambiguador morfossintáctico) atribui uma probabilidade para cada uma das possíveis etiquetas de cada token e, com base na análise das terminações, tenta saber que etiquetas são possíveis nas formas desconhecidas. Esta classificação é realizada de acordo com a aprendizagem realizada num corpus de treino etiquetado.

2.7. Anotação morfossintáctica

FreeLing permite utilizar dous métodos de anotação morfossintáctica: um modelo probabilístico com base em HMM e um método híbrido (*relax*) que combina informação estatística com restrições linguísticas definidas manualmente (Padró, 1998).

O modelo híbrido é —apesar de ligeiramente mais lento— de maior precisão do que o estatístico, mas requer a criação manual das restrições para cada variedade, pelo que no treino dos módulos para português e galego foi utilizado o modelo HMM.

Existem dous factores cruciais —para além do dicionário— no desempenho de um etiquetador morfossintáctico probabilístico: um deles é o tamanho —e qualidade— do corpus de treino. Quanto maior for o corpus, melhor será o modelo aprendido pelo sistema durante o treino (embora com certos limites (Banko e Brill, 2001)). O outro é o conjunto de etiquetas:

se este for muito complexo, a informação fornecida pelo etiquetador será maior, mas a sua precisão mais baixa. Pelo contrário, se o *tagset* for reduzido, a sua precisão será mais elevada, mas pode correr-se o risco de que a informação obtida não seja suficiente para os objectivos do *PoS-tagger*.

Um dos propósitos da adaptação de FreeLing para português e galego foi o de utilizá-lo como etiquetador morfossintáctico base para o analisador sintáctico DepPattern (Secção 1.4). Tendo em conta que este sistema requer informação morfológica (género, número, pessoa, tempo e modo verbal, etc.), e com base nos *tagsets* utilizados nas outras línguas analisadas por FreeLing, decidiu-se utilizar as recomendações propostas pelo Grupo EAGLES (Leach e Wilson, 1996). O *tagset* definido para português europeu contém 255 *tags*, enquanto para galego se empregaram 277 etiquetas.[7]

O *tagset* utilizado contém informação morfossintáctica detalhada, mas não todos estes dados são utilizados propriamente pelo *PoS-tagger*; este usa unicamente os dous primeiros elementos da etiqueta, sendo os restantes extraídos do léxico. O primeiro elemento da etiqueta (*D*, Determinante, *P*, Pronome, *N*, Nome, etc.) indica a categoria morfossintáctica; o segundo (*D* Demonstrativo, *P*, Possessivo, etc., variando em função do primeiro elemento) refere a subclasse da categoria à qual pertence. O resto de entradas das etiquetas varia em função da categoria principal, e englobam aspectos como o possuidor (singular ou plural) dos possessivos, o grau (aumentativo ou diminutivo) dos nomes, o caso dos pronomes ou informação sobre modo, pessoa e número dos verbos.

Avaliação

A precisão da anotação morfossintáctica é crucial para subsequentes tarefas de PLN, sobretudo naquelas formas que contêm ambiguidade e que podem provocar maior índice de erros em etapas posteriores.

De modo geral, considera-se que a *baseline* para esta tarefa se situa em 90%, e que o estado-da-arte supera o 97% nos melhores resultados. Contudo, têm surgido algumas críticas à avaliação destas ferramentas, com base no tipo de texto utilizado durante o processo. Comparando diferentes avaliações de *PoS-taggers* sobre textos de diversas procedências (blogues,

[7] A estas quantidades são acrescentados 24 *tags* de símbolos de pontuação (atribuídos não pelo *PoS-tagger*, mas pelo identificador de pontuação). A Tabela A.1 contém o formato do conjunto de etiquetas utilizado. Note-se que, para manter a compatibilidade com outros *tagsets* de FreeLing, os elementos que não sejam precisos em português e galego serão marcados com um <0> (veja-se como exemplo os valores semânticos dos nomes, que ocupam os elementos 5 e 6).

jornais digitais e outros *sites*) e tipologias (literário, científico, jornalístico, etc.), e não em texto com condições mais homogéneas, a precisão desce abaixo de 93%, e apresenta grandes níveis de variação em função do género textual (Giesbrecht e Evert, 2009).

Neste trabalho, a avaliação do processo de anotação morfossintáctica foi realizada dividindo o número de tokens cuja etiqueta foi correctamente atribuída pelo número total de tokens do texto. Esta tarefa, aparentemente trivial, pode apresentar problemas derivados do alinhamento entre o *gold-standard* (o corpus de referência, corrigido manualmente) e o texto etiquetado automaticamente. Este último pode conter um número diferente de tokens do que o *gold-standard*, devido à tokenização ou à identificação de nomes próprios, locuções, etc. Assim, a forma *Presidente Mário Soares*, pode ser analisada como um único nome próprio (*Presidente_Mário_Soares*), pode ser divida em dous elementos (*Presidente / Mário_Soares*), ou em três (*Presidente / Mário / Soares*). Para tratar estes casos, o sistema de avaliação tem três parâmetros de execução, com o seguinte funcionamento:

– *NoTok*: Se são detectados erros de divisão (*split*): *Presidente_Mário_Soares NP* versus *Presidente NP / Mário NP / Soares NP*, unicamente é avaliado o *tag* do primeiro token, pelo que é contabilizado um acerto. Este método considera que os erros de tokenização não devem ser levados em conta na avaliação do *PoS-tagger*.

– *Tok*: Se houver diferenças de tokenização, são contabilizados todos os erros (no caso anterior, três). Note-se que, no caso de que a tokenização e a atribuição da etiqueta em palavras com mais de um token sejam correctas, é marcado um único acerto.

– *NoLoc*: Este tipo de avaliação ignora todos os tokens que não estiverem alinhados; no exemplo referido, não seria contabilizado nenhum erro nem acerto.

Como foi dito na Secção 2.3, o Bosque apresenta alguma inconsistência na anotação das locuções e outras expressões multipalavra, facto que devemos ter em conta na consideração dos resultados destas avaliações. Por esta razão, foi gerada uma outra versão do corpus de teste, na qual se realizou *split* de todos os elementos que continham mais de um token. Assim, executando o *PoS-tagger* sem identificação de locuções nem de nomes próprios compostos, a saída é um texto alinhado perfeitamente com este novo *gold-standard*, pelo que a avaliação resulta mais simples. Assim, um quarto método (*OnlyTag*) tem em conta unicamente os erros e acertos, evitando diferenças de tokenização entre os corpora avaliados. Neste caso, na avaliação das locuções ou dos nomes próprios compostos são contabilizados todos os tokens,

Avaliação	**Português**		**Galego**	
	Tag Completo	*SingleTags*	*Tag Completo*	*SingleTags*
NoTok	94,79	96,01	97,70	98,04
Tok	94,47	95,73	97,19	97,56
NoLoc	95,04	96,26	97,72	98,07
OnlyTag	94,32	95,54	97,50	97,91

Tabela 2.1: Precisão dos *PoS-taggers* em português europeu e galego.

pelo que uma locução bem etiquetada somará mais acertos do que com outros métodos. Esta distorção, contudo, é compensada de alguma maneira pelos casos em que a ferramenta falha, nos quais também são contabilizados um maior número de erros.

Os quatro métodos referidos avaliam a precisão do *PoS-tagger* com o *tagset* definido (Tabela A.1). Uma vez que este contém informação muito pormenorizada, e varia notoriamente em relação aos utilizados noutros trabalhos, foi realizada também uma avaliação de cada uma das execuções com um *tagset* mais reduzido (*SingleTags*). Para este fim, unicamente se tiveram em conta os dous primeiros elementos das etiquetas (categoria e tipo, salvo para os verbos, em que se avalia também um terceiro elemento: o modo), ignorando assim informação que pode ser inferida por outros meios.[8] Desta maneira, e apesar de os resultados não poderem ser directamente comparáveis com os de outros trabalhos (devido a diferenças não apenas no *tagset*, mas também nos corpora de treino e de teste, entre outras), estes dados são obtidos em condições similares às de outras análises. Ao mesmo tempo, esta avaliação demonstra a importância do *tagset* no desempenho de um etiquetador morfossintáctico.

A Tabela 2.1 mostra os resultados das diferentes avaliações do *PoS-tagger* para português e galego. Em português, os valores são a média de cinco execuções sobre extractos aleatórios de quase 10.000 tokens, com o sistema treinado nos restantes 130.000, salvo para o método *OnlyTag*, treinado sobre 90.000 tokens e avaliado sobre perto de 50.000.

Em galego, o treino foi realizado sobre o corpus completo (quase 238.000 tokens), e a avaliação sobre um corpus extraído de jornais electrónicos e etiquetado manualmente, de 6.200 tokens.

Os resultados das várias avaliações dos dous sistemas indicam que, actualmente, FreeLing consegue realizar análises morfossintácticas de textos de diversa procedência com desempenhos próximos do estado-da-arte, situado entre $\approx 95\%$ e $\approx 97\%$ em função da variedade linguística e da avaliação (veja-se a Secção 2.2).

[8] Este *tagset* pode ver-se na Tabela A.2 (página 159), também sem os símbolos de pontuação.

Entre as duas variedades linguísticas, as diferenças de precisão são notórias, mas os resultados não devem ser directamente confrontados. A este respeito devemos notar, por um lado, os recursos utilizados; enquanto o *PoS-tagger* de português foi treinado sobre extractos de 130.000 tokens, para galego usou-se um texto mais de 100.000 tokens maior, pelo que é esperável que o desempenho seja superior (Banko e Brill, 2001). Em relação a isto, note-se que os resultados mais baixos da avaliação do português foram com o método *OnlyTag*, treinado sobre um corpus mais pequeno. Por outro lado, é importante destacar também as características dos corpora de teste utilizados para a avaliação. O português foi avaliado sobre extractos do próprio Bosque 8.0, com mais ruído —e de maior tamanho— do que o corpus de avaliação de galego (mais consistente e com menos ruído). Estas diferenças de desempenho sugerem, como Giesbrecht e Evert (2009), que as características do texto a etiquetar influenciam decisivamente a qualidade da etiquetagem.

2.7.1. Diferentes variedades do português

O modelo de anotação morfossintáctica apresentado para português foi treinado (e avaliado) com recursos específicos da variedade padrão de Portugal. Contudo, as estratégias de extracção de informação avaliadas nesta tese pretendem utilizar como fontes corpora diversos extraídos da Web, que podem estar escritos em diferentes variedades desta língua.

Para além disso, o Acordo Ortográfico de 1990 (AO), que tenta unificar as normas ortográficas das diferentes variedades nacionais do português, está a ser implantado em vários dos países de língua oficial portuguesa.[9] Assim, alguns dos maiores jornais do Brasil e Portugal já utilizam a nova ortografia desde 2010 (e.g., *Diário de Notícias* ou *Jornal de Notícias*, em Portugal, ou *Folha de São Paulo* no Brasil), enquanto outros não o fazem (e.g., o português *Público*). Portanto, na Web em português coexistem hoje em dia textos escritos em diferentes ortografias e em diferentes variedades nacionais (cujas diferenças são principalmente lexicais e sintácticas).

Tendo isto em conta, esta secção avalia a utilização de diferentes anotadores morfossintácticos, treinados com várias combinações de recursos de português europeu, brasileiro e do Acordo Ortográfico de 1990. As avaliações são realizadas com um novo corpus da língua portuguesa, que inclui textos de diferentes variedades linguísticas, tipologias textuais e normas ortográficas.

[9] `http://pt.wikipedia.org/wiki/Acordo_Ortografico_de_1990`

Variedade	*Tamanho*	*Vocabulário*
Brasil	11.460	3.137
Portugal	13.987	3.637
Angola	4.180	1.403
Moçambique	5.517	1.700
Wikipedia	17.187	4.003
Total	52.331	9.873

Tabela 2.2: Tamanho (em número de tokens) e vocabulário (número de pares token-*tag* diferentes) do Corpus-Web e dos subcorpora.

Recursos

Para além dos recursos de português europeu (PE) apresentados nos testes anteriores (Secção 2.3), foram utilizados os seguintes corpora e dicionários:

Corpora

– Mac-Morpho (PB): para português do Brasil utilizou-se o corpus Mac-Morpho, que contém 1.167.183 formas.[10]

– Corpus-Web: para a avaliação, criou-se um novo corpus que inclui diferentes tipologias textuais, variedades linguísticas e normas ortográficas, gerado para representar de alguma maneira a Web em português e cuja anotação morfossintáctica foi revista e corrigida manualmente. O corpus tem mais de 52.000 tokens e inclui textos das seguintes fontes: três jornais portugueses, dous brasileiros, um de Angola e um de Moçambique. Para além disso, contém textos da Wikipedia em português,[11] que à sua vez inclui diferentes variedades. A Tabela 2.2 mostra os pormenores deste novo recurso.

Note-se que os jornais de Moçambique e Angola, e um dos jornais de Portugal não empregam o AO, mas os outros jornais portugueses e brasileiros sim. Além disso, Moçambique e Angola utilizaram historicamente a ortografia do português europeu, embora tenham diferenças lexicais e sintácticas. O corpus da Wikipedia contém textos do Brasil e Portugal, tanto com ortografia anterior ao AO, como posterior.

[10] `http://www.nilc.icmc.usp.br/lacioweb/corpora.htm`
[11] `http://pt.wikipedia.org`

Léxicos

– PB_Dict: como dicionário de PB, utilizou-se o léxico DELAF_PB, que contém 878.651 formas e 61.095 lemas.[12]

– PEB_Dict: um novo dicionário composto foi gerado, com base nos dicionários de PE e PB. Este novo recurso criou-se unificando todos os triplos (token-lema-*tag*) dos dicionários PE e PB. Nos casos em que os dicionários mostravam inconsistências (principalmente em palavras funcionais), preferiu-se a versão do português europeu. O dicionário PEB contém uns 1.254.000 triplos token-lema-*tag* e uns 1.179.000 pares token-*tag*, de 112.000 lemas diferentes.[13]

– AO+_Dict: o último dos recursos utilizados foi um novo dicionário composto, baseado no dicionário PEB e num novo dicionário específico do Acordo Ortográfico. Assim, o léxico AO+ contém formas que não são correctas no AO (anteriores a ele), tendo uns 1.277.000 triplos token-lema-*tag* e $\approx$ 1.200.000 pares token-*tag*. O número de lemas é de 119.000.

É importante referir que os processos de fusão dos diferentes léxicos implicaram um aumento na ambiguidade do anotador morfossintáctico, dado que nos léxicos fusionados algumas formas pertencem a um maior número de triplos token-lema-*tag* do que nos dicionários simples.

Para a realização destes testes, foi criado um novo *tagset* (Tabela A.3, página 160), mais simples do que os utilizados anteriormente, com 23 etiquetas, 24 *tags* de pontuação, 5 de expressões numerais e 1 para datas e horas. Uma vez que não é utilizado qualquer método de tokenização (a entrada do *PoS-tagger* é o texto já tokenizado), mantêm-se duas etiquetas específicas para contracções ambíguas. Todos os recursos foram automaticamente adaptados a este *tagset*, reduzindo-se em $\approx$ 50.000 o número de triplos de cada dicionário.

Modelos

Para avaliar o desempenho de diferentes etiquetadores nos vários corpora disponíveis, foram treinados os seguintes modelos:

[12] `http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html`

[13] O dicionário PEB é menor do que o dicionário PE (Secção 2.3) porque a conversão a um *tagset* reduzido (mostrado a seguir) reduziu o número de entradas token-lema-*tag*.

– Modelos específicos para português europeu (`PEtag`) e português do Brasil (`PBtag`), utilizando unicamente os recursos próprios de cada variedade.

– Modelos combinados, com o fim de avaliar tanto a sua precisão na anotação do Corpus-Web como nos corpora com normas ortográficas anteriores ao AO.

Para treiná-los, foi utilizado o módulo HMM de FreeLing.

O modelo `PEtag` foi treinado aproximadamente no 83% ($\approx$ 120.000 tokens) do corpus Bosque (PE), e avaliado no restante 17% ($\approx$ 23.000).[14]

O modelo `PBtag` utiliza $\approx$ 79% para o treino, e $\approx$ 21% para teste. Uma vez que o corpus PB (Mac-Morpho) é muito maior do que o PE (Bosque), foram também seleccionados dous subcorpora, com o fim de obter um conjunto equilibrado para realizar mais combinações: uma versão reduzida do corpus de treino (com $\approx$ 150.000 tokens) e uma versão reduzida do corpus de teste ($\approx$ 23.000 tokens), obtendo assim conjuntos de treino e teste de tamanhos mais próximos dos utilizados em português europeu.

O modelo `ALLtag` utiliza os corpora PE e PB para treinar, e o dicionário PEB_Dict. `ALLtag+` foi treinado com os mesmos corpora, mas utilizando o dicionário AO+_Dict.

Por último, os modelos `PEBtag` foram treinados com o corpus PE e a versão reduzida do corpus de treino de PB. `PEBtag` e `PEBtag+` também diferem no dicionário utilizado: o primeiro usou o dicionário PEB_Dict, enquanto o modelo `PEBtag+` empregou o dicionário AO+_Dict.

Todos os conjuntos de treino e teste foram seleccionados de modo aleatório, e os corpora de teste nunca foram utilizados durante o processo de aprendizagem.

Avaliação

Os testes avaliam tanto o desempenho dos *PoS-taggers* nos corpora PE, PB e Corpus-Web, como em cada um dos subcorpora deste último: Angola (AN), Moçambique (MO), Brasil (BP_AO), Portugal (PE_AO) e Wikipedia (Wiki). A média total de anotação foi calculada substituindo o corpus PB pela sua versão reduzida, para minorar o desvio nos resultados.

A Tabela 2.3 contém os resultados (precisão) dos diferentes modelos avaliados. À diferença dos experimentos anteriores (Tabela 2.1), estes testes utilizam unicamente o módulo *PoS-tagger* (com a entrada tokenizada, excepto as contracções ambíguas), pelo que só se mostra um tipo de avaliação.

[14] O número de tokens varia em relação aos testes da Secção 2.7 devido a diferenças de tokenização.

Modelo	*BP*	*EP*	*AN*	*MO*	*PB_AO*	*PE_AO*	*Wiki*	*Web*	*Total*
`PBtag`	95,96	96,03	97,06	96,39	96,35	96,88	95,52	96,28	96,13
`PEtag`	95,35	**97,46**	**98,18**	**97,76**	**97,29**	**97,80**	96,25	**97,20**	**96,85**
`ALLtag`	**96,07**	96,94	97,30	96,91	96,68	97,18	96,50	96,83	96,64
`ALLtag+`	**96,07**	96,94	97,30	96,91	96,68	97,21	**96,53**	96,86	96,65
`PEBtag`	95,74	97,04	97,37	97,06	96,97	97,28	96,43	96,92	96,65
`PEBtag+`	95,74	97,04	97,37	97,06	96,97	97,31	96,45	96,93	96,66

Tabela 2.3: Precisão de 6 *PoS-taggers* em diferentes corpora de teste. *Web* é a *micro-average* (veja-se a Secção 4.3) dos resultados de *AN*, *MO*, *PB_AO*, *PE_AO* e *Wikipedia*. O resultados de *Total* são as médias de todos os resultados, excepto *PB*, substituído pela sua versão reduzida.

Os modelos `PBtag` e `PEtag` obtiveram $95,96\%$ e $97,46\%$ nos seus respectivos conjuntos de teste, mas a sua precisão desceu $1,4\%$ e $0,6\%$ (respectivamente) na análise da outra variedade. Nos mesmos corpora (PB e PE), o desempenho dos modelos combinados dependeu da distribuição do conjunto utilizado para a aprendizagem. Assim, os etiquetadores `ALLtag` (com maior quantidade de dados de PB) funcionam melhor no corpus PB, enquanto a precisão dos modelos `PEBtag` é maior ao analisar textos em PE.

Se compararmos as duas variantes dos modelos `ALLtag` e `PEBtag`, com as `PBtag` e `PEtag`, os sistemas combinados mostram-se mais compensados na etiquetação dos conjuntos PE e PB.

Em relação ao dicionário, o impacto do léxico AO+_Dict nestes corpora foi nulo, dado que nem PB nem PE contêm textos com a ortografia do AO.

Na anotação do Corpus-Web, o modelo `PEtag` continua a ser o melhor em cada um dos subcorpora, excepto no da Wikipedia. A este respeito, cabe referir que a consistência entre o léxico e o corpus de treino desta variedade é maior (Secção 2.3), e que boa parte do Corpus-Web está escrita seguindo a ortografia PE (antes do Acordo Ortográfico).[15]

De modo geral, os modelos `PEBtag` funcionam melhor do que os `ALLtag` (excepto no corpus da Wikipedia), mas não melhoram os resultados do modelo `PEtag`.

Os resultados no Corpus-Web mostram que o dicionário AO+_Dict tem um impacto baixo, mas positivo, na anotação. O seu efeito é só visível naqueles textos cuja ortografia tem mais mudanças derivadas da utilização do AO (PE_AO e Wikipedia), tendo pequenas melhoras ($\approx 0,3$) se se utilizar a versão alargada do dicionário, que inclui formas do Acordo Ortográfico.

[15] Lembre-se que os textos que formam AN, MO e um de PE_AO utilizam essa ortografia, pelo que os resultados seguem tendências similares às do corpus PE.

Contudo, apesar de que os novos dicionários aumentam a ambiguidade do etiquetador, o seu impacto foi positivo em todos os testes.

Em suma, é preciso apontar que a consistência entre o corpus de treino e o dicionário foi crucial nos testes realizados, sendo o modelo `PEtag` o de maior precisão. À parte disso, o desvio existente entre as diferentes variedades linguísticas, tanto nos conjuntos de treino como de teste, também teve impacto nos resultados. Finalmente, os testes realizados mostraram que os novos dicionários têm uma influência positiva na etiquetação de textos anteriores e posteriores ao Acordo Ortográfico da língua portuguesa.

2.7.2. Estratégias de correcção

A análise dos erros dos diferentes *PoS-taggers* indica que muitos dos problemas destes modelos derivam da dispersão dos dados morfossintácticos ou léxico-semânticos. Os modelos estatísticos têm, em geral, melhor desempenho do que os etiquetadores baseados em regras, apesar de os primeiros serem treinados com pouca informação de carácter linguístico. Assim, a utilização de regras de correcção sobre a saída de um etiquetador pode melhorar a precisão da anotação.

Nesta secção, é apresentado um compilador de gramáticas que gera *parsers* orientados à correcção da anotação morfossintáctica. O *parser* utiliza como entrada a saída de um *PoS-tagger*, e gera uma nova saída em diferentes formatos, um dos quais, idêntico ao do etiquetador inicial. Assim, incluindo gramáticas básicas com regras de correcção, é possível realizar um processamento que melhore o texto anotado.

Alguns trabalhos como o já referido Padró (1998) utilizam restrições inseridas manualmente no processo de aprendizagem de *PoS-taggers* probabilísticos. Outros, como Finger (2000), implementam no próprio etiquetador estratégias de correcção similares à aqui apresentada.

Método

O procedimento utilizado para a correcção de erros de anotação morfossintáctica baseia-se no compilador de gramáticas incluído em DepPattern (Secção 1.4). As gramáticas utilizam um formalismo específico que permite modificar a informação linguística de cada token (tipo, categoria, género, número, etc.), e estabelecer dependências sintácticas entre eles.

Assim, o método de correcção consiste na inclusão de regras que modifiquem ou adicionem informação linguística, definindo o padrão léxico-sintáctico de erros sistemáticos.

Um exemplo deste tipo de regras é o mostrado na Figura 2.1, que corrige a etiqueta atribuída ao token *a* quando aparece à esquerda de um nome masculino.

Single: DET<token:[Aa]> [NOUN<gender:M>]
Corr: tag:PS, type:P, lemma:a

Figura 2.1: Exemplo de regra de correcção.

"Single" indica que a regra é individual (não de dependência sintáctica), sendo o elemento fora dos parênteses rectos aquele que vai ser corrigido, enquanto o que está dentro é o contexto de aplicação. No exemplo, um determinante feminino (DET) ocorre antes de um nome (NOUN) com género masculino, pelo que a regra de correcção (Corr) substitui a sua categoria por PS (preposição), e o seu lema por `a`.

Uma regra como esta poderia corrigir casos como "a nível nacional", analisado previamente como segue (com formato <token lema *TAG*>, e o *tagset* definido na Tabela A.1):

```
a o DA0FS0                          a a SPS00
nível nível NCMS000        →        nível nível NCMS000
nacional nacional AQ0CS0            nacional nacional AQ0CS0
```

O formalismo permite especificar padrões de aplicação maiores, incluindo elementos facultativos, disjunções, conjunções e outro tipo de expressões regulares. Também pode estabelecer previamente dependências sintácticas, o que possibilita a simplificação de algumas das regras, aumentando deste modo a abrangência do corrector.

AdjunctLeft: ADJ NOUN
Agreement: gender, number

Figura 2.2: Exemplo de regra de dependência sintáctica.

A regra da Figura 2.2 estabelece uma dependência entre um adjectivo (ADJ) e um nome com concordância em género e número, sendo o núcleo da dependência o elemento à direita. Aplicando esta regra antes da regra de correcção mostrada acima, é possível lidar com padrões <DET ADJ NOUN>, porque o adjectivo é agora dependente do nome.

Depois de aplicar gramáticas de correcção, o tempo de execução do processo de anotação incrementa-se de 11.500 a 8.500 tokens por segundo, pelo que o custo computacional do *parser* não é alto.

Número	*Token*	*Tag correcto*	*Tag atribuído*
86	que	CS	PR0CN000
81	a	SPS00	DA0FS0
41	um	Z	DI0MS0
40	a	DA0FS0	SPS00
38	que	PR0CN000	CS
37	uma	Z	DI0FS0
24	o	PD0MS000	DA0MS0

Tabela 2.4: Erros mais frequentes do *PoS-tagger* em $\approx$ 50.000 tokens.

Erros do etiquetador morfossintáctico

A Tabela 2.4 mostra os erros mais comuns encontrados numa avaliação dos resultados do *PoS-tagger* de PE (Tabela 2.1). A anotação do token *que* produziu 141 erros, principalmente entre conjunção (CS) e pronome relativo (PR0CN000), com 126 erros. A forma *a* teve 153 erros (entre determinante, pronome demonstrativo e pessoal, preposição e nome comum), enquanto *o* foi incorrectamente etiquetado em 67 casos. A anotação de *um* e *uma* (numerais ou determinantes) também foi um dos erros mais frequentes.

Numa primeira análise dos erros de anotação, alguns deles revelam-se como facilmente corrigíveis através de regras (morfo)sintácticas. A não concordância em género ou número entre determinante e nome é um exemplo deste tipo de erros. Assim, a etiquetação de tokens específicos como estes pode ser melhorada com regras básicas. Como exemplo, a regra de correcção exemplificada acima (Figura 2.1) melhora a precisão da anotação de *a* como preposição em 33% (corrige 29 de 87 erros), e só gera 3 novos erros (de 46). Este é um padrão muito rígido, pelo que a regra produz poucos erros (algum deles provocado por erros de etiquetação do nome, e não pela própria regra). Outros contextos, como aqueles em que aparece *que* como conjunção ou pronome necessitam um processamento mais complexo.

Testes

A modo de teste, foi implementado um conjunto de regras de correcção, em particular dos tokens *a*, *o* e *que*, nos contextos de erro mais frequentes. As regras foram escritas manualmente, com base numa análise semiautomática dos erros. O *parser* de correcção foi testado em cinco corpora de 10.000 tokens cada, analisando o seu desempenho em cada execução.

Token		*Erros*			*Erros*		**Melhora**
	Tag correcto	*Antes*	*Depois*	*Tag correcto*	*Antes*	*Depois*	
a	SPS00	87	39	DA0FS0	46	27	50,38%
o	PD0MS000	29	10	DA0MS0	24	19	45,28%
que	CS	87	59	PR0CN000	39	33	26,98%

Tabela 2.5: Resultado das regras aplicadas.

No caso de *a*, a regra mostrada acima foi melhorada, ampliando o seu contexto de aplicação: *a* é agora anotado como preposição antes de nomes e adjectivos plurais e masculinos ou de expressões numerais seguidas de nomes e adjectivos masculinos, entre outros contextos.

De modo similar, as regras de *o* e *que* também incluem contextos complexos, embora não tenham sido adicionadas excepções que possam evitar corrigir falsos negativos, tais como algumas estruturas fixas.

As regras de correcção do segundo caso (*o*) lidam com estruturas diferentes de <DET NOUN>, modificando a etiqueta de determinante para pronome (e.g., antes de um pronome relativo ou da preposição *de*). A regra contrária (de pronome relativo a determinante) é aplicada antes de contextos interrogativos específicos.

Finalmente, o caso de *que* foi mais problemático, já que a análise de erros requer um processamento mais profundo. As regras unicamente lidam com algumas expressões comuns, tais como *uma vez que*, *para que*, etc. À parte destas, foram incluídas regras que modificam a etiqueta de pronome para conjunção em contextos comparativos (*melhor/pior do que...*) e em frases completivas depois da preposição *de*.

A Tabela 2.5 mostra os resultados da aplicação do conjunto de regras avaliado, que indicam que a implementação de regras simples melhora a anotação dos tokens alvo entre $\approx 27\%$ e $\approx 50\%$.

Apesar de que os testes são uma avaliação preliminar de um método de correcção *PoS-tagging*, os resultados indicam que alguns dos erros mais frequentes produzidos por etiquetadores estatísticos podem ser corrigidos por um conjunto pequeno de regras com base linguística.

Depois da aplicação do *parser* de correcção, a precisão do etiquetador morfossintáctico viu-se incrementada em $\approx 1,1\%$ no mesmo corpus de 50.000 tokens.

2.8. Conclusões

Este capítulo apresentou o desenvolvimento e adaptação de diversos módulos de processamento da linguagem natural para português e galego. Os módulos formam parte de uma etapa de processamento inicial, necessária para análises posteriores tais como as diferentes estratégias para a extracção de relações apresentadas em capítulos posteriores desta tese.

Assim, as principais contribuições deste capítulo são as seguintes:

– Módulos de segmentação de orações para português e galego

– Módulos de tokenização para português e galego

– Módulos de análise morfológica com lematização para português e galego

– Módulos de anotação morfossintáctica para português e galego

Para além disso, a adaptação e implementação dessas ferramentas implicou também outras contribuições, como a adaptação para o *standard* EAGLES de corpora e léxicos de português e galego (bem como para outros *tagsets* mais simples).

Foi também criado um corpus de teste com anotação morfossintáctica para galego e um corpus de português que contém textos de várias tipologias textuais, variedades nacionais e normas ortográficas desta língua.

A este respeito, foram avaliadas diferentes combinações de léxicos e corpora para a etiquetação morfossintáctica da Web em português.

Finalmente, apresentou-se uma estratégia com base linguística, de desenvolvimento simples e de aplicação rápida, que permite corrigir alguns dos erros mais frequentes dos etiquetadores morfossintácticos estatísticos.

Em suma, o trabalho descrito neste capítulo possibilita a realização de várias etapas iniciais de processamento da linguagem natural, que permitirão que módulos posteriores de análise semântica sejam aplicados.

CAPÍTULO 3

RECONHECIMENTO DE ENTIDADES MENCIONADAS

3.1. Introdução

Diversas tarefas do PLN, tais como a extracção de relações ou os sistemas de resposta a perguntas precisam da execução prévia de ferramentas que sejam capazes de reconhecer, em texto, entidades como pessoas ou localizações, por exemplo. O processo de identificação e de classificação das entidades é conhecido como Reconhecimento de Entidades Mencionadas (REM —ou NER, do inglês *Named Entity Recognition*), e faz parte das tarefas de extracção de informação. Embora alguns sistemas realizem o reconhecimento em um único processo, o REM pode ser dividido em duas subtarefas: a identificação e a classificação das próprias entidades.

A primeira das tarefas referidas (identificação) consiste na detecção automática de Entidades Mencionadas (EM) em texto livre:

"**José_Souto** foi ver o **Celta_de_Vigo** a **Balaídos**"

A segunda (classificação) tem como objectivo atribuir às entidades identificadas uma classe semântica previamente definida (pessoa, organização, data, quantidade, etc.). Assim, o resultado da aplicação de um classificador de EM no exemplo anterior poderia ser o seguinte:

"**José_Souto**$_{\text{PESSOA}}$ foi ver o **Celta_de_Vigo**$_{\text{ORGANIZAÇÃO}}$ a **Balaídos**$_{\text{LOCALIZAÇÃO}}$"

Na presente tese, esta classificação pode ser utilizada pelos sistemas de ER com o fim de seleccionar os argumentos que pertençam a uma determinada classe. Assim, se extrairmos exemplos de uma relação como `éPresidenteDe`, a utilização de sistemas REM permite-nos escolher só aqueles casos em que o primeiro argumento seja uma entidade da classe "pessoa".

Diversas ferramentas REM foram avaliadas em conferências como as já referidas Conference on Computational Language Learning (CoNLL) ou as Automatic Content Extraction (ACE). Para a língua portuguesa, o desenvolvimento de sistemas de reconhecimento de entidades foi promovido por duas edições da conferência HAREM.[1] Para espanhol existem também diversas ferramentas de REM, entre as quais se destaca Carreras *et al.* (2002), avaliado como o melhor sistema da conferência CoNLL 2002, e cuja implementação tem sido portada para FreeLing (Atserias *et al.*, 2006). Em relação ao galego, até à realização deste trabalho não conhecíamos nenhuma ferramenta dedicada ao reconhecimento de entidades mencionadas nesta língua.

Tendo isto em conta, e com o propósito de manter FreeLing como sistema base de processamento para este trabalho, o presente capítulo apresenta (i) a adaptação de dous sistemas de identificação EM em português e galego, (ii) a adaptação de um classificador de EM estatístico para português e (iii) a implementação de um classificador de EM, com base em recursos e em regras, para português e galego. Adicionalmente, descreve-se a adaptação e implementação de módulos de reconhecimento de expressões numéricas, quantidades, datas e horas para português e galego.

Os sistemas de identificação de EM (um deles com máquinas de estados finitos e o outro de aprendizagem automática) e o classificador estatístico são diferentes módulos de FreeLing, enquanto o classificador com base em recursos e regras é independente, mas utiliza a saída dos sistemas de identificação do próprio FreeLing. A avaliação das diferentes ferramentas implementadas resultou em desempenhos similares (e nalguns casos, superiores) aos mesmos sistemas para outras línguas, bem como aos de outras ferramentas com objectivos semelhantes.

A seguinte secção (3.2) faz uma revisão do trabalho relacionado. Depois, na Secção 3.3 mostram-se e avaliam-se os diferentes módulos de identificação e classificação de nomes próprios. A seguir, a Secção 3.4 apresenta os módulos adicionais de reconhecimento de entidades

[1] `http://www.linguateca.pt/harem/`

de base numérica (datas e horas, quantidades, etc.), bem como a sua avaliação. Finalmente, as conclusões são expostas na Secção 3.5.

Este capítulo baseia-se na publicação Garcia *et al.* (2012), incorporando também dados de Gamallo e Garcia (2011) (devidamente referidos), bem como alguns resultados não publicados.

3.2. Trabalho relacionado

Esta secção apresenta brevemente aqueles trabalhos e avaliações conjuntas dedicadas ao inglês (por ser a língua para a qual mais recursos e desenvolvimento existem), bem como ao português e espanhol, dado que, como foi dito, desconhecemos outros trabalhos de reconhecimento de entidades mencionadas para galego.

As conferências MUC-6 e MUC-7, realizadas em 1995 e 1998 respectivamente e focadas na análise do inglês, foram as primeiras avaliações de sistemas REM. Nas MUC definiram-se três grandes classes de entidades: "timex" (datas e horas), "numex" (expressões numéricas) e "enamex" (que continha nomes próprios referidos a organizações, pessoas e localizações). Os melhores resultados da MUC-7 obtiveram valores da medida F de 93,39% no total da classificação (Mikheev *et al.*, 1998). Outros encontros como os já referidos ACE (e também a própria MUC-7) realizaram diferentes avaliações tendo em conta também outro tipo de tarefas de extracção.

As *shared task* das conferências CoNLL 2002 e 2003 incluíram avaliações de sistemas de classificação independentes da língua (espanhol e holandês em 2002 e inglês e alemão em 2003), para entidades "enamex" (categoria à qual foi adicionada a classe "miscelânea" para classificar entidades diferentes de organizações, pessoas e localizações). Nestas avaliações, os melhores sistemas obtiveram valores da medida F de 72% (alemão), 88% (inglês), 77% (holandês) e 81% (espanhol). Como foi dito, este último sistema é a base dos módulos de classificação de entidades mencionadas de FreeLing.

Para a língua portuguesa realizaram-se duas avaliações conjuntas de reconhecimento de entidades mencionadas —HAREM (Santos e Cardoso, 2007) e Segundo HAREM (Mota e Santos, 2008)—, com resultados que variaram desde valores da medida F de 60% até 85% em função do tipo de avaliação, mais ou menos rígida. Estes resultados, porém, não são directamente comparáveis com outros sistemas e avaliações, já que as directrizes de classificação diferem notoriamente das de outras conferências.

Detendo-nos nas características dos próprios sistemas de reconhecimento, pode afirmar-se que a tendência dominante de desenvolvimento destes recursos é a combinação de regras e de máquinas de estados finitos para a identificação de expressões "timex/numex", e de modelos estatísticos —de aprendizagem automática— para o tratamento de entidades "enamex".

Contudo, existem métodos baseados em regras (Bick, 2006) para a classificação de entidades "enamex" e modelos híbridos (Ferreira *et al.*, 2007), ambos desenhados para a língua portuguesa.

Os sistemas probabilísticos são habitualmente treinados de modo supervisionado, pelo que precisam de corpora etiquetados manualmente (por exemplo Finkel *et al.* (2005) para o inglês, Carreras *et al.* (2002) para o espanhol ou Ferrández *et al.* (2007) para o português). Estas ferramentas, por sua vez, utilizam diferentes algoritmos (ou combinações deles) como Conditional Random Fields, AdaBoost, máquinas de vectores de suporte (referidos como SVM, do inglês *Support Vector Machines*) ou modelos ocultos de Markov (HMM), entre outros.

A dificuldade de obtenção de recursos de qualidade para o treino dos diferentes modelos inspirou várias abordagens de classificação não-supervisionada (ou semisupervisionada). Assim, o aumento de fontes semiestruturadas de fácil acesso (como Freebase[2] ou DBpedia[3]) permite a obtenção de recursos e de corpora potencialmente aplicáveis no treino destes modelos. Neste sentido, alguns trabalhos recentes propõem estratégias que tiram proveito de fontes como a Wikipedia para melhorar os sistemas de classificação e extracção (Mika *et al.*, 2008). De modo similar, Nothman *et al.* (2008) utiliza as ligações internas da Wikipedia para anotar automaticamente entidades em texto não estruturado, empregado posteriormente no treino de modelos estatísticos.

Por último, em Gamallo e Garcia (2011) é apresentado um classificador semântico de nomes próprios para o português que utiliza um conjunto de regras e grandes listas de entidades obtidas de modo (semi)automático. Este é um dos dous classificadores de nomes próprios (entidades "enamex") para português descritos neste trabalho, assim como o sistema base do classificador para galego.

[2] `http://www.freebase.com`
[3] `http://www.dbpedia.org`

3.3. Reconhecimento de nomes próprios

A análise dos nomes próprios apresentada neste capítulo divide-se em duas tarefas: (i) a identificação e (ii) a classificação semântica. A identificação consiste na detecção correcta das fronteiras de um nome próprio ("Museo_do_Pobo_Galego" —identificado como um único nome próprio— *versus* *"Museo do Pobo_Galego", onde erroneamente se detectam dous nomes próprios). A classificação consiste na atribuição à entidade de uma etiqueta que denote uma classe semântica previamente definida.

O primeiro dos processos foi realizado através de dous módulos de FreeLing, com o fim de avaliar o desempenho de cada um deles de modo independente. Em relação à classificação, tanto o módulo AdaBoost de FreeLing como o já referido sistema baseado em regras e recursos foram utilizados em português, sendo este último também implementado para galego.

3.3.1. Identificação

Os dous módulos utilizados para a identificação de nomes próprios foram *basic* e BIO. O primeiro (*basic*) consiste numa máquina de estados finitos que detecta sequências de palavras que começam por maiúsculas, e numa lista de palavras funcionais (*de*, *por*, etc.) que podem ocupar uma posição intermédia em nomes próprios compostos.[4] Esta estratégia identifica expressões como "John Lennon" ou "Universidade de Vigo", e em combinação com um desambiguador morfossintáctico (veja-se o Capítulo 2), detecta com alta precisão se um token em posição inicial de oração é ou não é um nome próprio ("Café com leite" *versus* "Café_Starbucks em Barcelona").

Este método não precisa de um corpus anotado de aprendizagem, sendo tanto a adaptação para português e galego como a execução rápidas. Contudo, existem casos em que o identificador falha sistematicamente, uma vez que não são atribuídos valores de probabilidade para cada um dos elementos que podem formar o nome próprio: assim, tanto a expressão "Ministério_de_Educação" como *"[nessa altura chegou] Sarkozy_de_Roma" são analisadas como um único nome próprio.

[4] As máquinas de estados finitos (ou *finite-state-automata*, FST) são modelos matemáticos compostos de um conjunto finito de estados abstractos. Em cada momento, a máquina encontra-se em um único estado, e muda para outro se se dar uma condição previamente definida (Black, 2013). Na identificação de nomes próprios compostos, a máquina verifica se cada um dos tokens começa ou não por maiúscula, se entre eles se encontram palavras funcionais, etc.

