INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Reconhecimento de Entidades Mencionadas em Língua Portuguesa: Locais, Pessoas, Organizações e Acontecimentos

Luís Carlos da Silva Romão

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

Júri

Presidente:	Doutor Ernesto José Marques Morgado
Orientador:	Doutor Nuno João Neves Mamede
Co-orientador:	Doutora Maria Luísa Torres Ribeiro Marques da Silva Coheur
Vogal:	Doutora Irene Pimenta Rodrigues

Setembro 2007

Agradecimentos

Gostaria de agradecer a todas as pessoas que, directa ou indirectamente, contribuíram para a realização desta tese de mestrado, especialmente ao meu orientador e co-orientadora, o Professor Nuno Mamede e a Professora Luísa Coheur, por toda a disponibilidade e apoio demonstrado. Gostaria também de agradecer a Cristina Mota e Caroline Hagège pela indispensável ajuda técnica, e a toda a equipa do L2F.

Por último, uma palavra especial de apreço para Ana Mendes, Ana Guimarães, João Loureiro e Telmo Machado, colegas de trabalho, por todo o *feedback*, sugestões e apoio demonstrado.

Lisboa, 17 de Novembro de 2007

Luís Carlos da Silva Romão

Resumo

A tarefa de reconhecimento de entidades mencionadas (REM) é uma subtarefa da área de extracção da informação que tem como objectivo a localização em textos de língua natural de elementos atómicos referentes a entidades específicas e sua posterior classificação em categorias predefinidas. Este documento analisa e compara várias estratégias adoptadas para a realização desta tarefa e descreve um sistema de reconhecimento de entidades mencionadas para a língua portuguesa que identifica entidades que são locais, pessoas, organizações e acontecimentos e as classifica de acordo com uma hierarquia de classificação, utilizando uma abordagem orientada à língua portuguesa, manual, e baseada exclusivamente em listas de palavras e regras, quer contextuais, quer baseadas na estrutura das entidades. O sistema foi avaliado segundo os critérios de avaliação da edição do fórum de avaliação HAREM de 2005, apresentando, em relação aos sistemas concorrentes, resultados no geral acima da média e obtendo o melhor resultado na tarefa de identificação de organizações e na classificação global em alguns cenários de avaliação.

Abstract

Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify atomic elements in natural language text into predefined categories. This document analyzes and compares several different strategies used in NER and describes a named entity system for the Portuguese language that identifies entities that are locations, people, organizations or events and classifies them according to a classification hierarchy, using a Portuguese-oriented, manual approach, based solely on lexicons and manual rules, either contextual or based on the entity's structure. The system was evaluated according to the criteria defined by HAREM, a named entity recognition evaluation forum for the Portuguese language, and its results were in general above average when compared to other participant systems, obtaining the best results in the identification of organizations and the best global results in several of the classification evaluation scenarios.

Palavras-chave
Keywords

Palavras-chave

Reconhecimento de Entidades Mencionadas (REM)

Locais

Pessoas

Organizações

Acontecimentos

Keywords

Named Entity Recognition (NER)

Locations

People

Organizations

Events

Índice

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Motivação

O reconhecimento de entidades mencionadas (REM)[1] é uma subtarefa da área de extracção de informação cujo objectivo se prende com a localização e classificação de elementos átomicos num texto, tais como nomes de pessoas, organizações, locais, expressões temporais, quantidades ou valores monetários. Estes elementos contêm geralmente um nome próprio e referem-se a uma entidade específica. Como exemplo, na frase 1.1 as entidades mencionadas encontram-se sublinhadas:

Frase 1.1: *O Pedro comprou uma Coca-Cola em Lisboa no Festival de Jazz.*

A identificação e classificação de entidades mencionadas é útil para várias aplicações na área do processamento da língua natural, nomeadamente sistemas de pergunta-resposta, em que perguntas do tipo "Quem...?" , "Onde...?" ou "Quando...?" , por exemplo, contêm necessariamente uma entidade mencionada como parte central da resposta. Por outro lado, o reconhecimento de entidades mencionadas permite também obter informação estruturada a partir de informação não estruturada (e.g. textos retirados da Internet), podendo a identificação das entidades mencionadas ser também útil em tarefas de procura de informação. Em adição, este reconhecimento pode também ser aplicável a domínios como os da bioinformática e biologia molecular, na identificação de nomes de compostos químicos, moléculas ou proteínas.