O segundo dos módulos de identificação de nomes próprios de FreeLing tenta corrigir estes e outro tipo de erros implementando o método estatístico BIO (também conhecido como IOB). Esta estratégia de aprendizagem supervisionada precisa de um corpus de treino anotado cujos nomes próprios sejam divididos em B (*begin*, início) e I (*inside*, dentro), para além dos tokens que não formam parte de um nome próprio (O: *outside*, fora). O corpus de aprendizagem, bem como um conjunto de atributos lexicais (que incluem listas de nomes próprios frequentes, palavras funcionais, etc.), permitem treinar um classificador que detecte as fronteiras das entidades "enamex", em função das probabilidades de cada token ser B, I ou O.

Foram treinados cinco modelos diferentes, em função da frequência dos atributos no corpus de treino, do modelo BIO com o algoritmo AdaBoost (Carreras *et al.*, 2002) (os cinco modelos utilizaram diferentes grupos de atributos do corpus: desde aqueles com frequência superior a 1% até todos os extraídos). Tanto para português como para galego, foram utilizados os mesmos corpora que para treinar os módulos *PoS-tagging* (Secção 2.3): uns 138.000 tokens e 7.300 nomes próprios em português e $\approx$ 240.000 tokens e $\approx$ 11.800 entidades "enamex" em galego.

3.3.2. Classificação

A classificação de nomes próprios é o processo que consiste na atribuição, depois de identificadas as fronteiras de um nome próprio, de uma etiqueta semântica previamente estabelecida. Apesar de existirem tarefas que requerem uma classificação mais detalhada, as etiquetas utilizadas na ferramenta aqui apresentada são as "enamex", amplamente empregadas no reconhecimento de entidades desde a avaliação MUC-6. Estas etiquetas diferenciam três classes principais: PER (pessoa), ORG (organização) e LOC (localização), às quais desde a conferência CoNLL 2002 se acrescentou MISC (outra, ou miscelânea) para classificar as entidades que não pertencem a nenhum dos tipos anteriores.

A classificação de determinados nomes próprios de acordo com as etiquetas estabelecidas provocou algumas diferenças tanto nas várias edições das avaliações referidas como noutras (Segundo HAREM, por exemplo). Dous dos principais problemas que surgem na classificação de nomes próprios são a polissemia e a metonímia. Neste sentido, determinados nomes de países, cidades, etc. podem ser classificados (para além de LOC), como ORG ("**Bélgica** assinou o Tratado de Roma"), como PER ("**Vigo** opõe-se à destruição do sector naval"), etc., em função dos critérios de etiquetação utilizados.

Neste trabalho, tanto na implementação do sistema de classificação, como na anotação manual dos corpora de treino e teste, só foi considerada a homonímia, ignorando-se portanto as interpretações metonímicas das entidades mencionadas (que podem ser identificadas em processos posteriores de análise).[5]

Como foi dito na Secção 3.2, existem diversas estratégias para o desenvolvimento de sistemas de classificação de EM: métodos que utilizam regras e listas externas de entidades, modelos estatísticos que precisam de corpora de treino anotados e outras estratégias de aprendizagem automática menos supervisionadas.

Nesta secção são apresentados dous modelos diferentes de classificação: primeiro, o classificador AdaBoost incluído em FreeLing (Carreras *et al.*, 2002), treinado para português com uma versão do corpus Bosque 8.0 (Secção 2.3) com anotação semântica adicionada manualmente.[6] A seguir, a estratégia baseada em regras e recursos, implementada para português em Gamallo e Garcia (2011) e adaptada para galego no presente trabalho.

O sistema estatístico utiliza classificadores AdaBoost multi-classe (com uma classe para cada uma dos quatro tipos de "enamex"). A janela utilizada é de -3/+3 tokens, pelo que o classificador analisa os atributos dos três elementos anteriores e posteriores à entidade com o fim de decidir a sua classe. Os atributos utilizados são lexicais (tokens e lemas), morfossintácticos (categoria morfossintáctica e constituinte sintáctico), morfológicos (prefixos e sufixos), e *gazetteers* e *trigger words*.

Os *gazetteers* são listas de entidades conhecidas, já classificadas em função do seu tipo (e.g., "Jerry Garcia": PER, "Santiago de Compostela": LOC). As *trigger words* são palavras que co-ocorrem num segmento do texto com uma entidade, sugerindo a sua classificação numa determinada classe (por exemplo "empresa: ORG", "amigo": PER, etc.).

O sistema de regras e recursos também necessita, para além de um conjunto de heurísticas de classificação semântica, de *trigger words* e de listas de *gazetteers* de grande tamanho. Com o fim de se obterem estes dous tipos de recursos de modo (semi)automático, foi utilizada a seguinte estratégia:

Para extrair as *trigger words*, procura-se na árvore de categorias da Wikipedia (da língua alvo) um conjunto de categorias que sejam subclasses de pessoas, organizações e localizações. Este processo realiza-se seleccionando categorias que contenham as palavras "pessoa", "organização" ou "lugar" (e sinónimos) como núcleo da categoria (p. ex., "Organizações

[5] Veja-se Gamallo e Garcia (2011) para uma discussão mais pormenorizada.

[6] Uma vez que não existiam corpora disponíveis com tamanho suficiente para treinar modelos supervisionados para galego, o módulo de classificação de EM de FreeLing só foi adaptado para português.

internacionais"). Depois, escolhem-se os hipónimos de cada categoria e seleccionam-se os núcleos ("Associação Europeia de Livre Comércio" → "associação"; "Comissão Sismológica Europeia" → "comissão", etc.). As listas obtidas neste processo são depois utilizadas como *trigger words* e como palavras semente para a obtenção dos *gazetteers*.

Os *gazetteers*, também obtidos a partir da Wikipedia, são extraídos da seguinte forma: primeiro verifica-se, para cada artigo da Wikipedia, se as suas categorias contêm como núcleo alguma das palavras semente obtidas no processo anterior (p. ex. "sindicato"). Se assim for, o título do artigo (que é uma entidade mencionada) é extraído e classificado em função da classe da palavra. Assim, o título "Associação Internacional de Trabalhadores" é incluído na lista de organizações porque se inclui na categoria "Sindicatos anarquistas", e "sindicato" foi previamente classificada como *trigger word* para a classe ORG. Para além desta estratégia, também são extraídas entidades que contenham uma *trigger word* conhecida nos campos "tipo" e "ocupação" das *infoboxes* da Wikipedia.

Uma vez obtidos os recursos necessários para a aplicação do classificador, este utiliza o seguinte algoritmo de etiquetação:

- Se a entidade aparece só numa das listas de *gazetteers*, é classificada de acordo com a classe da lista.

- Se for uma forma homónima (aparece em várias listas) ou desconhecida, procura-se nos três tokens anteriores e posteriores à entidade para verificar se correspondem a alguma *trigger word* conhecida (p. ex., "o meu *amigo* Anselmo"). Se houver várias *trigger words*, é preferida a mais próxima (ou a que se encontrar antes da entidade). O algoritmo contém um conjunto de regras que evitam a classificação, entre outros fenómenos, de complementos preposicionais (em função das preposições que possam existir entre uma *trigger word* e uma entidade). Assim, a entidade "Banco de Portugal" no contexto "fundador do Banco de Portugal" não será classificada como pessoa apesar de ter a *trigger word* (da classe PER) *fundador*.

- Se a entidade não é conhecida nem tem *trigger words* próximas, analisa-se a sua forma para verificar se contém *trigger words* internas ("*Museo* do Pobo Galego") ou se é acrónimo.

- Finalmente, se as regras anteriores não conseguem atribuir uma etiqueta, um último conjunto de heurísticas decide entre a classificação como MISC ou como ORG, em

função tanto da forma do nome próprio como do contexto morfossintáctico mais próximo.

Uma vez que a estratégia é fortemente dependente de recursos externos (e que estes não são muito abundantes, nomeadamente em galego), na secção de avaliação são realizados testes com diferentes listas de *gazetteers*.

3.3.3. Testes e avaliação

Nesta secção são apresentados os diversos testes realizados para avaliar tanto os identificadores como os classificadores semânticos de entidades "enamex".

Identificação

O primeiro conjunto de testes analisou o desempenho dos dous identificadores de nomes próprios (*basic* e BIO) em português e galego. Para avaliar cada um dos modelos das duas línguas, foram utilizados duas selecções aleatórias de 40.000 tokens (uma para cada idioma) dos corpora referidos (Secção 3.3.1). Os conjuntos restantes ($\approx$ 200.000 tokens em galego e $\approx$ 100.000 tokens em português) empregaram-se para treinar os modelos estatísticos (BIO).

A avaliação foi realizada de acordo com as directrizes da CoNLL, tendo em conta a anotação tanto dos tokens B como dos I. A Tabela 3.1 mostra os resultados da avaliação do módulo *basic*, bem como o melhor dos modelos estatísticos treinados. A precisão é a percentagem de entidades correctamente classificadas pelo sistema, e o *recall* a percentagem de entidades do corpus que foram correctamente classificadas. Finalmente, a medida F é a média harmónica entre a precisão e o *recall*.

Os melhores resultados para português foram obtidos seleccionando aqueles atributos com mais de três ocorrências no corpus de treino, enquanto em galego o melhor modelo BIO utilizou todos os atributos obtidos no processo de aprendizagem (contudo, as diferenças máximas entre os diferentes modelos BIO foram de 0,5% na medida F).

Os resultados desta avaliação mostram que o método probabilístico obtém entre $4,85$ (português) e $4,75$ (galego) pontos percentuais (na medida F) mais do que os modelos baseados em máquinas de estados finitos. Os modelos BIO têm melhores resultados de *recall* do que os *basic*, mas é na precisão onde se destaca a superioridade deste método, obtendo diferenças de mais de 7% em cada língua. Contudo, note-se que os resultados de *basic*, que não pre-

Modelo	**Português**			**Galego**		
	Prec	*Rec*	*F1*	*Prec*	*Rec*	*F1*
basic	84,05	88,58	86,26	87,72	92,31	89,96
BIO	92,54	89,52	**91,01**	94,86	94,77	**94,81**

Tabela 3.1: Precisão, *recall* e medida F dos identificadores FST e estatísticos de entidades "enamex" em galego e português.

cisa de corpus de treino e que tem uma execução mais rápida, superam 86% de medida F em português e estão próximos de 90% em galego.

Classificação

Os seguintes testes foram dedicados a conhecer o funcionamento dos sistemas de classificação semântica dos nomes próprios. Para português foram comparados o modelo estatístico com o método de regras e recursos, sendo este último também avaliado em galego.

Português

Os diferentes modelos para português foram avaliados utilizando várias listas de *gazetteers*, obtidas seguindo o método exposto na Secção 3.3.2. A Tabela 3.2 contém o número de *trigger words* e dos três conjuntos de *gazetteers* utilizados: "es" são as listas de entidades do classificador para espanhol de FreeLing. As listas "infobox" foram extraídas das *infoboxes* da Wikipedia, enquanto "categ" contém entidades obtidas da árvore de categorias da Wikipedia.

O modelo estatístico foi treinado com um conjunto (seleccionado de modo aleatório) de 87.000 tokens do corpus Bosque. Para a avaliação (tanto do classificador estatístico como do baseado em regras e recursos), foram utilizados cinco corpora de diferentes domínios, anotados manualmente seguindo o critério de desambiguação de homónimos (excepto o corpus *harem*, etiquetado para a conferência HAREM tendo em conta vários tipos de metonímia e um maior número de categorias):

– *bosque* (A): os restantes 53.000 tokens do corpus Bosque (género jornalístico, domínio geral).

– *wiki* (B): 30.000 tokens de artigos da Wikipedia em português (género enciclopédico, domínio geral).

Listas	LOC	ORG	PER
trigger words	82	57	320
gazetteers es	7.312	2.263	2.598
gazetteers es+infobox	23.732	4.586	17.600
gazetteers es+infobox+categ	58.305	13.599	64.735

Tabela 3.2: Número de *trigger words* e dos conjuntos de *gazetteers* para as três classes de entidades "enamex" diferentes de MISC nos testes em português.

– *europarl* (C): 30.000 tokens da versão portuguesa do corpus Europarl[7] (género formal, domínio político).

– *brasil* (D): 24.000 tokens do corpus brasileiro do European Corpus Initiative Multilingual Corpus I (ECI/MCI)[8] (género técnico, domínio político, variedade brasileira).

– *harem* (E): 70.000 tokens do corpus da conferência HAREM (Mota e Santos, 2008) (género e domínio gerais). A anotação foi adaptada às quatro categorias "enamex" utilizadas no presente trabalho.

A Tabela 3.3 contém os resultados dos dous métodos de classificação nos cinco corpora referidos, bem como os valores médios (*macro-average*, veja-se a Secção 4.3). Cada modelo foi executado com três conjuntos diferentes de *gazetteers* (*es*, *infobox* e *categ*), e sem nenhum (*nulo*). Os resultados foram obtidos usando o sistema de avaliação da conferência CoNLL, unicamente avaliando as entidades correctamente identificadas pelo identificador de entidades BIO. Em termos de eficiência, o modelo de regras foi aproximadamente um 40% mais rápido do que o sistema estatístico.

Se se compararem as duas estratégias de classificação, os resultados obtidos não indicam que uma delas seja claramente melhor do que a outra. Por um lado, os modelos supervisionados funcionam melhor em três dos cinco corpora utilizados (*bosque* (A), *europarl* (C) e *harem* (E)), sendo os resultados no corpus *bosque* esperáveis devido a que contém o mesmo tipo de documentos do que o corpus de treino. Por outro lado, os sistemas de regras tiveram melhores resultados tanto no corpus *wiki* (B) como no *brasil* (D), utilizando todos os *gazetteers*. O bom desempenho deste sistema no corpus *wiki* (91,8%) pode dever-se ao facto de os recursos serem extraídos principalmente da própria Wikipedia. Em média, o sistema de regras teve valores

[7] `http://www.statmt.org/europarl`
[8] `http://www.elsnet.org/eci.html`

Cp	Recursos											
	nulo			**es**			**es+infobox**			**es+infobox+categ**		
	Prec	*Rec*	*F1*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
A	32,7	32,4	32,6	54,5	54,3	54,4	69,5	69,3	69,4	74,5	74,2	74,3
	74,3	74,1	74,2	75,2	75,1	75,2	77,3	77,0	77,2	77,8	77,5	**77,7**
B	32,2	32,1	32,1	47,9	47,9	47,9	75,0	75,0	75,0	91,8	91,8	**91,8**
	79,5	79,5	79,5	79,7	79,7	79,7	83,0	83,0	83,0	87,6	87,6	87,6
C	61,4	61,5	61,4	72,7	72,9	72,8	74,4	74,5	74,5	75,2	75,4	**75,3**
	50,0	50,1	50,1	52,5	52,6	52,5	62,0	62,1	62,1	63,2	63,2	63,2
D	45,8	45,8	45,8	47,9	47,9	47,9	48,2	48,2	48,2	73,7	73,7	73,7
	76,8	76,8	76,8	78,4	78,4	**78,4**	75,8	75,8	75,8	75,8	75,8	75,8
E	28,9	23,7	26,1	43,7	35,8	39,4	59,1	48,4	53,2	64,7	53,0	58,3
	62,3	51,2	56,2	61,7	50,6	55,6	65,7	53,9	59,2	66,7	54,8	**60,2**
M	40,2	39,1	39,6	53,4	51,8	52,5	65,2	63,1	64,0	76,0	73,6	**74,7**
	68,6	66,4	67,4	69,5	67,3	68,3	72,7	70,4	71,4	74,2	71,8	72,9

Tabela 3.3: Resultados (*Prec*isão, *Rec*all e medida F, *F1*) dos classificadores de EM nos cinco corpora (Cp) em português apresentados, e valores médios (M), em quatro execuções: sem recursos externos (nulo) e com os três conjuntos de *gazetteers* referidos. A linha superior de cada corpus contém os resultados do classificador de regras e recursos e a linha inferior do sistema estatístico.

Classe	**Regras**			**Estatístico**		
	Prec	*Rec*	*F1*	*Prec*	*Rec*	*F1*
LOC	76,36	54,19	63,40	84,76	57,42	68,46
PER	86,31	88,78	87,53	88,10	93,67	90,80
ORG	79,11	68,20	73,25	81,25	69,73	75,05
MISC	32,56	51,38	39,86	35,19	52,29	42,07

Tabela 3.4: Resultados de classificação dos quatro tipos de entidades "enamex", obtidos pelo sistema de regras e recursos e pelo modelo estatístico no corpus em português *bosque* (A), com todos os *gazetteers*.

de medida F ligeiramente superiores aos do modelo supervisionado (74,7% *versus* 72,9%). Em relação aos baixos resultados dos classificadores no corpus *harem*, estes podem dever-se a que este corpus foi anotado seguindo directrizes diferentes às dos outros corpora (tanto de treino como de teste), bem como às regras de desambiguação do sistema de regras e recursos.

Uma vez que o sistema não probabilístico é fortemente dependente dos *gazetteers*, os resultados sem a utilização destes recursos são baixos, mas melhoram à medida que o tamanho das listas aumenta (de $39,6\%$ a $74,7\%$). Porém, o sistema supervisionado utiliza, para além dos *gazetteers*, diferentes atributos extraídos no processo de treino, sendo as listas um pequeno factor no modelo de decisão. Assim, o seu desempenho tem uma variação menor em relação ao tamanho dos *gazetteers* (de $67,4\%$ a $72,9\%$).

Classe	*Precisão*	*Recall*	*Medida F*
LOC	80,87	77,82	79,32
PER	77,49	88,62	82,68
ORG	83,71	81,07	82,37
MISC	81,48	64,23	71,84
Média	80,53	80,53	80,53

Tabela 3.5: Resultados do classificador estatístico treinado com todos os corpora de português (excepto o *harem*), avaliado sobre um subconjunto do corpus *bosque*. Modelo disponibilizado em FreeLing.

Por último, a Tabela 3.4 mostra os resultados individuais de classificação de cada classe de entidade nos mesmos sistemas (no corpus *bosque*, e utilizando todos os *gazetteers*). Os valores mais altos obtiveram-se nas entidades PER, enquanto que a classe MISC teve os resultados mais baixos. Isto pode dever-se, por um lado, a que a classe MISC é muito heterogénea. Por outro lado, a classificação desta categoria foi realizada sem listas de *gazetteers*.

Depois de se realizarem estes testes, treinou-se um classificador para português com maior quantidade de corpus de treino, com o fim de ser disponibilizado na distribuição de FreeLing. Para isso, utilizaram-se uns 120.000 tokens do corpus Bosque completo, o corpus *wiki*, o *europarl* e o *brasil*.[9]

Para avaliar este modelo, utilizou-se a selecção restante do corpus bosque ($\approx$ 20.000 tokens). Os resultados podem ver-se na Tabela 3.5. Embora a avaliação deste modelo não seja tão pormenorizada, o aproveitamento dos diferentes recursos permitiu treinar um classificador para português com melhor desempenho.

Galego

Para avaliar o sistema de classificação de regras e recursos em galego, foi anotada manualmente uma selecção aleatória de uns 20.000 tokens (com 924 nomes próprios) do corpus utilizado para a avaliação do sistema de identificação de entidades "enamex". Tendo em conta que o sistema é dependente de listas externas, e que o tamanho da Wikipedia em galego é notoriamente menor do que o de outras línguas (além de que muitos dos nomes próprios podem ser independentes da língua), foram realizados quatro testes diferentes:

Para o primeiro teste utilizaram-se unicamente os *gazetteers* extraídos da versão galega (gl) da Wikipedia; no segundo e no terceiro juntaram-se a estes os extraídos da versão por-

[9]Não foi utilizado o corpus *harem*, dado que devido às diferenças nos critérios de etiquetação, os resultados usando este recurso foram inferiores.

Listas	LOC	ORG	PER
trigger words	74	47	405
gazetteers gl	4.395	717	9.650
gazetteers pt	33.485	16.378	59.424
gazetteers es	63.468	21.551	88.342

Tabela 3.6: Número de *trigger words* e dos conjuntos de *gazetteers* para as três classes de entidades "enamex" diferentes de MISC nos testes em galego.

Gazetteers	*Precisão*	Recall	*Medida F*
gl	62,15	62,84	62,49
gl+pt	74,11	74,94	**74,52**
gl+es	59,80	60,47	60,14
gl+pt+es	68,31	69,08	68,69

Tabela 3.7: Resultados do classificador de regras e recursos em galego, em função dos conjuntos de *gazetteers* utilizados.

tuguesa (pt) e da espanhola (es), respectivamente; finalmente, levou-se a cabo uma avaliação com as listas de entidades de três versões da Wikipedia (galega, portuguesa e espanhola).[10] Em todas as avaliações foi utilizado o mesmo conjunto de *trigger words*, obtidas da Wikipedia em galego.

A Tabela 3.6 contém o número de entidades de cada uma das listas de *gazetteers*, bem como o número de *trigger words* utilizadas. Na Tabela 3.7 podemos ver os resultados dos diferentes testes de classificação de nomes próprios. Estes testes foram realizados utilizando o modelo BIO para a identificação das entidades. Avaliações preliminares em que foi usado o método *basic* tiveram resultados com valores da medida F $\approx 3\%$ mais baixos.

O primeiro conjunto de resultados mostra que a utilização de *gazetteers* extraídos unicamente da Wikipedia em galego não é suficiente para conseguir um bom desempenho de um sistema baseado em regras e recursos, obtendo valores da medida F de 62,49%. A utilização de listas de entidades em espanhol e português melhora, portanto, a qualidade do sistema (68,69%). Note-se, contudo, que o aumento da medida F é superior com as listas gl+pt (menor do que o conjunto gl+pt+es), pelo que se infere que os *gazetteers* extraídos da Wikipedia em espanhol podem ter algum tipo de ruído (o que explicaria também os valores do teste gl+es).

[10] O número de *gazetteers* apresenta pequenas variações em relação aos utilizados nos testes anteriores (para português), devido a que os recursos para galego foram extraídos de uma versão mais recente da Wikipedia.

Classe	*Número*	*Precisão*	Recall	*Medida F*
LOC	280	81,89	83,20	82,54
PER	121	61,31	77,06	68,29
ORG	438	73,79	80,00	76,77
MISC	85	62,50	7,94	14,08

Tabela 3.8: Resultados do classificador de regras e recursos para cada classe "enamex" em galego (e número de entidades de cada classe).

Assim, os resultados do sistema apresentado com os *gazetteers* do galego e do português ultrapassam 74% de medida F.

A Tabela 3.8 mostra os resultados individuais de cada classe "enamex". A análise destes dados indica-nos que os tipos LOC e ORG (e, em menor medida, PER) têm bons resultados tanto em termos de precisão como de *recall*. Contudo, o desempenho do classificador em cada classe difere dos resultados em português (onde PER obtinha os melhores valores). Estes dados podem dever-se à dependência dos recursos externos e à relação destes com o corpus de teste. Por último, os valores de *recall* da classe MISC são notoriamente mais baixos. Do mesmo modo que na avaliação em português, é preciso referir que tanto o tipo destas entidades quanto a sua contextualização são mais heterogéneas do que as restantes.

Uma vez que alguns dos erros de classificação do sistema foram provocados por erros anteriores na identificação dos nomes próprios, foi realizado um último teste assumindo uma entrada óptima no sistema de classificação. Esta última avaliação (utilizando os *gazetteers* gl+pt), que só analisa aquelas entidades correctamente reconhecidas pelo identificador de nomes próprios, teve um valor final da medida F de 80,44%.

É preciso ter em conta que em nenhum dos testes as listas de *gazetteers* tiveram algum tipo de revisão nem filtragem. Neste sentido, a adaptação das listas portuguesas para galego (e vice-versa) pode ser uma boa estratégia de melhoramento do sistema (Malvar *et al.*, 2010). Além disso, a aplicação de algum tipo de filtragem e/ou revisão sobre os *gazetteers*, assim como a utilização de listas com maior número de entidades (como as do inglês, por exemplo), podem contribuir para o aumento da precisão do classificador semântico aqui proposto.

Em termos gerais, os resultados obtidos pelos distintos sistemas apresentados nesta secção não são facilmente comparáveis com os de sistemas concebidos para outras línguas, devido tanto às características dos corpora de teste, como aos próprios objectivos de cada um dos módulos de reconhecimento.

Assim, em relação aos módulos de identificação de nomes próprios, os resultados de medida F foram similares aos obtidos para outras línguas (Carreras *et al.*, 2002).

Para além disso, a comparação entre os resultados do classificador semântico de nomes próprios é mais complexa: por um lado, os objectivos de vários classificadores costumam ser diferentes, em função do número de tipos e subtipos de entidades que pretendam classificar. Por outro lado, o tamanho e tipologia do corpus de teste é também muito variável, bem como a anotação de entidades potencialmente ambíguas. Tendo isto em conta, e observando que, por exemplo, os melhores sistemas das avaliações CoNLL (2002 e 2003) diferem em mais de 16 pontos percentuais (72,41% para o alemão e 88,76% para o inglês), os resultados obtidos com diferentes métricas e corpora não podem ser directamente comparáveis. O mesmo acontece se observarmos os resultados das avaliações do Segundo HAREM, cujos valores foram obtidos utilizando tipos e subtipos diferentes na classificação de entidades. Nesta avaliação, a métrica mais próxima da realizada no presente capítulo é o "Cenário Selectivo 2", que inclui as categorias "local" (com dous subtipos: "humano" e "físico"), "organização", "pessoa" e "tempo", na qual o sistema XIP-L2F/Xerox_3 obteve valores da medida F de 63,26%.

3.4. Reconhecimento de entidades de base numérica

Esta secção apresenta os módulos de reconhecimento de entidades "timex" e "numex", incluídos na *suite* FreeLing e adaptados para português e galego. Inclui, também, uma avaliação destes módulos em galego.

3.4.1. Numerais

O primeiro tipo de entidades de base numérica são as expressões numerais. O reconhecedor aplica-se depois do tokenizador, pelo que utiliza uma entrada já dividida em elementos individuais como palavras e sinais de pontuação.

Este módulo é composto por um conjunto de máquinas de estados finitos que detectam expressões numerais em vários formatos: numérico ("7,4", "325.275") e extenso ("trezentos vinte cinco mil duzentos e setenta e cinco", "um milhão e meio"), assim como outras formas lexicais como "dezenas", "milhares", "terços", etc. Além da identificação, o módulo normaliza as entidades, atribuindo um lema numérico a cada uma das expressões reconhecidas ("24", "vinte e quatro", "duas dúzias" $\rightarrow$ 24).

3.4.2. Datas

O módulo seguinte realiza o reconhecimento automático de datas e horas, precisando do reconhecedor de numerais para identificar algumas das expressões. O módulo é composto também por um conjunto de máquinas de estados finitos (específicas para português e galego), que identificam e normalizam datas e horas em formatos diferentes.

Este módulo reconhece formas como horas, dias da semana (e as suas partes: "meio-dia", "manhã", "tarde", "madrugada"), meses, séculos, anos, etc., que podem aparecer de modo individual ("Maio", "12h24") ou em diferentes combinações ("sete da manhã", "segunda-feira, vinte e sete de Julho de mil novecentos e oitenta", "Janeiro de 1968", etc.). As máquinas de estados finitos identificam também outro tipo de expressões comuns como "o passado mês de Julho" ou "as sete e um quarto da tarde" (adaptadas, em português, às ortografias anterior e posterior ao Acordo Ortográfico de 1990).

Uma vez identificadas as expressões que contêm uma data e/ou uma hora, o módulo realiza uma normalização, atribuindo-lhes uma etiqueta que segue os *standards* propostos pelo Grupo EAGLES (Leach e Wilson, 1996), com o formato: [DIA:DD/MM/AAAA:hh.mm:xm] (cujos campos se separam por ":" e que incluem (i) o nome do dia semana, (ii) o dia, mês e ano, (iii) as horas e minutos e (iv) a divisão entre am/pm, respectivamente).

3.4.3. Quantidades

O reconhecedor de quantidades depende do reconhecedor de expressões numerais e consiste num conjunto de máquinas de estados finitos ao qual se acrescenta um ficheiro externo com etiquetas e expressões regulares relativas a quantidades, unidades monetárias, longitudes, etc.

As expressões identificadas por este módulo são também variadas e em diferentes formatos: são reconhecidos rácios e percentagens ("dous terços", "3,5%", "nove por cento", etc.) assim como quantidades físicas ("sete quilómetros por hora", "1.500 toneladas", etc.) ou monetárias ("doze milhões de euros", "7.000 escudos", etc.).

O sistema reconhece actualmente uns 320 tipos de unidades diferentes (moedas, distâncias, velocidades, pesos, temperaturas, etc.) em perto de 900 contextos diferentes. Depois de identificadas, as entidades recebem uma etiqueta normalizada que atribui o tipo (peso, moeda, etc.) e o valor de cada uma delas.

3.4.4. Testes e avaliação

Para conhecer o desempenho das adaptações dos reconhecedores de entidades "timex" e "numex" foi realizado um pequeno conjunto de testes com um corpus anotado em galego.

Os testes têm como objectivo a realização de uma avaliação preliminar dos reconhecedores de expressões numerais, datas e quantidades sobre texto real. Com este fim, foram seleccionadas aleatoriamente 10 notícias do jornal em galego *Galicia Hoxe* (de todas as secções), criando um corpus de aproximadamente 10.000 tokens, com 270 entidades de base numérica etiquetadas manualmente.

Para conhecer o desempenho dos reconhecedores foram realizadas duas avaliações diferentes, em função do critério de anotação utilizado. A primeira (*Dura*), faz uma anotação estrita de cada uma das entidades, tendo em conta conhecimento externo, e não a forma das entidades. Assim, num título como "Aforro de millón e medio no gasto", a expressão "millón e medio" é anotada como "moeda", uma vez que do conteúdo da notícia (ou de conhecimento externo) se infere o seu significado. Do mesmo modo, numerais como "2009" ("De acordo co crecemento medio de 2009") ou "19.099" ("Os salarios máis baixos atópanse en Canarias (18.926 euros), en Estremadura (19.099)") são anotados como "data" e "moeda", respectivamente.

A segunda avaliação (*Branda*) tem em conta só aquele tipo de anotação que os módulos adaptados realizam, e que está directamente relacionado tanto com a forma da expressão, como com o contexto léxico-semântico mais próximo. Neste sentido, expressões isoladas como "2009" ou "19.099" são anotadas como "número" excepto se o seu contexto incluir evidências de pertencerem a outra classe de entidades ("ano 2009" ou "19.099€", por exemplo).

A Tabela 3.9 mostra os resultados das avaliações referidas, tendo em conta a etiquetação de cada tipo de entidades bem como o desempenho geral dos reconhecedores.

Para além dos números e datas (reconhecidas pelos módulos do mesmo nome), as percentagens, moedas e unidades foram analisadas pelo reconhecedor de quantidades. As principais diferenças entre as avaliações *Dura* e *Branda* têm a ver com a classificação de expressões numéricas em contextos ambíguos, às quais o sistema atribui a etiqueta número. Assim, entre as duas avaliações, a anotação de numerais passa de 63% a 93%, a de moedas de 63% a 100%, e a de datas de 73% a 95%. Tendo em conta as propriedades dos módulos adaptados, assim como a ambiguidade de expressões como as referidas nos parágrafos anteriores, a avaliação *Branda* dos reconhecedores de numerais, quantidades e datas mostra que o desempenho destes módulos se situa em redor de 94% ($\approx$ 70% na avaliação *Dura*). Contudo, estes resultados

Entidade	**Dura**				**Branda**			
	Núm.	*Prec*	*Recall*	*F1*	*Núm.*	*Prec*	*Recall*	*F1*
Números	111	68,75	59,46	63,77	160	97,24	89,81	93,38
Percentagens	16	93,75	93,75	93,75	16	93,75	93,75	93,75
Moedas	38	63,16	63,16	63,16	24	100	100	100
Unidades	24	83,33	83,33	83,33	22	95,24	90,91	93,02
Datas	81	96,00	59,26	73,29	48	95,83	95,83	95,83
Total	270	77,23	64,08	70,04	270	96,85	92,14	94,43

Tabela 3.9: Resultados dos módulos adaptados de reconhecimento de entidades "timex" e "numex" em galego (e número de entidades), em duas avaliações: *Dura* e *Branda*.

só podem entender-se como preliminares, uma vez que o corpus de teste não tem um tamanho suficiente para os considerar definitivos.

3.5. Conclusões

O presente capítulo descreveu a implementação e adaptação de diferentes módulos de reconhecimento de entidades mencionadas em português e em galego.

Primeiro, foram adaptados e avaliados dous sistemas de identificação de nomes próprios: baseados em (i) máquinas de estados finitos e em (ii) estratégias supervisionadas.

A seguir foi apresentado um método de classificação semântica de nomes próprios, que funciona através de um conjunto de regras e de recursos extraídos (semi)automaticamente. O desempenho deste método foi comparado com o do classificador probabilístico de FreeLing, que foi treinado e disponibilizado para português.

Finalmente, foram também adaptados diferentes módulos de reconhecimento de expressões numéricas e de quantidades, e foi criado um novo módulo de reconhecimento de datas e horas para português e galego.

Em relação à identificação dos nomes próprios, os resultados dos testes indicam que os sistemas estatísticos têm valores de medida F $\approx 5\%$ maiores do que os sistemas baseados em máquinas de estados finitos.

A respeito da classificação semântica, as diferentes avaliações não provaram que uma das duas estratégias seja melhor do que a outra na análise do português. Contudo, o sistema de regras e recursos proposto, bem como os módulos adaptados de reconhecimento de entidades de base numérica, obtêm resultados próximos dos valores obtidos pelos reconhecedores de

avaliações com métricas similares, tais como as *shared task* das conferências CoNLL (Tjong Kim Sang e de Meulder, 2003).

Assim, as principais contribuições do presente capítulo são as seguintes (para português e galego):

– Disponibilização de módulos de identificação de nomes próprios baseados em máquinas de estados finitos.

– Disponibilização de módulos de identificação de nomes próprios baseado em classificadores AdaBoost.

– Disponibilização de um módulo de classificação de nomes próprios baseado em classificadores AdaBoost (para português).

– Implementação de sistemas de classificação de nomes próprios baseados em regras e recursos, disponibilizados em Gamallo *et al.* (2014).

– Adaptação e implementação de reconhecedores de expressões numéricas, de quantidades, datas e horas.

– Adição da anotação manual das entidades "enamex" ao corpus Bosque 8.0.

– Novo corpus com anotação manual de entidades "enamex" em galego.

Tenha-se em conta que todos os módulos adaptados estão incluídos em FreeLing, e que o sistema de classificação de regras e recursos disponibiliza-se sob licenças livres.

Os sistemas REM apresentados neste capítulo (dependentes das ferramentas descritas no Capítulo 2) permitem classificar semanticamente diversos tipos de entidades, pelo que facilitam a aplicação das estratégias para a extracção de relações mostradas em capítulos seguintes.

Parte II

Estratégias para a Extracção de Relações

CAPÍTULO 4

EXTRACÇÃO DE RELAÇÕES. REVISÃO

4.1. Introdução

Este capítulo faz uma revisão de diferentes abordagens que têm sido utilizadas para a realização de extracção de relações, mostrando os trabalhos mais importantes de cada uma delas, bem como as métricas de avaliação mais comuns, utilizadas nesta tese.

Primeiro, são apresentadas diferentes aproximações à ER em função do número de relações extraídas pelo sistema. A este respeito, são mostrados também vários métodos que visam reduzir o esforço da construção manual de corpora de treino ou da introdução de pares ou de padrões semente. A seguir, é feita uma revisão daqueles artigos que trabalharam especificamente com extracção de relações biográficas, por serem estas as mais relacionadas com este trabalho. Depois, inclui-se uma secção que analisa a informação linguística utilizada pelas várias estratégias de extracção. Mais à frente mostram-se alguns trabalhos dedicados especificamente à extracção em português, espanhol e galego. Finalmente, apresentam-se as métricas de avaliação utilizadas nos testes sobre a extracção de relações, bem como as conclusões deste capítulo.

4.2. Trabalho relacionado

Domínio fechado

A extracção de relações em domínio fechado consiste na construção de extractores para um conjunto finito de relações. Assim, definida uma relação como `LocaldeMorte`, um

sistema poderia extrair pares relacionados como: *John Lennon – Nova Iorque* ou *Manuel Fraga Iribarne – Madrid*. Uma vez que o processo de adaptação para novas relações pode ser custoso, este tipo de extracção restringe-se à selecção de poucas relações semânticas.

O primeiro trabalho que aplicou padrões (*pattern-matching*) para identificar pares relacionados semanticamente foi Hearst (1992). Neste trabalho utiliza-se um pequeno conjunto inicial de padrões de superfície para obter relações de hiperonímia (e.g., "*hiperónimo*, such as *hipónimo*"). Para aumentar o conjunto de padrões é comum o uso de estratégias de *bootstrapping*, que consistem na procura iterativa de novos padrões que contenham pares já conhecidos (obtidos no processo anterior). A este tipo de técnica, Brin (1998) acrescenta uma avaliação dos novos padrões descobertos, com o fim de seleccionar só aqueles que tenham boa precisão.

Snowball é uma ferramenta que também aprende padrões de extracção de modo iterativo (Agichtein e Gravano, 2000). Este sistema extrai novos padrões e avalia a sua qualidade para obter, sem intervenção humana, pares relacionados semanticamente. O trabalho apresentado em Ravichandran e Hovy (2002) utiliza pares de sistemas de reposta a perguntas para extrair padrões automaticamente. KNOWITALL também extrai exemplos de relações semânticas de modo não supervisionado, ao aprender novas regras de extracção utilizando *bootstrapping* (Etzioni *et al.*, 2004). Uma outra ferramenta que faz ER na Web é *Espresso*: este sistema começa utilizando padrões com muita abrangência mas pouca precisão, e a seguir aproveita a Web para filtrar aqueles padrões com menor precisão (Pantel e Pennacchiotti, 2006). Em Bunescu e Mooney (2007) utilizam-se como pares semente exemplos positivos e negativos das relações alvo, que permitem obter dados suficientes para posteriormente treinar classificadores.

Domínio aberto

Mais recentemente, foram propostas novas estratégias para extracção de relações com o objectivo tanto de minimizar o esforço da obtenção de padrões e de corpora anotado como de aumentar significativamente o número de relações a extrair.

A supervisão-distante (*distant-supervision*) aproveita grandes repositórios de pares já classificados (bases de dados, ontologias, etc.) para etiquetar de modo automático orações que contenham esses pares (como positivas), ou que contenham pares não consistentes com a base de dados (como negativas) (Mintz *et al.*, 2009). Assim, dos pares *Sergey Brin – Google* e *Paul Allen – Microsoft* (da relação `fundadorDe`) poderiam extrair-se as seguintes orações:

– Larry Page e *Sergey Brin* fundaram a *Google* (positivo)

– *Paul Allen* foi um dos fundadores da *Microsoft* (positivo)

– *Sergey Brin* afirmou que a proposta da *Microsoft* é desanimadora (negativo)

– *Paul Allen* processa *Google*, Apple e outras empresas (negativo)

Essas orações são depois utilizadas para treinar classificadores estatísticos. Ao existirem recursos que contêm exemplos de centenas ou milhares de relações diferentes, a supervisão-distante permite construir corpora anotado com todas essas relações.

LUCHS aplica estratégias de supervisão-distante utilizando como entrada fontes com ruído, como as *infoboxes* da Wikipedia (Hoffmann *et al.*, 2010), conseguindo minimizar o impacto do ruído e extrair exemplos de mais de 5.000 relações.

Em Bollegala *et al.* (2010), algoritmos de *clustering* sequencial são aplicados para agrupar diferentes padrões léxico-semânticos, utilizados para extrair relações. Assim, se muitos pares de entidades (e.g., *Adobe* e *Macromedia*) aparecem frequentemente em padrões diferentes ("*Adobe* acquires *Macromedia*", "when *Adobe* bought *Macromedia*", "*Adobe* buys *Macromedia*", etc.), estes padrões estarão potencialmente a expressar a mesma relação. Uma vez agrupados os padrões, estes podem ser utilizados para obter novos exemplos da mesma relação.

Extracção de informação aberta

A Extracção de Informação Aberta (*Open Information Extraction*, OIE), apresentada em Banko *et al.* (2007), é um novo paradigma que consiste na extracção de triplos de base verbal (ou proposições) sem necessidade de se especificarem previamente as relações desejadas. Assim, um sistema de extracção de informação aberta obtém triplos de dous argumentos e uma relação (*arg1*, `rel`, *arg2*) que descrevem proposições presentes no texto:

> *"O chefe da oposição boicotou as eleições em Maio depois de ter sido acusado de corrupção."*

– *O chefe da oposição* `boicotou` *as eleições*

– *O chefe da oposição* `boicotou_as_eleições_em` *Maio*

– *O chefe da oposição* `boicotou_as_eleições_depois_de` *ter sido acusado de corrupção*

A OIE permite extrair sem necessidade de intervenção prévia exemplos de um número infinito de relações. Porém, as relações extraídas podem ser demasiado específicas.

Trabalhos posteriores continuaram a melhorar os sistemas de extracção de informação aberta utilizando corpora anotados e sintaxe superficial para construir o extractor (Fader *et al.*, 2011; Etzioni *et al.*, 2011). Outras estratégias para a OIE também foram propostas, como *Woe*, que utiliza informação sintáctica (de dependências) obtida da análise da Wikipedia para gerar corpora de treino anotado.

Para além destes métodos, foram também apresentados sistemas de OIE que não precisam corpus de treino, extraindo as proposições através de regras de base sintáctica. Em Gamallo *et al.* (2012) é apresentado *DepOE*, um sistema de extracção de informação aberta (multilíngue) que aplica regras de extracção sobre a saída dos *parsers* de dependências de DepPattern.[1] Um outro sistema de características similares é ClauseIE (Corro e Gemulla, 2013), que obtém mediante regras resultados superiores aos modelos estatísticos referidos acima.

Extracção de relações biográficas

Vários artigos e conferências dedicaram-se à extracção de relações de carácter biográfico, mais relacionadas com o tipo de extracções realizadas nesta tese.

Mann (2002) descreve um método para criar ontologias de nomes próprios orientadas a sistemas de resposta a perguntas, que consiste na selecção de nomes comuns que precedem nomes próprios ("o *compositor*$_{\text{nome comum}}$ *Wolfang Amadeus Mozart* $_{\text{nome próprio}}$"). Inspirados neste trabalho, Fleischman *et al.* (2003) treinam algoritmos de aprendizagem automática para classificar exemplos de estruturas "nome comum – nome próprio".

Em Jijkoun *et al.* (2004) avaliam-se padrões sintácticos superficiais para a obtenção de pares de resposta a perguntas. Os resultados deste trabalho mostram que a análise sintáctica melhora o desempenho dos sistemas de resposta a perguntas, apesar de que a extracção é menos precisa do que a obtida através de padrões de superfície.

Garera e Yarowsky (2009) apresentam diferentes abordagens para extrair factos biográficos. O objectivo do trabalho consiste em incluir nos extractores informação estrutural de textos biográficos (posição dos factos no documento, extracção de conhecimento explícito,

[1] Este sistema será utilizado durante os testes de OIE desta tese no Capítulo 8.

etc.). Deste modo consegue-se melhorar o desempenho dos padrões de superfície para a extracção de relações.

BioSnowball (Liu *et al.*, 2010) é uma ferramenta orientada à compilação de informação pessoal desde a Web, informação que é depois utilizada para gerar páginas biográficas similares à Wikipedia.

Outros trabalhos também realizaram extracção automática de conhecimento biográfico, embora este tipo de extracção não fosse o seu objectivo prioritário. Assim, LEILA (Suchanek *et al.*, 2006) é um sistema que extrai exemplos de relações semânticas utilizando análise sintáctica. Pasca *et al.* (2006) também obtêm pares pertencentes à relação `DatadeNascimento` mediante a generalização de padrões encontrados com um pequeno conjunto de pares semente.

Finalmente, duas conferências incluíram recentemente tarefas relacionadas com a extracção de factos biográficos: a Knowledge Base Population (KBP) Slot Filling Track (da Text Analysis Conference, TAC)[2] e a Person Attribute Extraction (da Web People Search Evaluation Campaign, WePS)[3], ambas desde 2009. O objectivo destas tarefas consiste na extracção de atributos de um conjunto predefinido de relações biográficas. Em função do cenário, algumas tarefas incluem *clustering* de páginas *web*, resolução de correferência e extracção de relações de múltiplos documentos.

Conhecimento linguístico

Diferentes abordagens para a extracção de relações estudaram a incorporação de conhecimento linguístico (desde *pattern-matching* simples a modelos estatísticos complexos), com o objectivo de melhorar a extracção de pares relacionados semanticamente.

Assim, os primeiros sistemas de extracção de relações (Hearst, 1992) utilizaram padrões lexicais simples (p. ex., "such as") para obter os pares. Depois, outros trabalhos implementaram estratégias de generalização de padrões, orientadas a aumentar a sua abrangência. Foram empregados vários algoritmos de generalização, que tentam seleccionar o conteúdo lexical mais relevante ou encontrar padrões similares que representem a mesma relação. Assim, Finkelstein-Landau e Morin (1999) utilizaram o *longest common string*; Ravichandran e Hovy (2002) aplicaram o *suffix tree* e Ruiz-Casado *et al.* (2005) calcularam a *edit distance* entre os diferentes padrões encontrados.

[2] `http://www.nist.gov/tac/data/index.html`
[3] `http://nlp.uned.es/weps/`

O já referido Mann (2002) inclui etiquetas morfossintácticas nos padrões para extrair relações como `Profissão`. De modo similar, *BioSnowball* (Liu *et al.*, 2010) utiliza uma combinação de sequências lexicais e de etiquetas morfossintácticas para representar os padrões.

Uma novidade de Suchanek *et al.* (2006) foi a utilização da sintaxe, através da aplicação da *link grammar*, um tipo de análise sintáctica similar à análise de dependências. De modo semelhante, Akbik e Broß (2009) aplicam padrões de dependências sintácticas para enriquecer uma wiki semântica através da Wikipedia.

Outros trabalhos como Bunescu e Mooney (2005), Snow *et al.* (2005) ou Nguyen *et al.* (2007) também criaram diferentes modelos estatísticos com base nos resultados da análise sintáctica. Mais recentemente, Yan *et al.* (2009) fizeram extracção de relações combinando padrões de superfície extraídos da Web com padrões de dependências obtidos da análise sintáctica da Wikipedia.

RGAI (Nagy e Farkas, 2010) foi o sistema com os melhores resultados na tarefa Attribute Extraction da WePS-3. O método empregado consta de duas partes: primeiro, extraem-se parágrafos que contenham informação lexical relevante para uma relação específica (por exemplo, os tokens *nasceu* ou *aniversário* para `DatadeNascimento`). Depois, um classificador extrai pares que pertencem a essa relação utilizando expressões regulares, comparação de padrões e outras heurísticas adaptadas a cada relação.

Os melhores resultados da KBP Slot Filling task de 2011 foram obtidos por Sun *et al.* (2011), trabalho que treina classificadores mediante supervisão-distante. Os atributos utilizados no processo de aprendizagem obtêm-se da sequência de tokens (palavras que aparecem antes e depois das entidades candidatas), da classe semântica das entidades (e.g., pessoa, organização, etc.) e da árvore de dependências: são utilizados tanto o caminho de dependências mais curto entre o par de entidades (*shortest dependency path*) como a cabeça sintáctica e o dependente de cada entidade.

A propósito da efectividade dos diferentes atributos com informação linguística nos processos de extracção de relações, alguns trabalhos analisaram o seu impacto no corpus ACE (que contém anotação para ER genérica). Kambhatla (2004) e Zhou *et al.* (2005) mostram que a incorporação de atributos baseados em conhecimento linguístico melhora o desempenho dos classificadores. Contudo, Jiang e Zhai (2007) concluíram que a adição de atributos mais complexos pode influir negativamente na qualidade dos extractores.

Outras estratégias que utilizam *kernels* (e combinações de *kernels*) também foram propostas para melhorar as tarefas de ER no mesmo corpus ACE (Zhao e Grishman, 2005; Zhang *et al.*, 2006; Zhou *et al.*, 2009; Nguyen *et al.*, 2009).

Extracção de relações em português, espanhol e galego

A maior parte dos sistemas apresentados nos pontos anteriores realizam ER em inglês. Contudo, existem vários trabalhos de extracção de relações em português e espanhol, para além de algum sistema multilíngue que funciona sobre textos em galego.

Especificamente para português, foram vários os trabalhos que se focaram em diferentes tipos de relações. Freitas (2007) avalia a efectividade de padrões inspirados em Hearst (1992) para a obtenção de pares de hipónimos e hiperónimos. Oliveira *et al.* (2008) também lidam com a extracção de diferentes relações entre conceitos —como a já mencionada `hiperonímia/hiponímia`, `parte_de`, `inclusão`, etc.— num dicionário de português. Derivado deste trabalho, o projecto Onto.PT pretende construir automaticamente uma ontologia lexical para esta língua (Oliveira e Gomes, 2010). Outros trabalhos como Costa e Branco (2012) identificam eventos e expressões temporais em texto livre, enquanto Collovini (2014) aplica o algoritmo Conditional Random Fields para a obtenção de relações no domínio das organizações.

A avaliação Segundo HAREM incluiu uma tarefa sobre extracção de relações genéricas entre entidades mencionadas (ReRelEM) (Mota e Santos, 2008). Apresentaram-se três sistemas: REMBRANDT, que utiliza conhecimento extraído da Wikipedia para realizar REM, e um conjunto de heurísticas e de regras gramaticais para extrair relações (Cardoso, 2008); SEI-Geo, que aplica padrões do tipo "such as" (e outras estratégias como *trigger words* e verbos) para estabelecer relações entre localizações (Chaves, 2008); e SeRELeP, que também tenta identificar relações mediante a aplicação de regras simples em entidades mencionadas reconhecidas previamente (Bruckschen *et al.*, 2008). Em relação à extracção de relações biográficas, Soares *et al.* (2011) avaliam vários algoritmos e atributos para classificar orações (mas não pares de entidades) que contenham esse tipo de relações.

Em espanhol, Sánchez-Cuadrado *et al.* (2003) utilizam padrões sintáctico-semânticos para a construção de um *thesaurus* de zoologia. Sierra *et al.* (2008) propõem o uso de padrões verbais para extrair relações de (i) `hiperonímia` e `hiponímia`, (ii) `sinonímia` e (iii) exemplos de `individualidade` (`quantidade/massa` e `membro/grupo`). Outros

trabalhos como Soler e Alcina (2008) também aplicam padrões lexicais para obter exemplos da relação `parte-tudo` (no domínio da cerâmica).

Em Aguado de Cea *et al.* (2008) apresenta-se um método (e uma ferramenta) para reutilizar padrões de desenho de ontologias para o enriquecimento destes recursos (mediante a associações dos padrões a novos padrões léxico-sintácticos). Este trabalho analisa também a extracção em textos em inglês e alemão.

Por último, o sistema multilíngue de extracção de informação aberta *DepOE* (Gamallo *et al.*, 2012) também realiza OIE em português, espanhol e galego. Até ao momento, é o único trabalho que conhecemos sobre extracção de relações em galego.

Apresentadas as diferentes tipologias e abordagens para a extracção de relações, bem como os atributos que vêm sendo utilizados, podemos situar o trabalho realizado nesta tese como segue: em relação ao domínio das extracções, estas enquadram-se em domínio fechado, sendo de carácter enciclopédico (e especificamente biográfico). Contudo, o Capítulo 8 apresenta um conjunto de avaliações de extracção de informação aberta, cujas extracções foram restringidas a aqueles triplos cujo primeiro argumento identifica uma entidade pessoa.

Em relação às abordagens utilizadas para a ER, esta tese aplica estratégias de supervisão-distante cujas instâncias são os pares relacionados (Capítulo 5), aproximações supervisionadas que classificam individualmente cada padrão (Capítulo 6), bem como abordagens baseadas em regras sintáctico-semânticas (Capítulo 7).

Finalmente, para extrair relações em português, galego e espanhol, no presente trabalho são empregados —para além das regras sintáctico-semânticas já referidas— diferentes tipos de atributos linguísticos que vão desde simples tokens, lemas e *PoS-tags* a padrões generalizados ou combinações complexas de dependências sintácticas.

4.3. Métricas de avaliação

De modo geral, os sistemas de extracção de relações são avaliados utilizando métricas *standard*. A avaliação realiza-se utilizando os resultados de classificação que cada sistema produz nos conjuntos de teste, através da matriz de confusão da Figura 4.1.

Aqui, os verdadeiros positivos e os verdadeiros negativos são os exemplos do conjunto de teste correctamente classificados pelo sistema como positivos e negativos, respectivamente. Os falsos positivos e os falsos negativos são os erros produzidos por classificar como positivos exemplos negativos e como negativos exemplos positivos, respectivamente.

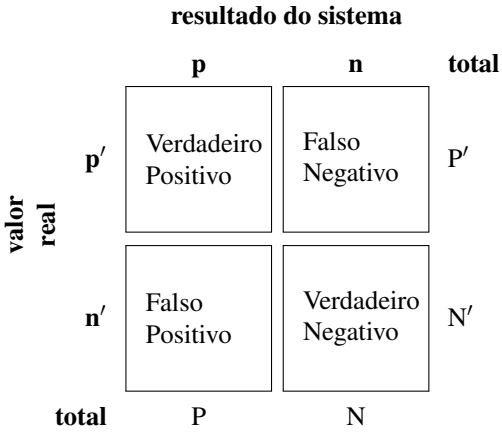


Figura 4.1: Matriz de confusão utilizada para as avaliações de ER.

Precisão: para calcular a precisão é geralmente utilizada a seguinte fórmula, que divide as classificações positivas correctas pelo número total de decisões correctas do sistema:

$$precisão = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ positivos} \tag{4.1}$$

***Recall*:** os resultados de *recall* obtêm-se dividindo as classificações positivas correctas pelo número total de exemplos correctos no conjunto de teste:

$$recall = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ negativos} \tag{4.2}$$

Medida F: finalmente, para obter a medida F (*F1* ou *F-score*) calcula-se a média harmónica entre a precisão e o *recall*:

$$medida\ F = 2 \cdot \frac{precisão \cdot recall}{precisão + recall} \tag{4.3}$$

Para obter os valores médios de alguns resultados foram calculadas tanto a *micro-average* como a *macro-average*. A *micro-average* calcula-se construindo uma tabela global com os resultados de cada relação, computando a seguir a precisão, *recall* e medida F do total. A *macro-average* é a média individual dos resultados de cada relação.

Para além destas medidas, a avaliação de alguns processos requer métricas específicas, que são apresentadas oportunamente nos respectivos testes.

4.4. Conclusões

Este capítulo fez uma revisão de diferentes estratégias existentes para a extracção de relações semânticas de texto livre. Foram apresentadas técnicas de extracção em domínio fechado, bem como outras alternativas orientadas à extracção de um maior número de relações, assim como a extracção de informação aberta.

Para além disso, vários trabalhos focados na extracção de relações biográficas foram apresentados, seguidos de uma análise da utilização de informação de carácter linguística pelos próprios extractores.

Depois, foi mostrada uma breve panorâmica dos principais trabalhos que lidaram com a extracção de relações nas línguas alvo desta tese: o português, o espanhol e o galego. Assim mesmo, as técnicas aplicadas neste trabalho foram situadas em relação à revisão realizada.

Finalmente, foram apresentadas as métricas de avaliação *standard* utilizadas nos testes de extracção de relações.

CAPÍTULO 5

EXTRACÇÃO DE RELAÇÕES MEDIANTE SUPERVISÃO-DISTANTE

5.1. Introdução

A grande quantidade de dados que existe na Web faz com que muita da informação disponível seja redundante, aparecendo em diferentes fontes, línguas e formatos. Um sistema de extracção de relações pode aproveitar-se desta abundância de conteúdos para obter informação sobre entidades de maneira fiável (Mann, 2002).

Este capítulo apresenta um conjunto de testes para a extracção de relações semânticas em domínio fechado que tenta tirar partido da redundância da informação na Web. A estratégia utilizada consiste na construção de classificadores que analisem os diferentes contextos em que um par relacionado semanticamente aparece.

Assim, dada uma relação semântica e um par candidato (e.g., `Profissão`, *Billie Holiday – cantora*), o sistema analisa um conjunto de orações em que o par candidato aparece e classifica-o como positivo ou negativo para a relação, em função dos contextos em que ocorreu.

Com o fim de minimizar o esforço de construção de um corpus de treino anotado que contenha orações positivas e negativas para a relação alvo, foi aplicada uma estratégia de supervisão-distante (Mintz *et al.*, 2009), que obtém de modo automático corpora anotados. Dada a escassez de recursos na Web para o galego, este capítulo não contém testes de extracção para esta língua, sendo as avaliações realizadas unicamente em português e espanhol.

Para construir os classificadores foram utilizadas técnicas de aprendizagem automática e avaliados diferentes atributos que representam as estruturas linguísticas em que os pares ocorrem.

Os resultados das avaliações indicaram que a generalização de atributos de base pseudo-sintáctica permite criar classificadores com bom desempenho para a estratégia de extracção proposta. Contudo, é preciso referir que o método de construção de corpora mediante a supervisão-distante produz resultados variáveis em função dos recursos utilizados e da relação alvo.

A seguinte secção descreve o método de obtenção de corpus anotado mediante supervisão-distante. Depois, a Secção 5.3 apresenta os atributos utilizados para treinar os classificadores supervisionados. A Secção 5.4 contém os testes realizados, e na 5.5 mostram-se alguns dos problemas da extensão do método proposto a outras relações. Finalmente, as conclusões são apontadas na Secção 5.6.

O conteúdo deste capítulo foi publicado nos trabalhos Garcia e Gamallo (2011b,c), que contêm avaliações para português e espanhol, respectivamente.

5.2. Método

Para obter o corpus anotado foi utilizada, do seguinte modo, a estratégia de supervisão-distante:

Primeiro foi obtido um ficheiro *dump* (imagem) da Wikipedia para cada língua (português e espanhol).[1] Cada imagem foi convertida a texto plano, e foram eliminadas as marcas de formatação e as ligações externas.

Para cada relação semântica (nos testes, `Profissão`), foram obtidos pares já classificados das *infoboxes* da Wikipedia na mesma língua: por exemplo, *Fernando Pessoa – poeta*, *Fernando Pessoa – escritor*, etc. (com uma precisão de $\approx 95\%$).

A seguir, foram extraídas do texto livre da Wikipedia todas as orações que continham um nome de pessoa e uma profissão conhecidas (presentes nos pares extraídos das *infoboxes*). As orações foram classificadas como *positivas* quando os dous termos coincidiam com um par conhecido da lista inicial, e como *negativas* se o par não existia na mesma lista.

Depois foram aplicados os módulos de lematização e de anotação morfossintáctica de FreeLing. Em espanhol, FreeLing também foi empregado para o reconhecimento de entida-

[1] `http://dumps.wikimedia.org/`

des mencionadas, enquanto que o sistema de regras apresentado no Capítulo 3 foi utilizado para o REM em português. As dependências sintácticas foram geradas nas duas línguas por DepPattern.

Finalmente, os termos alvo (nome próprio e profissão) foram substituídos por **X** e **Y**, respectivamente, sendo as orações divididas em três contextos: *anterior*, *intermédio* e *posterior*, em função da sua posição em relação a **X** e **Y**.

Uma vez que o processo foi realizado sem revisão manual, este produziu anotação de falsos positivos (p. ex., "*Linus Torvalds* discutiu com um *engenheiro de software*": **positivo**) e falsos negativos ("*Fernando Pessoa* foi um *crítico literário*": **negativo**, porque o atributo *crítico literário* não aparece na *infobox*). A revisão manual de um conjunto de orações anotadas (utilizadas como corpus de teste nas avaliações) mostrou que a anotação automática teve perto de 80% de precisão na relação `Profissão`.

5.3. Atributos

Cada uma das orações analisadas mediante o processo anterior representa uma *estrutura linguística* que contém toda a informação necessária para os sistemas de extracção. Uma estrutura linguística pode ser concebida como um espaço que incorpora vários tipos de conhecimento, e do qual são extraídos os atributos utilizados pelos classificadores.

Cada estrutura linguística contém o contexto dos termos relacionados, sendo **X** o nome de pessoa e **Y** a profissão. Inclui-se na própria estrutura linguística o contexto anterior ao primeiro termo, o contexto intermédio, e o posterior. Estes contextos têm uma dimensão máxima de 12 tokens (para o intermédio) e de 3 para os contextos anterior e posterior, tendo sido estas janelas seleccionadas empiricamente durante a realização de testes preliminares.

Na Figura 5.1, as colunas 1, 2, 3, e 4 representam a posição, token, lema e categoria morfossintáctica (veja-se o *tagset* na Tabela A.3), respectivamente. Uma vez que a estrutura linguística também contém informação sintáctica de dependências, a coluna 5 identifica o núcleo do token actual, e a coluna 6 mostra a função sintáctica.[2] Esta estrutura está inspirada no formato utilizado nas conferências CoNLL, definido em Lin (2003).

As estruturas linguísticas obtidas de cada uma das orações anotadas foram utilizadas para extrair os atributos necessários para o treino dos classificadores. Foram utilizados quatro tipos de atributos:

[2] Aqui, a etiqueta *subj* significa sujeito; *punct*, pontuação; *adjn*, adjunto; *cprep*, complemento preposicional; *term*, termo; *spec*, especificador; *attr*, atributo e *modif* modificador.

Oração: Kimberley Deal (nascida em 10 de Junho de 1961) é uma cantora americana

Polaridade: Kimberley_Deal `Profissão` cantora: **positivo**

Estrutura linguística:

pos	*token*	*lema*	*PoS-tag*	*núcleo*	*etiqueta*
0	**X**	Kimberley_Deal	PESSOA	6	subj
1	(	(	Fa	2	punct
2	nascida	nascer	VB	0	adjn
3	em	em	PS	2	cprep
4	10_de_Junho_de_1961	10/06/1961	DATA	3	term
5	)	)	Fc	2	punct
6	é	ser	V	-	-
7	uma	um	DT	8	spec
8	**Y_Pr**	cantor	NC	6	attr
9	americana	americano	AD	8	modif

Figura 5.1: Exemplo de uma oração com a relação `Profissão`, o par relacionado e a polaridade, e a sua estrutura linguística.

Padrões básicos:

O primeiro tipo de atributos utiliza toda a informação presente na estrutura linguística, excepto dous elementos: informação de dependências e alguns lemas. Os padrões básicos só contêm lemas de verbos, nomes comuns e preposições, dado que em testes preliminares esta selecção produzia melhores resultados do que o uso de todos os lemas ou nenhum deles. Este facto sugere que os verbos, os nomes comuns e as preposições contêm a informação mais relevante na representação dos contextos léxico-sintácticos dos termos relacionados. Um exemplo de um padrão básico é o seguinte:

Oração: Kimberley Deal é uma cantora americana

Padrão: <**X** ser_VB DT **Y** AD>

Devido à rigidez deste tipo de atributos, é preciso referir que precisam de uma grande quantidade de corpus de aprendizagem, porque pequenas variações em pontuação ou modificação adverbial ou adjectival geram atributos diferentes. Portanto, a dispersão de dados é crucial aqui.

Generalização de padrões:

Com o fim de minimizar o problema da dispersão de dados, foi aplicado um algoritmo baseado na similaridade entre padrões básicos que os generaliza, aumentando assim a sua abrangência. Para generalizar dous padrões, primeiro verifica-se se são similares, e depois são removidas aquelas unidades não partilhadas entre eles (Ruiz-Casado *et al.*, 2005). A similaridade (*Dice_lcs*) entre dous padrões p_1 e p_2 define-se através do *longest common string* (a cadeia de caracteres comum mais longa) e da métrica *Dice*, do seguinte modo:

$$Dice\_lcs(p_1, p_2) = \frac{2 * lcs(p_1, p_2)}{longitude(p_1) + longitude(p_2)} \tag{5.1}$$

onde $lcs(p_1, p_2)$ é o tamanho do *longest common string* entre os padrões p_1 e p_2, enquanto $longitude(p_i)$ representa o tamanho do padrão p_i. Isto significa que a similaridade entre dous padrões é a função do seu *longest common string* e das suas longitudes.

Uma vez calculada a similaridade entre dous padrões p_1 e p_2, extrai-se o *longest common string* só se p_2 é o padrão mais similar de p_1 e o valor de similaridade é maior do que um limite específico (no testes aqui descritos, 0,75). O *longest common string* de dous padrões é considerado a sua generalização.

Saco de lemas e *PoS-tags*:

Uma alternativa à utilização de padrões como atributos, é o uso de elementos mais pequenos, que aumentam a abrangência dos classificadores. Estes elementos podem ser tokens (nas estratégias *bag-of-words*) ou lemas, entre outros.

Para construir os classificadores, foram utilizadas combinações de lemas e *PoS-tags*, pelo que da oração mostrada nos exemplos anteriores ("**X** é uma **Y** americana") seriam extraídos os seguintes atributos: <ser_VB>, <DT>, <AD> (mais uma vez, só alguns lemas foram seleccionados, nomeadamente aqueles das categorias com maior informação semântico-sintáctica nas restrições de selecção entre dependências: verbos, nomes comuns e preposições). A utilização de exemplos negativos durante o processo de treino é aqui mais importante, dado que o classificador deve aprender que lemas são os mais importantes em cada relação para tomar a decisão correcta.

Dependências sintácticas:

A informação sintáctica foi obtida com DepPattern, que identifica as dependências mais frequentes entre os termos relacionados. Novamente, este tipo de atributos só inclui os lemas dos verbos, nomes comuns e preposições.

De cada estrutura linguística, foram seleccionados como atributos dous tipos de dependências: (i) dependências entre os dous termos relacionados (**X** ou **Y**) e (ii) dependências entre um dos dous termos relacionados e uma entidade do contexto (anterior, intermédio ou posterior). Por exemplo, da oração "**X** é uma **Y** americana", as dependências seleccionadas seriam as seguintes: <subj:ser_VB;**X**>, <attr;ser_VB;**Y**>, <spec;**Y**;DT> e <modif;**Y**;AD>.

Cada atributo é um triplo formado pela etiqueta de dependência, o núcleo e o dependente. Só foram escolhidas dependências que contivessem, no mínimo, um dos termos relacionados (**X** ou **Y**). A informação seleccionada, portanto, corresponde-se com o contexto de dependências locais dos termos alvo.

5.4. Testes e avaliação

Nos diferentes testes realizados foi avaliado tanto o desempenho individual dos atributos como várias combinações deles, para classificar exemplos da relação `Profissão`.

Os testes levaram-se a cabo com o *software* WEKA (Witten e Frank, 2005), utilizando SMO (*Sequential Minimal Optimization*, algoritmo de optimização para treinar máquinas de vectores de suporte, SVM) (Platt, 1999), que teve em testes preliminares melhores resultados do que classificadores baseados em Naive Bayes e em árvores de decisão.

Os corpora de treino foram obtidos das versões em português e em espanhol da Wikipedia (Maio de 2010), utilizando a estratégia de supervisão-distante apresentada na Secção 5.2. Para cada língua foram extraídos uns 50.000 pares das *infoboxes*, com os quais se obtiveram dous conjuntos de $\approx$ 500.000 orações, classificadas automaticamente como positivas ou negativas para a relação `Profissão`. Para treinar os diferentes modelos, seleccionaram-se aleatoriamente 2.000 orações de cada língua. Para as avaliações, extraíram-se e revisaram-se manualmente 700 orações para cada língua (também aleatórias e diferentes das utilizadas para o treino). As métricas de avaliação utilizadas são as apresentadas na Secção 4.3.

Resultados

Os classificadores individuais foram treinados utilizando os tipos de atributos definidos na Secção 5.3:

- *pattern-all* e *pattern-mid* utilizam os padrões básicos como atributos. O primeiro contém os três contextos de cada oração (anterior, intermédio e posterior) enquanto *pattern-mid* foi treinado unicamente com os padrões intermédios.
- *pattern_gen-mid* utiliza os padrões intermédios generalizados como atributos.
- *bow-all* e *bow-mid* foram construídos utilizando os sacos de lemas e *PoS-tags*: *bow-all* com os três contextos, e *bow-mid* unicamente com os intermédios.
- *dep-all* e *dep-mid* são os modelos criados com os atributos baseados em dependências sintácticas.

Modelos individuais:

O primeiro conjunto de testes avaliou sete classificadores (para cada língua), construídos utilizando unicamente um tipo de atributos. Os resultados (Tabela 5.1) indicam que os melhores atributos são aqueles baseados em padrões léxico-sintácticos generalizados: *pattern_gen-mid* com valores de medida F de 78% e 83% em português e espanhol, respectivamente. Os resultados dos classificadores *pattern-all* são muito mais baixos devido ao seu pouco *recall*. Uma vez que os contextos anterior e posterior têm muita variação, os modelos *pattern-mid* mostraram um melhor desempenho. Em relação aos classificadores gerados com sacos de lemas e *PoS-tags*, estes têm resultados divergentes em função da língua analisada: enquanto em espanhol superam o 71% de medida F, em português só o modelo *bow-all* obtém resultados satisfatórios (71% *versus* 44% de *bow-mid*). Finalmente, os modelos baseados em dependências sintácticas têm um melhor comportamento quando treinados com os três contextos (anterior, intermédio e posterior).

Similaridade e combinações de atributos:

Na análise das diferenças entre os modelos individuais foi calculado o coeficiente de similaridade *Dice*, para conhecer se os erros e acertos de cada classificador no corpus de teste foram ou não nos mesmos exemplos. De modo geral, um coeficiente *Dice* alto significa que

Modelo	**Português**			**Espanhol**		
	Prec	*Rec*	*F1*	*Prec*	*Rec*	*F1*
pattern-all	91,66	2,65	5,16	**100**	12,99	23,00
pattern-mid	**94,90**	58,45	72,34	98,04	56,50	71,68
pattern_gen-mid	93,26	66,90	**77,90**	97,72	**72,60**	**83,31**
bow-all	74,14	**67,87**	70,87	83,02	62,15	71,08
bow-mid	74,13	31,15	43,87	88,28	59,60	71,16
dep-all	79,21	48,79	60,38	86,62	65,82	74,80
dep-mid	76,92	41,06	53,54	86,21	35,31	50,10

Tabela 5.1: Precisão, *recall* e medida F de 7 classificadores baseados em atributos individuais em português e espanhol. Relação `Profissão`.

existem poucas decisões correctas tomadas em exemplos diferentes, enquanto um coeficiente baixo implica que existem mais decisões correctas tomadas em exemplos diferentes. Assim, só pares de atributos com valores baixos do coeficiente *Dice* foram combinados, uma vez que a probabilidade de que sejam complementares é maior.

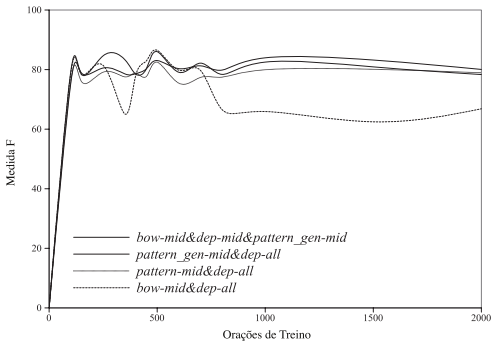
Curvas de aprendizagem:

Para além de treinar modelos combinados, foram geradas curvas de aprendizagem das melhores combinações, com o fim de conhecer a quantidade de dados necessários para atingir o melhor desempenho.

A Figura 5.2 mostra a medida F dos classificadores criados com as melhores combinações de atributos em diferentes estágios de aprendizagem (em função do número de orações utilizadas para o treino).

Em português, a adição dos atributos *dep-all* aos padrões intermédios generalizados conseguiu o melhor desempenho, aumentando em quase 5 pontos a medida F ($\approx 84\%$) ao utilizar um corpus de 1.500 exemplos durante o treino. Em espanhol, a melhor combinação foi obtida unicamente com atributos extraídos do contextos intermédio (*bow-mid+dep-mid+pattern_gen-mid*), atingindo os 88% de medida F. Esta combinação, contudo, também obteve bons resultados em português, atingindo 82% de medida F.

Em relação às curvas de aprendizagem, pode observar-se que estas se estabilizam quando o corpus de treino atinge ≈ 1.000 exemplos, pelo que não são precisos muitos mais dados para melhorar o desempenho.



(a) Português (b) Espanhol

Figura 5.2: Medida F *versus* tamanho do corpus de treino (de 0 a 2000 orações) das 4 melhores combinações de atributos em português e espanhol.

Em suma, as diferentes combinações baseadas numa análise de similaridade permitem treinar classificadores com melhores resultados tanto em precisão como em *recall* do que aqueles criados com atributos individuais.

5.5. Extensão a novas relações

Uma vez realizados os diferentes testes com a relação `Profissão`, a mesma estratégia de construção de corpora mediante supervisão-distante foi aplicada para a obtenção de dados de outras relações biográficas, como `LocaldeNascimento`, `DatadeNascimento`, `LocaldeMorte` e `DatadeMorte`.

Contudo, os novos corpora gerados não tiveram a mesma qualidade do que os criados para a relação `Profissão`, cuja precisão tinha sido de cerca de 80%. As diferenças de qualidade na utilização da estratégia de supervisão-distante deveram-se principalmente a três factores:

– Relação alvo: diferentes relações semânticas geram corpora com grandes variações de qualidade. Assim, as diferenças de precisão dos corpora de `DatadeNascimento` e de `DatadeMorte` foram de $\approx 30\%$.

– Especificidade dos pares: em função do recurso utilizado para a obtenção dos pares classificados (e.g., *infoboxes*, Freebase, etc.), estes podem ser muito ou pouco específicos (em função da relação), para além de poderem ter maior ou menor quantidade de

ruído. Por exemplo, um dos pares obtidos para a relação `LocaldeNascimento` em espanhol foi *Francisco Franco – Espanha*, cujo segundo atributo é pouco específico. Este tipo de pares provocaram grandes quantidades de falsos positivos durante a extracção, o que resultou em conjuntos de orações anotadas com uma precisão de $\approx 15\%$, o que não permite treinar classificadores de qualidade.

– Coesão entre os pares e o corpus: se os pares classificados e o corpus de extracção não pertencem ao mesmo —ou similar— domínio, a estratégia pode produzir grandes quantidades de dados incorrectos (como já apontado em Riedel *et al.* (2010)). Nos nossos testes, a utilização de pares de `DatadeNascimento` de Freebase para a extracção da Wikipedia reduziu a precisão em mais de 20% (embora obtendo mais do dobro de orações anotadas). Aqui, a coesão não depende só do segundo argumento (a data, local ou profissão, por exemplo), mas também do primeiro argumento (o nome de pessoa), dado que estes podem aparecer de muitas formas diferentes ("Lennon", "John Lennon", "John W. Lennon", etc.), como se mostrará no Capítulo 8.

Estes factores provocaram que a construção de classificadores para outras relações não tivesse os mesmos resultados do que para `Profissão`, apesar de que as tendências no impacto dos atributos fossem similares.

Assim, com o fim de ampliar o número de relações semânticas analisadas pelos classificadores, foram avaliadas outras estratégias de extracção, tais como classificadores supervisionados com corpora de treino corrigidos manualmente (Capítulo 6) e métodos baseados em regras obtidas de modo semiautomático (Capítulo 7).

5.6. Conclusões

Este capítulo apresentou uma estratégia de construção de classificadores para a extracção de exemplos da relação `Profissão` em português e espanhol.

Os classificadores foram treinados com corpora obtidos automaticamente mediante técnicas de supervisão-distante.

A estratégia de extracção consiste na classificação de cada par ("pessoa – profissão") como positivo ou negativo para a relação alvo, em função dos contextos linguísticos em que aparece.

Em relação ao desempenho, foram avaliados diferentes atributos de base linguística. A este respeito, a utilização do algoritmo *longest common string* para a generalização de padrões

léxico-sintácticos gerou os melhores atributos para o treino dos sistemas. Para além disso, a combinação de diferentes atributos —realizada mediante medidas de similaridade— permitiu construir modelos com um melhor equilíbrio entre precisão e *recall*.

Contudo, a extensão da estratégia proposta a outras relações não foi tão satisfatória como com `Profissão`, devido a algumas deficiências na aplicação do método de supervisão-distante, pelo que diferentes estratégias para a extracção de relações serão propostas nos capítulos seguintes.

Capítulo 6

Extracção de Relações mediante Classificadores Supervisionados

6.1. Introdução

Como mostrou o Capítulo 4, nos últimos anos foram utilizadas várias estratégias (como a supervisão-distante) para minimizar o esforço de anotar manualmente corpora, ao lado de abordagens de extracção não supervisionadas. Contudo, um dos principais métodos para a extracção com alta qualidade de relações semânticas em texto livre continua a ser o uso de classificadores supervisionados. Estes classificadores são treinados com orações que contêm pares das relações alvo, e utilizam diversos tipos de atributos: desde informação ortográfica ou lexical até atributos mais complexos como etiquetas morfossintácticas ou outros elementos com diferente grau de conhecimento sintáctico e semântico.

Para criar classificadores supervisionados de alta qualidade, alguns trabalhos seleccionam os atributos para um tipo específico de padrões ou de relações, construindo sistemas de extracção muito precisos (Fleischman *et al.*, 2003). Outros trabalhos têm mostrado que alguns atributos (e combinações) funcionam bem com determinadas relações, mas o seu desempenho reduz-se em cenários diferentes (Agichtein, 2005). Contudo, não tem havido muitas pesquisas sobre a importância da selecção dos atributos para a extracção de relações, sendo normalmente só focadas na análise do inglês (Kambhatla, 2004; Zhou *et al.*, 2005).

A este respeito, a avaliação da efectividade de vários tipos de conhecimento linguístico no processo de extracção de relações é importante não só para esta tarefa, mas também para

a linguística teórica e computacional em geral. Esta avaliação permite conhecer o melhor modo de representar estruturas linguísticas que contenham relações semânticas entre dous elementos.

O objectivo do presente capítulo é realizar uma avaliação de diferentes atributos para a extracção de relações biográficas em português e espanhol, mediante a utilização de classificadores supervisionados.

Para isso, é feita uma avaliação sistemática da efectividade de vários tipos de informação linguística, analisando atributos genéricos que têm sido utilizados em diversos trabalhos de extracção de relações, obtidos de diferentes níveis de conhecimento linguístico.

À diferença da abordagem utilizada no capítulo anterior, os sistemas construídos aqui classificam como positivo ou negativo cada padrão (que contém um par), e não cada par em função do conjunto de padrões em que aparece.

Em relação aos corpora, foram construídos dous recursos (um para português e outro para espanhol) com as seguintes relações biográficas: `LocaldeNascimento`, `LocaldeMorte`, `DatadeNascimento`, `DatadeMorte` e `Profissão`. Para além disso, foi realizada uma análise pormenorizada dos padrões de cada relação, com o objectivo de saber como são expressas estas relações biográficas em português e em espanhol. Os dados iniciais para a construção dos corpora foram obtidos mediante supervisão-distante, mas a anotação foi posteriormente corrigida de modo manual, pelo que a classificação é considerada supervisionada. Sobre os corpora resultantes foram aplicados algoritmos de aprendizagem automática, com o fim de treinar e avaliar vários classificadores cuja única diferença reside no tipo de atributos linguísticos com que foram construídos.

Os resultados de diversos testes mostram que a utilização da lematização —para além de conhecimento semântico básico obtido através do REM— melhora o desempenho dos classificadores baseados em *sacos de palavras*. Para além disso, a informação pseudo-sintáctica (representada por bigramas de lemas) pode utilizar-se para evitar processos computacionalmente custosos como a análise sintáctica que, por sua vez, não melhorou significativamente o desempenho dos extractores.

O conteúdo deste capítulo foi publicado no artigo Garcia e Gamallo (2013), e organiza-se como segue: a Secção 6.2 mostra o processo de construção dos corpora e uma análise das estruturas linguísticas que contêm. A seguir, a Secção 6.3 centra-se na descrição dos atributos, nomeadamente no diferentes tipos de conhecimento linguístico utilizado pelos classificado-

res. Depois, a Secção 6.4 contém os diversos testes realizados em português e em espanhol, enquanto as conclusões do presente capítulo se encontram na Secção 6.5.

6.2. Corpora

Esta secção apresenta as principais características dos corpora utilizados nos testes, bem como o seu processo de construção mediante a estratégia de supervisão-distante. Foram criados dous corpora, um para português e outro para espanhol.

Construção dos corpora

Construir manualmente corpora anotados é um processo custoso, mas necessário para treinar modelos estatísticos que dependem de dados de alta qualidade. Para minimizar o esforço de construção e anotação, foi aplicada a técnica de supervisão-distante apresentada no capítulo anterior, mas ampliada com um maior número de relações e de pares iniciais, o que permitiu obter recursos de mais tamanho.

Como foi visto, esta estratégia possibilita a obtenção dados de qualidade para algumas relações e conjuntos de dados (Hoffmann *et al.*, 2010). Contudo, trabalhos como Riedel *et al.* (2010) (ou alguns testes levados a cabo no capítulo anterior, veja-se a Secção 5.5) indicaram que a supervisão-distante pode produzir grande quantidade de ruído se a base de conhecimento (da qual se obtêm os pares) e o corpus (de onde se extraem as orações) não pertencem ao mesmo —ou similar— domínio.

Para evitar este ruído, na construção dos corpora utilizados neste capítulo foram aplicados um conjunto de filtros e restrições, escolhendo unicamente os nomes de profissões mais comuns e só os nomes de pessoas e localizações que coincidissem exactamente com os dos pares extraídos. Finalmente, as orações classificadas mediante supervisão-distante foram corrigidas posteriormente de modo manual, sendo portanto o treino dos classificadores supervisionado.

É preciso referir que a aplicação deste método para línguas diferentes do inglês não é sempre possível, uma vez que os atributos de muitas relações são específicos para cada idioma, e as principais bases de conhecimento são desenhadas para o inglês.[1]

Assim, os pares que não dependem da língua (aqueles que contêm datas) foram obtidos de Freebase e da DBpedia em inglês. Os outros conjuntos de pares (dependentes de língua)

[1] A modo de exemplo, na altura da criação destes corpora, a DBpedia para inglês continha 433.042 exemplos de `DatadeNascimento`, enquanto as versões para português e espanhol tinham 47.460 e 1.243, respectivamente.

obtiveram-se das *infoboxes* da Wikipedia em português e espanhol (e das incipientes versões da DBpedia nestas línguas). Foram empregados os seguintes conjuntos de pares:

- `DatadeNascimento` e `DatadeMorte`: 460.703 pares (independentes da língua)
- `LocaldeNascimento`: 45.588 (pt) e 8.952 (es)
- `LocaldeMorte`: 11.664 (pt) e 1.319 (es)

Para a relação `Profissão`, só foram utilizados os nomes de profissões mais comuns (aqueles com mais de 20 ocorrências na lista de pares semente), para minimizar a extracção de ruído. Assim, empregaram-se 68 profissões para português, e 96 para espanhol.