Embora seja aparentemente uma tarefa simples, o reconhecimento de entidades mencionadas vê-se confrontado com vários desafios: as entidades podem ser difíceis de encontrar, e uma vez encontradas, difíceis de classificar, dependendo ainda esta classificação da finalidade do sistema. Por exemplo, nas frases *"O estado das finanças públicas é lastimável"* e *"O estado da Califórnia foi ganho pelos Democratas"*, a palavra *estado* só se refere a uma entidade mencionada no segundo caso. Por outro lado, em frases como *"A Igreja é contra o aborto"* e *"A Igreja da Luz é um edifício renascentista"*, embora a palavra *Igreja* seja nos dois casos uma entidade mencionada, a sua classificação difere: na primeira frase refere-se a uma instituição (a Igreja Católica Apostólica Romana), mas na segunda a um local físico.

[1] Em inglês *Named Entity Recognition* (NER).

As categorias escolhidas para um determinado sistema de REM dependem ainda da sua finalidade. Se a classificação geográfica, por exemplo, é importante numa determinada área (e.g., um sistema de pergunta-resposta sobre viagens), então essas categorias tenderão a ser mais refinadas do que noutros sistemas em que esta classificação não é tão relevante.

Pretende-se desenvolver um sistema de reconhecimento de entidades mencionadas que identifique e classifique entidades em textos de língua portuguesa de acordo com os critérios de identificação e categorização definidos na secção 3.4. Em termos gerais, pretende-se classificar as entidades em quatro categorias distintas: locais, pessoas, organizações e acontecimentos. Cada uma das categorias está subdividida em vários tipos, representando cada tipo um nível de classificação mais específico dentro das mesmas.

Este sistema tem como objectivo auxiliar o funcionamento de um sistema de pergunta-resposta (Mendes, 2007), que se baseia nas entidades mencionadas presentes num texto para determinar a resposta a perguntas do tipo "Quem...?" ou "Onde...?" , que contêm necessariamente uma entidade mencionada do tipo *Pessoa* e *Local*, respectivamente. Este sistema pretende participar no fórum de avaliação CLEF (*CLEF - Cross-Language Evaluation Forum*, n.d.), na categoria de resposta automática a perguntas.

Em adição, pretende-se submeter o sistema de reconhecimento de entidades mencionadas ao próximo fórum de avaliação HAREM (*HAREM - Avaliação de Reconhecimento de Entidades Mencionadas*, n.d.), o único fórum de avaliação de reconhecimento de entidades mencionadas para a língua portuguesa.

1.2 Estratégia

A tarefa de identificação e classificação será efectuada através de técnicas de processamento de língua natural (por oposição a métodos estatísticos), fazendo uso da ferramenta XIP[1], inserida numa cadeia de processamento mais vasta, como descrito em maior detalhe no Capítulo 3.

A identificação e reconhecimento das entidades segue uma abordagem orientada à língua portuguesa, manual, e baseada exclusivamente em regras, quer contextuais, quer baseadas na estrutura das entidades. Utilizam-se também listas de palavras, que são criadas com base em informação recolhida de *corpus* de texto jornalístico.

[1] *Xerox Incremental Parser*

1.3 Roteiro

No Capítulo 2 é feita uma descrição e comparação das principais estratégias usadas, tanto para a língua portuguesa como para outras línguas, no reconhecimento de entidades mencionadas. É explicitada a arquitectura do sistema usado no reconhecimento, assim como descrito o método de trabalho (Capítulo 3), seguindo-se uma análise mais pormenorizada relativa à implementação das estratégias definidas anteriormente (Capítulo 4). São então descritos os critérios de avaliação e apresentados os resultados, comparando-os com aqueles obtidos por sistemas semelhantes (Capítulo 5), assim como as conclusões que se podem retirar do trabalho realizado e quais os melhoramentos e adições a efectuar no futuro (Capítulo 6).