Os corpora utilizados para a extracção de orações foram os seguintes: as versões em português e espanhol da Wikipedia (de 700mb e $1,6$gb respectivamente) e 1gb (pt) e 225mb (es) de textos jornalísticos (do jornal *Público* em português e *El País* em espanhol). A extracção final teve rácios Wikipedia/jornal de $90\%/10\%$ (pt) e $93\%/7\%$ (es).

Os corpora obtidos foram analisados com as ferramentas descritas nos Capítulos 2 e 3, construindo-se uma *estrutura linguística* para cada oração (veja-se a Figura 5.1 na página 70). Só se seleccionaram orações cujas entidades coincidissem exactamente com as presentes nas listas de pares, e que tivessem sido classificadas pelos sistemas REM com a mesma classe (pessoa, localização, etc.), excepto as datas, que foram seleccionadas se, no mínimo, o ano era o mesmo do que nos pares classificados.

A classificação automática (positiva ou negativa) das orações para as relações alvo foi corrigida manualmente, bem como a classificação de outras entidades que aparecessem nas orações. Finalmente, substituíram-se os elementos dos pares por **X** (PESSOA) e por **Y** (**Y_Loc**, **Y_Dat** ou **Y_Pr** para localizações, datas e profissões, respectivamente).

Ambos os corpora têm anotação das cinco relações referidas: `LocaldeNascimento`, `DatadeNascimento`, `LocaldeMorte`, `DatadeMorte`, e `Profissão`. O tamanho é de 268.469 e 152.817 tokens em português e espanhol, respectivamente.

A Tabela 6.1 mostra o número de pares etiquetados para cada relação e língua. Os pares positivos foram também utilizados como negativos para relações diferentes com entidades da mesma classe. Assim, orações como "PESSOA nasceu em LOCALIZAÇÃO" (positiva para `LocaldeNascimento`) utilizou-se como exemplo negativo para `LocaldeMorte`.[2]

[2]Excepto algumas excepções —não tidas em conta—, onde uma oração pode ser positiva e negativa para uma mesma relação: "PESSOA nasceu e faleceu em LOCALIZAÇÃO".

Relação	**Português**		**Espanhol**	
	Pos.	*Neg.*	*Pos.*	*Neg.*
`LocaldeNascimento`	1.312	1.186	493	1.149
`DatadeNascimento`	552	510	419	662
`LocaldeMorte`	879	5.040	563	833
`DatadeMorte`	402	835	570	771
`Profissão`	1.277	740	1.595	715
Total	4.422	8.311	3.640	4.130

Tabela 6.1: Número de pares candidatos para cada relação e língua nos corpora anotados. *Pos.* são os pares positivos e *Neg.* os negativos para cada relação semântica.

Propriedades linguísticas dos corpora

Antes de extrair os atributos para treinar os classificadores, foi realizada uma análise pormenorizada dos padrões que contêm as relações alvo em ambos os corpora. Primeiro, criou-se um histograma do número de tokens que aparecem entre as duas entidades candidatas. Depois, obtiveram-se os padrões mais comuns para cada relação e língua.

A Figura 6.1 contém um histograma que sintetiza o número de tokens intermédio (de 0 a 18) entre as entidades relacionadas em português e espanhol.

Apesar de que a distância entre as entidades difere em cada relação (facto que mostrará a análise dos padrões mais frequentes), o histograma da Figura 6.1 evidencia que as relações semânticas ocorrem com maior frequência entre entidades próximas (com o melhor rácio entre entidades separadas por menos de 8 tokens, variando em função da língua e da relação). O número de pares negativos aumenta à medida que o tamanho da janela cresce, aparecendo cada vez menos pares positivos.

A propósito dos padrões em que as relações biográficas ocorrem, a Tabela 6.2 inclui os dous melhores padrões por relação e língua, tendo em conta tanto a sua percentagem no conjunto de orações positivas como a sua precisão. Os padrões aparecem simplificados pela utilização de disjunções e pela omissão de alguns elementos opcionais.

A taxonomia apresentada na Tabela 6.2 mostra que as relações alvo ocorrem frequentemente em padrões biográficos específicos de alta precisão, que contêm em parênteses factos biográficos sobre a pessoa mencionada previamente. Esta tabela também indica que algumas relações têm uma dependência forte desse tipo de padrões (os quais representam entre 36% e 49% dos pares positivos), enquanto outras ocorrem numa maior variedade de estruturas.

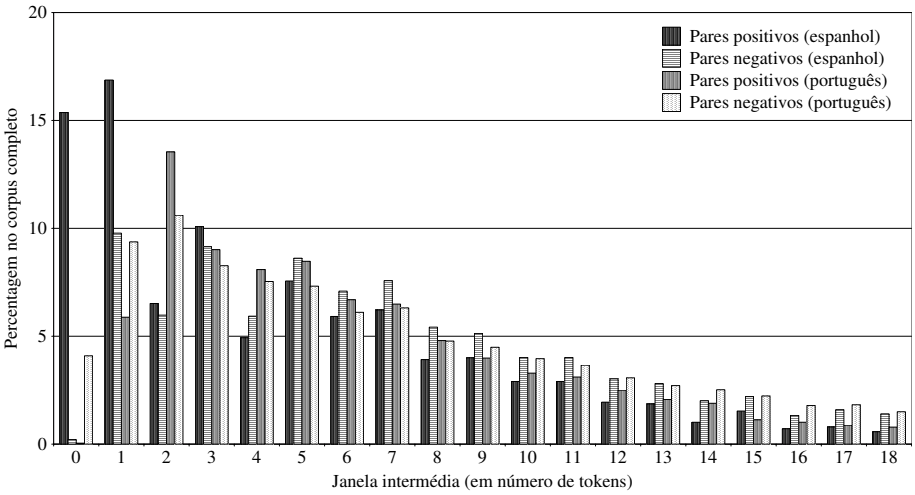


Figura 6.1: Histograma de pares positivos/negativos *versus* contexto intermédio em número de tokens. Os valores são a *micro-average* das cinco relações analisadas.

Contudo, é preciso apontar que, com a excepção dos padrões de alta precisão já referidos, outras estruturas são muito ambíguas, e a sua utilização para identificar uma relação específica pode depender de elementos lexicais que ocorrem fora da janela intermédia.

Estas análises mostram a importância da selecção adequada da janela intermédia, bem como dos atributos para representar os padrões. Note-se que os padrões que representam relações biográficas não se baseiam só em estruturas lexicais e sintácticas, mas também contêm uma grande variedade de sinais de pontuação que representam informação linguística relevante.

6.3. Atributos

Com o fim de avaliar a efectividade do conhecimento linguístico para a extracção de relações, os atributos foram organizados em função da complexidade da análise, começando desde os mais simples. Esta secção apresenta e discute os diferentes atributos utilizados para treinar os classificadores.

Líng.	*Rel.*	**Padrão**	**%**	**Prec**
Pt	DN	**X (** {NP ,|; NP ,|;} **Y_Dat**	30,4	96
		X {(|,} **nasceu|nascido** {em NP ,} **em** {o}|**a Y_Dat**	13,4	97
	LN	**X** {,} **nasceu|nascido** {em a|o NP ,} **em o Y_Loc**	49,3	98
		X {,} **(** {Dat , NP , em} **Y_Loc**	14,2	94
	DM	**X (** {NP , NP ,} **Dat** {-|;|,} {NP ,} **Y_Dat**	20,7	100
		X {,} **morreu|faleceu em Y_Dat**	2,5	100
	LM	**X** {,} **(NP ,** {NP ,} **Dat** {-|; Dat} **Y_Loc**	21,1	99
		X {,} **morreu|faleceu** {em Dat} **em** {o|a} **Y_Loc**	2,2	100
	Pr.	**Y_Pr** {,} {de a NP} **X**	27,1	99
		X ,|é|foi {um|uma} **Y_Pr**	13,5	97
Es	DN	**X (** {NP ,|; NP ,|} **Y_Dat**	45,8	94
		X {(|,} **nació|nacido** {en NP} **el Y_Dat**	6,4	100
	LN	**X** {,} **(** {NP|Dat ,} **Y_Loc**	39,8	98
		X nació|nacido {el Dat ,} **en Y_Loc**	9,3	100
	DM	**X (** {NP , NP ,} **Dat** {-|;|, NP ,} **Y_Dat**	24,2	95
		X falleció|murió {en NP} **en Y_Dat**	3,5	95
	LM	**X** {,} **(NP ,** {NP ,} **DAT** {-|;|, Dat ,} **Y_Loc**	34,0	100
		X {,} **falleció|murió** {en Dat} **en Y_Loc**	1,2	100
	Pr.	**Y_Pr** {de NP ,} **X**	36,0	99
		X ,|es|fue {un|una} **Y_Pr**	5,2	96

Tabela 6.2: Melhores padrões (em função da precisão e da frequência de exemplos positivos) por língua e relação. Os elementos em parênteses rectos são —cada um deles— opcionais e as barras verticais representam disjunção. Por motivos de espaço, os padrões mostram-se simplificados, omitindo-se algumas flexões verbais e outros elementos opcionais. *DN* refere-se a `DatadeNascimento`; *LN*, `LocaldeNascimento`; *DM*, `DatadeMorte`; *LM*, `LocaldeMorte` e *Pr.* a `Profissão`. *%* é a percentagem de cada padrão no conjunto de orações positivas de cada relação. *Prec* é a sua precisão no corpus.

Restrições semânticas:

Antes de apresentar os diferentes atributos, é necessário apontar que a selecção das entidades candidatas restringiu-se às classes semânticas específicas para cada relação. Assim, **X** será sempre um nome próprio de pessoa e **Y** será data nas relações `DatadeNascimento` e `DatadeMorte`, localização em `LocaldeNascimento` e `LocaldeMorte`, e será um nome de profissão em `Profissão`. O impacto destas restrições será analisada através da avaliação de uma *baseline* que não utiliza esta classificação semântica.

6.3.1. Atributos primários

A primeira categoria de atributos não provém estritamente da análise linguística, mas inclui informação sobre a posição de **X** e **Y** na oração. Os atributos primários consistem nas seguintes informações (o seu valor provém do exemplo mostrado na Figura 5.1, na página 70):

– Posição absoluta de **X** na oração: 0

– Posição absoluta de **Y** na oração: 8

– Direcção da relação: **X_Y** (1) ou **Y_X** (2): 1

– Distância (em número de tokens) entre as duas entidades: 7

Apesar de que estes atributos não representam directamente a relação semântica, são amplamente utilizados na literatura, e testes preliminares mostraram que a sua utilização é positiva em combinação com outros atributos de carácter linguístico (definidos a seguir), os quais não incluem informação explícita sobre a posição das entidades. Assim, estes atributos —não utilizados nos testes do Capítulo 5— serão combinados com outras categorias que incluem conhecimento lexical, morfossintáctico, pseudo-sintáctico ou sintáctico.

6.3.2. Lexicais

Estes atributos utilizam elementos lexicais presentes na estrutura linguística. Foram avaliados dous tipos:

– *Tok*: incluem os tokens que ocorrem entre as duas entidades e alguns dos contextos anterior e posterior (definidos mais abaixo)

– *Lem*: igual que *Tok*, mas utilizando os lemas

Os tokens são atributos comuns utilizados para caracterizar estruturas linguísticas que contêm relações semânticas. A sua utilização como *sacos de palavras* adiciona informação lexical importante. Veja-se um exemplo dos atributos *Tok* extraídos do mesmo exemplo:

Oração: **X** (nascida em 10 de Junho de 1961) é uma **Y_Pr** americana

Tok: <(>, <nascida>, <DATA>, <)>, <é>, <uma>, <americana>

É importante referir que os atributos *Tok* e *Lem* não são simples *sacos de palavras*. *Tok* e *Lem* incluem a identificação das fronteiras de entidades complexas (como nomes próprios: por exemplo *Kimberley_Deal*) e expressões temporais (*10 de Junho de 1961*), devido à aplicação de diferentes módulos de reconhecimento de entidades mencionadas.

Em línguas com alto grau de flexão como o português ou o espanhol, a utilização de tokens pode provocar dispersão de dados. Tenha-se em conta, por exemplo, que nestas línguas cada verbo contém umas 50 formas diferentes. Um modo de generalizar estes atributos é a utilização de lemas, com os quais o exemplo anterior seria representado da seguinte maneira:

Lem: <(>, <nascer>, <DATA>, <)>, <ser>, <um>, <americano>

Repare-se que estes atributos também serão extraídos de orações similares, como "**X** (nascido em 1948) foi um **Y** americano". Assim, o uso de *Lem* reduz a dispersão de dados, apesar de perder alguma informação morfológica presente na flexão nominal e nas formas verbais.

O emprego da estratégia de *sacos de palavras* (contendo tokens, lemas e outros atributos) para representar os padrões implica a definição de dous patamares. Primeiro, os contextos anterior e posterior de **X** e **Y**, com o fim de caracterizar as relações definidas fora da janela intermédia: "A cidade de nascimento de **X** é **Y_Loc**". Se o classificador só utiliza os elementos intermédios, este tipo de relações não poderão ser extraídas. Porém, se estes contextos forem muito amplos, o classificador poderá ser treinado com ruído e elementos irrelevantes.

Segundo, o tamanho da janela intermédia (entre **X** e **Y**) também deve ser definida. Se esta for muito pequena (4, 5, 6 tokens) o número de atributos e a complexidade dos padrões reduz-se, mas o classificador terá baixa abrangência, já que muitas orações vão ser excluídas (como mostrou a Secção 6.2). O aumento do tamanho da janela incrementará a abrangência do sistema, aumentando também o número de atributos. Estas janelas serão definidas empiricamente na Secção 6.4.1.

6.3.3. Morfossintácticos

O conjunto de atributos pode ser ampliado incluindo informação morfossintáctica, por exemplo *PoS-tags*. Estas etiquetas podem ser adicionadas aos classificadores de duas maneiras: (i) enriquecendo os atributos baseados em lemas (*Lem_PoS*, como foi mostrado na Secção 5.3) ou adicionando um novo nível de abstracção linguística (*PoS*), que representaria uma estrutura linguística através das categorias morfossintácticas em vez das unidades lexi-

cais. Assim, incluindo estes atributos no classificador poderia melhorar-se o reconhecimento de padrões linguísticos. Os *PoS-tags* obtidos da oração de exemplo seriam os seguintes:

Oração: **X** (nascida em 10 de Junho de 1961) é uma **Y_Pr** americana

Lem_PoS: <Fa>, <nascer_VB>, <em_PS> <DATA>, <Fc>, <ser_VB>, <um_DT>, <americano_AD>

PoS: <Fa>, <VB>, <PS> <DATA>, <Fc>, <VB>, <DT>, <AD>

6.3.4. Pseudo-sintácticos

A análise sintáctica fornece informação sobre a função dos diferentes elementos da estrutura linguística, mas a utilização de analisadores automáticos pode trazer problemas como um elevado custo computacional ou a geração de ruído. Para além disso, línguas diferentes do inglês (como as analisadas nesta tese) podem não ter *parsers* disponíveis.

Tendo em conta estas assunções, é interessante avaliar alguns atributos que, de algum modo, representem informação pseudo-sintáctica mediante a codificação da posição de diferentes elementos numa oração. Assim, nos testes também foram utilizadas sequências dous lemas adjacentes (*Bigramas* de lemas) e de três (*Trigramas*):

Oração: **X** (nascida em 10 de Junho de 1961) é uma **Y_Pr** americana

Bigramas: <X_(>, <(_nascer>, <nascer_em>, <em_DATA>, <DATA_)>, <)_ser>, <ser_um>, <um_Y>, <Y_americano>

Trigramas: <X_(_nascer>, <(_nascer_em>, <nascer_em_DATA>, <em_DATA_)>, <DATA_)_ser>, <)_ser_um>, <ser_um_Y>, <um_Y_americano>

Outros atributos de carácter pseudo-sintáctico também foram avaliados, como padrões léxico-sintácticos intermédios e longos (utilizados no Capítulo 5). Contudo, várias avaliações provaram que não eram úteis em nenhum dos classificadores supervisionados para português e espanhol, dada a dispersão de dados nos corpora e a abordagem de classificação utilizada aqui. A principal razão da diferença de desempenho em relação aos testes do capítulo anterior deriva da estratégia de construção dos classificadores: enquanto no capítulo anterior os atributos de cada uma das instâncias eram os diferentes padrões em que um par ocorria, agora os atributos são extraídos unicamente da oração a classificar. Assim, é compreensível que um conjunto de padrões permita classificar um par com maior precisão do que um único padrão.

6.3.5. Sintácticos

Finalmente, o último conjunto de atributos codifica directamente a informação sintáctica. Este tipo de informação foi utilizada nos últimos anos por vários trabalhos, utilizando *parsers* para extrair as estruturas sintácticas que contêm relações semânticas.

A sintaxe codifica, num nível linguístico profundo, a função de cada elemento numa oração. Assim, as mesmas funções sintácticas podem ser extraídas de orações com estruturas de superfície muito diferentes. Os seguintes exemplos partilham a mesma estrutura sintáctica, apesar de a sua forma diferir notoriamente:

> *Exemplo 1*: **X** (nascida em 10 de Junho de 1961) é uma $\mathbf{Y}_{\text{(cantora)}}$ americana
>
> *Exemplo 2*: Muitos jornalistas musicais disseram que **X** foi provavelmente a melhor $\mathbf{Y}_{\text{(cantora)}}$ da América

Repare-se que nas orações anteriores, **X** é o sujeito do verbo *ser*, enquanto **Y** (*cantora*) é o atributo (sintáctico) do mesmo verbo.

Para representar esta informação como atributos, foram avaliadas três estratégias:

– *Deps*: dependências sintácticas individuais

– *SDPs*: caminho de dependências mais curto (entre **X** e **Y**)

– *Pths*: caminho de dependências completo

As dependências sintácticas (já utilizadas na Secção 5.3) são triplos que representam ligações entre dous elementos linguísticos relacionados (núcleo e dependente) através de uma etiqueta sintáctica. A primeira parte das dependências define a função sintáctica, sendo o núcleo e o dependente a segunda e a terceira, respectivamente.

> *Oração*: **X** (nascida em 10 de Junho de 1961) é uma **Y_Pr** americana
>
> *Deps*: <subj;ser;**X**>, <punct;nascer;Fa>, <adjn;**X**;nascer>, <cprep;nascer;em>, <term;em;DATA>, <punct;nascer;Fc>, <spec;**Y**;um>, <attr;ser;**Y**>, <modif;**Y**;americano>

Deste ponto de vista, as dependências são atributos similares aos bigramas de lemas, mas construídos através de (e fornecendo) informação sintáctica. Note-se que no exemplo anterior,

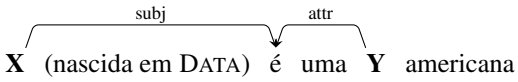


Figura 6.2: Exemplo do caminho de dependências mais curto entre **X** e **Y**.

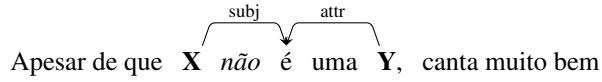


Figura 6.3: Exemplo do caminho de dependências mais curto entre **X** e **Y** que não representa uma relação semântica. Neste exemplo, informação crucial como a negação perde-se.

o primeiro e o penúltimo triplo são cruciais para definir a relação `Profissão`, enquanto os outros contêm informação extra. Esta observação inspirou o seguinte tipo de atributos, o caminho de dependências mais curto (SDP, do inglês *shortest dependency path*).

O SDP representa o caminho mais curto entre duas entidades numa árvore de dependências. A Figura 6.2 mostra o caminho o caminho de dependências mais curto entre **X** e **Y** na oração *Exemplo 1*. Assim, o SDP desta oração pode representar-se mediante o seguinte atributo: <subj;ser;**X**//attr;ser;**Y**>.

Apesar de que o SDP é frequentemente considerado como uma das melhores estratégias de representação uma relação entre dous elementos numa oração, alguns trabalhos têm indicado que informação importante (como a negação) pode perder-se utilizando o caminho de dependências mais curto (Wu e Weld, 2010). A Figura 6.3 mostra uma oração com o mesmo SDP que a anterior, mas que não contém a mesma relação semântica.

Assim, uma ligação diferente entre as entidades, utilizando o caminho de dependências completo (*Pths*, que inclui os dependentes dos elementos nucleares do *SDP*), foi avaliada. A seguir mostra-se um exemplo deste terceiro tipo de atributos sintácticos:

> *Oração*: **X** (nascida em 10 de Junho de 1961) é uma **Y_Pr** americana
>
> *Pths*: <subj;ser;**X**//punct;nascer;Fa//adjn;**X**;nascer//cprep;nascer;em//term;em;DATA//punct;nascer;Fc//spec;**Y**;um//attr;ser;**Y**//modif;**Y**;americano>

Este último tipo de atributos adicionará, provavelmente, informação irrelevante em muitos casos, mas também pode incluir elementos cruciais ignorados pelos *SDPs*. Como nas anteriores categorias, os atributos sintácticos foram avaliados de modo individual e em várias combinações.

6.4. Testes e avaliação

Esta secção mostra os resultados de diversos testes realizados para conhecer a efectividade de cada um dos atributos definidos acima.

Primeiro, apresentam-se alguns testes relacionados com o tamanho da janela de tokens entre as entidades. A seguir são os expostas várias avaliações de classificadores de aprendizagem automática treinados com os atributos linguísticos já referidos.

Os atributos são analisados utilizando duas estratégias: primeiro, avaliando individualmente o desempenho de cada um deles. Depois, utilizando uma abordagem *bottom-up*, começando com os atributos mais básicos e sistematicamente adicionado informação mais complexa.

Os testes foram realizados utilizando a implementação do algoritmo máquinas de vectores de suporte (SVM) de *libsvm* (Chang e Lin, 2011). Para avaliar as 10 relações (5 em cada língua), foi utilizada a estratégia de um-contra-todos.

Cada classificador foi treinado utilizando 80% dos dados, e avaliado no 20% restante (ambos os conjuntos seleccionados aleatoriamente). Para ajustar os parâmetros de SVM, realizou-se uma validação cruzada de 10 iterações.

6.4.1. Tamanho das janelas

O primeiro teste consiste na avaliação de diferentes tamanhos das janelas para a tarefa de extracção. Especificamente, este teste tenta responder as seguintes perguntas: qual é a melhor janela intermédia entre **X** e **Y**? E quais são as melhores janelas para o contexto anterior à primeira entidade e posterior à segunda?

Esta avaliação foi realizada com classificadores compostos de atributos *Primários* e *Tok* (Secção 6.3), em todas as relações e línguas.

Primeiro, os classificadores foram treinados utilizando uma janela intermédia de 2 tokens, e acrescentando 2 tokens mais em cada iteração, até atingir uma janela de 20 elementos. Nestes testes utilizaram-se 4 tokens nos contextos anterior e posterior às entidades.

A Figura 6.4 mostra os valores (*micro-average*) de precisão e *recall* dos 5 classificadores para português e espanhol.[3] O gráfico mostra que a utilização de uma janela intermédia de mais de 16 tokens não melhora a precisão dos sistemas. Contudo, uma janela maior permite os classificadores analisar mais orações, pelo que o *recall* aumenta. Lembre-se que a Figura 6.1

[3] Utilizando as métricas de avaliação definidas na Secção 4.3.

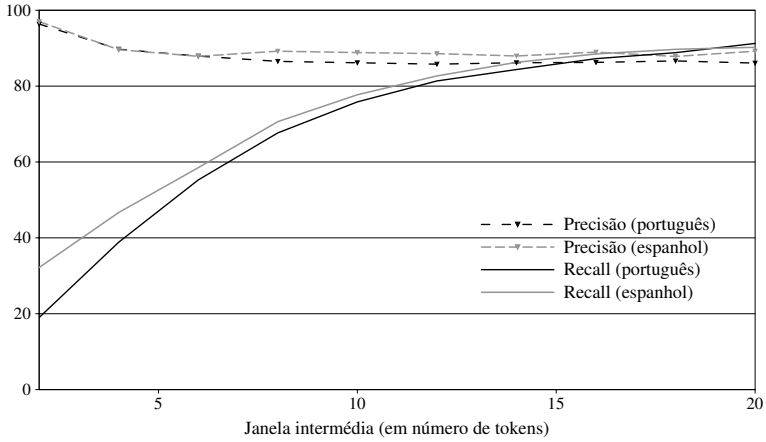


Figura 6.4: Precisão e *recall* (*micro-average*) de cinco classificadores para português e espanhol (um por relação) *versus* tamanho da janela intermédia (de 2 a 20, em número de tokens).

mostrou que os padrões positivos ocorrem mais frequentemente com janelas intermédias reduzidas (< 8). Seguindo os resultados da Figura 6.4, nos subsequentes testes será utilizada uma janela intermédia de 16 tokens.

Foram realizados mais testes para conhecer o melhor tamanho das janelas anterior e posterior. Como foi referido, o contexto de duas entidades **X** e **Y** consiste, para além da janela intermédia, da janela anterior à primeira entidade e e posterior à segunda. Utilizaram-se diferentes tamanhos destas janelas (de $0_{(\text{anterior})}/16_{(\text{intermédio})}/0_{(\text{posterior})}$ a $8/16/8$).

Porém, os resultados destes testes mostraram que modificar o número de elementos nos contextos anterior e posterior afecta muito ligeiramente o desempenho dos sistemas. A este respeito, mudar o tamanho destas janelas de 0 a 8 causou diferenças máximas de 1% de medida F, não produzindo tendências similares em cada um dos classificadores avaliados.

Em relação aos padrões, é importante mencionar que as novas estruturas obtidas utilizando janelas anterior e posterior de maior tamanho são mais difíceis de representar do que os contextos mais pequenos. Assim, os resultados médios sugerem que a utilização de janelas de 3 tokens antes e depois das entidades alvo é suficiente para obter um equilíbrio entre o desempenho e a eficiência, tendo em conta o número de atributos (lembre-se que as janelas utilizadas nos testes do Capítulo 5 tinham o mesmo tamanho).

6.4.2. Efectividade dos atributos

Para conhecer a efectividade dos diferentes atributos foi seguida a seguinte estratégia: primeiro utilizaram-se classificadores construídos com um só tipo de atributos (*Tok*, *Lem*, etc.). Assim, os resultados destes testes sugeriram que tipo de atributos representam melhor o nível linguístico a que pertencem (lexical, sintáctico, etc.). A seguir, várias combinações de atributos foram avaliadas, treinando modelos com os melhores atributos de cada nível de abstracção linguística com o fim de reduzir a redundância. Lembre-se que os atributos *Primários* (Secção 6.3.1) foram incorporados em todos os classificadores.

Restrição semântica e subespaços contextuais

Antes do conjunto principal de testes foram gerados dous modelos *baseline* com o objectivo de (i) conhecer a importância da restrição semântica das entidades candidatas e (ii) determinar o melhor modo de representar os elementos das janelas anterior e posterior no vector de atributos.

Como foi mostrado nas Secções 6.2 e 6.3, a abordagem utilizada neste trabalho utiliza conhecimento semântico para restringir os pares candidatos a serem analisados (classificando unicamente pares "pessoa – localização" para as relações `LocaldeNascimento` e `LocaldeMorte`, e não para `Profissão`, por exemplo). Assim, duas *baselines* foram utilizadas para avaliar o impacto do reconhecimento de entidades mencionadas no processo de aprendizagem.

Os classificadores *Baseline_1* (*B1*) utilizam mais pares candidatos do que os outros modelos, uma vez que não tiram proveito da classificação semântica dos nomes próprios. Assim, pares negativos (que não podem pertencer à relação testada: por exemplo, "localização – número" para `DatadeNascimento`) foram automaticamente adicionados aos conjuntos de treino e teste. O número de candidatos de cada subcorpus é assim $5,2$ (pt) e $4,9$ (es) vezes maior do que os corpora *standard*.

Os sistemas *Baseline_2* (*B2*) diferem de *B1* na restrição semântica dos candidatos, uma vez que *B2* aplica um filtro semântico para seleccionar os pares candidatos. Tanto os classificadores *B1* como os *B2* foram treinados com os atributos *Tok* extraídos de contextos $3/16/3$.

As duas primeiras colunas das Tabelas 6.3 e 6.4 contêm os resultados das duas *baselines* em português e em espanhol, respectivamente. Os resultados mostram-se para cada relação, com valores de *macro* e *micro-average*.[4]

A comparação entre os modelos *B1* e *B2* demonstra a efectividade da utilização da filtragem semântica, a qual incrementa em mais de 25% a medida F (*micro-average*) do processo de classificação. Esta restrição semântica permite que os modelos não classifiquem pares que normalmente não podem pertencer às relações alvo. Assim, melhoram-se os resultados tanto na precisão como no *recall*, excepto na relação `DatadeMorte`, cuja precisão desce $\approx 16\%$, embora com um incremento notório do *recall* ($\approx 60\%$).

Os modelos *B2* e *Lx1* foram treinados com o mesmo conjunto de atributos, mas o último difere das *baselines* no modo de representar os contextos anterior e posterior no vector de atributos. À diferença das *baselines* —que unificam os três espaços no mesmo vector—, o classificador *Lx1* utiliza três subespaços (*anterior*, *intermédio* e *posterior*), pelo que as diferenças entre *B2* e *Lx1* (Tabelas 6.3 e 6.4) devem-se ao modo como estes elementos se organizam no vector.

Novamente, o desempenho de todos os classificadores melhora com a representação de atributos proposta. Os subespaços contextuais permitem que os classificadores representem melhor as estruturas nas quais a relação acontece, identificando a importância da ordem relativa dos elementos lexicais em cada padrão.

As diferenças entre os resultados de *B2* e *Lx1* seguem tendências similares: em geral, a medida F melhora entre 2% e 4% no novo cenário. Porém, as relações cujo atributo (**Y**) é uma data têm uma melhora maior (entre 9% e 10% em português e entre 5% e 9% em espanhol). Como mostraram as análises dos padrões (Secção 6.2), estas relações ocorrem frequentemente em estruturas biográficas que contêm datas de nascimento e morte, pelo que a utilização de subespaços ajuda os classificadores a desambiguar entre muitos pares candidatos para `DatadeNascimento`/`DatadeMorte`. Assim, esta estratégia será aplicada no treino de todos os modelos seguintes.

Níveis lexical e morfossintáctico

O seguinte teste pretende conhecer qual é o melhor modo de representar as unidades lexicais nas orações (Secção 6.3.2). Para isto foram avaliados três modelos: *Lx1* (como foi

[4] DN significa `DatadeNascimento`; LN, `LocaldeNascimento`; DM, `DatadeMorte`; LM, `LocaldeMorte` e Pr significa `Profissão`. Adicionalmente, *Ma* e *Mi* indicam os valores de *macro* e *micro-average*, respectivamente.

Rel		Modelo									
		B1	*B2*	*Lx1*	*Lx2*	*Ms2*	*BiG*	*TrG*	*Dep*	*Pth*	*SDP*
DN	*P*	69.3	83.8	93.4	93.8	88.8	91.3	81.3	70.3	59.2	60.7
	R	33.8	87.4	95.2	95.2	95.7	96.1	98.6	80.2	94.7	94.7
	F	45.5	85.6	94.3	**94.5**	92.1	93.7	89.1	74.9	72.9	74.0
LN	*P*	86.5	94.7	96.8	94.6	93.3	95.7	95.4	93.5	60.1	98.2
	R	72.3	91.4	94.1	95.7	92.2	95.7	96.9	96.1	96.5	63.3
	F	78.7	93.0	95.5	95.2	92.7	95.7	**96.1**	94.8	74.1	77.0
DM	*P*	92.3	73.9	82.5	87.0	88.5	82.3	85.0	58.8	0	100
	R	16.6	68.3	78.0	78.6	69.0	83.5	74.5	53.1	0	13.1
	F	28.1	71.0	80.1	82.6	77.5	**82.9**	79.4	55.8	0	23.2
LM	*P*	68.8	89.3	95.7	95.8	97.9	97.3	97.5	93.8	100	98.5
	R	16.6	85.5	87.6	88.5	71.0	86.1	81.6	81.6	4.2	19.9
	F	26.8	87.4	91.5	**92.0**	82.3	91.4	88.8	87.2	8.1	33.2
Pr	*P*	80.3	84.5	88.1	89.0	90.1	83.6	82.6	86.5	73.1	74.8
	R	76.8	92.1	92.8	92.3	91.9	96.1	96.5	91.7	82.0	83.3
	F	78.5	88.1	90.4	90.6	**91.0**	89.4	89.0	89.0	77.7	78.9
Ma	*P*	79.4	85.2	91.3	92.0	91.7	90.0	88.4	80.6	58.5	86.4
	R	43.2	85.0	89.5	90.1	83.9	91.5	89.6	80.5	55.7	54.9
	F	51.5	85.0	90.3	**91.0**	87.1	90.6	88.5	80.4	46.6	57.2
Mi	*P*	79.8	86.3	91.6	92.1	91.8	89.6	87.7	84.0	65.6	76.1
	R	49.0	87.2	90.6	91.0	85.2	92.3	91.0	84.4	59.9	59.0
	F	60.8	86.8	91.1	**91.6**	88.4	90.9	89.3	84.2	62.6	66.5

Tabela 6.3: Precisão, *recall* e Medida F por relação e modelo em português. As abreviaturas da relações (*Rel*) estão expandidas na nota de rodapé 4.

dito, construído com atributos *Tok*), *Lx2* (com *Lem*) e *Ms2* (baseado em *PoS*), que contém informação morfossintáctica.[5]

Os resultados destes modelos (Tabelas 6.3 e 6.4) mostram, de acordo com as hipóteses colocadas na Secção 6.3.2, que a utilização de lemas em vez de tokens (*Lx2 versus Lx1*) generaliza os padrões ao reduzir a dispersão de dados na aprendizagem dos classificadores.

Esta generalização provoca melhoras leves em todos as relações nas duas línguas avaliadas ($\approx 0,5\%$), excepto em `DatadeMorte` em espanhol. Contudo, esta melhora (a respeito de *Lx1*) não é estatisticamente significante.[6] Uma análise pormenorizada dos resultados indica que os modelos com lemas como *sacos de palavras* classifica correctamente alguns candi-

[5] Os atributos *Lem_PoS* também foram avaliados, mas os resultados são omitidos porque foram quase idênticos aos modelos *Lx2*.

[6] A significância estatística foi calculada com o *teste t*, onde *p-valor* < 0.05.

Rel		**Modelo**									
		B1	*B2*	*Lx1*	*Lx2*	*Ms2*	*BiG*	*TrG*	*Dep*	*Pth*	*SDP*
BD	*P*	63.2	82.0	96.3	97.0	86.2	92.4	85.3	58.8	100	83.3
	R	25.8	89.8	93.4	95.2	86.2	94.0	97.0	70.1	1.8	12.0
	F	36.6	85.7	94.8	**96.1**	86.2	93.2	90.8	63.9	3.5	20.9
BP	*P*	74.5	90.0	90.8	94.6	88.4	91.8	85.9	86.1	43.1	100
	R	38.8	82.7	90.8	89.8	85.7	90.8	86.7	75.5	95.9	32.7
	F	51.0	86.2	90.8	**92.2**	87.1	91.3	86.3	80.4	59.5	49.2
DD	*P*	100	86.2	91.8	90.0	83.0	91.4	93.8	64.3	66.7	100
	R	9.4	79.0	84.1	83.7	77.6	84.6	77.1	55.6	0.9	10.4
	F	17.1	82.4	87.8	86.7	80.2	**87.9**	84.6	59.7	1.8	18.6
DP	*P*	0	93.1	95.7	98.0	97.4	97.5	96.0	90.4	100	92.9
	R	0	88.7	93.9	93.0	87.8	92.5	89.2	92.5	2.8	6.1
	F	0	90.9	94.8	**95.4**	92.4	94.9	92.5	91.4	5.5	11.5
Pr	*P*	87.9	90.4	93.1	93.5	93.9	91.5	89.7	87.4	75.3	84.5
	R	70.6	92.3	93.1	92.3	92.3	97.3	97.9	96.3	99.5	78.4
	F	78.3	91.3	93.1	92.9	93.1	**94.3**	93.6	91.6	85.7	81.3
Ma	*P*	65.1	88.3	93.5	94.6	89.8	92.9	90.1	77.4	77.0	92.1
	R	28.9	86.5	91.1	90.8	85.9	91.8	89.6	78.0	40.2	27.9
	F	36.6	87.3	92.3	**92.6**	87.8	92.3	89.5	77.4	31.2	36.3
Mi	*P*	84.6	89.0	93.5	94.2	91.3	92.5	90.2	80.5	68.8	85.9
	R	49.0	88.5	91.6	91.2	87.9	93.5	92.1	84.1	55.0	43.6
	F	62.1	88.7	92.6	92.6	89.6	**93.0**	91.2	82.3	61.1	57.9

Tabela 6.4: Precisão, *recall* e medida F por relação e modelo em espanhol. As abreviaturas estão expandidas na nota de rodapé 4.

datos com formas verbais menos frequentes (como por exemplo o particípio feminino plural "nacidas" em espanhol).

Os modelos *Ms2* foram construídos com um nível maior de generalização, ao substituir a informação lexical dos padrões pela sua representação morfossintáctica.

De novo, os padrões biográficos (que contêm muitos signos de pontuação) são relativamente bem cobertos pelos atributos *PoS*, mas a remoção da informação lexical implica uma descida geral no desempenho destas classificadores (de $\approx 3\%$), excepto em `Profissão`, onde representar os adjectivos (e outras classes de palavras pouco relevantes para a definição da relação) através da sua categoria morfossintáctica simplifica o conjunto de atributos dos classificadores.

Estes testes confirmam a validade dos atributos baseados em lemas (*Lem*) para a representação das unidades lexicais. As seguintes avaliações medem a efectividade dos modelos baseados em atributos pseudo-sintácticos.

Nível pseudo-sintáctico

Foram criados dous modelos diferentes utilizando informação pseudo-sintáctica: *BiG*, com bigramas de lemas, e *TrG*, com trigramas de lemas.

Apesar de que os resultados mostram variação entre as diferentes relações e línguas, em geral os bigramas de lemas têm um melhor comportamento do que os trigramas. Em português, os trigramas funcionaram melhor em `LocaldeNascimento`, devido à grande quantidade de padrões nos quais as entidades alvo podem ser cobertas por um único atributo <**X_(_Y_Loc**>. De modo similar, os resultados de `Profissão` tiveram menos variação, uma vez que os padrões desta relação são frequentemente bem representados tanto por bigramas como por trigramas de lemas.

Porém, os melhores modelos pseudo-sintácticos não melhoram nitidamente os classificadores *Lx2*, cujos resultados são ligeiramente melhores em quase todas as relações em português. Contudo, os valores (*micro-average*) de *BiG* em espanhol são superiores aos de *Lx2*, porque a melhora dos classificadores *BiG* só acontece em `Profissão` e `DatadeMorte`. Neste caso, unicamente as melhoras de *BiG* para `Profissão` foram estatisticamente significantes.

Nível sintáctico

Por último, foram comparadas três estratégias diferentes para avaliar a efectividade da análise sintáctica na extracção de relações semânticas: modelos *Dep*, que contêm dependências sintácticas como os atributos principais; *Pth*, com caminhos completos de dependências, e *SDP*, cujos classificadores se construíram com os caminhos de dependências mais curtos entre as duas entidades.

As três últimas colunas das Tabelas 6.3 e 6.4 contêm os resultados dos modelos construídos com informação sintáctica. Lembre-se que nalguns padrões os caminhos de dependências (tanto o mais curto como o completo) não foram extraídos. Este facto foi provocado por duas razões principais: (i) falta de pontuação nalgumas orações —o que implicou erros no *parsing*— e (ii) abrangência dos analisadores, que não conseguiram estabelecer algumas de-

pendências entre as entidades. Assim, 40% (pt) e 54% (es) dos dados não têm atributos *Pths*, enquanto o *SDP* não foi extraído de 54% (pt) e 63% (es) dos corpora.

Portanto, os valores de *recall* destes classificadores são notoriamente baixos (excepto em `Profissão` e `DatadeNascimento` em português, cujos padrões foram melhor analisados pelos *parsers*).

Em geral, as dependências individuais não têm valores altos de precisão, salvo quando contêm conhecimento não ambíguo (por exemplo, <subj;nascer;**X**> e <em;nascer;**Y**>) pelo que os resultados médios são piores do que os modelos *B2* e *BiG*.

Os classificadores *Pth* e *SDP* dependem nitidamente da abrangência dos analisadores e da tipologia dos padrões. Os valores de precisão destes modelos superam *Lx2* e *BiG* nalgumas relações (com resultados de precisão de 100% nalguns testes). Contudo, outros padrões que contêm *Pths* e *SDP* ambíguos (ou que não têm estes atributos) provocaram desempenhos mais baixos.

Entre estes dous modelos, *SDP* parece ter melhor precisão do que *Pth*, apesar de que ambos apresentam problemas de *recall* (até em orações cujos atributos *Pths* e *SDPs* foram bem extraídos).

Da realização deste teste podemos inferir que a utilização isolada de atributos baseados em análise sintáctica não é adequada para a construção de sistemas de extracção de relações (tendo em conta a precisão e a abrangência dos analisadores). Contudo, a alta precisão de alguns modelos *Pth* e *SDP* sugere que esta informação pode ser positiva em combinação com outros atributos baseados em análises lexicais e/ou pseudo-sintácticas.

6.4.3. Combinações de atributos

Para optimizar as combinações de atributos em termos de desempenho e de eficiência, elementos diferentes de cada nível linguístico (lexical, pseudo-sintáctico e sintáctico) foram combinados com o fim de identificar atributos com informação redundante e de construir os melhores classificadores.

Primeiro, os melhores atributos lexicais (*Lem*) e pseudo-sintácticos (*Bigramas*) foram combinados para treinar modelos *L/B*. A seguir, *Lem* também foram utilizados para criar classificadores com atributos *Pths* (*L/P*).[7] Finalmente, combinaram-se três níveis diferentes

[7] *Pths* e *SDPs* tiveram um comportamento muito similar nos testes individuais, mas os primeiros produziram resultados ligeiramente melhores nos modelos combinados ($\approx 0,3\%$ melhor do que *SDPs*).

Relação		**Modelo**					
		Português			**Espanhol**		
		L/B	*L/P*	*L/B/S*	*L/B*	*L/P*	*L/B/S*
`DatadeNascimento`	*Precisão*	92.2	93.4	91.4	97.0	97.6	97.0
	Recall	97.1	95.7	97.1	95.8	95.2	95.8
	F1	**94.6**	94.5	94.2	**96.4**	96.4	**96.4**
`LocaldeNascimento`	*Precisão*	96.1	95.0	95.8	95.7	95.8	94.6
	Recall	94.9	96.1	96.1	90.8	91.8	89.8
	F1	95.5	95.5	**95.9**	93.2	**93.8**	92.2
`DatadeMorte`	*Precisão*	81.5	86.8	82.8	91.5	89.7	92.0
	Recall	84.8	81.4	82.8	85.1	85.1	86.5
	F1	83.1	**84.0**	82.8	88.1	87.3	**89.2**
`LocaldeMorte`	*Precisão*	97.3	96.4	96.6	97.6	96.6	97.6
	Recall	87.0	88.5	86.7	93.9	93.0	93.4
	F1	91.9	**92.3**	91.4	**95.7**	94.7	95.4
`Profissão`	*Precisão*	88.4	89.7	90.8	94.5	93.1	94.4
	Recall	93.6	92.1	93.4	94.8	93.1	95.3
	F1	91.0	90.9	**92.1**	94.7	93.1	**94.9**
Macro-average	*Precisão*	91.1	92.3	91.5	95.2	94.5	95.1
	Recall	91.5	90.8	91.2	92.1	91.6	92.2
	F1	91.2	**91.5**	91.3	**93.6**	93.0	**93.6**
Micro-average	*Precisão*	91.5	92.5	92.2	94.9	93.9	94.9
	Recall	91.0	91.4	91.8	92.9	91.9	93.2
	F1	91.7	91.9	**92.0**	93.9	92.9	**94.0**

Tabela 6.5: Precisão, *recall* e medida F por relação, língua e modelo. Os classificadores combinam atributos de diferentes níveis de análise linguística: *L*, lexical (`Lem`); *B*, pseudo-sintáctica (`Bigramas`); *P* e *S*, sintáctica (`Pths` e `SDPs`, respectivamente).

de análise linguística nos classificadores *L/B/S*, utilizando a melhor combinação possível dos atributos fornecidos por estas análises: *Lem*, *Bigramas* e *SDPs*.

A Tabela 6.5 contém os resultados destas três combinações para cada relação e língua, para além dos valores médios.

O primeiro modelo (*L/B*), que combina atributos baseados em análises lexicais e pseudo-sintácticas, teve resultados estatisticamente melhores do que os mesmos atributos utilizados individualmente. Os classificadores para português melhoraram $0,6\%$ e $0,8\%$ os modelos *Lx2* e *BiG*, respectivamente. Em espanhol, a melhora foi de $1,27\%$ e $0,8\%$. Para além disso, esta combinação melhorou tanto em precisão como em *recall* todos os classificadores (com uma única excepção: `LocaldeMorte` em português).

Assim, os atributos lexicais e pseudo-sintácticos, apesar de construídos com a mesma informação de base (lemas), são adequados para serem combinados em classificadores estatísticos, sem necessidade de utilizar ferramentas de maior custo computacional como os *parsers*.

A seguinte combinação de atributos (*L/P*) foi realizada utilizando conhecimento lexical (*Lem*) e sintáctico (*Pths*). Em português, o caminho de dependências completo ajudou a superar o desempenho dos modelos *Lx2* em todas as relações (com uma melhora de $\approx 0,4\%$ nos valores *micro-average* da medida F). Estes modelos também tiveram melhor desempenho (embora só $0,2\%$) do que as combinações anteriores (*L/B*).

Em espanhol, a utilização de informação sintáctica junto com atributos baseados em lemas causou ligeiras melhoras ($0,2\%$) nos classificadores, quando comparados com o uso isolado de lemas. Contudo, os resultados destas combinações não superam os modelos de bigramas de lemas (*BiG* e *L/B*). É preciso lembrar, contudo, que o *parser* de português extraiu mais atributos *Pths* e *SDPs* do que o espanhol, como foi referido na Secção 6.4.2.

Finalmente, a terceira combinação (*L/B/S*) analisou o uso dos atributos *Tok*, *Bigramas* e *SDPs* nos mesmos classificadores.

Em português, esta última combinação tem melhor desempenho do que os modelos *L/B* ($\approx 0,3\%$) e *L/P* ($0,07\%$), especialmente naquelas relações onde *SDP* tinha atingido bons resultados ($> 75\%$ de medida F): `LocaldeNascimento` e `Profissão`. Contudo, os classificadores *L/B/S* não melhoram o desempenho dos modelos anteriores para `DatadeMorte` e `LocaldeMorte`, obtendo de facto piores resultados (*micro-average*) do que os modelos *L/B* e *L/P*.

Em espanhol, a adição de atributos sintácticos não causou melhoras notórias nos classificadores, embora os resultados médios são significativamente melhores do que aqueles obtidos pelos sistemas criados com um único tipo de atributos.

Em suma, este último conjunto de testes demonstrou que a combinação de diferentes atributos que não produzem redundância de informação melhora o desempenho geral dos classificadores. Isto significa que atributos complementares permitem construir melhores classificadores. Tendo em conta que para línguas como o português ou o espanhol as ferramentas de análise sintáctica não são abundantes, pode ser interessante avaliar este tipo de atributos com analisadores mais maduros.

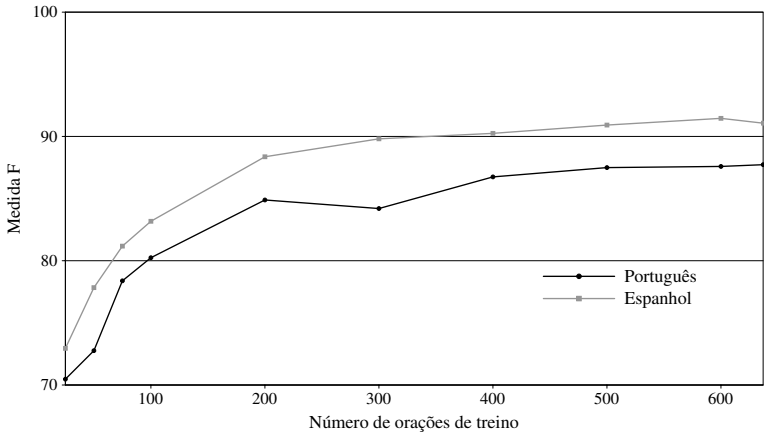


Figura 6.5: Medida F (*micro-average*) de cinco classificadores para português e espanhol *versus* dados de treino (de 25 a 637 orações). O teste foi realizado em 200 exemplos aleatórios.

6.4.4. Curva de aprendizagem

A propósito da quantidade de dados de treino necessários para criar os classificadores, a Figura 6.5 mostra a curva de aprendizagem (valores *micro-average*) dos classificadores *L/B/S* em português e espanhol.

Os testes foram realizados utilizando conjuntos aleatórios de 25, 50, 75, 100, 200, 300, 400, 500, 600 e 637 orações de treino.[8] As curvas indicam que os classificadores têm picos de desempenho com perto de 500 orações, obtendo melhoras muito pequenas ao utilizar mais dados de aprendizagem. Contudo, outros testes com relações para as quais existem mais dados (`LocaldeNascimento` e `Profissão`) sugerem que os classificadores começam a produzir um sobre-ajuste com perto de 1.400 orações, momento em que a medida F não melhora (ou até piora, nalgum caso).

Por último, na Figura 6.5 também se pode observar como a curva de aprendizagem segue tendências similares em cada língua.

6.4.5. Análise de erros

Com o objectivo de conhecer as principais fontes de erros produzidas pelos classificadores, foi realizada uma análise pormenorizada dos falsos positivos e negativos dos resultados

[8] 637 porque é o número de orações de treino da relação com menos dados, `DatadeNascimento` (pt).

de *L/B/S*, tendo em conta todos os erros de todas as relações em português e em espanhol. Os erros foram classificados de acordo com a sua tipologia. Para além disso, calculou-se a percentagem (*micro-average*) de cada tipo de erro nas duas línguas:

Ambiguidade e erros dos *parsers* (≈43%)

Este tipo de erros ocorreu quando a informação codificada pelos atributos lexicais e/ou pseudo-sintácticos não foi suficiente para representar a relação entre as duas entidades, e o caminho de dependências mais curto (que poderia desambiguar o padrão) não foi extraído ou é incorrecto. A seguinte oração (falso positivo para `LocaldeNascimento`) exemplifica esta categoria de erros:

Oração: [...] filho de PESSOA e de **X** nasceu no **Y_Loc** [...]

Padrão: [PESSOA e de]$_{\text{anterior}}$ **X** nascer em o **Y_Loc** [...]$_{\text{posterior}}$

Padrões sem informação explícita (≈25%)

Outra classe compõe-se de erros que ocorreram quando o padrão (ou até a oração inteira) não continha informação explícita sobre a relação alvo, o que produziu um falso negativo. O seguinte exemplo mostra um padrão onde a relação `DatadeNascimento` entre **X** e **Y_Dat** não é explícita, e a informação lexical principal encontra-se fora do escopo dos atributos extraídos:

Oração: Teve sete filhos: PESSOA em DATA, PESSOA em DATA, **X** em **Y_Dat** [...]

Padrão: [em DATA ,]$_{\text{anterior}}$ **X** em **Y_Dat** [...]$_{\text{posterior}}$

Padrões pouco frequentes (≈15%)

Outros pares positivos foram classificados erroneamente (falsos negativos) quando apareceram em padrões pouco frequentes (com menos de dous exemplos no corpus de treino). Alguns destes casos podem ser classificados correctamente se contêm atributos relevantes (como o caminho de dependências), mas de modo geral produziram falsos negativos:

Oração: [...] **X**, que era apenas um homem da zona rural de **Y_Loc** [...]

Padrão: **X**, que ser apenas um homem de a zona rural de **Y_Loc**

Caminho de dependências mais curto ambíguo (≈14%)

Como foi mostrado durante a avaliação, os atributos baseados no caminho de dependências mais curto têm alta precisão. Contudo, podem afectar negativamente se forem ambíguos em exemplos positivos e negativos. O seguinte par de orações contém, respectivamente, um falso negativo e um verdadeiro positivo para `Profissão` em espanhol, representados pelo mesmo *SDP*:

Oração 1: **X** y un **Y_Pr** fueron dos de los homenajeados [...]

Oração 2: **X**, poeta y **Y_Pr**, recibió el patrocinio de [...]

SDPs: <coord;y;**X**//coord;y;**Y_Pr**>

Neste tipo de cenário, os modelos *L/B/S* classificaram as duas orações como positivas para `Profissão`.

Para alem disso, os *SDPs* extraídos de orações biográficas —que normalmente contêm dependências não lexicais entre entidades e a pontuação circundante— não têm alta precisão, uma vez que aparecem indistintamente tanto em padrões positivos como negativos.

Outros erros (≈3%)

Finalmente, outros erros com diferentes origens foram causados por (i) problemas de lematização e (ii) erros no reconhecimento de entidades mencionadas.

Um dos poucos exemplos de erros produzidos pela lematização inclui o particípio "morto", incorrectamente lematizado como "matar" (em vez de "morrer"). Neste caso, um par de `LocaldeMorte` como o seguinte foi classificado como negativo para essa relação:

Oração: **X**, morto em **Y_Loc** [...]

Padrão: **X** , matar em **Y_Loc**

Outros erros menos frequentes dependeram da ferramenta REM, que pode classificar como nome comum um nome de pessoa, ou vice-versa. No seguinte exemplo, "Belén" (localização) foi incorrectamente classificada como nome comum, pelo que não foi uma entidade candidata para a relação `LocaldeMorte`: "**X** (LOCALIZAÇÃO, DATA - Belén, DATA)".

6.4.6. Discussão

Esta secção discute brevemente os resultados dos testes realizados, tendo em conta tanto as hipóteses formuladas durante a apresentação dos atributos (Secção 6.3) como os trabalhos relacionados analisados no Capítulo 4.

Primeiro, é preciso referir que a filtragem dos pares candidatos pela sua classe semântica permite que os sistemas melhorem em medida F em mais de 25% (*micro-average*), aumentando tanto a precisão como o *recall*. Contudo, à diferença de outros trabalhos (Kambhatla, 2004; Zhou *et al.*, 2005), a classificação semântica de outros nomes próprios que co-ocorrem nos mesmos padrões não contribuiu para um melhor desempenho. Para além disso, a utilização de vários subespaços para a representação dos diferentes contextos melhorou o funcionamento de todos os classificadores.

É importante notar que, depois de as entidades candidatas serem correctamente reconhecidas por um sistema REM, classificadores simples (que contêm só atributos *Primários* e lexicais) são suficientes para atingir bons valores de precisão e *recall* (sobre 91% e 92% em português e espanhol, respectivamente). Assim, pode inferir-se que (i) é factível construir ontologias de nomes próprios de alta qualidade com pouco esforço manual e que (ii) abordagens que utilizam classificadores supervisionados são úteis para a tarefa de extracção de relações de domínio biográfico.

Em relação à efectividade dos atributos, os primeiros testes mostraram que a generalização das unidades lexicais através da lematização (*Lem*) melhora o desempenho dos classificadores. Este facto está em concordância com os testes realizados por Agichtein (2005), que decidiu utilizar palavras (tokens) e *stems* (raízes). O uso de etiquetas morfossintácticas, contudo, não produziu nenhuma melhora nos testes realizados (salvo nalguns padrões cuja dependência da informação lexical é mínima, como "**Y_Pr X**" (já utilizados em Mann (2002)).

A respeito dos atributos pseudo-sintácticos, os testes realizados neste capítulo mostraram que os bigramas de lemas foram aqueles que melhores representaram este nível de análise. Os trigramas de lemas também tiveram bom desempenho em contextos nos quais as entidades estão próximas, mas dependem de outros atributos para representar melhor as relações. Para além disso, os padrões léxico-sintácticos tiveram um impacto baixo nos classificadores, à diferença da estratégia apresentada no Capítulo 5.

Apesar de que a análise sintáctica pode ser computacionalmente custosa e menos precisa do que outras tarefas de PLN (lematização, REM), fornece informação útil. Assim, e embora os atributos sintácticos não funcionassem bem individualmente, a sua combinação com in-

formação lexical e pseudo-sintáctica permitiu que melhorasse o desempenho dos sistemas de classificação. Contudo, é importante referir que alguns dos caminhos de dependências mais curtos podem ser uma fonte de erros se representarem entidades ligadas por pontuação e/ou coordenação, e não por unidades lexicais.

Os testes realizados com várias combinações de atributos indicaram que a utilização de informação complementar, obtida de diferentes níveis de representação linguística, permite construir melhores classificadores com menos informação redundante. A este respeito, as melhores combinações (tanto para português como para espanhol), foram aquelas baseadas em lemas, bigramas de lemas e *SDPs*. Para além disso, os testes realizados também mostraram que usando unicamente dados lexicais (*Lem*) e pseudo-sintácticos (*Bigramas*) é possível construir classificadores de alta precisão sem necessidade de aplicar análise sintáctica.

Em relação a isto, é importante lembrar que os *parsers* utilizados neste trabalho —mas também outros analisadores *estado-da-arte* treinados com corpora livres— não produziram árvores de dependências completos em quase a metade dos dados, pelo que os resultados só podem ser considerados como preliminares. Assim, o conhecimento fornecido pelos atributos sintácticos não adicionou sempre nova informação aos modelos, uma vez que alguns caminhos de dependências são similares aos bigramas de lemas.

Portanto, alguns dos resultados do presente capítulo são similares aos apresentados em Jijkoun *et al.* (2004) e Jiang e Zhai (2007), trabalhos que concluíram que atributos mais complexos poderiam piorar o desempenho dos classificadores (ou melhorar o *recall* descendo a precisão). Contudo, a informação sintáctica tem sido apontada como útil em trabalhos prévios de extracção de relações (Kambhatla, 2004; Zhou *et al.*, 2005; Mintz *et al.*, 2009). Por conseguinte, seria interessante analisar o desempenho dos atributos sintácticos com *parsers* e corpora diferentes.

Do dito até aqui podem extrair-se várias conclusões úteis para melhorar a extracção de entidades relacionadas semanticamente: primeiro, uma análise prévia dos padrões nos quais ocorrem as relações alvo pode ser útil para adaptar os atributos (e a sua representação) durante o desenho dos classificadores. A este respeito, trabalhos como Garera e Yarowsky (2009) introduziram diferentes estratégias desenhadas *ad-hoc* para a extracção de factos.

Segundo, os erros de lematização podem ser evitados analisando aquelas unidades lexicais que são cruciais para as relações desejadas (por exemplo, os verbos “morrer” ou “nascer” para relações relacionadas com o nascimento ou a morte).

Terceiro, se os padrões mostram uma dependência forte de sinais de pontuação ou de estruturas de coordenação, os atributos sintácticos podem tornar-se pouco úteis, uma vez que as análises dos *parsers* representam melhor as relações entre elementos lexicais.

Por último, e embora o desempenho dos classificadores avaliados é o suficientemente bom para extrair com precisão exemplos das relações desejadas, é preciso mais trabalho para conhecer se os métodos apresentados neste capítulo são aptos para a extracção de diferentes relações e textos (Grishman, 2010).

6.5. Conclusões

O trabalho apresentado neste capítulo avaliou diferentes atributos de base linguística para conhecer o seu impacto no desenvolvimento de classificadores supervisionados para a extracção de relações. Para além disso, analisou as melhores combinações de atributos evitando a incorporação de informação redundante no sistema de extracção.

Assim, complementa-se o trabalho apresentado no capítulo anterior, utilizando um maior conjunto de relações e de atributos, bem como uma abordagem de classificação diferente.

Os testes realizados mostraram, primeiro, que é preferível a utilização de lemas em vez de tokens, e que o conhecimento semântico é crucial para a filtragem dos candidatos. Para além disso, a combinação de atributos lexicais com informação pseudo-sintáctica como os bigramas de lemas também melhorou o desempenho dos classificadores. Finalmente, a análise sintáctica forneceu conhecimento útil para ser combinado com atributos provenientes de outros níveis de análise, apesar de que o benefício geral para o sistema não foi muito notório.

Tendo isto em conta, os resultados também destacaram que é possível criar classificadores de alta qualidade até para línguas com poucos recursos, dado que combinações de atributos obtidos mediante lematização são suficientes para treinar classificadores com alta fiabilidade.

Por último, o trabalho realizado neste capítulo também permitiu disponibilizar dous corpora (para português e espanhol) com anotação manual de 10 relações semânticas (5 para cada língua) de domínio biográfico, bem como diferentes análises dos contextos em que as entidades relacionadas ocorrem.

CAPÍTULO 7

EXTRACÇÃO DE RELAÇÕES COM BASE EM REGRAS

7.1. Introdução

Os sistemas de extracção de relações em domínio fechado podem dividir-se em dous grandes blocos em função da estratégia utilizada: para além dos métodos de aprendizagem automática (Capítulos 5 e 6), que treinam classificadores utilizando atributos que representam o espaço linguístico onde ocorrem as relações, as estratégias baseadas em regras transformam as mesmas estruturas em padrões ou regras de carácter léxico-sintáctico, que se aplicam sobre novos corpora para a extracção de pares relacionados.

A ER segue a assunção de que algumas regularidades linguísticas (e.g., padrões léxico-sintácticos) representam o mesmo tipo de conhecimento semântico. Contudo, um dos principais problemas deste tipo de abordagens é que pequenas variações em pontuação, modificação adjectival ou adverbial, etc. podem impedir encontrar os padrões adequados. Assim, um mesmo par de termos relacionados pode ocorrer numa variedade, teoricamente infinita, de orações (no seguinte exemplo, representando a relação `LocaldeNascimento`):

– "*López Bouza* nasceu na localidade galega de *Ferrol*"

– "*López Bouza* nasceu em *Ferrol*"

– "*López Bouza* nascia no município de *Ferrol*"

– "*López Bouza*, nascido na localidade corunhesa de *Ferrol*"

Tanto as estratégias de aprendizagem automática como os sistemas baseados em regras evitam o problema da dispersão (i) utilizando um maior número de exemplos de aprendizagem —e utilizando técnicas de de lematização (como mostrou o Capítulo 6)—, ou (ii) aplicando *parsers* que identificam os constituintes e as suas funções sintácticas em cada oração. Porém, a obtenção de dados de treino de alta qualidade não é sempre factível, já que pode ser necessário um esforço manual de anotação ou de correcção (como foi visto no Capítulo 6). Além disso, os *parsers* para línguas diferentes do inglês podem não ser disponibilizados de modo livre, ou fazer análises parciais ou inferiores ao estado-da-arte.

No presente capítulo apresenta-se uma nova técnica de extracção de relações que simplifica os contextos linguísticos mediante uma análise sintáctica parcial. Esta simplificação permite que regras genéricas de extracção, baseadas em padrões léxico-sintácticos, melhorem a sua abrangência na obtenção de pares relacionados.

Os padrões são seleccionados de modo semiautomático em função da sua precisão, evitando assim o possível ruído gerado na obtenção de corpus mediante supervisão-distante. Depois, os padrões mais precisos são generalizados aplicando o algoritmo *longest common string* (já utilizado no Capítulo 5), sendo estes padrões genéricos adicionados como regras sintáctico-semânticas numa gramática de dependências.

O método proposto é avaliado, para português e espanhol, com duas relações semânticas do domínio biográfico, tanto em conjuntos de teste como nas versões completas da Wikipedia para estas línguas e em textos jornalísticos. Os resultados mostram que a utilização de análise sintáctica parcial permite que o sistema melhore o seu *recall* mantendo os altos valores de precisão das técnicas baseadas em regras.

A Secção 7.2 formula a motivação da estratégia apresentada. Depois, a Secção 7.3 descreve a técnica de compressão de orações, enquanto a Secção 7.4 mostra o modo de obtenção dos padrões e das regras de extracção. Finalmente, a Secção 7.5 contém os testes realizados e a 7.6 as conclusões deste capítulo.

O conteúdo deste capítulo foi publicado em Garcia e Gamallo (2011a,e).

7.2. Motivação

O método de extracção proposto neste capítulo segue uma premissa que sugere que determinadas construções linguísticas expressam o mesmo tipo de conhecimento, como relações

semânticas ou ontológicas (Aussenac-Gilles e Jacques, 2006; Aguado de Cea *et al.*, 2009). Para além disso, baseia-se na seguinte assunção:

> *As relações semânticas podem expressar-se do mesmo modo do que as dependências sintácticas*

Uma relação semântica presente numa oração pode, frequentemente, ser representada através de uma ligação de dependências entre duas entidades, inclusive se houver elementos com informação extra que podem tornar a oração muito complexa do ponto de vista linguístico. Esta informação extra não representa a relação semântica, mas pode especificar o significado dos termos relacionados, ou introduzir conhecimento não relevante para a relação. Entre os padrões mais frequentes que expressam cada relação, podem encontrar-se variações do mesmo padrão *base*, do qual diferem pela existência de modificadores, adjuntos, ou estruturas de coordenação, por exemplo. Uma vez que estes padrões têm uma alta precisão, torna-se crucial encontrar um modo de os fazer mais genéricos, para ampliar a sua abrangência. Com este fim, é utilizada a seguinte estratégia:

1. Compressão da oração: é aplicada uma gramática parcial que estabelece dependências sintácticas entre os elementos com informação adicional (modificadores, adjuntos, pontuação, etc.). A gramática mantém unicamente os núcleos das dependências, produzindo assim uma estrutura linguística simplificada.

2. Extracção de padrões: são extraídos padrões léxico-sintácticos, posteriormente simplificados através da aplicação do algoritmo *longest common string* (página 71). Finalmente, os padrões simplificados são transformados em regras genéricas de extracção semântica, adicionadas a uma gramática de dependências.

A combinação de regras *standard* de dependências com regras genéricas de extracção semântica permite que o sistema incremente a abrangência sem perder precisão.

7.3. Análise parcial para a compressão de orações

Um dos pontos de maior importância da estratégia apresentada consiste na simplificação de orações com o objectivo de extrair informação mais facilmente. Para isto, são utilizadas gramáticas e *parsers* de DepPattern.

As gramáticas básicas desta *suite* contêm regras para diferentes fenómenos linguísticos, desde a modificação nominal a estruturas mais complexas como coordenação ou aposição. Porém, o processo de simplificação só requer um certo tipo de dependências, aquelas que comprimem uma oração mantendo o seu significado básico. Assim, seguindo outras estratégias para a compressão de orações (Molina *et al.*, 2011), foram modificadas as gramáticas por defeito, utilizando unicamente aquelas regras que identificam os seguintes constituintes subordinados e satélites:

– Pontuação (pontos de interrogação, aspas, vírgulas, etc.)

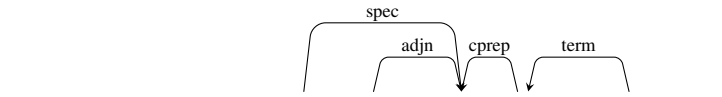
– Coordenação de nomes comuns e adjectivos

– Frases nominais, adverbiais e adjectivais

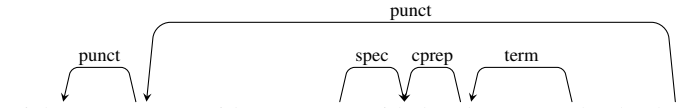
– Complementos preposicionais, perífrases verbais e aposições

– Orações negativas (onde o verbo herda a informação de negação)

Uma vez executado o *parser*, os dependentes identificados por essas regras são eliminados. Assim, é obtida uma estrutura comprimida que não contém satélites nem modificadores. Nos exemplos 7.1 e 7.2 podem ver-se dous casos de análise parcial, onde os elementos no início das setas são os dependentes e os núcleos ocupam a posição final.[1]



(7.1)



(7.2)

Tendo em conta que só os núcleos de cada dependência se mantêm, o processo de compressão das duas orações anteriores vai produzir uma única estrutura simplificada. Note-se

[1] Aqui, *spec* significa especificador; *adjn*, adjunto; *cprep*, complemento preposicional; *term*, termo e *punct* pontuação.

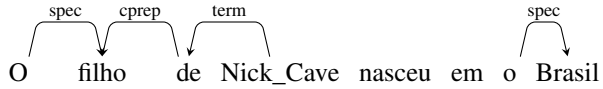
que os núcleos de frases nominais com entidades mencionadas em complementos preposicionais ("cidade de LOC", "região do LOC", etc.) herdam a informação a informação dos nomes próprios dependentes, pelo que nos exemplos, "cidade" representa uma localização específica:

<**Nick_Cave** *nasceu/nascido em* **cidade**>

As regras semânticas genéricas são depois aplicadas sobre estas estruturas simplificadas. Por exemplo:

> Se um nome de pessoa é Núcleo, um nome de localização é Dependente e o verbo *nascer* é um Relator, os dous nomes (pessoa e localização) encontram-se numa relação de `LocaldeNascimento`

Esta regra pode ser proposta para cobrir tanto os dous exemplos anteriores como muitos outros. Para além disso, a análise sintáctica previne a aplicação desta regra em orações como 7.3, onde o núcleo da primeira frase nominal não é um nome de pessoa, mas um nome comum (*filho*).



(7.3)

Assim, neste tipo de orações (e noutros como orações negativas), as regras semânticas não extrairão pares incorrectos como `LocaldeNascimento`, *Nick Cave – Brasil*, mas permitirão extrair o local de nascimento de "o filho de Nick Cave".

<**filho** *nasceu em* **Brasil**>

O formalismo gramatical de DepPattern permite que os analisadores mantenham os dependentes de uma regra depois da sua aplicação. Deste modo, se quisermos adicionar vários conjuntos de regras para extrair diferentes relações, o sistema só precisaria de uma única execução sobre o corpus.

Em suma, a compressão de orações realizada através da análise sintáctica parcial simplifica as estruturas linguísticas mas mantém a sua informação básica. Assim, a adição de regras semânticas genéricas (convertidas desde padrões léxico-sintácticos) no fim das gramáticas de dependências, permite que o *parser* incremente a abrangência do processo de extracção.

Oração: "Nick Cave nasceu na cidade de Warracknabeal"

Polaridade: Nick Cave `LocaldeNascimento` *Warracknabeal*, **positivo**

Padrão: <**X** nasceu_VB em_PS DT cidade_NC de_PS **Y**>

Figura 7.1: Exemplo de uma oração, a polaridade dos termos relacionados e o padrão léxico-sintáctico.

7.4. Obtenção dos padrões e das regras

Esta secção apresenta o método de extracção dos padrões léxico-sintácticos e a estratégia para gerar as regras genéricas.

Obtenção dos padrões

Seguindo a assunção de que muitos dos casos em que ocorre uma relação semântica são representados por padrões léxico-sintácticos similares, o propósito aqui é obter exemplos desses padrões e extrair deles as suas estruturas base (sem a informação adicional), para as transformar em regras semânticas. Para automatizar este processo, emprega-se a estratégia de supervisão-distante apresentada no Capítulo 5 (Secção 5.2). Este processo permite-nos obter orações anotadas automaticamente como a mostrada na Figura 7.1.

O processo de obtenção de padrões é realizado sem revisão humana, pelo que são obtidos casos de *falsos positivos* e *falsos negativos*. Como foi mostrado na Secção 5.5, a qualidade da extracção pode variar em função da relação alvo e dos recursos utilizados (pares relacionados e corpora). A seguinte secção ilustra como uma selecção dos padrões de maior confidência permite evitar a obtenção de padrões de baixa precisão, minimizando assim o ruído produzido pela estratégia de supervisão-distante.

Generalização de padrões

O seguinte método é aplicado para a criação de padrões genéricos, que são transformados posteriormente em regras de extracção de alta precisão:

1. Primeiro, são seleccionados todos os padrões do tipo "**X**[...]**Y**" e escolhidos os mais precisos em função do seu valor de confidência. Este valor obtém-se da seguinte maneira: calcula-se a frequência positiva e negativa de cada padrão. Depois, a frequência negativa é restada da positiva, sendo os padrões ordenados pelo valor de confidência

Padrões obtidos: <**X** nascer_VB em_PS **Y**>,
<**X** nascer_VB em_PS a_DT cidade_NC de_PS **Y**>,
<**X** nascer_VB em_PS NP Fc **Y**>,
<**X** Fc nascer_VB em_PS **Y**>,
<**X** nascer_VB CC residir_VB em_PS **Y**>, [...]
Padrão genérico: <**X** nascer_VB em_PS **Y**>
Regra: NP<**tp:P**> VB<l:nascer> [PS<l:em>] **NP**<**tp:L**>

Tabela 7.1: Exemplo de generalização de padrões para a relação `LocaldeNascimento` em português. Mostram-se alguns dos padrões obtidos, o padrão genérico e a regra de extracção. Na regra, "tp" significa tipo, (onde P = pessoa e L = localização), e "l", lema.

resultante. Finalmente, os n padrões com maior valor são seleccionados ($n = 20$ nos testes). O mesmo processo é feito para os padrões "**Y**[...]**X**".

2. Depois, é aplicado o algoritmo *longest common string* para a generalização dos padrões, sendo o *longest common string* de dous padrões a sua generalização.

3. Os padrões generalizados que não formam parte dos 20 padrões iniciais são descartados, pelo que o resultado é um pequeno conjunto de padrões de alta confidência (veja-se como exemplo a Tabela 7.1).

4. Os padrões genéricos obtidos são adicionados como blocos de regras a uma gramática, que já contém um conjunto de regras de dependências para compressão de orações. Nas novas regras, a entidade **X** considera-se o núcleo e **Y** o dependente. Este último processo é realizado manualmente, permitindo verificar se o processo automático gerou alguma regra incorrecta.

5. Finalmente, a gramática é compilada num *parser*, o qual se aplica num corpus para obter os triplos "**X** `relação` **Y**".

A Tabela 7.1 mostra um exemplo do processo de generalização de padrões, incluindo os melhores padrões extraídos, o seu padrão genérico e a regra de extracção. A aplicação do algoritmo *longest common string* sobre os melhores padrões permite obter um conjunto pequeno de regras de alta qualidade de modo pouco supervisionado. As regras, adicionadas no fim de uma gramática de análise parcial, extraem pares pertencentes à relação desejada. Note-se que as regras incluem também informação semântica, que será obtida no processo de extracção através dos sistemas REM apresentados no Capítulo 3.

7.5. Testes e avaliação

Para avaliar a estratégia proposta, foram realizados três tipos de testes: primeiro, comparou-se o método de regras com duas *baselines* num corpus com exemplos da relação `Profissão` corrigidos manualmente, em espanhol. O sistema de regras foi avaliado de duas maneiras: (i) utilizando uma grande quantidade inicial de pares relacionados e (ii) com um pequeno conjunto de pares semente.

Depois, foram executados dous *parsers* com regras de extracção para `Profissão` e `LocaldeNascimento` em português e espanhol, nas versões completas das respectivas Wikipedias.

Finalmente, os *parsers* em português também foram aplicados num corpus jornalístico, para analisar o seu desempenho numa tipologia textual diferente.[2]

Para a realização dos testes foram extraídos $\approx$ 10.000 pares iniciais para cada relação e língua (português e espanhol) das *infoboxes* da Wikipedia. A seguir, identificaram-se perto de 20.000 orações com um nome de pessoa e (i) uma profissão (para a relação `Profissão`) ou (ii) uma localização (para `LocaldeNascimento`), classificadas automaticamente como positivas ou negativas mediante a estratégia de supervisão-distante. Finalmente, seleccionaram-se aleatoriamente conjuntos de 2.000 orações para cada relação e língua, e um conjunto adicional de 200 para `Profissão`. Este último corpus foi utilizado para avaliar o desempenho do extractor com poucos dados de entrada.

Para a avaliação num conjunto fechado (primeiro teste, em espanhol), foram seleccionadas aleatoriamente 1.000 orações (diferentes das dos corpora anteriores), cuja polaridade foi manualmente corrigida.

Resultados

O primeiro teste tem como objectivo comparar o desempenho do método baseado em regras com duas *baselines* (em espanhol): *Baseline_1* selecciona todas as orações positivas (sem ocorrências negativas) do conjunto inicial de 2.000, e substitui os nomes próprios pelo seu *PoS-tag*. A seguir, faz *pattern-matching* desses padrões no conjunto de teste. *Baseline_2* utiliza as selecções de 2.000 orações para treinar classificadores binários (empregando a mesma estratégia de classificação que no Capítulo 6), representando cada oração com os elementos

[2] A estratégia descrita neste capítulo não foi aplicada nos corpora apresentados no Capítulo 6 (Garcia e Gamallo, 2013) porque os presentes testes tinham sido realizados anteriormente (Garcia e Gamallo, 2011a,e).

Número	*Regra*
1	**NC<tp:Occ> NP<tp:P>**
2	**NP<tp:P>** VB<l:ser> [DT] **NC<tp:Occ>**
3	**NC<tp:Occ>** CONJ<tp:S> NP<tp:P> Fc **NP<tp:P>**
4	**NC<tp:Occ>** CONJ<tp:S> **NP<tp:P>**
5	**NC<tp:Occ>** [Fc] **NP<tp:P>**
6	**NC<tp:Occ>** N<tp:P> Fc **NP<tp:P>**
7	**NP<tp:P>** Fc [AD] **NC<tp:Occ>**
8	**NP<tp:P> NC<tp:Occ>**

Tabela 7.2: Regras obtidas semiautomaticamente para a relação `Profissão` em espanhol. As regras 1 e 2 foram incorporadas ao sistema *Regras_1*, enquanto *Regras_2* contém as 8 regras da tabela. NP<tp:P> em negrito é o nome de pessoa extraído, enquanto NC<tp:Occ> é o nome de profissão (podendo ser também uma estrutura coordenada de nomes de profissão: "o *escritor, músico e cantor*"). Os elementos em parênteses rectos são opcionais, pelo que algumas regras poderiam ser unificadas. Assim mesmo, no exemplo as regras omitem valores como género e número, que bloqueiam a extracção de pares sem concordância.

Modelo	*Precisão*	*Recall*	*Medida F*
Baseline_1	100	5,80	10,10
Baseline_2	44,51	42,54	43,50
Regras_1	99,02	55,80	71,38
Regras_2	99,16	65,20	**78,70**

Tabela 7.3: Precisão, *recall* e medida F de duas *baselines* e dos dous modelos de regras na relação `Profissão` em espanhol.

token_PoS-tag como atributos. Para criar os classificadores utilizou-se (como no Capítulo 5) a implementação do algoritmo SMO de WEKA.

Para avaliar a estratégia de regras, foram construídos dous sistemas: o primeiro extraindo as regras do conjunto de 200 orações (*Regras_1*, com só duas regras de extracção), e o segundo utilizando as 2.000 orações (*Regras_2*, com oito regras). A Tabela 7.2 mostra as regras utilizadas por estes sistemas. O *parser* só extrai as 15 profissões mais comuns das *infoboxes* da Wikipedia, pelo que a avaliação só contém as extracções que incluam estes 15 nomes.

A Tabela 7.3 mostra os resultados dos quatro sistemas descritos no corpus de teste. A *Baseline_1* (*pattern-matching*) teve uma precisão de 100%, mas devido ao baixo valor de *recall*, a medida F é de só 10%. Pequenas variações nas estruturas linguísticas incrementam a sua dispersão, pelo que os padrões iniciais não coincidem com muitos dos encontrados no corpus de teste. A *Baseline_2* teve um melhor desempenho, mas devido ao ruído no corpus de

Língua	*Relação*	*Precisão*	*Pares Extraídos*
Espanhol	`Profissão`	85,35	241.323
	`LocaldeNascimento`	95,56	13.083
Português	`Profissão`	93,86	17.281
	`LocaldeNascimento`	90,34	5.762

Tabela 7.4: Precisão e número de pares únicos extraídos para cada relação nas Wikipedias espanhola e portuguesa.

treino (não corrigido), produziu um grande número de falsos positivos. Este facto provocou que os valores de precisão não atingissem 45%.

Os sistemas de regras tiveram um desempenho nitidamente superior às *baselines* propostas. *Regras_1*, com só duas regras genéricas, obteve 55% de *recall*, mantendo a alta precisão do modelo de *pattern-matching*. A utilização de mais dados permitiu obter 8 regras genéricas, pelo que o sistema *Regras_2* aumentou o *recall* em mais de 10% sem diminuir a precisão.

Uma vez que as orações utilizadas no teste tinham sido filtradas com uma pequena lista de nomes de profissão, foram realizadas novas extracções para conhecer o desempenho do sistema proposto em condições reais. Assim, o sistema *Regras_2* foi executado para realizar extracções em toda a Wikipedia (em português e em espanhol). Foram incluídas 7 regras para `Profissão` no *parser* de português, e 4 regras para `LocaldeNascimento` em cada *parser* (português e espanhol). A informação semântica obtida mediante REM só foi utilizada nas regras para `LocaldeNascimento` que não incluíam verbos (p.e., *nacer/nascer*).

Antes de avaliar a extracção em toda a Wikipedia, foram eliminados da saída dos extractores os tokens com menos de 3 caracteres ou com números. Os pares de `Profissão` foram filtrados com nomes de profissões presentes nas *infoboxes* de cada língua (250 em português e 500 em espanhol). Para avaliar a relação `LocaldeNascimento` utilizou-se a saída completa das regras genéricas. Em todos os casos, foram analisados 50 pares aleatórios e calculada a média aritmética da extracção.

A Tabela 7.4 contém os resultados das duas extracções nas Wikipedias portuguesa e espanhola, com uma única execução para cada língua. É preciso referir que o tamanho de cada Wikipedia era de $3,2$gb em espanhol e de $1,8$ em português.

Em espanhol foram extraídos mais de 241.000 pares únicos de `Profissão`, e mais de 13.000 casos diferentes de `LocaldeNascimento`. Os valores de precisão da primeira relação foram piores do que os obtidos nos testes anteriores (85% *versus* 99%). Contudo, uma análise mais profunda dos resultados mostra que muitos dos erros produzidos nesta extracção deveram-se a processos anteriores (nomeadamente à identificação de nomes próprios), pelo

Relação	*Precisão*	*Pares Extraídos*
`Profissão`	84,54	41.669
`LocaldeNascimento`	84,67	11.842

Tabela 7.5: Precisão e número de pares únicos para cada relação da extracção no jornal português *Público* com o sistema *Regras_2*.

que a precisão das regras é provavelmente maior. Os resultados de `LocaldeNascimento` tiveram uma precisão maior, embora o número de extracções foi muito menor do que na anterior relação (13.083 *versus* 241.323).

Em português, o sistema extraiu mais de 17.000 e 5.700 pares únicos de `Profissão` e `LocaldeNascimento` respectivamente. Uma vez que as regras de extracção em espanhol e português foram muito similares, as diferenças entre as duas línguas podem dever-se a várias razões: por um lado, o tamanho da Wikipedia em espanhol, que é quase o dobro. Por outro lado, o número de nomes de profissão também era menor em português do que em espanhol. Contudo, as extracções em português tiveram uma alta precisão ($90\% - 93\%$).