2 Estado da Arte

2.1 Introdução

Desde o final dos anos noventa que se tem vindo a assistir a um interesse crescente na identificação de entidades mencionadas, particularmente em aplicações relacionadas com língua natural, biologia molecular e bioinformática.

Existem vários fóruns de avaliação internacionais dedicados a este domínio, entre os quais se destacam o MUC (*Message Understanding Conferences*) (*MUC - Message Understanding Conferences*, n.d.), CoNLL (*Computational Natural Language Learning*) (*CoNLL - Computational Natural Language Learning*, n.d.), ACE (*Automatic Content Extraction*) (*ACE - Automatic Content Extraction*, n.d.) e o HAREM (Avaliação de Reconhecimento de Entidades Mencionadas) (*HAREM - Avaliação de Reconhecimento de Entidades Mencionadas*, n.d.), para a língua portuguesa.

O fórum de avaliação MUC foi o primeiro em que se realizou uma avaliação de reconhecimento de entidades mencionadas, em 1995, e apresenta uma divisão em três categorias: i) pessoas, organizações e locais (ENAMEX); ii) valores e expressões temporais (TIMEX); iii) valores e expressões numéricas (NUMEX). Os domínios de texto são restritos, resumindo-se, por exemplo, a artigos sobre acidentes aéreos no MUC-7.

O fórum de avaliação CoNLL apresenta uma divisão em quatro categorias: i) pessoas; ii) locais; iii) organizações; iv) miscelânea. Os sistemas concorrentes têm de ser independentes da língua e incluir um componente de aprendizagem.

O fórum de avaliação ACE apresenta uma divisão em cinco categorias: i) pessoas; ii) organizações; iii) locais; iv) entidades geopolíticas; v) infra-estruturas. É orientado às línguas inglesa, árabe e chinesa, tendo as categorias sido estendidas em 2005 para incluir também veículos e armas. Os *corpora* usados incluem artigos jornalísticos e textos retirados da *Internet*.

O fórum de avaliação HAREM, para a língua portuguesa, apresenta uma divisão extensiva em 41 categorias e subcategorias e utiliza como *corpora* uma colecção de texto jornalístico, literário, entrevistas, *Internet* e correio electrónico.

Os sistemas de reconhecimento de entidades mencionadas estão assentes em diferentes estratégias, desde técnicas baseadas em gramáticas ao uso de modelos estatísticos.

Neste capítulo descrevem-se diferentes estratégias de reconhecimento de entidades mencionadas, começando por dois exemplos de sistemas que são independentes da língua (Secção 2.2) e dois sistemas mais específicos orientados a línguas particulares (Secção 2.3). Estudam-se também as abordagens tomadas pelos sistemas que tiveram o melhor desempenho nos fóruns de avaliação anteriormente referidos (Secção 2.4) e efectua-se uma comparação entre as diferentes estratégias abordadas (Secção 2.5).

2.2 REM independente da língua

Muitos dos sistemas de reconhecimento de entidades mencionadas usam recursos específicos à língua que pretendem tratar, não sendo por isso aplicáveis a línguas diferentes. Contudo, têm havido alguns estudos no sentido de usar o mesmo sistema de reconhecimento para diferentes idiomas.

Apresentam-se em seguida dois desses sistemas, que usam duas abordagens distintas para a tarefa em questão: memorização simples (Secção 2.2.1) e pistas contextuais e morfológicas (Secção 2.2.2).

2.2.1 REM usando memorização simples

O estudo de Palmer e Day (Palmer & Day, 1997) descreve um sistema de reconhecimento de entidades mencionadas independente da língua, isto é, sem qualquer conhecimento das línguas sobre o qual opera. Este sistema baseia-se na análise automática das cadeias de caracteres que compõem os textos, não utilizando por isso nem listas de palavras nem informação sobre a segmentação das partes do discurso.