Note-se que tanto `LocaldeNascimento` como `Profissão` são relações de carácter biográfico, pelo que é esperável que recursos enciclopédicos como a Wikipedia contenham muitos exemplos destas relações. Porém, uma vez que um dos objectivos do presente trabalho é extrair informação de diferentes fontes, foi aplicado o mesmo *parser* (*Regras_2*) para a extracção destas duas relações num corpus jornalístico (em $1,2$gb do *Público*, jornal português de domínio geral).

A Tabela 7.5 contém os resultados desta última extracção, cuja avaliação foi feita da mesma maneira que as realizadas na Wikipedia. O número de extracções é o dobro do que as realizadas na Wikipedia (cujo corpus tinha um tamanho similar). A precisão, contudo, foi entre 6% e 9% mais baixa, sendo de $\approx 84\%$ para as duas relações. Mais uma vez, muitas das extracções incorrectas deveram-se a erros produzidos por módulos anteriores.

7.6. Conclusões

Neste capítulo apresentou-se um novo método de extracção de relações baseada em regras obtidas de modo pouco supervisionado. Antes da aplicação das regras de extracção, utilizou-se uma técnica de compressão de texto que usa análise parcial de dependências, simplificando as estruturas linguísticas e favorecendo um aumento na abrangência das regras de extracção.

Para a obtenção das regras empregou-se uma estratégia de supervisão-distante. O ruído produzido por este processo foi minimizado ao seleccionarem-se só os padrões de maior confidência, posteriormente generalizados e adicionados como regras semânticas a uma gramática de dependências.

Diferentes avaliações em português e em espanhol mostraram que o método mantém a alta precisão dos sistemas de *pattern-matching*, incrementando notoriamente os valores de *recall*. Assim, a estratégia utilizada permite criar de modo simples regras de extracção de alta qualidade, pelo que pode ser um método promissor para a construção rápida de sistemas de extracção de relações de domínio fechado.

Parte III

Resolução de Correferência e Extracção de Informação Aberta

Capítulo 8

Resolução de Correferência de Entidades Pessoa para a Extracção de Informação Aberta

8.1. Introdução

Quando se produz um discurso, seja este oral ou escrito, diversos conceitos são normalmente expressos de diferentes maneiras sem que a referência à mesma entidade discursiva se perca. Assim, uma pessoa como "Ayrton Senna da Silva" pode ser referida, para além de pelo próprio nome (ou variantes como "Ayrton Senna" ou "Senna"), por um pronome pessoal ("Ele"), por uma frase nominal ("o piloto brasileiro") ou por um pronome relativo ("que"), entre outras unidades linguísticas. Quando diferentes expressões (menções) referem à mesma entidade discursiva encontram-se numa relação de correferência (Recasens e Martí, 2010).[1]

Resolver a correferência entre as diferentes menções é uma tarefa crucial para diversas aplicações do processamento da linguagem natural, como a sumarização textual (Steinberger *et al.*, 2007) ou a extracção de informação (Banko e Etzioni, 2008).

Especificamente para a extracção de informação, as entidades pessoa (isto é, que referem a uma pessoa) foram aquelas a que foi dedicado um maior esforço desde diversas perspectivas. Avaliações como a Knowledge Base Population (da conferência TAC) ou a Person Attribute Extraction (da WePS), tarefas como a Personal Name Matching (Cohen *et al.*, 2003), ou

[1] Neste trabalho, uma menção é cada uma das expressões que referem a uma pessoa, e uma entidade é o grupo de todas as menções que referem à mesma pessoa no texto (Recasens e Martí, 2010).

diferentes trabalhos em extracção de relações de entidades pessoa (Mann, 2002; Suchanek, 2009) são alguns exemplos da sua importância.[2]

Este capítulo tem como objectivo analisar o impacto da resolução de correferência na extracção de relações. Para isso, é apresentado um sistema de resolução de correferência de entidades pessoa, avaliado em vários cenários e com diferentes métricas.

São também apresentados três corpora (um para cada uma das línguas alvo desta tese) com anotação correferencial das entidades pessoa. Estes corpora permitem, do ponto de vista linguístico, compreender melhor como o discurso se organiza a nível semântico-referencial (Gordon e Hendrick, 1998). Do ponto de vista do PLN, são úteis para avaliar sistemas de resolução de correferência. Assim, são também utilizados para conhecer a eficácia da ferramenta de resolução referida anteriormente.

Finalmente, são realizados vários testes que mostram a utilidade da aplicação de sistemas de resolução de correferência antes da execução de ferramentas de extracção de informação aberta (OIE).

A Secção 8.2 faz uma revisão do trabalho relacionado. Depois, o sistema de resolução de correferência descreve-se na Secção 8.3. A seguir (Secção 8.4), são apresentados os corpora criados, sendo as avaliações mostradas na Secção 8.5. A Secção 8.6 inclui os testes de resolução de correferência aplicada à extracção de informação aberta, enquanto as conclusões deste capítulo se encontram na Secção 8.7.

As publicações Garcia e Gamallo (2011d, 2014a,b,c) foram utilizadas como base para a redacção do presente capítulo.

8.2. Trabalho relacionado

Esta secção faz uma breve revisão de diversos trabalhos relacionados com a resolução de correferência. Primeiro, mostram-se as principais abordagens que foram utilizadas. A seguir, apresentam-se diferentes corpora e directrizes de anotação correferencial, úteis para a etiquetação dos corpora realizada no presente capítulo.

Abordagens

A resolução de correferência (e de anáfora) é uma tarefa de longa trajectória dentro do processamento da linguagem natural, pelo que foi o tema principal de muitos trabalhos. Os

[2] Veja-se a Secção 4.2 para uma revisão mais pormenorizada.

diversos sistemas de resolução de correferência podem agrupar-se com base em duas distinções: (i) abordagens por "par de menções" (*mention-pair*) *versus* abordagens "centradas na entidade" (*entity-centric*) e (ii) modelos de aprendizagem automática *versus* modelos de regras.

Por um lado, os sistemas *mention-pair* classificam duas menções num texto como correferentes ou não utilizando um vector de atributos obtido desse par de menções. Por outro lado, as abordagens *entity-centric* decidem se uma menção (ou uma entidade parcial) pertence a uma outra entidade parcial utilizando atributos de outras menções das mesmas entidades (parciais).[3]

Geralmente, os classificadores de aprendizagem automática utilizam corpora anotados para treinar sistemas de resolução de correferência supervisionados. Estes modelos usam exemplos etiquetados para aprender preferências e restrições (McCarthy e Lehnert, 1995; Soon *et al.*, 2001; Ng e Cardie, 2002; Sapena *et al.*, 2013), enquanto os modelos não supervisionados utilizam técnicas de *clustering* para a resolução de correferência (Haghighi e Klein, 2007; Ng, 2008).

As estratégias baseadas em regras empregam conjuntos de regras e de heurísticas para encontrar o melhor elemento ao qual ligar cada menção (Lappin e Leass, 1994; Baldwin, 1997; Mitkov, 1998; Bontcheva *et al.*, 2002; Raghunathan *et al.*, 2010; Lee *et al.*, 2013). Este último sistema baseia-se numa abordagem multi-passe que resolve primeiro as ligações *fáceis*, aumentando depois o *recall* com mais regras que aproveitam a informação obtida das ligações entre menções já resolvidas. Inspirado nesse trabalho, Stoyanov e Eisner (2012) apresentaram *EasyFirst*, sistema que utiliza corpora anotados para aprender o grau de dificuldade das ligações de correferência.

A propósito das línguas alvo desta tese, diversos trabalhos lidaram com a resolução de correferência em português (Paraboni, 1997; Chaves e Rino, 2007; Cuevas e Paraboni, 2008). Coelho e Carvalho (2005) adaptaram o algoritmo de Lappin e Leass (1994) para esta língua, e de Souza *et al.* (2008) apresentaram uma abordagem supervisionada para resolver a correferência entre frases nominais.

Para espanhol, Palomar *et al.* (2001) publicaram um conjunto de restrições e preferências para a resolução da anáfora pronominal. Recasens e Hovy (2009) analisaram o impacto de diversos atributos na resolução de correferência supervisionada, implementados posteriormente em Recasens e Hovy (2010). A disponibilização de um grande corpus anotado para

[3] As entidades parciais são conjuntos de menções da mesma entidade.

o espanhol (Recasens e Martí, 2010) permitiu que outros modelos supervisionados fossem adaptados para esta língua (Recasens *et al.*, 2010; Sapena *et al.*, 2013).

Em relação à análise do galego, não conhecemos ferramentas específicas que lidem com a resolução de correferência ou de anáfora neste idioma.

Outras áreas relacionadas, como o já referido *matching* de nomes de pessoa, abordam a resolução de correferência de nomes pessoais ao ligar variantes que têm como referente a mesma pessoa (Cohen *et al.*, 2003).

O sistema apresentado neste capítulo utiliza uma estratégia similar a Lee *et al.* (2013), mas adapta —e adiciona— alguns módulos para entidades pessoa, e enriquece outros com heurísticas baseadas em conhecimento linguístico, como a análise de catáforas e restrições sintácticas.

Corpora

O interesse na obtenção de corpora com anotação correferencial tornou-se patente nas Message Understanding Conferences (MUC) (Fisher *et al.*, 1995; Chinchor e Hirschmann, 1997), que começaram a desenvolver directrizes de anotação e a construir corpora para o inglês com este tipo de informação.

Outras avaliações como as Anaphora Resolution Exercise (ARE) focaram a sua atenção na resolução da anáfora pronominal e da correferência de frases nominais (Orasan *et al.*, 2008). Alguns trabalhos anteriores como Mitkov *et al.* (2000) tinham continuado a desenvolver corpora e ferramentas de anotação.

Com base na anotação das conferências MUC, Hoste (2005) propôs um esquema de etiquetação para o holandês, utilizado no projecto COREA (Bouma *et al.*, 2007).

Recasens e Martí (2010) definiram novas directrizes de anotação para a correferência em espanhol e catalão, excluindo algumas relações consideradas em trabalhos anteriores, como a *parte-tudo*, *bound anaphora* ou *bridging reference*. Este trabalho também publicou corpora com anotação correferencial, e inspirou a avaliação SemEval-2010 Task #1: Coreference Resolution in Multiple Languages (Recasens *et al.*, 2010). Para além de espanhol e catalão, esta avaliação também disponibilizou corpora para outras línguas como o inglês, o alemão, o holandês e o italiano.[4]

[4] `http://stel.ub.edu/semeval2010-coref/`

Para português, Collovini *et al.* (2007) publicaram Summ-it, um corpus de português do Brasil orientado à sumarização automática, cuja anotação correferencial segue as directrizes da MUC.

Do mesmo modo que em relação às ferramentas, até à realização deste trabalho não conhecíamos nenhum corpus com anotação correferencial para galego.

Devido à escassez de recursos para galego e português, este capítulo publica três corpora de características similares para estas duas línguas e para espanhol, que permitem conhecer o funcionamento da correferência de entidades pessoa nestas línguas, bem como avaliar sistemas de resolução. As directrizes de anotação baseiam-se nas apresentadas em Recasens e Martí (2010).

8.3. Sistema de anotação

A ferramenta de resolução de correferência apresentada neste trabalho (LinkPeople) está inspirada no Stanford Deterministic Coreference Resolution System (Raghunathan *et al.*, 2010; Lee *et al.*, 2013), utilizando uma arquitectura multi-passe que aplica sequencialmente um conjunto de módulos de resolução. Os módulos são executados começando por aqueles de maior precisão, seguindo pelos que aumentam o *recall*.

O sistema adiciona novos filtros baseados em informação linguística, tanto para menções catafóricas (cujo referente está mais à frente) como anafóricas (com o referente numa posição anterior): inclui um módulo de alta precisão que procura menções catafóricas de frases nominais e de pronomes pessoais e elípticos. A introdução deste módulo baseia-se na afirmação de que as frases nominais finitas não são sempre anafóricas (Vieira e Poesio, 2000). Adicionalmente, LinkPeople aplica um conjunto de restrições sintácticas no módulo de resolução pronominal, aumentando a sua precisão ao bloquear aquelas ligações que não satisfazem as restrições (Mitkov, 1998; Palomar *et al.*, 2001; Chaves e Rino, 2007).

Arquitectura de LinkPeople

LinkPeople baseia-se em dous princípios essenciais: (i) a abordagem centrada nas entidades e (ii) a arquitectura multi-passe. Por um lado, a abordagem *entity-centric* permite que o sistema utilize todos os atributos da entidade quando uma menção é avaliada. Por outro lado, a arquitectura multi-passe enriquece a entidade (com mais atributos) em cada iteração.

Quem foi $_1$[o cantor dos Beatles]$_1$. $_2$[O músico John Winston Ono Lennon]$_1$ foi um dos fundadores dos Beatles. Com $_3$[Paul McCartney]$_2$, $_4$[∅]$_1$ formou uma dupla de compositores. $_5$[Lennon]$_1$ nasceu em Liverpool, único filho de $_6$[Julia]$_3$ e $_7$[Alfred Lennon]$_4$. $_{8/9}$[Os pais]$_{3/4}$ d$_{10}$[ele]$_1$ chamaram-$_{11}$[no]$_1$ $_{12}$[John Winston Lennon]$_1$. Em 1971, $_{13}$[Lennon]$_1$ atinge o sucesso com o álbum Imagine. $_{14}$[O músico]$_1$ foi assassinado em 1980.

Figura 8.1: Exemplo de um texto com anotação correferencial de entidades pessoa. As menções aparecem entre parênteses rectos. Os números à esquerda correspondem-se com as *ids* das menções, enquanto o número da direita é a *id* da entidade.

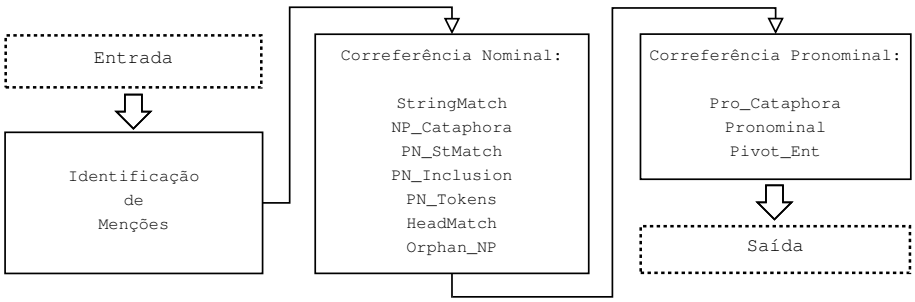


Figura 8.2: Arquitectura do sistema.

Assim, os passes subsequentes aproveitam a informação fornecida pelos módulos prévios de resolução de correferência.

A Figura 8.1 contém um texto com anotação correferencial de entidades pessoa. Este excerto é utilizado para mostrar o funcionamento do sistema. A entrada de LinkPeople necessita ser pré-processada por ferramentas de PLN, que adicionem *PoS-tags*, REM e análise de dependências. No presente trabalho, esta informação foi obtida com as diferentes ferramentas apresentadas nos capítulos anteriores.

Módulos de resolução de correferência

A Figura 8.2 resume a arquitectura do sistema, que começa no processo de identificação de menções. Depois, um conjunto de módulos de resolução nominal e pronominal é aplicado. Os módulos de maior precisão executam-se no início, enquanto os outros módulos incrementam o *recall* com base nos atributos extraídos nos passes anteriores.

Na primeira etapa, um módulo específico identifica as menções que referem a uma entidade pessoa utilizando a informação fornecida pelo etiquetador morfossintáctico e pelo REM, bem como aplicando estratégias básicas para a identificação de pronomes elípticos e de fra-

ses nominais: primeiro identificam-se nomes de pessoa e frases nominais que incluam nomes de pessoa que não formem parte de uma frase preposicional ("o piloto Ayrton Senna" *versus* "a casa do Ayrton Senna"). Depois, procuram-se frases nominais finitas cujo núcleo possa referir a uma pessoa (por exemplo, "o cantor"). Finalmente, seleccionam-se determinantes e pronomes possessivos (singulares) e aplicam-se regras básicas de identificação de pronomes relativos, pessoais e elípticos (em posição inicial de oração e depois de frases adverbiais e preposicionais) (Ferrández e Peral, 2000). Neste ponto, cada menção pertence a uma entidade diferente. Os atributos que pode ter cada entidade são: género, número, núcleo(s) da frase nominal, núcleo(s) do nome pessoal e nome pessoal completo.

Uma vez identificadas as menções, os módulos de resolução aplicam-se de modo sequencial. A execução de cada passe utiliza a seguinte estratégia (excepto nalguma regra, explicada mais abaixo): as menções são percorridas desde o início do texto, e cada menção é *seleccionada* se (i) não é a primeira menção do texto e (ii) é a primeira menção da sua entidade. Uma vez seleccionada uma menção, o sistema procura menções *candidatas* para trás, com o fim de encontrar um antecedente apropriado (nos testes o antecedente foi procurado em todo o texto). Se se encontrar um antecedente, as menções são fusionadas na mesma entidade (partilhando desde esse momento os atributos). A seguir, a próxima menção seleccionada é avaliada.

Para além da identificação das menções, o versão actual de LinkPeople contém os seguintes módulos:

StringMatch (StM): este passe faz *matching* estrito da cadeia completa das duas menções (a seleccionada e a candidata). No exemplo (Figura 8.1), as menções 13 e 5 são agrupadas por este módulo.

NP_Cataphora (NP_C): este módulo verifica se a primeira menção do texto —no primeiro parágrafo— é uma frase nominal que não contenha um nome de pessoa. Se assim for, considera-se uma menção catafórica, e o sistema procura na seguinte oração um nome de pessoa com função de sujeito. Nesse caso, as menções são ligadas se têm concordância de género e número. As menções 1 e 2 do exemplo cumprem estes requisitos, pelo que são fusionadas. Note-se que, no fim deste passe, esta entidade tem como núcleos da frase nominal as palavras "cantor" e "músico", e "John Winston Ono Lennon" como nome pessoal completo. Este módulo também encontra algumas estruturas fixas de sinonímia através de caminhos de dependências, tais como "PESSOA$_a$, também conhecida como PESSOA$_b$".

PN_StMatch (PN_St): nesta etapa, o sistema procura menções que partilhem o nome de pessoa completo, embora os seus núcleos sejam diferentes (ou se uma delas não tem núcleo). "O músico John Lennon" e "John Lennon" (caso que não está na Figura 8.1) seria um exemplo.

PN_Inclusion (PN_I): aqui, o sistema verifica se o nome próprio completo (na entidade) da menção seleccionada inclui o nome próprio da menção candidata (também na entidade), ou vice-versa. No exemplo, a menção 5 liga-se à 2 neste passe. Repare-se que a menção 7 não se agrupa à menção 5, porque o nome próprio completo da entidade a que pertence a menção 5 é "John Winston Ono Lennon", que não é compatível com "Alfred Lennon". Além disso, a menção 13 não é seleccionada por este módulo porque não é a primeira menção da entidade da qual faz parte.

PN_Tokens (PN_T): este módulo separa em tokens o nome próprio completo da entidade a que pertence a menção seleccionada, e verifica se o nome próprio completo (na entidade) da menção candidata contém todos os tokens na mesma ordem, ou vice-versa (excepto algumas palavras vazias, como "Sr.", "Jr.", etc.). Uma vez que o par "John Winston Ono Lennon – John Winston Lennon" é compatível, as menções 12 e 5 fusionam-se.

HeadMatch (HM): nesta etapa, o sistema verifica se a menção seleccionada e a candidata partilham os núcleos (ou os núcleos das entidades a que pertencem). Na Figura 8.1, a menção 14 liga-se à menção 13.

Orphan_NP (Orph): este último módulo de resolução de correferência nominal aplica regras baseadas em resolução pronominal a frases nominais órfãs. Assim, uma frase nominal finita é marcada como órfã se nesta etapa ainda é um *singleton* (uma menção que não tem expressões correferenciais) e não contém um nome de pessoa. Uma frase nominal órfã liga-se ao nome de pessoa anterior com o qual tenha concordância em género e número. No exemplo, as menções 8 e 9 ligam-se a 7 e 6.

Pro_Cataphora (Pro_C): de modo similar a NP_Cataphora, este módulo verifica se o texto começa com um pronome pessoa (ou elíptico). Neste caso, o módulo analisa se a seguinte oração contém um nome próprio compatível.

Pronominal (PRO): este é o módulo *standard* de resolução de correferência pronominal. Para cada pronome seleccionado, verifica se as menções nominais candidatas satisfazem as restrições sintácticas e morfossintácticas (inspiradas em Palomar *et al.* (2001)). Estas restrições classificam-se em conjuntos dedicados a cada tipo de pronome, que bloqueiam a ligação entre as menções se alguma delas é violada. Entre elas existem: um pronome objecto (directo ou indirecto) não pode correferir com o seu sujeito (menção 11 *versus* 8 e 9); um pronome pessoal não pode correferir com uma menção dentro de uma frase preposicional (menção 4 *versus* menção 3), o núcleo de uma frase preposicional não pode correferir com um elemento que o c-comande (menção 10 *versus* menções 8 e 9), um possessivo não pode correferir com a frase nominal à qual pertence ou um pronome prefere como seu antecedente uma frase nominal em posição de sujeito (menções 10 e 11 *versus* menções 6 e 7). Assim, na Figura 8.1, o pronome elíptico (menção 4) liga-se à menção 2, e as menções 10 e 11 à menção 5. Este módulo só procura candidatos na mesma oração e na anterior à menção seleccionada.

Pivot_Ent: este último módulo só é aplicado se existem menções pronominais órfãs (não ligadas a nenhum nome próprio ou frase nominal). Primeiro, o módulo verifica se o texto contém uma entidade central, que é o nome de pessoa mais frequente num texto, cuja frequência seja no mínimo 33% maior do que a segunda pessoa com mais ocorrências. Se existir uma entidade central, todos os pronomes órfãos são ligados a esta entidade. Se não, cada pronome é ligado à anterior menção nominal (sem nenhum tipo de restrição).

Os módulos StringMatch, PN_StMatch e HeadMatch já existiam no sistema apresentado em Lee *et al.* (2013) —com ligeiras diferenças—, enquanto os passes Pronominal e Pivot_Ent se inspiram no trabalho de Palomar *et al.* (2001). Por outro lado, PN_Inclusion e PN_Tokens aproveitam algumas heurísticas já avaliadas em Garcia e Gamallo (2011d), sendo NP_Cataphora, Pro_Cataphora (inspirados em Vieira e Poesio (2000)) e Orphan_NP módulos originais adicionados a LinkPeople.

8.4. Corpora anotados

Esta secção apresenta três corpora com ligações de correferência de entidades pessoa em português, espanhol e galego. Os corpora foram desenhados de acordo a dous objectivos: (i) fornecer recursos para conhecer como a correferência funciona nestas línguas e (ii) avaliar o funcionamento de sistemas de resolução de correferência. Os corpora são disponibilizados

Língua	*Tipologia*	*Documentos*	*Tokens*
Português	Jornal	91	34k
	Wikipedia	6	17k
Espanhol	Jornal	27	18k
	Wikipedia	12	28k
Galego	Jornal	28	17k
	Wikipedia	29	25k
Total	Jornal	146	70k
	Wikipedia	47	71k
	Total	193	141k

Tabela 8.1: Tamanho dos corpora em número de documentos de tokens por língua e tipologia textual.

livremente em dous formatos diferentes, pelo que podem ser ampliados e melhorados de modo colaborativo.

Os corpora

Os textos utilizados como fonte para a construção dos corpora foram compilados da Internet em 2012, tentando cobrir diferentes tipologias textuais e variedades linguísticas. Entre as primeiras, contêm textos jornalísticos e enciclopédicos (Wikipedia). Em relação às variedades linguísticas, incluem-se textos de Portugal, Brasil, Moçambique e Angola (pt), Espanha e Argentina (es) e Galiza (gl). Para além disso, os artigos da Wikipedia podem pertencer a variedades diferentes de português e espanhol.

A Tabela 8.1 mostra o tamanho dos três corpora em número de documentos e em tokens, tendo em conta a tipologia textual e a língua.[5]

Cada corpus contém entre 43k e 51k tokens ($\approx$ 142k tokens, no total). Uma vez que os corpora se focam principalmente em entidades pessoa, a distribuição entre textos jornalísticos e enciclopédicos (geralmente com mais informação sobre pessoas) é de $\approx 35\%/65\%$, excepto em português, devido ao interesse adicional em obter corpora desta língua em diferentes variedades (tanto nacionais como antes e depois do Acordo Ortográfico). Note-se que o corpus de português foi também utilizado para avaliar *PoS-taggers* no Capítulo 2 (referido como Corpus-Web na Secção 2.7.1).

[5] As estatísticas foram computadas com a versão 0.2 dos corpora. Revisões posteriores podem incluir variações em relação a estes resultados.

Unidade Linguística	*Exemplo*
Nome de Pessoa	"*Ayrton Senna* cursou o primário nos Colégios Santana...
Frase Nominal	"Uma semana depois, *o piloto brasileiro* não conseguir tempo...
Pron. Zero	"∅ Começou a competir oficialmente nas provas de kart...
Pron. Clítico	"Isso *o* deixou empatado com Nigel Mansell"
Pron. Relativo	"[...] vinte pontos de diferença para Senna, *que* estava com zero"
Pron. Pessoal	"*Ele* sentia-se frustrado por..."
Pron. Demonstrativo	"É bem provável que *essa* seja a esposa escolhida"
Pron. Indefinido	"*Ambos* chegaram à F1..."
Possessivo	"Senna começou *sua* carreira competindo...

Tabela 8.2: Tipos de unidades correferenciais (e exemplos em português).

Para construir estes recursos, primeiro foram seleccionados aleatoriamente artigos jornalísticos e enciclopédicos —de pessoas— de diferentes fontes da Internet (tendo em conta a sua variedade linguística). Os textos foram tokenizados, lematizados e anotados morfossintacticamente por FreeLing, que também foi utilizado para REM em espanhol. Este último processo foi realizado em português e em galego com o sistema de regras apresentado no Capítulo 3. Depois, aplicou-se DepPattern para enriquecer os corpora com dependências sintácticas. Finalmente, a anotação correferencial foi adicionada manualmente por dous linguistas, seguindo o formato da SemEval-2010 Task #1 (Recasens *et al.*, 2010).

Directrizes de anotação

Diferentes expressões que referem à mesma entidade do discurso foram anotadas como correferentes quando entre elas existia uma relação de identidade de referente. As expressões predicativas, apositivas e parentéticas também foram marcadas (com etiquetas especiais), apesar de que não são consideradas correferentes por alguns autores (Recasens e Martí, 2010). A anotação foi realizada de modo individual, apesar de que algumas entidades aparecem em diferentes artigos e línguas.

A primeira coluna da Tabela 8.2 mostra os tipos de unidades linguísticas candidatos a serem expressões correferenciais. As Frases Nominais (FNs) com um único token (por exemplo, um pronome) ou que contenham unicamente um nome próprio, são classificadas pela sua categoria morfossintáctica (pronome pessoal, nome de pessoa, etc.) e não pela sua categoria sintáctica (frase nominal). Assim, na Tabela 8.2, Nome de Pessoa refere todos os nomes

	Português		Espanhol		Galego		Total		
	J	*W*	*J*	*W*	*J*	*W*	*J*	*W*	*Total*
N. de Pessoa	31,4	34,5	24,0	26,7	28,0	30,9	28,2	30,1	29,3
Frase Nominal	24,3	11,5	12,8	11,9	21,0	8,5	19,7	10,5	14,4
Pron. Zero	26,6	26,0	34,0	32,7	30,7	36,2	29,9	32,5	31,4
Pron. Clítico	3,7	6,9	13,4	8,9	6,5	8,6	7,7	8,3	8,0
Pron. Relativo	3,6	1,7	2,6	2,3	3,4	2,0	3,2	2,0	2,5
Pron. Pessoal	4,1	8,1	2,3	2,5	2,2	0,7	3,1	3,1	3,1
Pron. Demons.	0,1	0	0,1	0,2	0	0,2	0,1	0,2	0,1
Pron. Indef.	0,4	0,4	0,2	0,2	0,2	0,1	0,3	0,2	0,2
Possessivo	5,8	11,1	10,5	14,5	8,1	12,9	7,9	13,5	10,9
Menções totais	2.418	1.561	1.826	2.634	925	2.631	5.169	6.826	11.995

Tabela 8.3: Distribuição e número total de menções em função do tipo, língua e tipologia textual (onde *W* é Wikipedia e *J* jornais).

de pessoa ocupando uma frase nominal completa, sem especificadores. Deste modo, Frase Nominal só inclui FNs com um mínimo de duas unidades linguísticas: um núcleo e um especificador ("Ayrton Senna" é classificado como nome pessoal, enquanto "o piloto Ayrton Senna" ou "o piloto brasileiro" classificam-se como frases nominais).

A anotação foi realizada unicamente quando as menções referiam a entidades pessoa identificadas nalguma parte do texto (isto é, com, pelo menos, uma ocorrência de um nome de pessoa).

Para além dos exemplos da Tabela 8.2, os corpora também incluem anotações de citações. Assim, os pronomes pessoais anafóricos que aparecem em expressões entre aspas são ligados às entidades às quais referem: "Senna disse [...]: "*Eu* acho que"...".

Estatísticas

Para cada corpus foram calculadas várias estatísticas relacionadas com a distribuição das expressões correferenciais. A Tabela 8.3 contém a percentagem de cada tipo de menção nos corpora, mostrando que os nomes de pessoa (como frases nominais completas) e os pronomes elípticos (zero) são as expressões mais frequentes para referir-se às entidades pessoa, com perto de 30% das menções cada uma. As frases nominais e os possessivos também ocuparam mais de 10% (14% e 11% respectivamente). Pelo contrário, a frequência dos pronomes demonstrativos e indefinidos é escassa, com valores médios de entre $0,1\%$ e $0,2\%$, respectivamente.

	Português		Espanhol		Galego		Total		
	J	*W*	*J*	*W*	*J*	*W*	*J*	*W*	*Total*
1 Menção	30,6	50,3	35,9	44,9	16,0	52,3	29,5	49,3	40,3
2 Menções	17,0	17,3	21,5	18,0	24,8	18,4	19,5	18,0	18,7
3 Menções	11,5	9,6	8,2	16,7	11,2	7,1	10,6	11,0	10,8
> 3 Menções	40,1	22,8	34,4	20,5	48,0	22,2	40,4	21,8	30,2
Tamanho da Ent. (Nom.)	3,1	3,6	3,3	3,1	3,0	2,6	3,1	3,1	3,1
Tamanho da Entidade	5,6	7,9	9,1	8,3	6,7	6,6	7,2	7,6	7,4
Maior Entidade (Nom.)	35	218	26	146	20	121			
Maior Entidade	145	674	173	651	57	273			
Mais Frequente	50,2	62,5	59,3	64,5	53,4	69,9	54	66,1	60,1

Tabela 8.4: Distribuição das entidades em função do número de menções nos corpora (acima). Tamanho médio das entidades e tamanho (em número de menções) da maior entidade de cada língua (centro): valores de todas as menções e das nominais (Nom., que só inclui FNs e nomes próprios). Distribuição da entidade mais frequente de cada texto (abaixo).

A Tabela 8.3 também indica que a distribuição das ligações de correferência nas três línguas analisadas é similar. Os pronomes zero são menos frequentes em português do que em espanhol (tanto em textos jornalísticos como enciclopédicos). Por outro lado, as frases nominais são menos utilizadas para referir-se a entidades pessoa em espanhol do que em português. Em média, os valores da distribuição de menções em galego situam-se entre as outras duas línguas analisadas.

A propósito do tamanho das entidades (o número de menções de cada entidade nos corpora), a Tabela 8.4 inclui a distribuição das entidades pessoa anotadas (linhas superiores). As entidades com uma só menção (*singletons*) são as mais frequentes (40%), enquanto $\approx 30\%$ das entidades têm mais de três menções.

As linhas centrais da mesma tabela contêm o tamanho médio das entidades, bem como o tamanho da maior entidade de cada corpus. Os valores foram calculados para todas as menções e unicamente para as menções nominais (frases nominais e nomes de pessoa). O tamanho médio das entidades é similar em todos os corpora, com valores médios de $7,4$ e $3,1$ para todas as menções e só para as nominais, respectivamente.

Em relação à maior entidade de cada corpus, é importante notar a diferença entre os textos jornalísticos e os enciclopédicos, que têm como tópico uma pessoa específica. Assim, alguns textos da Wikipedia em português e em espanhol têm entidades com mais de 600 menções,

enquanto os artigos jornalísticos não excedem as 200. Em galego, os valores da Wikipedia são mais baixos devido ao tamanho dos artigos, que são geralmente mais curtos (Tabela 8.1).

A linha inferior da Tabela 8.4 mostra a percentagem (*micro-average*) das menções da entidade mais frequente em cada texto, que representa a entidade principal de que se está a falar. Os valores dos textos jornalísticos são menores do que os enciclopédicos, uma vez que estes artigos se centram habitualmente numa só pessoa. Note-se que, em média, mais de 60% das menções referem a uma única entidade em cada texto. Isto indica que grande parte dos documentos jornalísticos e enciclopédicos tratam sobre temas relacionados com uma entidade central.

Alguns destes resultados diferem dos dados extraídos de outros corpora com anotação de outro tipo de entidades. A este respeito, Márquez *et al.* (2013) obtêm 86% de *singletons* em inglês e 75% e 70% em espanhol e catalão respectivamente. As diferenças devem-se, principalmente, a que nos recursos aqui apresentados só foi anotado um tipo de entidades.

Formato

Os corpora são distribuídos em dous formatos diferentes: (i) o formato por defeito, inspirado na SemEval-2010 Task #1 (Recasens *et al.*, 2010), e (ii) em formato *brat*, uma ferramenta de código aberto que permite que o anotador visualize o texto de um modo eficiente.[6]

O formato por defeito contém doze colunas com a seguinte informação: (1) posição do token na oração, (2) token, (3) lema, (4) *PoS-tag* de FreeLing, (5) *PoS-tag* básico, (6) género, (7) número, (8) núcleo sintáctico, (9) etiqueta sintáctica, (10) classe da entidade mencionada, (11) tipo de correferência e (12) anotação correferencial.

A anotação correferencial contém, para cada menção, a *id* da entidade a que pertence, bem como a posição de início e fim de cada menção, indicada com parênteses de abertura e fecho (Figura 8.3). Quando um token faz parte de mais de uma menção, o símbolo '|' separa as *ids* de cada entidade.

Os corpora distribuídos em SemEval-2010 contêm mais informação linguística do que os apresentados neste trabalho (como a anotação de roles semânticos), para além de pequenas diferenças nas etiquetas morfossintácticas e sintácticas. À parte disso, a principal diferença na anotação destes dous recursos tem a ver com a pontuação. Em SemEval-2010, a anotação das entidades inclui a pontuação que as circunda, enquanto nos recursos aqui descritos só os

[6] `http://brat.nlplab.org`

pos	*token*	...	*PoS*	...	*et. sint.*	...	*corref*
1	O	...	DET	...	SpecL	...	(1
2	filho	...	NOUN	...	SubjL	...	_
3	de	...	PS	...	CprepR	...	_
4	o	...	DET	...	SpecL	...	(2
5	piloto	...	NOUN	...	Term	...	_
6	brasileiro	...	ADJ	...	AdjnR	...	2)|1)
7	foi	...	VERB	...	ROOT	...	_

Figura 8.3: Exemplo de uma anotação correferencial das FNs "O filho de o piloto brasileiro".

tokens que contêm palavras são considerados parte da entidade (excepto as entidades cuja forma inclui pontuação no interior).

8.5. Testes e avaliação

Esta secção inclui diferentes avaliações do sistema de resolução de correferência apresentado, bem como uma análise pormenorizada dos erros por ele produzidos.

LinkPeople foi avaliado nas três línguas alvo do presente trabalho, utilizando os corpora descritos na secção anterior. Uma vez que algumas das anotações destes recursos não foram corrigidas manualmente (*PoS-tags*, informação sintáctica, etc.), a avaliação seguiu a *regular setting* (utilizando a nomenclatura de SemEval-2010). Assim mesmo, não foram utilizados recursos externos (dicionários de géneros de nomes próprios, WordNet, etc.), seguindo o *closed setting*. Em relação à identificação das menções, foram realizadas duas avaliações: a primeira com as menções já identificadas (*gold mentions*), e a segunda com identificação automática das menções (*system mentions*). Para esta última avaliação, utilizou-se o módulo básico de identificação de menções apresentado na Secção 8.3.[7]

Com o fim de comparar os resultados do sistema aqui apresentado, foram também avaliadas quatro *baselines* conhecidas: (i) *Singletons* (Stons), onde cada menção pertence a uma entidade diferente; (ii) *All_in_One* (AOne), onde todas as menções pertencem à mesma entidade; (iii) *HeadMatch* (HMb), que agrupa na mesma entidade aquelas menções que partilhem o núcleo da frase nominal e classifica cada pronome como *singleton*, e (iv) *HeadMatch_Pro*

[7]Excepto para os pronomes elípticos, que foram obtidos como *gold mentions* para preservar o alinhamento necessário para computar os resultados. Os testes mostrados na Secção 8.6 simulam um cenário real.

(HMP), igual à anterior, mas ligando cada pronome à menção nominal anterior com concordância de género e número.[8]

Foram utilizadas cinco métricas de avaliação: MUC (Vilain *et al.*, 1995), B^3 (Bagga e Baldwin, 1998), $CEAF_{entity}$ (Luo, 2005), BLANC (Recasens e Hovy, 2011) e CoNLL (Co) (Pradhan *et al.*, 2011), que é a média aritmética das três primeiras medidas referidas. Os resultados foram obtidos com os *scripts* utilizados em SemEval-2010 (para BLANC) e ConLL 2011 (para as outras métricas).

Resultados

A Tabela 8.5 contém os resultados das quatro *baselines* e do sistema apresentado na configuração *gold mentions*, enquanto os resultados de *system mentions* se mostram na Tabela 8.6. O primeiro bloco de cada língua inclui os resultados das *baselines*. As linhas centrais mostram os valores dos diferentes módulos de LinkPeople adicionados incrementalmente (veja-se a Figura 8.2). As primeiras nove linhas (StM $>$ PRO) incluem duas regras de defeito para classificar as menções não analisadas pelos módulos em activo: (i) as menções nominais não analisadas são *singletons* e (ii) os pronomes ligam-se à menção nominal anterior com concordância de género (excepto os pronomes analisados por PRO, neste modelo). Para além disso, os sistemas PRO não restringem o número de orações anteriores quando procuram antecedentes. O último modelo (LinkP, o resultado de todos os módulos de LinkPeople) inclui uma restrição de distância no passe Pronominal (veja-se a Secção 8.3), pelo que combina o módulo Pronominal com Pivot_Ent.

Como esperado, as *baselines Singletons* e *HeadMatch* obtêm resultados baixos em quase todas as métricas e línguas (os valores de *Singletons* em MUC são nulos porque esta medida não recompensa a identificação correcta de *singletons*). Porém, os modelos *All_in_One* obtiveram resultados razoavelmente bons nalgumas métricas (MUC e B^3). As diferenças entre os valores destes testes e os obtidos em SemEval-2010 devem-se à existência (neste trabalho) de um único tipo de entidades (pessoas). Como foi visto, os textos jornalísticos e enciclopédicos focam-se habitualmente em só uma ou duas pessoas (isto é, existe um número menor de entidades em cada texto), pelo que a precisão é maior em *All_in_One* e menor em *Singletons*.

Como foi mostrado em Recasens e Hovy (2010), as *baselines HeadMatch_Pro* obtêm bons resultados nas três línguas analisadas, e em todas as métricas: $\approx 60\%$ e $\approx 67\%$ de F1

[8] Devido a diferenças de língua e de formato, outros sistemas de resolução de correferência (Raghunathan *et al.*, 2010; Sapena *et al.*, 2013, por exemplo) não puderam ser utilizados nesta avaliação.

Lg	*Mod*	**MUC**			**B³**			**CEAF$_e$**			**BLANC**			**Co**
		R	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*	*F1*
Pt	Stons	-	-	-	15,0	100	26,1	65,3	10,9	18,7	50,0	29,0	36,7	14,9
	AOne	93,8	85,5	**89,4**	94,8	47,5	63,3	11,9	78,1	20,7	50,0	21,0	29,1	57,8
	HMb	26,5	93,9	41,3	22,2	97,9	36,2	72,3	16,1	26,4	53,6	78,5	44,2	34,6
	HMP	76,0	91,2	82,9	46,0	85,8	59,9	76,7	49,2	59,9	68,5	80,0	68,1	67,6
	StM	69,8	91,5	79,2	38,8	88,7	54,0	78,1	40,5	53,3	64,7	79,2	62,9	62,2
	NP_C	70,4	91,4	79,6	39,2	88,5	54,3	78,3	41,5	54,3	64,7	79,2	62,9	62,7
	PN_St	72,8	91,9	81,3	40,9	88,3	55,9	79,3	44,7	57,2	65,0	79,2	63,4	64,8
	PN_I	77,1	92,5	84,1	50,5	87,5	64,0	81,9	52,7	64,1	71,1	81,0	71,2	70,8
	PN_T	77,3	92,5	84,2	50,8	87,5	64,3	82,0	53,0	64,4	71,1	81,0	71,3	71,0
	HM	79,7	92,3	85,6	53,6	85,5	65,9	81,3	58,3	67,9	71,5	80,7	71,7	73,1
	Orph	83,4	91,8	87,4	58,1	82,7	68,3	81,4	70,2	75,4	71,6	80,3	71,9	77,0
	ProC	83,4	91,8	87,4	58,1	82,7	68,3	81,4	70,3	**75,5**	71,6	80,3	72,0	77,0
	PRO	81,8	91,7	86,4	59,1	83,9	69,3	82,7	66,5	73,7	76,0	83,7	76,7	76,5
	LinkP	82,7	92,7	87,4	65,8	84,5	**74,0**	84,4	67,9	75,2	83,6	85,4	**84,2**	**78,9**
Es	Stons	-	-	-	10,9	100	19,7	69,5	8,7	15,4	50,0	29,4	37,0	11,7
	AOne	91,7	88,4	**90,0**	92,6	51,3	66,0	6,4	83,0	11,9	50,0	20,6	29,2	55,9
	HMb	20,7	94,2	34,0	15,4	98,0	26,6	75,4	11,9	20,6	51,3	74,6	39,9	27,0
	HMP	78,2	90,7	84,0	35,3	81,2	49,2	72,9	51,5	60,4	59,3	74,7	55,5	64,5
	StM	73,9	90,7	81,4	30,1	83,7	44,3	73,9	41,6	53,3	58,6	75,6	54,1	59,7
	NP_C	74,1	90,7	81,5	30,2	83,7	44,4	73,9	42,0	53,6	58,6	75,6	54,1	59,8
	PN_St	75,4	91,0	82,5	31,2	83,1	45,4	73,8	44,1	55,2	58,6	75,4	54,3	61,0
	PN_I	78,8	91,7	84,8	39,3	82,2	53,1	75,9	52,8	62,3	62,0	76,7	59,6	66,7
	PN_T	79,0	91,7	84,9	40,0	82,1	53,8	76,0	53,3	62,7	62,6	76,3	60,5	67,1
	HM	80,5	92,0	85,9	41,7	80,9	55,1	75,6	57,3	65,2	63,1	75,0	61,4	68,7
	Orph	81,1	91,9	86,1	42,3	80,5	55,5	75,4	59,8	66,7	63,2	75,0	61,6	69,4
	ProC	82,3	91,9	86,8	43,2	79,6	56,0	74,6	64,1	68,9	63,0	74,7	61,4	70,6
	PRO	82,6	92,4	87,2	46,0	80,8	58,7	77,5	65,8	71,2	66,8	77,9	66,2	72,4
	LinkP	84,1	94,1	88,8	62,9	84,8	**72,2**	83,4	71,0	**76,7**	81,7	84,9	**82,6**	**79,2**
Gl	Stons	-	-	-	14,6	100	25,4	71,7	11,0	19,1	50,0	28,4	36,3	14,8
	AOne	96,6	86,0	91,0	97,1	53,9	69,3	9,0	82,7	16,2	50,0	21,6	30,1	58,8
	HMb	21,1	90,5	34,2	20,2	97,5	33,5	74,1	14,3	24,0	51,3	74,7	39,1	30,6
	HMP	81,9	89,8	85,7	44,1	83,6	57,7	70,0	53,5	60,6	61,3	76,5	57,9	68,0
	StM	77,1	90,6	83,3	36,5	86,7	51,4	75,1	45,5	56,6	58,9	76,9	53,7	63,8
	NP_C	77,6	90,7	83,6	37,2	86,7	52,1	75,2	46,2	57,3	59,2	77,0	54,3	64,3
	PN_St	79,0	90,9	84,6	39,1	86,2	53,8	75,6	48,8	59,3	59,7	77,0	55,1	65,9
	PN_I	83,1	91,5	87,1	46,7	85,3	60,4	76,7	57,8	66,0	62,5	77,5	59,5	71,1
	PN_T	83,3	91,5	87,2	48,2	85,3	61,6	76,9	58,6	66,5	63,2	77,9	60,5	71,8
	HM	84,6	91,6	87,9	49,8	84,4	62,6	76,8	62,0	68,6	63,4	77,5	60,8	73,1
	Orph	84,7	91,3	87,9	49,9	83,9	62,6	76,8	63,2	69,4	63,3	77,3	60,8	73,3
	ProC	84,7	91,3	87,9	49,1	83,9	62,6	76,8	63,2	69,4	63,3	77,3	60,8	73,3
	PRO	86,9	92,5	89,6	60,7	86,8	71,4	82,8	72,2	77,1	73,6	82,0	73,9	79,4
	LinkP	89,0	94,6	**91,7**	72,9	88,4	**79,9**	87,6	76,6	**81,7**	82,7	85,8	**83,4**	**84,4**

Tabela 8.5: Resultados de LinkPeople (*gold mentions*) comparados com as *baselines* em Português (Pt), Galego (Gl) e Espanhol (Es). *LinkP* são os resultados da execução do sistema completo.

Lg	*Mod*	**MUC**			**B^3**			**$CEAF_e$**			**BLANC**			**Co**
		R	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*	*F1*
Pt	Stons	-	-	-	12,9	88,2	22,5	62,5	11,1	18,8	50,0	25,6	33,9	13,8
	AOne	75,6	73,1	74,3	67,1	37,7	48,3	9,5	61,3	16,4	50,0	24,4	32,8	46,3
	HMb	22,0	89,6	35,3	19	86,1	31,1	68,4	15,7	25,6	53,3	75,4	40,8	30,7
	HMP	65,4	79,4	71,7	41,5	68,9	51,8	69,2	54,2	60,8	70,1	77,7	68,4	61,4
	StM	61,4	79,4	69,2	35,0	71,6	47,0	70,8	46,0	55,8	65,1	75,9	61,5	57,3
	NP_C	61,4	79,4	69,2	35,0	71,6	47,0	70,8	46,1	55,8	65,1	75,9	61,5	57,4
	PN_St	63,7	79,9	70,9	36,6	70,8	48,3	71,6	50,7	59,4	66,2	76,2	63,1	59,5
	PN_I	67,9	81,1	73,9	45,5	69,7	55,1	74,0	61,9	67,4	72,3	78,3	71,1	65,5
	PN_T	68,1	81,1	74,0	45,8	69,7	55,3	74,0	62,4	67,7	72,3	78,3	71,2	65,7
	HM	68,8	81,0	74,4	47,3	68,7	56,0	73,6	64,9	69,0	72,8	78,3	71,8	66,5
	Orph	68,8	81,0	74,4	47,3	68,7	56,0	73,6	64,9	69,0	72,8	78,3	71,8	66,5
	ProC	68,8	81,0	74,4	47,3	68,7	56,0	73,6	64,9	69,0	72,8	78,3	71,8	66,5
	PRO	69,0	81,2	74,6	47,9	69,2	56,6	74,0	65,4	69,4	73,4	78,6	72,5	66,9
	LinkP	69,9	82,0	**75,5**	55,8	69,2	**61,8**	76,6	68,8	**72,5**	84,4	84,5	**84,4**	**69,9**
Es	Stons	-	-	-	9,2	90,2	16,7	63,3	8,3	14,7	50,0	25,8	34,0	10,5
	AOne	77,5	78,8	78,1	69,0	44,2	53,9	5,7	73,5	10,5	50,0	24,2	32,6	47,5
	HMb	17,6	89,5	29,5	13,1	88,1	22,7	68,3	11,2	19,2	51,1	71,0	36,6	23,8
	HMP	68,2	81,5	74,2	31,4	68,1	43,0	62,6	52,0	56,8	60,0	70,7	54,7	58,0
	StM	65,1	81,7	72,5	26,5	70,4	38,5	63,3	42,1	50,6	59,0	71,6	52,7	53,8
	NP_C	65,2	81,6	72,5	26,5	70,4	38,5	63,3	42,2	50,7	59,0	71,6	52,7	53,9
	PN_St	66,3	81,8	73,2	27,2	69,4	39,0	63,1	45,0	52,6	59,2	71,3	53,2	54,9
	PN_I	69,6	82,8	75,6	34,2	68,2	45,5	64,6	55,0	59,4	62,8	72,8	58,8	60,2
	PN_T	69,7	82,8	75,7	35,0	68,2	46,2	64,7	55,5	59,8	63,5	72,7	60,1	60,6
	HM	70,1	82,4	75,8	35,4	67,4	46,4	64,2	58,5	61,2	63,8	72,3	60,5	61,1
	Orph	70,1	82,4	75,8	35,4	67,4	46,4	64,2	58,5	61,2	63,8	72,3	60,5	61,1
	ProC	70,1	82,4	75,8	35,4	67,4	46,4	64,2	58,5	61,2	63,8	72,3	60,5	61,1
	PRO	70,2	82,3	75,8	35,9	67,3	46,8	64,6	59,4	61,9	64,0	72,0	61,0	61,5
	LinkP	72,8	85,3	**78,5**	50,4	70,7	**58,9**	73,5	68,3	**70,8**	75,6	76,4	**75,5**	**69,4**
Gl	Stons	-	-	-	12,7	84,0	22,0	67,5	10,0	17,4	50,0	24,6	32,9	13,1
	AOne	83,1	71,1	76,6	75,9	40,9	53,2	7,4	67,1	13,3	50,0	25,4	33,7	47,7
	HMb	17,4	84,6	28,8	17,4	81,5	28,6	69,3	12,4	21,0	50,1	70,4	35,2	26,2
	HMP	72,4	73,7	73,0	38,0	62,3	47,2	61,4	52,5	56,6	61,2	72,4	55,4	59,0
	StM	68,7	73,8	71,2	31,5	65,1	42,4	66,4	45,3	53,8	58,6	72,6	50,6	55,8
	NPC	69,0	73,9	71,4	32,0	65,0	42,9	66,5	46,0	54,4	58,9	72,7	51,1	56,2
	PN_St	70,7	74,0	72,3	34,4	64,1	44,8	66,6	50,4	57,4	60,4	73,1	53,6	58,1
	PN_I	74,6	75,2	74,9	42,8	63,0	50,9	66,7	60,1	63,2	64,9	74,8	60,6	63,0
	PN_T	74,9	75,2	75,1	43,5	63,0	51,5	66,9	61,0	63,8	65,2	74,9	61,0	63,4
	HM	75,0	75,1	75,0	44,0	62,8	51,7	66,8	62,0	64,3	65,1	74,4	61,1	63,7
	Orph	75,0	75,1	75,0	44,0	62,8	51,7	66,8	62,0	64,3	65,1	74,4	61,1	63,7
	ProC	75,0	75,1	75,0	44,0	62,8	51,7	66,8	62,0	64,3	65,1	74,4	61,1	63,7
	PRO	75,0	75,0	75,0	44,2	62,7	51,8	67,0	62,4	64,6	65,3	74,4	61,4	63,8
	LinkP	78,5	78,5	**78,5**	65,0	67,0	**66,0**	78,6	73,4	**75,9**	86,0	86,3	**85,9**	**73,5**

Tabela 8.6: Resultados de LinkPeople (*system mentions*) comparados com as *baselines* em Português (Pt), Galego (Gl) e Espanhol (Es). *LinkP* são os resultados da execução do sistema completo.

BLANC e CoNLL, respectivamente, em *gold mentions* e $\approx 59,5\%$ nas mesmas métricas em *system mentions*.

As diferenças entre os cenários *gold* e *system mentions* apreciam-se já nos resultados das *baselines* e dos primeiros módulos do sistema, com resultados entre 1% e mais de 10% piores quando aplicado o módulo básico de identificação de menções.

Em relação aos diferentes passes de LinkPeople, o desempenho dos primeiros módulos de *matching* depende da distribuição e da estrutura dos nomes próprios e das frases nominais nos corpora. A este respeito, PN_StMatch funciona bem em todos os contextos. Porém, PN_Inclusion destaca-se dentro dos módulos nominais, ao aumentar em mais de 5% (BLANC e CoNLL) o desempenho do modelo anterior, tanto em *gold* como em *system mentions*. Isto deve-se ao grande aumento em *recall*, unido à alta precisão deste módulo.

É importante ressaltar que a adição de alguns módulos parece melhorar não só o *recall*, mas também a precisão geral do sistema, apesar de que a ordem de aplicação prioriza a precisão em favor do *recall*. Isto é devido à execução das duas regras de defeito: uma vez que o sistema utiliza mais módulos, a quantidade de fusões de entidades (parciais) aumenta. Assim, a precisão incrementa-se porque as novas fusões restringem ligações incorrectas realizadas pelas duas regras de defeito nos modelos anteriores.

O módulo HeadMatch é o primeiro que lida com menções sem nome próprio (excepto algumas regras aplicadas em NP_Cataphora, com baixo *recall*). Graças aos atributos extraídos nos passes prévios, este módulo melhora os resultados de todos os modelos nas três línguas.

O desempenho de Orphan_NP e Pro_Cataphora também depende do corpus e da métrica de avaliação, para além do processo de identificação de menções. As menções analisadas por estes módulos são algumas frases nominais finitas sem nomes próprios, que foram detectadas muito poucas vezes pelo módulo básico de identificação. Assim, tanto Orphan_NP como Pro_Cataphora não variam os resultados no cenário *system mentions*. Em *gold mentions*, Pro_Cataphora produz uma perda de $0,2\%$ em espanhol em BLANC (mas melhora $1,1\%$ em CoNLL). Pela sua parte, Orphan_NP faz com que o sistema não classifique como *singletons* algumas menções, facto que favorece o desempenho dos módulos pronominais. De modo similar, Pro_Cataphora previne que o seguinte módulo seleccione incorrectamente menções catafóricas.

O módulo *standard* de resolução da correferência pronominal também incrementa o desempenho de todos os sistemas (com uma única excepção: os resultados em *gold mentions* em português com a métrica CoNLL, cujos valores aumentaram notoriamente com a aplicação do

módulo Orphan_NP). Os resultados deste módulo em *system mentions* também são positivos, embora a sua melhora seja menor devido à dificuldade em identificar unicamente os pronomes que refiram a entidades pessoa.

Finalmente, uma das principais contribuições para o desempenho positivo de LinkPeople é a combinação do módulo Pronominal com Pivot_Ent. Esta combinação reduz o escopo do primeiro módulo, reforçando deste modo o impacto das restrições sintácticas. Para além disso, Pivot_Ent procura entidades pessoa proeminentes no texto, e liga os pronomes órfãos a estas entidades. Nas três línguas, esta melhora é notoriamente superior quando calculada através da métrica BLANC.

A última linha de cada língua mostra os resultados actuais de LinkPeople nos três corpora, com valores *macro-average* de $\approx 83\%$ e $\approx 81\%$ (*gold mentions*) e $\approx 82\%$ e $\approx 71\%$ (*system mentions*) em BLANC e CoNLL, respectivamente. Os resultados finais em *system mentions* na métrica BLANC são notoriamente melhores do que nos módulos anteriores, sendo a diferença em relação ao cenário *gold mentions* mínima. Isto deve-se a que a medida BLANC avalia unicamente as ligações de correferência (e as não correferenciais), mas não o processo de identificação de menções. Este processo teve valores de F1 de $85,9\%$ (pt), $87,9\%$ (es) e $85,8\%$ (gl).[9]

Análise de erros

Para conhecer os principais tipos de erros produzidos pelos módulos de resolução de correferência, foram seleccionados aleatoriamente 150 erros (50 de cada língua) dos resultados de LinkPeople na avaliação *gold mentions*. Cada erro foi analisado, procurando a origem que o provocou, e classificado de acordo à sua tipologia. Esta secção mostra as diferentes tipologias de erros e alguns exemplos, organizados pela sua frequência de ocorrência nos resultados (a primeira percentagem entre parênteses é a frequência média, seguida dos valores para português, espanhol e galego).[10] Os exemplos são casos reais de menções incorrectamente analisadas (ou pares de menções que pertencem à mesma entidade), com algumas simplificações para facilitar a compreensão:[11]

[9] Lembre-se que o cenário *system mentions* utilizou os pronomes elípticos de *gold mentions*, pelo que os resultados num cenário real seriam mais baixos. Do mesmo modo, esta versão de LinkPeople não inclui regras específicas para o tratamento de menções plurais nem recursos externos para a identificação de menções pessoais nem de extracção de género, facto que ilustra a margem de melhora que tem o sistema.

[10] Os resultados de 0% nalgumas línguas e categorias não significam que essas línguas não possam ter erros dessas tipologias, mas que não apareceram devido ao número de erros analisados.

[11] Alguns exemplos utilizam o formato de anotação mostrado na Figura 8.1.

Ligações não realizadas entre frases nominais e/ou nomes próprios (46%: 58% / 48% / 32%)

Esta categoria inclui algumas tipologias de erros que diferem no tipo de conhecimento e análise requerida pelo sistema para ligar correctamente duas menções:

Núcleos sinónimos (35,3%: 48% / 26% / 32%): O tipo mais frequente de ligações não realizadas produziu-se por menções da mesma entidade cujos núcleos são sinónimos:

Menção A: "El *joven*"
Menção B: "el *muchacho*"

Conhecimento externo (do mundo real) (6%: 0% / 18% / 0%): Esta classe inclui menções da mesma entidade que não partilham atributos lexicais, normalmente porque referem a entidades conhecidas no mundo real:

Menção A: "la presidenta"
Menção B: "Cristina Kirchner"

Aqui, a frase nominal "la presidenta" é utilizada para referir-se a "Cristina Kirchner", mas as menções não foram ligadas porque o sistema não utiliza recursos que definam a Cristina Kirchner como uma *presidenta*.

Conhecimento semântico (2,7%: 4% / 4% / 0%): A ausência de outro tipo de conhecimento semântico, como pares de hipónimos-hiperónimos, também produz ligações perdidas como a que segue:

Menção A: "o escocês"
Menção B: "o britânico"

Modificadores dos núcleos (1,3%: 4% / 0% / 0%): Os modificadores internos de alguns núcleos também podem produzir ligações perdidas, como no exemplo seguinte, onde uma menção não contém o modificador *adjunto*:

Menção A: "o *ministro*"
Menção B: "o *ministro-adjunto*"

Diferenças ortográficas (0,7%: 2% / 0% / 0%): Alguns nomes pessoais podem aparecer escritos de modo diferente no mesmo texto:

Menção A: "André *Villas-Boas*"
Menção B: "André *Villas Boas*"

Erros produzidos por análise automática incorrecta (15,3%: 2% / 22% / 22%)

Uma vez que não toda a anotação dos corpora foi corrigida manualmente, alguns erros produzidos pelos analisadores automáticos (sintácticos e morfossintácticos) provocam ligações perdidas e incorrectas entre algumas menções:

Erros nas restrições sintácticas (10,7%: 0% / 16% / 16%): Os pronomes objecto (directo ou indirecto), bem como algumas frases nominais etiquetadas erroneamente, não são cobertos por algumas restrições sintácticas, provocando ligações incorrectas, por exemplo, entre um pronome e a frase nominal que é sujeito do próprio pronome. No corpus em espanhol, a frase nominal "el moyanista" (em "le transmitió el moyanista") foi analisada como objecto directo (e não como sujeito), pelo que um pronome posterior foi ligado a ela incorrectamente.

Género incorrecto (2,7%: 2% / 2% / 4%): O género de alguns nomes e adjectivos pode ser etiquetado incorrectamente, causando ligações erradas ou provocando ligações perdidas. Por exemplo, a palavra "atleta" (que pode ser masculina ou feminina) etiquetada como masculina em galego bloqueou uma ligação do pronome "ela".

Núcleo incorrecto (2%: 0% / 4% / 2%): Os erros de análise morfossintáctica (normalmente entre nomes e adjectivos) também produzem incorrecções na análise de dependências, o que implica extracções erróneas dos núcleos das frases nominais:

Menção: "el jugador alemán"
Núcleo extraído: *alemán (em vez de *jugador*)

Ligações perdidas por anáfora pronominal de longa distância (11,3%: 14% / 2% / 18%)

Este tipo de erros aparece quando a distância entre uma menção pronominal e o seu antecedente nominal supera a distância permitida pelo módulo de análise (nos testes, entre duas e quatro orações, em função do módulo), e o antecedente não é a entidade proeminente.

Erros na correferência de citações (10%: 10% / 6% / 14%)

Uma outra categoria de erros inclui menções dentro de citações. Estas menções podem referir ao falante (primeira pessoa) ou a uma terceira pessoa da citação:

Primeira pessoa (4,7%: 6% / 2% / 6%): A 1ª pessoa de uma citação deve ser ligada ao falante, e não sempre à menção anterior, como incorrectamente fez LinkPeople no seguinte exemplo (note-se que o pronome elíptico também poderia ser um pronome de 3ª pessoa):

"Si $\emptyset_{1ª}$ tuviera que redactar [...]", resumió Lezcano$_{\text{Falante}}$.

Terceira pessoa (5,3%: 4% / 4% / 8%): As 3as pessoas de uma citação não devem ser ligadas ao falante:

Gustavo$_{\text{Falante}}$: "Cuando yo$_{1ª}$ me fui, él$_{3ª}$ dejó Boca."

Neste caso, o sistema ligou incorrectamente o pronome "él" com a menção "Gustavo".

Ligações incorrectas entre menções plurais (5,3%: 4% / 8% / 4%)

A resolução da correferência das menções plurais foi realizada através de ligações básicas às menções anteriores, o que produziu classificações incorrectas. Além disso, algumas menções plurais incluem entidades com diferentes géneros (por exemplo, "amigos" pode referir a entidades masculinas e femininas, mas o género gramatical da palavra é masculino nas três línguas analisadas):

$_1$[Hulk]$_1$, $_2$[Moutinho]$_2$ e $_3$[Álvaro Pereira]$_3$ na lista de compra de $_4$[Villas-Boas]$_4$ [...]. $_{5/6/7}$[O trio do F.C. Porto]$_{2/3/*4}$ [...].

Neste exemplo, a menção plural ("O trio do F.C. Porto") liga-se às três menções nominais anteriores com concordância de género, pelo que é realizada uma ligação incorrecta entre as menções 7 e 4.

Erros provocados por concordância de género incorrecta (4,7%: 4% / 6% / 4%)

Algumas frases nominais que referem à mesma entidade podem ter géneros diferentes, causando assim ligações perdidas ou incorrectas:

Menção A: "la víctima" (feminino)
Menção B: "el muchacho" (masculino)

Note-se que estes erros não são provocados por análises incorrectas dos *PoS-taggers* nem dos *parsers*, à diferença dos referidos acima como "Género incorrecto".

Erros produzidos por Pivot_Ent e restrições sintácticas (4,6%: 6% / 8% / 0%)

As restrições sintácticas, apesar de precisas, podem bloquear algumas ligações correctas. Isto pode provocar (i) uma análise incorrecta do discurso ou (ii) a aplicação de Pivot_Ent, que ligará um pronome à entidade mais frequente, ligação que pode ser errada:

$_1$[El escritor]$_1$ tuvo que visitar a $_2$[Martín]$_2$ en el hotel. Según $_3$[$\emptyset$]$_{*1}$ dijo [...]

Aqui, o sujeito elíptico de *dijo* é *Martín*, mas a ligação é bloqueada por uma restrição sintáctica: o antecedente de um pronome elíptico com função de sujeito deve ser um sujeito. Assim, o sistema liga incorrectamente a menção 3 à menção 1.

Ligações incorrectas entre frases nominais com o mesmo núcleo (1,3%: 0% / 0% / 4%)

Num mesmo texto, diferentes entidades podem ser referidas por menções que partilhem o núcleo, o que pode provocar erros na resolução da correferência, em função da sua posição e dos seus atributos. Uma frase nominal como "o presidente" pode referir-se a duas pessoas como "o presidente da Academia" e "o presidente do Governo".

Ligações incorrectas produzidas por erros de módulos anteriores (0,7%: 0% / 0% / 2%)

Os primeiros módulos podem produzir algumas ligações incorrectas que causam erros nas análises subsequentes. Por exemplo, no corpus galego, NP_Cataphora agrupou erroneamente a frase nominal "o alcalde" com o nome próprio "Dorribo". A seguir, HeadMatch ligou a menção "Dorribo" com "o alcalde Orozco", criando uma entidade incorrecta com duas pessoas diferentes (Dorribo e Orozco).

Erros produzidos por estruturas fixas da língua (0,7%: 2% / 0% / 0%)

Outros erros menos frequentes incluem algumas estruturas fixas como o seguinte possessivo catafórico:

Por $_1$[sua]$_1$ parte, $_2$[Cristina]$_{*2}$ [...]

Os resultados desta análise de erros fornecem informação importante para o trabalho futuro. Assim, incluir algum tipo de conhecimento semântico (por exemplo, sinónimos), melhorar a resolução de correferência pronominal ou implementar regras específicas para as citações poderia resolver muitos dos erros mais frequentes causados pelo sistema.

8.6. Correferência e extracção de informação aberta

Esta secção analisa o impacto da resolução de correferência na tarefa de extracção de relações, através da utilização de um sistema multilíngue de extracção de informação aberta.

À diferença das aproximações de domínio fechado para a extracção de informação (apresentadas na Parte II desta tese), que dependem de um conjunto finito de relações semânticas predefinidas, os sistemas de extracção de informação aberta realizam extracções não supervisionadas de todo o tipo de relações de base verbal (Banko *et al.*, 2007).

Por exemplo, de uma oração como "Obikwelu arrecadou a medalha de ouro dos 100 metros nos Jogos da Lusofonia de 2009", um sistema de OIE poderia obter a seguinte informação estruturada (com dous argumentos e uma relação de base verbal em cada extracção):

OIE$_1$: *Obikwelu*$_{\text{Arg1}}$ `arrecadou` *a medalha de ouro dos 100 metros*$_{\text{Arg2}}$

OIE$_2$: *Obikwelu*$_{\text{Arg1}}$ `arrecadou_a_medalha_de_ouro_dos_100_metros_em` *os Jogos da Lusofonia de 2009*$_{\text{Arg2}}$

Porém, muitas das menções de cada entidade pessoa que aparecem num texto são diferentes, pelo que a extracção final pode não ser semanticamente completa. Assim, do mesmo texto que o exemplo anterior,[12] um sistema de OIE poderia extrair relações como estas:

OIE$_3$: *Francis Obiorah Obikwelu*$_{\text{Arg1}}$ é_um *atleta*$_{\text{Arg2}}$

OIE$_4$: *que*$_{\text{Arg1}}$ reside_em *Lisboa*$_{\text{Arg2}}$

OIE$_5$: *Ele*$_{\text{Arg1}}$ decidiu_correr_por *Portugal*$_{\text{Arg2}}$

Por um lado, estas extracções não incluem os referentes dos pronomes ("que", "Ele"), pelo que o conhecimento extraído pode não ser semanticamente útil. Por outro lado, a extracção não indica que "Obikwelu", "Francis Obiorah Obikwelu" e "Ele" referem à mesma entidade, enquanto "que" refere a outra pessoa (neste caso, à velocista nigeriana Mercy Nku).

Tendo isto em conta, a aplicação de um sistema de resolução de correferência antes do processo de extracção pode melhorar o resultado de duas maneiras: (i) aumentando o *recall* ao desambiguar a referencialidade dos pronomes e (ii) enriquecendo a extracção final ao ao agrupar as diferentes menções nominais e pronominais de cada entidade.

Testes

Com o fim de medir o impacto de LinkPeople no processo de extracção de informação aberta, a versão mais recente de *DepOE* (Gamallo *et al.*, 2012), um sistema multilíngue de OIE, foi executada na saída do sistema de resolução de correferência.

Para realizar os testes, foi compilado um novo corpus para cada uma das três línguas alvo, contendo cada recurso 5 artigos da Wikipedia e 5 artigos jornalísticos.

DepOE foi aplicado duas vezes: primeiro, utilizando como entrada os três corpora em texto plano (DepOE). Depois, executando-se na saída de LinkPeople (DepOE+).

LinkPeople aplicou-se utilizando o cenário *system mentions* e sem nenhum tipo de recursos externos. Empregou-se um módulo básico de identificação de pronomes elípticos —com um *parser* de DepPattern— que procura por este tipo de unidades em posição inicial de oração e depois de frases adverbiais e preposicionais. Toda a informação linguística foi obtida com as ferramentas de PLN apresentadas na Parte I desta tese.

Para calcular a precisão de *DepOE*, 300 triplos que contivessem como mínimo uma entidade pessoa como um dos argumentos foram seleccionados aleatoriamente e revistos de modo

[12] http://pt.wikipedia.org/wiki/Francis_Obikwelu

Oração:	"Debutó en la Tercera división"	"Anderson viajou por Europa"
DepOE	∅	*Anderson* `viajou_por` *Europa*
DepOE+	*Ander Herrera* `debutó_en` *la 3ª división*	*Wes Anderson* `viajou_por` *Europa*

Tabela 8.7: Exemplos de extracção de DepOE e DepOE+ em espanhol (esquerda, cuja tradução para português é "Debutou na terceira divisão") e português (direita). A extracção de DepOE+ em espanhol extrai um novo triplo —não extraído por DepOE— de uma oração com sujeito elíptico, enquanto o argumento do exemplo em português é enriquecido com o nome próprio completo (e ligado a outras menções no mesmo texto).

Língua	**DepOE**			**DepOE+**			**E**
	W	*J*	*P*	*W*	*J*	*P*	
Português	82	133	39%	111	155	56%	75%
Espanhol	47	82	49%	80	86	58%	84%
Galego	168	114	49%	221	115	54%	77%

Tabela 8.8: Resultados de duas execuções de *DepOE*. *W* and *J* são o número de extracções dos artigos da Wikipedia e jornalísticos, respectivamente. *P* é a precisão da extracção, e *E* o enriquecimento fornecido por LinkPeople.

manual (100 para cada língua). Na primeira execução (sem resolução de correferência), as extracções com pronomes pessoais nos argumentos não foram computadas, sendo consideradas como não especificadas semanticamente. Assim, o maior número de extracções na segunda execução (DepOE+) deve-se à identificação de pronomes pessoais (incluindo elípticos). A coluna central da Tabela 8.7 contém um exemplo de uma nova extracção obtida graças à resolução de correferência.

LinkPeople também relacionou menções nominais com diferente forma (coluna direita da Tabela 8.7), enriquecendo a extracção ao permitir que o sistema OIE agrupe informação variada sobre a mesma entidade. Uma estimação deste enriquecimento calculou-se assim: de todos os triplos correctos (dos revistos), verificou-se se a menção pessoal no argumento tinha sido correctamente resolvida por LinkPeople. Estes casos foram divididos pelo número total de triplos correctos, sendo estes resultados considerados como o valor de *enriquecimento*.

A Tabela 8.8 contém os resultados das execuções de DepOE e de DepOE+. DepOE+ conseguiu extrair $22,7\%$ mais triplos que o modelo básico, e a sua precisão aumentou em $10,6\%$. Estes resultados mostram que a melhora foi maior na Wikipedia, pelo facto de que a maior entidade (pessoa) nos textos enciclopédicos é maior do que a dos textos jornalísticos. Além disso, as páginas da Wikipedia contêm geralmente mais pronomes anafóricos referentes às entidades pessoa. Finalmente, a última coluna da Tabela 8.8 inclui a percentagem de enrique-

cimento da extracção depois da utilização de LinkPeople. Apesar de que estes resultados não são uma avaliação directa da OIE, sugerem que a informação extraída por *DepOE* é até $\approx 79\%$ melhor quando é obtida depois da aplicação de um sistema de resolução de correferência.

8.7. Conclusões

Este capítulo centrou-se na resolução de correferência de entidades pessoa e na combinação desta tarefa com a extracção de informação aberta.

Com este fim foi criado um sistema de resolução de correferência multilíngue para entidades pessoa, com uma arquitectura multi-passe e uma abordagem centrada nas entidades que inclui um conjunto de restrições sintácticas no módulo de resolução pronominal.

Para a avaliação do sistema, foram criados três corpora (um para cada uma das línguas analisadas) com anotação correferencial das entidades pessoa. Diversos testes nestes corpora mostraram que a ferramenta tem um bom desempenho em diferentes cenários e em cada uma das línguas.

O último conjunto de testes também mostrou que a combinação da resolução de correferência com a extracção de informação aberta aumenta o *recall* da extracção, ao permitir obter, por exemplo, novos triplos com pronomes pessoais e elípticos desambiguados. Além disso, a agrupação das diferentes menções nominais e pronominais em entidades enriquece a extracção, facilitando uma melhor organização posterior da informação obtida automaticamente.

Note-se que um sistema de resolução de correferência também pode ser aplicado em estratégias de extracção em domínio fechado, tais como as avaliadas na Parte II desta tese. No nosso caso, a inexistência de sistemas de resolução (para as línguas analisadas) durante a realização dos testes em domínio fechado impediu conhecer o seu impacto neste tipo de tarefas.

Em conclusão, o trabalho realizado no presente capítulo sugere que a incorporação da resolução de correferência no processo de extracção de informação aberta emerge como uma estratégia promissora para a extracção de informação estruturada a partir de texto livre.

Capítulo 9

Conclusões e Trabalho Futuro

Este capítulo sumariza as principais conclusões obtidas em cada uma das partes desta tese, bem como as contribuições fruto de todo o trabalho realizado. Finalmente, apontam-se direcções para o trabalho futuro.

9.1. Principais conclusões

Parte I

A primeira parte da tese focou-se na adaptação e no desenvolvimento de recursos e de ferramentas para diferentes tarefas do processamento da linguagem natural em português e em galego.

Assim, o Capítulo 2 descreveu a adaptação dos módulos de tokenização, de segmentação de orações, de análise morfológica com lematização e de *PoS-tagging* para as duas línguas referidas, bem como de diferentes corpora e *tagsets* para a anotação morfossintáctica.

A seguir, o Capítulo 3 apresentou o processo de adaptação de ferramentas de reconhecimento de entidades mencionadas, assim como uma estratégia de classificação semântica que utiliza regras e recursos obtidos de modo semiautomático.

As contribuições da Parte I permitiram processar e anotar corpora em diferentes línguas com informação necessária para aplicar as diferentes estratégias de extracção de relações utilizadas nas Partes II e III deste trabalho.

Parte II

A segunda parte da tese começou fazendo uma revisão do estado-da-arte das tarefas de extracção de relações (Capítulo 4), sendo algumas das estratégias apresentadas implementadas para português e espanhol nos capítulos subsequentes.

Assim, o Capítulo 5 avaliou uma estratégia de supervisão-distante para a obtenção de corpora anotados de modo semiautomático, posteriormente utilizados para treinar sistemas de extracção de relações que aproveitam a redundância de informação na Web. Os resultados mostraram que a estratégia de supervisão-distante funciona bem com algumas relações, mas o seu desempenho reduziu-se ao estendê-la a novas relações.

Derivado das conclusões do capítulo anterior, o Capítulo 6 utilizou corpora de aprendizagem corrigidos manualmente para avaliar o impacto de vários atributos de carácter linguístico em classificadores supervisionados. Diferentes testes mostraram que a generalização da informação lexical (mediante lematização) em combinação com estruturas pseudo-sintácticas (especialmente bigramas de lemas) permite construir classificadores com muito bom desempenho na extracção de relações biográficas. Além disso, a utilização de informação sintáctica (mediante os caminhos de dependências mais curtos entre as entidades relacionadas) adiciona conhecimento útil para a extracção, embora a melhora causada por estes atributos não seja determinante.

Por último, o Capítulo 7 apresentou uma nova estratégia baseada em regras para a extracção de relações semânticas. Este método utiliza técnicas de compressão de orações para simplificar os contextos linguísticos nos quais extrair relações. Além disso, aproveita a supervisão-distante para obter —de modo semiautomático— regras de extracção de alta precisão. A combinação da simplificação textual com as regras de extracção permitiu aumentar o *recall* mantendo uma alta precisão, pelo que a estratégia permite criar de modo rápido sistemas de extracção com bom desempenho.

Parte III

A terceira e última parte da tese apresentou, num único capítulo (8), um sistema de resolução de correferência de entidades pessoa e uma análise do impacto desta tarefa na extracção de informação aberta.

Assim, o sistema apresentado utiliza uma abordagem centrada nas entidades e uma arquitectura multi-passe que lhe permite obter bons resultados em diferentes línguas e cenários.

Em relação à extracção de relações, levou-se a cabo um conjunto de testes que provaram a melhora dos resultados da extracção de informação aberta quando realizada depois de um sistema de resolução de correferência.

Em suma, na presente tese avaliaram-se diferentes estratégias para a extracção de relações semânticas de carácter biográfico de textos em português, espanhol e galego, objectivo que tinha sido definido como *principal* na introdução deste trabalho.

Os métodos não foram comparados individualmente, uma vez que cada um deles tem necessidades e finalidades específicas. Assim, avaliaram-se técnicas para minimizar o esforço na construção de corpora, útil tanto para treinar classificadores como para obter regras de alta precisão. O impacto de diferentes níveis de conhecimento de base linguística também foi analisado no treino de classificadores supervisionados. Para além disso, foi utilizada análise linguística para realizar compressão de orações com a finalidade de aumentar o *recall* das extracções. Finalmente, um sistema de extracção de informação aberta foi avaliado em combinação com uma nova ferramenta de resolução de correferência de entidades pessoa, aumentando tanto a precisão como o *recall* na extracção de grandes quantidades de informação relacionada semanticamente.

Por um lado, as várias avaliações mostraram que a informação linguística (nomeadamente a produzida pela classificação semântica e pela lematização) é crucial para a extracção de informação mediante técnicas de aprendizagem automática, embora o seu desempenho se possa ver afectado negativamente se os sistemas de análise produzem ruído (por exemplo, os *parsers* sintácticos).

Por outro lado, as abordagens baseadas em regras obtiveram uma alta qualidade de análise nalgumas tarefas, tais como na extracção de relações do Capítulo 7 ou na resolução de correferência e na extracção de informação aberta do Capítulo 8.

Para avaliar os vários métodos de extracção de relações, foi preciso um trabalho de desenvolvimento e de adaptação de diversas ferramentas e recursos de processamento da linguagem natural para as três línguas alvo (objectivos paralelos). Neste sentido, alguns dos sistemas apresentados revelaram-se como as primeiras ferramentas para a realização de várias tarefas de processamento da linguagem natural em galego, e as primeiras com licenças livres para alguns tipos de análise em português.

9.2. Contribuições

Apresentadas as principais conclusões da tese, esta secção compila as diferentes ferramentas e recursos desenvolvidos e adaptados durante a realização do trabalho. Depois, mostram-se as publicações produzidas em cada capítulo.

Ferramentas e recursos

Uma vez que esta tese teve um carácter essencialmente prático, o trabalho realizado em cada um dos capítulos derivou na disponibilização de diferentes ferramentas e recursos, referidos a seguir:

- Módulos de segmentação de orações para português e galego
- Módulos de tokenização para português e galego
- Módulos de análise morfológica para português e galego
- Módulos de análise morfossintáctica para português e galego
- Adaptação do corpus Bosque 8.0 ao *standard* EAGLES
- Adaptação do léxico LABEL-LEX (SW) ao *standard* EAGLES
- Corpora com anotação morfossintáctica para diferentes variedades de português e galego
- Léxicos (e ampliação de léxicos já existentes) com anotação morfossintáctica para português e galego
- Módulos de identificação de nomes próprios para português e galego
- Módulos de classificação de nomes próprios para português e galego
- Módulos de reconhecimento de expressões numéricas, quantidades e horas para português e galego
- Anotação de entidades mencionadas no corpus Bosque 8.0
- Corpus de teste com anotação manual de entidades mencionadas para galego

– Corpora com anotação de relações biográficas para português e espanhol

– Ferramenta de resolução de correferência de entidades pessoa para português, espanhol e galego

– Corpora com anotação correferencial de entidades pessoa para português, espanhol e galego

As ferramentas e recursos referidos, cujos desempenhos são próximos do estado-da-arte —ou similares a sistemas com os mesmos objectivos— são todos disponibilizados sob licenças livres (GPL 3), ou mantendo a licença original do recurso, no caso das adaptações.[1]

Publicações

Com excepção dos capítulos introdutório (1) e de revisão do estado-da-arte (4), para além deste último, os restantes capítulos produziram as seguintes publicações:

Capítulo 2

– Garcia, Marcos e Pablo Gamallo, 2010a. Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática. Revista para o Processamento Automático das Línguas Ibéricas*, 2(2): 59–67

– Garcia, Marcos e Pablo Gamallo, 2010b. Do processamento morfológico à análise sintáctica de corpora multilíngue. In *Actas del XXXIX Simposio Internacional de la Sociedad Española de Lingüística*. Santiago de Compostela.

– Garcia, Marcos e Pablo Gamallo, 2010c. Using Morphosyntactic Post-Processing to Improve PoS-tagging Accuracy. In *Proceedings of the 9th International Conference on Computational Processing of Portuguese Language (PROPOR 2010). Extended Activities*. Porto Alegre.

– Garcia, Marcos, Pablo Gamallo, Iria Gayo e Miguel A. Pousada Cruz, 2014. PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*, 53: 95–101.

[1] As contribuições são acessíveis em `http://gramatica.usc.es/~marcos/phd.html`

Capítulo 3

– Garcia, Marcos, Iria Gayo e Isaac González López, 2012. Identificação e Classificação de Entidades Mencionadas em Galego. *Estudos de Lingüística Galega*, 4: 13–25.

Capítulo 5

– Garcia, Marcos e Pablo Gamallo, 2011b. Evaluating Various Features on Semantic Relation Extraction. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov e Nikolai Mikolov, editores, *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, páginas 721–726. Hissar.

– Garcia, Marcos e Pablo Gamallo, 2011c. An Exploration of the Linguistic Knowledge for Semantic Relation Extraction in Spanish. In Patrick Saint-Dizier e Rutu Mehta-Melkar, editores, *Proceedings of the Joint Workshop FAM-LbR/KRAQ'11. Learning by Reading and its Applications in Intelligent Question-Answering at 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, páginas 7–12. Barcelona.

Capítulo 6

– Garcia, Marcos e Pablo Gamallo, 2013. Exploring the Effectiveness of Linguistic Knowledge for Biographical Relation Extraction. *Natural Language Engineering*, CJO 2013: 1–33. doi:10.1017/S1351324913000314.

Capítulo 7

– Garcia, Marcos e Pablo Gamallo, 2011a. Dependency-Based Text Compression for Semantic Relation Extraction. In Kuzman Ganchev Preslav Nakov, Zornitsa Kozareva e Jerry Hobbs, editores, *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition (IEKA 2011) at 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, páginas 21–28. Hissar.

– Garcia, Marcos e Pablo Gamallo, 2011e. A Weakly-Supervised Rule-Based Approach for Relation Extraction. In *Proceedings of the XIV Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2011). Workshop on Knowledge Extraction and Exploitation from Semi-structures Online Sources (KEESOS)*. La Laguna.

Capítulo 8

- Garcia, Marcos e Pablo Gamallo, 2011d. Resolución de Correferencia de Nombres de Persona para Extracción de Información Biográfica. *Procesamiento del Lenguaje Natural*, 47: 283–291.

- Garcia, Marcos e Pablo Gamallo, 2014a. Entity-Centric Coreference Resolution of Person Entities for Open Information Extraction. *Procesamiento del Lenguaje Natural*, 53: 25–32.

- Garcia, Marcos e Pablo Gamallo, 2014b. An Entity-Centric Coreference Resolution System for Person Entities with Rich Linguistic Information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 741–752. Dublin.

- Garcia, Marcos e Pablo Gamallo, 2014c. Multilingual corpora with coreference annotation of person entities. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, editores, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, páginas 3229–3233. European Language Resources Association. Reykjavik.

9.3. Trabalho futuro

Esta tese abordou a extracção de relações semânticas como uma tarefa mais dentro de outros processos do processamento da linguagem natural, desde a segmentação de orações à resolução de correferência. Neste trabalho trataram-se, portanto, muitas tarefas diferentes que são susceptíveis de melhora.

Tendo em conta todo o conjunto de processos aplicados para a extracção de relações nas diferentes línguas, torna-se necessária uma análise que identifique as principais fontes de erros de cada um dos módulos utilizados. Os resultados desta análise podem permitir focar os esforços em melhorar aquelas ferramentas cujas deficiências causam os erros mais importantes (tanto do ponto de vista quantitativo como qualitativo) em processos subsequentes.

Em relação à extracção de relações semânticas, as diferentes estratégias apresentadas na Parte II mostraram que existem aspectos a melhorar em função dos objectivos da extracção.

Assim, para a realização de extracções de um grande número de relações, é preciso avaliar novas estratégias de extensão da supervisão-distante a novos domínios. Para a extracção em domínio fechado, a adaptação dos classificadores (mediante a introdução de regras ou de atributos específicos) a relações semânticas concretas pode melhorar notoriamente o seu desempenho.

Por último, é preciso referir que a Parte III foi concebida já como um ponto de partida para o trabalho futuro da presente tese. A aplicação da resolução de correferência orientada à extracção de relações (ou à extracção de informação aberta) é uma combinação promissora dentro deste âmbito. Neste sentido, a utilização de estratégias de *clustering* para organizar o conhecimento extraído, bem como a extensão da resolução de correferência a vários documentos ou a utilização de técnicas de *entity-linking* pode permitir a obtenção de informação de maior qualidade. Além disso, a adaptação de ferramentas como as apresentadas nesta tese a ambientes *Big Data* torna-se imprescindível para realizar tarefas de processamento da língua natural na Web em tempo-real (Gamallo *et al.*, 2014; Abuín *et al.*, 2014).

Apêndice A

Tagsets utilizados na etiquetação morfossintáctica

Adjectivos

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Adjectivo	A
2	Tipo	Qualificativo	C
		Ordinal	O
3	Grau	Aumentativo	A
		Diminutivo	D
		Superlativo	S
4	Género	Masculino	M
		Feminino	F
		Comum	C
5	Número	Singular	S
		Plural	P
		Invariável	N

Advérbios

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Advérbio	R
2	Tipo	Geral	G
		Negativo	N

Determinantes

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Determinante	D
2	Tipo	Artigo	A
		Demonstrativo	D
		Indefinido	I
		Possessivo	P
3	Pessoa	1ª/2ª/3ª	1/2/3
4	Género	Masculino	M
		Feminino	F
		Comum	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariável	N
6	Possuidor	Singular	S
		Plural	P

Verbos

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Verbo	V
2	Tipo	Principal	M
3	Modo	Gerúndio	G
		Indicativo	I
		Imperativo	M
		Infinitivo	N
		Particípio	P
		Conjuntivo	S
4	Tempo	Futuro do Pretérito	C
		Mais-que-Perfeito	M
		Futuro	F
		Imperfeito	I
		Presente	P
		Perfeito	S
5	Pessoa	1ª/2ª/3ª	1/2/3
6	Número	Singular	S
		Plural	P

Conjunções

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Conjunção	C
2	Tipo	Coordenativa	C
3		Subordinativa	S

Preposições

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Aposição	S
2	Tipo	Preposição	P
3	Forma	Simples	S

Nomes

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Nome	N
2	Tipo	Comum	C
		Próprio	P
3	Género	Masculino	M
		Feminino	F
		Comum	C
4	Número	Singular	S
		Plural	P
		Invariável	N
7	Grau	Aumentativo	A
		Diminutivo	D

Pronomes

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Pronome	P
2	Tipo	Demonstrativo	D
		Exclamativo	E
		Indefinido	I
		Pessoal	P
		Relativo	R
		Interrogativo	T
		Possessivo	X
3	Pessoa	1ª/2ª/3ª	1/2/3
4	Género	Masculino	M
		Feminino	F
		Comum	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariável	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblíquo	O
7	Possuidor	Singular	S
		Plural	P

Numerais

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Numeral	Z

Interjeições

Elemento	*Atributo*	*Valor*	*Tag*
1	Categoria	Interjeição	I

Tabela A.1: Formato do *tagset* estreito (*standard* EAGLES).

Categoria	**PoS-tag**
Adjectivo Ordinal	AO
Adjectivo Qualificativo	AQ
Conjunção Subordinativa	CS
Conjunção Coordenativa	CC
Determinante Artigo	DA
Determinante Demonstrativo	DD
Determinante Indefinido	DI
Determinante Possessivo	DP
Interjeição	I
Nome Comum	NC
Nome Próprio	NP
Pronome Demonstrativo	PD
Pronome Exclamativo	PE
Pronome Indefinido	PI
Pronome Pessoal	PP
Pronome Relativo	PR
Pronome Interrogativo	PT
Pronome Possessivo	PX
Advérbio Geral	RG
Advérbio Negativo	RN
Preposição	SP
Verbo: Gerúndio	VG
Verbo: Modo Indicativo	VI
Verbo: Modo Imperativo	VM
Verbo: Infinitivo	VN
Particípio	VP
Verbo: Modo Conjuntivo	VS
Numeral	Z

Tabela A.2: *Tagset* largo (*SingleTags*).

Categoria	**PoS-tag**
Adjectivo	AD
Advérbio	AV
Conjunção Coordenativa	CC
Conjunção Subordinativa	CS
Determinante (Definido/Indefinido)	DT
Determinante Demonstrativo	DD
Determinante Possessivo	DP
Preposição	PS
Verbo	VB
Particípio	VP
Nome Comum	NC
Nome Próprio	NP
Pronome Demonstrativo	PD
Pronome Exclamativo	PE
Pronome Indefinido	PI
Pronome Pessoal	PP
Pronome Relativo	PR
Pronome Interrogativo	PT
Pronome Possessivo	PX
Interjeição	I
Numeral	Z
Contracções com a Preposição *de*	DC
Contracções com a Presposição *por*	PC
Pontuação	F*
Datas/Horas	W
Expressões numéricas/Quantidades	Z*

Tabela A.3: *Tagset* largo para *PoS-tagging* em diferentes variedade de português.

Bibliografia

Abuín, José Manuel, Juan Carlos Pichel, Tomás Fernández Pena, Pablo Gamallo e Marcos Garcia, 2014. Perldoop: Efficient Execution of Perl Scripts on Hadoop Clusters. In *Proceedings of the 2014 IEEE International Conference on Big Data (IEEE Big Data 2014)*. Washington DC.

Agichtein, Eugene, 2005. *Extracting Relations from Large Text Collections*. Dissertação de Doutoramento, Columbia University, New York.

Agichtein, Eugene e Luis Gravano, 2000. Snowball: Extracting Relations from Large Plain-Text Collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, páginas 85–94.

Aguado de Cea, Guadalupe, Asunción Gómez-Pérez, Elena Montiel-Ponsoda e Mari Carmen Suárez-Figueroa, 2008. Natural language-based approach for helping in the reuse of ontology design patterns. In *Knowledge Engineering: Practice and Patterns*, páginas 32–47. Springer-Verlag.

Aguado de Cea, Guadalupe, Asunción Gómez-Pérez, Elena Montiel-Ponsoda e Mari Carmen Suárez-Figueroa, 2009. Using linguistic patterns to enhance ontology development. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD 2009)*, páginas 206–213.

Aires, Raquel V. Xavier, 2000. *Implementação, adaptação, combinação e avaliação de etiquetadores para o Português do Brasil*. Dissertação de Mestrado, Instituto de Ciências Matemáticas, Universidade de São Paulo, São Paulo.

Akbik, Alan e Jürgen Broß, 2009. Wanderlust: Extracting Semantic Relations from Natural Language Text Using Dependency Grammar Patterns. In *Proceedings of the Workshop on*

Semantic Search (SemSearch 2009) at the 18th International World Wide Web Conference (WWW 2009). Madrid.

Aluísio, Sandra M., Gisele M. Pinheiro, Marcelo Finger, M. Graças Volpe Nunes e Stella E. Tagnin, 2003. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In *Proceedings of Corpus Linguistics*, páginas 14–21.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró e Muntsa Padró, 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association, Genoa.

Aussenac-Gilles, Nathalie e Marie-Paule Jacques, 2006. Designing and Evaluating Patterns for Ontology Enrichment from Texts. In *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management. Managing Knowledge in a World of Networks (EKAW 2006)*, Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence (LNCS/LNAI), páginas 158–165. Springer-Verlag.

Bagga, Amit e Breck Baldwin, 1998. Algorithms for scoring coreference chains. In *Proceedings of the Workshop on Linguistic Coreference at the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, volume 1, páginas 563–566.

Baldwin, Breck, 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, páginas 38–45. Association for Computational Linguistics.

Banko, Michel e Oren Etzioni, 2008. The Tradeoffs Between Open and Traditional Relation Extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, volume 8, páginas 28–36. Association for Computational Linguistics, Columbus.

Banko, Michele e Eric Brill, 2001. Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing. In *Proceedings of the Conference on Human Language Technology*, páginas 1–5. Association for Computational Linguistics.

Banko, Michele, Michael J Cafarella, Stephen Soderland, Matt Broadhead e Oren Etzioni, 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI 2007)*, páginas 2670–2676. Morgan Kaufmann Publishers Inc.

Barcala, Francisco Mario, Eva Maria Domínguez Noya, Pablo Gamallo Otero, Marisol López Martínez, Eduardo Miguel Moscoso Mato, Guillermo Rojo, María Paula Santalla del Río e Susana Sotelo Docío, 2007. A corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a in a Minority Language. In Zygmunt Vetulani, editor, *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 3rd Language & Technology Conference*, páginas 359–363. Wydawnictwo Poznaskie Sp. z o.o, Poznan.

Bick, Eckhard, 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Dissertação de Doutoramento, University of Aarhus, Denmark.

Bick, Eckhard, 2006. Functional aspects on Portuguese NER. In Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Language (PROPOR 2006)*, volume 3960 de *Lecture Notes in Computer Science (LNCS)*, páginas 260–263. Springer-Verlag.

Bird, Steven e T. Mark Ellison, 1994. One-level phonology: Autosegmental representations and rules as finite automata. *Computational Linguistics*, 20(1): 55–90.

Black, Paul E., 2013. Finite state machine. In Vreda Pieterse e Paul E. Black, editores, *Dictionary of Algorithms and Data Structures*. National Institute of Standards and Technology. Http://www.nist.gov/dads/HTML/finiteStateMachine.html.

Bollegala, Danushka Tarupathi, Yutaka Matsuo e Mitsuru Ishizuka, 2010. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, páginas 151–160. Association for Computing Machinery.

Bontcheva, Kalina, Marin Dimitrov, Diana Maynard, Valentin Tablan e Hamish Cunningham, 2002. Shallow Methods for Named Entity Coreference Resolution. In *Proceedings of the*

Workshop on Chaınes de références et résolveurs d'anaphores at Traitement Automatique des Langues Naturelles (TALN 2002).

Bouma, Gosse, Walter Daelemans, Iris Hendrickx, Véronique Hoste e Anne-Marie Mineur, 2007. The COREA-project. Manual for the annotation of coreference in Dutch texts. Relatório técnico, University Groningen.

Branco, António e João Silva, 2003. Contractions: breaking the tokenization-tagging circularity. In *Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language (PROPOR 2003)*, volume 2721 de *Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence (LNCS/LNAI)*, páginas 167–170. Springer-Verlag.

Branco, António e João Silva, 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, páginas 507–510. European Language Resources Association, Paris.

Brants, Thorsten, 2000. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*. Association for Computational Linguistics.

Brill, Eric, 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4): 543–565.

Brin, Sergey, 1998. Extracting patterns and relations from the World Wide Web. In *Proceedings of the WebDB Workshop at the 6th International Conference on Extending Database Technology (EDBT 1998)*, páginas 172–183. València.

Bruckschen, Mírian, José Guilherme Camargo de Souza, Renata Vieira e Sandro Rigo, 2008. Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. In Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo 14, páginas 247–260. Linguateca.

Bunescu, Razvan C. e Raymond J. Mooney, 2005. A Shortest Path Dependency Kernel for Relation Extraction. In Associacion for Computational Linguistics, editor, *Proceedings of*

the Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), páginas 724–731. Association for Computational Linguistics, Vancouver.

Bunescu, Razvan C. e Raymond J. Mooney, 2007. Learning to Extract Relations from the Web Using Minimal Supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, volume 45, páginas 576–583. Association for Computational Linguistics.

Cardoso, Nuno, 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. In Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo 11, páginas 195–211. Linguateca.

Carreras, Xavier, Isaac Chao, Lluís Padró e Muntsa Padró, 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, páginas 239–242. European Language Resources Association, Lisboa.

Carreras, Xavier, Lluís Màrquez e Lluís Padró, 2002. Named Entid Extraction Using AdaBoost. In *Proceedingf of the 6th Conference on Natural Language Learning (CoNLL 2002)*, volume 20, páginas 1–4. Association for Computational Linguistics.

Castells, Manuel, 1996. The Rise of the Network Society. In *The Information Age. Economy, Society and Culture*, volume 1. Blackwell, Cambridge, MA / Oxford, UK.

Chang, Chih-Chung e Chih-Jen Lin, 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3): 1–27.

Chaves, Amanda Rocha e Lucia Helena Machado Rino, 2007. A resolução de pronomes anafóricos do português com base em heurísticas que apontam o antecedente. In *VI Congresso de Pós-Graduação da UFSCar*, volume 2, páginas 1272–1273. São Carlos, São Paulo.

Chaves, Marciria Silveira, 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM. In Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, capítulo 13, páginas 231–245. Linguateca.

Chinchor, Nance e Lynette Hirschmann, 1997. MUC-7 Coreference Task Definition (Version 3.0). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, volume 7. Association for Computational Linguistics.

Coelho, Thiago Thomes e Ariadne Maria Brito Rizzoni Carvalho, 2005. Uma adaptação do algoritmo de Lappin e Leass para resolução de anáforas em português. In *III Workshop em Tecnologia da Informação e da Linguagem Humana–TIL. Proceedings of XXV Congresso da Sociedade Brasileira de Computação*, páginas 2069–2078.

Cohen, William W., Pradeep Ravikumar e Stepehn G. Fienberg, 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*, páginas 73–78.

Collovini, Sandra, 2014. *Extração de Relações de Domínio de Organizações para o Português*. Dissertação de Doutoramento, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

Collovini, Sandra, Thiago I. Carbonel, Juliana Fuchs Thiesen, Jorge C. Coelho, Lúcia Rino e Renata Vieira, 2007. Summ-it: Um corpus anotado com informações discursivas visandoa sumarização automática. In *Proceedings of the V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2007)*. Rio de Janeiro.

Corro, Luciano Del e Rainer Gemulla, 2013. ClausIE: clause-based Open Information Extraction. In *Proceedings of the 22nd international conference on World Wide Web (WWW 2013)*, páginas 355–366. Rio de Janeiro.

Costa, Francisco e António Branco, 2012. Extracting Temporal Information from Portuguese Texts. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, volume 7243 de *Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence (LNCS/LNAI)*, páginas 99–105. Springer-Verlag.

Cuevas, Ramon Ré Moya e Invandré Paraboni, 2008. A machine learning approach to Portuguese pronoun resolution. In *Proceedings of the XI Conferencia Iberamia de Inteligencia Artificial*, Advances on Artificial Intelligence, páginas 262–271. Springer-Verlag.

Cunha, Celso e Luís F. Lindley Cintra, 1984. *Nova gramática do português contemporâneo*. João Sá da Costa, Lisboa.

de Souza, José Guilherme C., Patrícia Gonçalves e Renata Vieira, 2008. Learning Coreference Resolution for Portuguese Texts. In *Computational Processing of the Portuguese Language*, páginas 153–162. Springer-Verlag.

Domínguez Noya, Eva María, 2013. *Etiquetaxe e desambiguación automáticas en galego: o sistema XIADA*. Dissertação de Doutoramento, Universidade de Santiago de Compostela.

Domínguez Noya, Eva María, Francisco Mario Barcala e Miguel Ángel Molinero, 2009. Avaliación dun etiquetador automático estatístico para o galego actual: Xiada. *Cadernos de Lingua*, 30/31: 151–193.

Eleutério, Samuel, Elisabete Ranchhod, Cristina Mota e Paula Carvalho, 2003. Dicionários Electrónicos do Português. Características e Aplicações. In *Actas del VIII Simposio Internacional de Comunicación Social*, páginas 636–642. Santiago de Cuba.

Etzioni, Oren, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld e Alexander Yates, 2004. Web-scale information extraction in KnowItAll. In *Proceedings of the 13th international conference on World Wide Web (WWW 2004)*, páginas 100–110. Association for Computing Machinery.

Etzioni, Oren, Anthony Fader, Janara Christensen, Stephen Soderland e Mausam, 2011. Open information extraction: The second generation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, páginas 3–10. AAAI Press.

Fader, Anthony, Stephen Soderland e Oren Etzioni, 2011. Identifying relations for Open Information Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, páginas 1535–1545. Association for Computational Linguistics, Edinburgh.

Ferreira, Eduardo, João Balsa e António Branco, 2007. Combining Rule-based and Statistical Methods for Named Entity Recognition in Portuguese. In *Actas da 5a Workshop em Tecnologias da Informação e da Linguagem Humana*. Sociedade Brasileira de Computação, Rio de Janeiro.

Ferrández, Antonio e Jesús Peral, 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL 2000)*, páginas 166–172. Association for Computational Linguistics, Hong Kong.

Ferrández, Óscar, Zornitsa Kozareva, Antonio Toral, Rafael Muñoz e Andrés Montoyo, 2007. Tackling HAREM's portuguese named entity recognition task with spanish resources. In Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, capítulo 11, páginas 137–144. Linguateca.

Finger, Marcelo, 2000. Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho-Brahe. In Maria das Graças Volpe Nunes, editor, *Actas do 5o Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000)*, páginas 141–154. ICMC/USP, São Paulo.

Finkel, Jenny Rose, Trond Grenager e Christopher Manning, 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, páginas 363–370. Association for Computational Linguistics, Ann Arbor.

Finkelstein-Landau, Michal e Emmanuel Morin, 1999. Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods. In *Proceedings of the International Workshop on Ontological Engineering on the Global Information Infrastructure*, páginas 71–80.

Fisher, Davir, Stepehn Soderland, Fangfang Feng e Wendy Lehnert, 1995. Description of the UMass system as used for MUC-6. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, páginas 127–140. Association for Computational Linguistics.

Fleischman, Michael, Eduard Hovy e Abdessamad Echihabi, 2003. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, páginas 1–7. Association for Computational Linguistics, Sapporo.

Freitas, Maria Cláudia de, 2007. *Elaboração automática de ontologias de domínio: discussão e resultados*. Dissertação de Doutoramento, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

Gamallo, Pablo, 2012. A Depurative Strategy for Dependency Parsing with Finite-State Transducers. In *Seminario da Rede Galega de Recursos Lingüísticos para unha Sociedade do Coñecemento (RELISCO)*. Universidade da Coruña.

Gamallo, Pablo e Marcos Garcia, 2011. A resource-based method for named entity extraction and classification. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, volume 7026/2011 de *Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence (LNCS/LNAI)*, páginas 610–623. Springer-Verlag.

Gamallo, Pablo, Marcos Garcia e Santiago Fernández-Lanza, 2012. Dependency-based Open Information Extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, páginas 10–18. Association for Computational Linguistics, Avignon.

Gamallo, Pablo, Marcos Garcia, Isaac González López, Marta Muñoz e Iria Gayo, 2013. Learning verb inflection using Cilenis conjugators. *The Eurocall Review*, 21(1): 12–19.

Gamallo, Pablo e Isaac González López, 2011. A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. *International Journal of Corpus Linguistics*, 16(1): 45–71.

Gamallo, Pablo, Juan Carlos Pichel, Marcos Garcia, José Manuel Abuín e Tomás Fernández Pena, 2014. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data. *Procesamiento del Lenguaje Natural*, 53: 17–24.

Garcia, Marcos e Pablo Gamallo, 2010a. Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática. Revista para o Processamento Automático das Línguas Ibéricas*, 2(2): 59–67.

Garcia, Marcos e Pablo Gamallo, 2010b. Do processamento morfológico à análise sintáctica de corpora multilíngue. In *Actas del XXXIX Simposio Internacional de la Sociedad Española de Lingüística*. Santiago de Compostela.

Garcia, Marcos e Pablo Gamallo, 2010c. Using Morphosyntactic Post-Processing to Improve PoS-tagging Accuracy. In *Proceedings of the 9th International Conference on Computational Processing of Portuguese Language (PROPOR 2010). Extended Activities*. Porto Alegre.

Garcia, Marcos e Pablo Gamallo, 2011a. Dependency-Based Text Compression for Semantic Relation Extraction. In Kuzman Ganchev Preslav Nakov, Zornitsa Kozareva e Jerry Hobbs, editores, *Proceedings of the Workshop on Information Extraction and Knowledge Acquisition (IEKA 2011) at 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, páginas 21–28. Hissar.

Garcia, Marcos e Pablo Gamallo, 2011b. Evaluating Various Features on Semantic Relation Extraction. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov e Nikolai Mikolov, editores, *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, páginas 721–726. Hissar.

Garcia, Marcos e Pablo Gamallo, 2011c. An Exploration of the Linguistic Knowledge for Semantic Relation Extraction in Spanish. In Patrick Saint-Dizier e Rutu Mehta-Melkar, editores, *Proceedings of the Joint Workshop FAM-LbR/KRAQ'11. Learning by Reading and its Applications in Intelligent Question-Answering at 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, páginas 7–12. Barcelona.

Garcia, Marcos e Pablo Gamallo, 2011d. Resolución de Correferencia de Nombres de Persona para Extracción de Información Biográfica. *Procesamiento del Lenguaje Natural*, 47: 283–291.

Garcia, Marcos e Pablo Gamallo, 2011e. A Weakly-Supervised Rule-Based Approach for Relation Extraction. In *Proceedings of the XIV Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2011). Workshop on Knowledge Extraction and Exploitation from Semi-structures Online Sources (KEESOS)*. La Laguna.

Garcia, Marcos e Pablo Gamallo, 2013. Exploring the Effectiveness of Linguistic Knowledge for Biographical Relation Extraction. *Natural Language Engineering*, CJO2013: 1–33. doi:10.1017/S1351324913000314.

Garcia, Marcos e Pablo Gamallo, 2014a. Entity-Centric Coreference Resolution of Person Entities for Open Information Extraction. *Procesamiento del Lenguaje Natural*, 53: 25–32.

Garcia, Marcos e Pablo Gamallo, 2014b. An Entity-Centric Coreference Resolution System for Person Entities with Rich Linguistic Information. In *Proceedings of COLING 2014, the*

25th International Conference on Computational Linguistics: Technical Papers, páginas 741–752. Dublin.

Garcia, Marcos e Pablo Gamallo, 2014c. Multilingual corpora with coreference annotation of person entities. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk e Stelios Piperidis, editores, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, páginas 3229–3233. European Language Resources Association, Reykjavik.

Garcia, Marcos, Pablo Gamallo, Iria Gayo e Miguel A. Pousada Cruz, 2014. PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*, 53: 95–101.

Garcia, Marcos, Iria Gayo e Isaac González López, 2012. Identificação e Classificação de Entidades Mencionadas em Galego. *Estudos de Lingüística Galega*, 4: 13–25.

Garera, Nikesh e David Yarowsky, 2009. Structural, transitive and latent models for biographic fact extraction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, páginas 300–308. Association for Computational Linguistics, Atenas.

Giesbrecht, Eugenie e Stefan Evert, 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*. Donostia.

Gordon, Peter C. e Randall Hendrick, 1998. The representation and processing of coreference in discourse. *Cognitive Science*, 22(4): 389–424.

Graña, Jorge, Francisco Mario Barcala e Jesús Vilares, 2002. Formal Methods of Tokenization for Part-of-Speech Tagging. In *Computational linguistics and intelligent text processing*, volume 2276/2002, páginas 123–144. Springer-Verlag.

Grishman, Ralph, 2010. The impact of task and corpus on Event Extraction Systems. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, páginas 2928–2931. European Language Resources Association, Valletta.

Haghighi, Aria e Dan Klein, 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting on Association for*

Computational Linguistics (ACL 2007), volume 45, páginas 848–855. Association for Computational Linguistics, Praga.

Hearst, Marti A., 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, volume 2, páginas 539–545. Association for Computational Linguistics, Morristown.

Hoffmann, Raphael, Congle Zhang e Daniel S. Weld, 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, páginas 286–295. Association for Computational Linguistics, Uppsala.

Hoste, Véronique, 2005. *Optimization issues in Machine Learning of Coreference Resolution*. Dissertação de Doutoramento, Universiteit Antwerpen.

Jiang, Jing e ChengXiang Zhai, 2007. A Systematic Exploration of the Feature Space for Relation Extraction. In *Proceedings of the Human Language Technology Conference (HLT/NAACL 2007)*, páginas 113–120. Association for Computational Linguistics, Rochester.

Jijkoun, Valentin, Maarten De Rijke e Jori Mur, 2004. Information Extraction for Question Answering: Improving Recall Through Syntactic Patterns. In *Proceedings of the 20th Internacional Conference on Computational Linguistics (COLING 2004)*, páginas 1284–1290. Genebra.

Jurafsky, Dan e James H. Martin, 2009. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2 edição.

Kambhatla, Nanda, 2004. Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, páginas 178–181. Association for Computational Linguistics, Barcelona.

Kübler, Sandra, Ryan McDonald e Joakim Nivre, 2009. *Dependency Parsing*. Morgan and Claypool Publishers.

Lappin, Shalom e Herbert J. Leass, 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4): 535–561.

Leach, Geoffrey e Andrew Wilson, 1996. Recommendations for the Morphosyntactic Annotation of Corpora. Relatório técnico, Expert Advisory Group on Language Engineering Standard (EAGLES).

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu e Dan Jurafsky, 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4): 885–916.

Lin, Dekang, 2003. Dependency-based Evaluation of MINIPAR. In *Treebanks: Building and Using Parsed Corpora*, volume 20, páginas 317–329.

Liu, Xiaojiang, Zaiqing Nie, Nenghai Yu e Ji-Rong Wen, 2010. BioSnowball: automated population of Wikis. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, páginas 969–978. Association for Computing Machinery.

Luo, Xiaoqiang, 2005. On Coreference Resolution Performance Metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, páginas 25–32. Association for Computational Linguistics, Vancouver.

Malvar, Paulo, José Ramom Pichel, Óscar Senra, Pablo Gamallo e Alberto García, 2010. Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. *Linguamática. Revista para o Processamento Automático das Línguas Ibéricas*, 2(2): 31–38.

Mann, Gideon S., 2002. Fine-Grained Proper Noun Ontologies for Question Answering. In *Proceedings of the 2002 Workshop on Building and using Semantic Networks (SemaNet 2002)*, volume 11. Association for Computational Linguistics, Taipei.

Marques, Nuno e Gabriel Lopes, 2001. Tagging with Small Training Corpora. In *Proceedings of the International Conference on Intelligent Data Analysis*, volume 2189 de *Lecture Notes in Computer Science. Lecture Notes on Artificial Intelligence (LNCS/LNAI)*, páginas 63–72. Springer-Verlag.

McCarthy, Joseph e Wendy G. Lehnert, 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, páginas 1050–1055. Montreal.

Mika, Peter, Massimiliano Ciaramita, Hugo Zaragoza e Jordi Atserias, 2008. Learning to tag and tagging to learn: A case study on Wikipedia. *IEEE Intelligent Systems*, 23(5): 26–33.

Mikheev, Andrei, Claire Grover e Marc Moens, 1998. Description of the LTG system used for MUC-7. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

Mintz, Mike, Steven Bills, Rion Snow e Dan Jurafsky, 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009)*, páginas 1003–1011. Association for Computational Linguistics.

Mitchell, Tom, 1997. *Machine Learning*. McGraw Hill.

Mitkov, Ruslan, 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (ACL/COLING 1998)*, volume 2, páginas 869–875. Association for Computational Linguistics, Montreal.

Mitkov, Ruslan, Richard Evans, Constantin Orasan, Catalina Barbu, Lisa Jones e Violeta Sotirova, 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, páginas 49–58. Lancaster.

Molina, Alejandro, Iria da Cunha, Juan-Manuel Torres-Moreno e Patricia Velazquez-Morales, 2011. La compresión de frases: un recurso para la optimización de resumen automático de documentos. *Linguamática. Revista para o Processamento Automático das Línguas Ibéricas*, 2(3): 13–27.

Moore, Robert J., 2011. Eric Schmidt's "5 Exabytes" Quote is a Load of Crap. Acessível em http://blog.rjmetrics.com/2011/02/07/eric-schmidts-5-exabytes-quote-is-a-load-of-crap/ (01/10/2014).

Mota, Cristina e Diana Santos, editores, 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas. O Segundo HAREM*. Linguateca.

Márquez, Lluís, Marta Recasens e Emili Sapena, 2013. Coreference resolution: an empirical study based on SemEval-2010 Shared Task 1. *Language Resources and Evaluation*, 47: 661–694.

Nagy, István e Richárd Farkas, 2010. Person Attribute Extraction from the Textual Parts of Web Pages. In *CLEF (Notebook Papers/LABs/Workshops)*.

Ng, Vincent, 2008. Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, páginas 640–649. Association for Computational Linguistics, Honolulu.

Ng, Vincent e Claire Cardie, 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, páginas 104–111. Association for Computational Linguistics, Philadelphia.

Nguyen, Dat P. T., Yukata Matsuo e Mitsuru Ishizuka, 2007. Relation Extraction from Wikipedia Using Subtree Mining. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI 2007)*, volume 2, páginas 1414–1420. AAAI Press.

Nguyen, Truc-Vien T., Alessandro Moschitti e Giuseppe Riccardi, 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, volume 3, páginas 1378–1387. Association for Computational Linguistics, Singapura.

Nothman, Joel, James R Curran e Tara Murphy, 2008. Transforming Wikipedia into Named Entity Training Data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, páginas 124–132. Hobart, Australia.

Oliveira, Hugo Gonçalo e Paulo Gomes, 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, páginas 199–211. IOS Press, Lisboa.

Oliveira, Hugo Gonçalo, Diana Santos, Paulo Gomes e Nuno Seco, 2008. PAPEL: a dictionary-based lexical ontology for Portuguese. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira e Paulo Quaresma, editores, *Computational Processing of*

the Portuguese Language, volume 5190 de *Lecture Notes in Artificial Intelligence (LNAI)*, páginas 31–40. Springer-Verlag.

Orasan, Constantin, Dan Cristea, Ruslan Mitkov e António Branco, 2008. Anaphora Resolution Exercise: an Overview. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association, Marrakech.

Padró, Lluís, 1998. *A Hybrid Environment for Syntax-Semantic Tagging*. Dissertação de Doutoramento, Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya, Barcelona.

Padró, Lluís, Miquel Collado, Samuel Reese, Marina Lloberes e Irene Castellón, 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. European Language Resources Association, Valletta.

Padró, Lluís e Evgeny Stanilovsky, 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. European Language Resources Association, Istambul.

Palomar, Manuel, Antonio Ferrández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Saiz-Noeda e Rafael Muñoz, 2001. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4): 545–567.

Pantel, Patrick e Marco Pennacchiotti, 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, páginas 113–120. Association for Computational Linguistics, Sydney.

Paraboni, Ivandré, 1997. *Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa*. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

Pasca, Marius, Dekang Lin, Jeffrey Bigham, Andrei Lifchits e Alpa Jain, 2006. Organizing and Searching the World Wide Web of Facts - Step One: The One-million Fact Extraction

Challenge. In *Proceedings of the 21st National Conference on Artificial Intelligence*, volume 2, páginas 1400–1405. AAAI Press, Boston.

Platt, John C., 1999. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods – Support Vector Learning*, páginas 185–208. MIT Press, Cambridge.

Pollard, Carl e Ivan A. Sag, 1994. *Head-driven phrase structure grammar*. University of Chicago Press, Chicago.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel e Nianwen Xue, 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, páginas 1–27. Association for Computational Linguistics, Portland.

Raghunathan, Kathik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky e Christopher Manning, 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, páginas 492–501. Association for Computational Linguistics, Massachusetts.

Ratnaparkhi, Adwait, 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, páginas 133–142. Association for Computational Linguistics.

Ravichandran, Deepak e Eduard Hovy, 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, páginas 41–47. Association for Computational Linguistics, Philadelphia.

Real Academia Galega e Instituto da Lingua Galega, 2004. *Normas Ortográficas e Morfolóxicas do Idioma Galego*. Editorial Galaxia.

Recasens, Marta e Eduard Hovy, 2009. A deeper look into features for coreference resolution. In *Anaphora Processing and Applications*, páginas 29–42. Springer-Verlag.

Recasens, Marta e Eduard Hovy, 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, páginas 1423–1432. Association for Computational Linguistics, Uppsala.

Recasens, Marta e Eduard Hovy, 2011. BLANC: Implementing the Rand Index for Coreference Evaluation. *Natural Language Engineering*, 17(4): 485–510.

Recasens, Marta e M. Antònia Martí, 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4): 315–345.

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio e Yannick Versley, 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, páginas 1–8. Association for Computational Linguistics, Uppsala.

Ribeiro, Ricardo, Luís C. Oliveira e Isabel Trancoso, 2003. Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. In *Proceedings of the 6th Conference on Computational Processing on the Portuguese Language (PROPOR 2003)*, páginas 143–150. Springer-Verlag.

Riedel, Sebastian, Limin Yao e Andrew McCallum, 2010. Modeling Relations and Their Mentions Without Labeled Text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, volume 6323 de *Lecture Notes in Artificial Intelligence*, páginas 148–163. Springer-Verlag.

Ruiz-Casado, Maria, Enrique Alfonseca e Pablo Castells, 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Proceedings of the Atlantic Web Intelligence Conference (AWIC 2005)*, volume 3528, páginas 380–386. Lodz.

Santos, Diana e Nuno Cardoso, editores, 2007. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca.

Sapena, Emili, Lluís Padró e Jordi Turmo, 2013. A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Computational Linguistics*, 39(4): 847–884.

Sierra, Gerardo, Rodrigo Alarcón, César Aguilar e Carme Bach, 2008. Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1): 74–98.

Silva, João, 2007. *Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*. Dissertação de Mestrado, Faculdade de Ciências da Universidade de Lisboa.

Snow, Rion, Daniel Jurafsky e Andrew Y. Ng, 2005. Learning Syntactic Patterns for Automatic Hypernym Discovery. *Advances in Neural Information Processing Systems*, 17: 1297–1304.

Soares, Sérgio, Bruno Martins e Pavel Calado, 2011. Extracting biographical sentences from textual documents. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, páginas 718–730. Lisboa.

Soler, Victoria e Amparo Alcina, 2008. Patrones léxicos para la extracción de conceptos vinculados por la relación parte-todo en español. *Terminology*, 14(1): 99–123.

Solla Portela, Miguel Anxo, 2010. Módulo de acentuación para o galego en Freeling. *Linguamática. Revista para o Processamento Automático das Línguas Ibéricas*, 2(3): 59–64.

Soon, Wee Meng, Hwee Tou Ng e Daniel Chung Yong Lim, 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4): 521–544.

Steinberger, Josef, Massimo Poesio, Mijail A. Kabadjov e Kael Jezek, 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management: an International Journal*, 43(6): 1663–1680.

Stoyanov, Veselin e Jason Eisner, 2012. Easy-first Coreference Resolution. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, páginas 2519–2534. Bombaim.

Suchanek, Fabian M., 2009. *Automated Construction and Growth of a Large Ontology*. Dissertação de Doutoramento, Saarland University.

Suchanek, Fabian M., Georgiana Ifrim e Gerhard Weikum, 2006. LEILA: Learning to Extract Information by Linguistic Analysis. In *Proceedings of the 2nd Workshop on Ontology*

Population (OLP2) at the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING/ACL 2006). Association for Computational Linguistics, Sydney.

Sun, Ang, Ralph Grishman, Wei Xu e Bonan Min, 2011. New York University 2011 system for KBP slot filling. In *Proceedings of the Text Analytics Conference (TAC 2011)*. National Institute of Standards and Technology, Gaithersburg.

Swoyer, Stephen, 2007. Unstructured Data: Attacking a Myth. Acessível em http://tdwi.org/articles/2007/09/05/unstructured-data-attacking-a-myth.aspx (01/10/2014).

Sánchez-Cuadrado, Sonia, Juan Lloréns, Jorge Morato e José A. Hurtado, 2003. Extracción automática de relaciones semánticas. In *Actas de la 2a Conferencia Iberoamericana en Sistemas, Cibernética e Informática (CISCI 2003)*, páginas 265–268. Orlando.

Tesnière, Lucien, 1959. *Eléments de syntaxe structurale*. Librairie C. Klincksieck.

Tjong Kim Sang, Erik F. e Fien de Meulder, 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans e Miles Osborne, editores, *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*, páginas 142–147. Association for Computational Linguistics, Edmonton.

Tufis, Dan e Oliver Mason, 1998. Tagging Romanian texts: a case study for QTAG, a language independent probabilistic tagger. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, volume 1, páginas 589–596. European Language Resources Association, Granada.

Vieira, Renata e Massimo Poesio, 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4): 539–593.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly e Lynette Hirschman, 1995. A model-theoretic coreference scoring scheme. In *Proceedings of Message Understanding Conference 6 (MUC-6)*, páginas 45–52. Association for Computational Linguistics.

Wan, Xiaojun, Jianfeng Gao, Mu Li e Binggong Ding, 2005. Person resolution in person search results: WebHawk. In *Proceedings of the 14th ACM international Conference on Information and Knowledge Management (CIKM 2005)*, páginas 163–170. Association for Computing Machinery, New York.

Witten, Ian e Eibe Frank, 2005. *Data mining: practical machine learning tools and techniques with Java implementations*. Elsevier North-Holland, Inc., San Francisco.

Wu, Fei e Daniel S. Weld, 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, páginas 118–127. Association for Computational Linguistics, Uppsala.

Yan, Yulan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang e Mitsuru Ishizuka, 2009. Unsupervised relation extraction by mining Wikipedia texts using information from the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJNLP 2009)*, volume 2, páginas 1021–1029. Association for Computational Linguistics, Singapura.

Zhang, Min, Jie Zhang, Jian Su e Guodong Zhou, 2006. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, páginas 825–832. Association for Computational Linguistics, Sydney.

Zhao, Shubin e Ralph Grishman, 2005. Extracting Relations with Integrated Information Using Kernel Methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, páginas 419–426. Association for Computational Linguistics, Ann Arbor.

Zhou, GuoDong, Jian Su, Jie Zhang e Min Zhang, 2005. Exploring Various Knowledge in Relation Extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, páginas 427–434. Association for Computational Linguistics, Ann Arbor.

Zhou, Guodong, Min Zhang, Dong Hong e Ji Qiaoming Zhu, 2009. Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, páginas 728–736. Association for Computational Linguistics, Singapura.

Lista de Figuras

Lista de Tabelas

A presente tese avalia diferentes estratégias para a extracção automática de relações semânticas de textos em português, espanhol e galego. São utilizadas técnicas de aprendizagem automática e sistemas baseados em regras, sendo analisado o impacto de diferentes níveis de conhecimento linguístico nas várias abordagens avaliadas. Com o objectivo de implementar os sistemas de extracção, foram também construídas diversas ferramentas para o processamento da linguagem natural nos três idiomas referidos: desde módulos de segmentação de orações e de tokenização, a sistemas de desambiguação morfossintáctica, de reconhecimento de entidades mencionadas e de resolução de correferência. Como resultado do trabalho realizado, disponibilizam-se novas ferramentas e recursos para o processamento automático de textos em português, espanhol e galego.