Pretende-se no sistema em questão anotar as entidades em três categorias distintas: TIMEX (frases e expressões temporais), NUMEX (frases e expressões numéricas) e ENAMEX (nomes próprios, locais e organizações).

Utilizaram-se *corpora* de seis línguas distintas (inglês, francês, espanhol, português, japonês e chinês), apresentando-se na tabela 2.1 o número de entidades e a sua distribuição por categoria para cada uma das línguas. Todos os seis *corpora* consistem numa colecção de artigos de jornal, embora o conteúdo varie de língua para língua. Por exemplo, o *corpus* francês contém várias edições completas do jornal *Le Monde* enquanto que os artigos em inglês e espanhol foram especificamente seleccionados para o MUC-6 (*MUC - Message Understanding Conferences*, n.d.), pelo que consistem maioritariamente de textos de conferências de imprensa.

As categorias TIMEX e NUMEX não ultrapassam os 20-30% do total de entidades mencionadas e são por outro lado as mais fáceis de anotar, já que podem ser descritas por um número reduzido de padrões. Após a análise dos *corpora* foi possível representar todos as entidades NUMEX nas seis línguas em apenas cinco padrões (e.g., *sequência de dígitos seguida de* %). Do mesmo modo, com apenas alguns

Língua	EM	TIMEX	NUMEX	ENAMEX
Chinês	4454	17.21%	1.8%	80.9%
Francês	2321	18.6%	3.0%	78.4%
Inglês	2242	10.7%	9.5%	79.8%
Japonês	2146	26.4%	4.0%	69.6%
Português	3839	17.7%	12.1%	70.3%
Espanhol	3579	24.6%	3.0%	72.5%

Tabela 2.1: Distribuição de entidades mencionadas por língua nos *corpora*.

padrões é possível reconhecer cerca de 95% das entidades TIMEX em qualquer uma das línguas. O estudo foi então essencialmente centrado nas categorias cuja anotação é mais difícil, as ENAMEX.

Os *corpora* das várias línguas são divididos em *corpora* de treino e teste, sendo que um *corpus* de teste contém aproximadamente 450 entidades ENAMEX e um *corpus* de treino contém as restantes entidades, sendo o seu número variável consoante a língua. O sistema memoriza as entidades mencionadas presentes no *corpus* de treino e usa essa informação para classificar os textos do *corpus* de teste.

Para medir o desempenho do sistema, considerou-se a taxa de transferência de vocabulário, isto é, a percentagem de entidades que ocorrem no *corpus* de treino que também aparecem no *corpus* de teste. As entidades mais frequentes no *corpus* de treino correspondem, consoante a língua, a uma percentagem entre os 20% (para o francês) e os 80% (para o chinês) das entidades encontradas no *corpus* de teste, ainda que seja de notar que, a partir de um determinado nível (variável em cada língua), a memorização de mais entidades não afecta significativamente o desempenho do sistema.

No entanto, usando apenas memorização, o desempenho tende a diminuir devido à ambiguidade, por exemplo, quando uma entidade aparece referida dentro de uma outra entidade (`New York` - local, `New York Yankees` - organização) ou quando uma cadeia de caracteres pode tanto ser entidade como não o ser, consoante a situação (`apple` - maçã, `Apple` - companhia).

Tendo em conta as transferências de vocabulário em cada língua, bem como a ocorrência de entidades de cada categoria e subcategoria, estimou-se um valor que deveria ser alcançável por qualquer sistema que pretendesse efectuar uma tarefa idêntica. No que diz respeito às entidades TIMEX e NUMEX, a experiência demonstra que se podem anotar correctamente cerca de 95% das ocorrências, devido ao baixo número de padrões. Quanto às entidades ENAMEX, verifica-se que, apesar de tudo, com um sistema muito simples de reconhecimento, podem conseguir-se resultados bastante elevados, que, consoante a língua, poderão alcançar uma cobertura de até 70%.

Regras contextuais podem melhorar os resultados sem a necessidade de um conhecimento linguístico extensivo. Tal como para as categorias TIMEX e NUMEX, também aqui muitas das entidades podem ser reconhecidas após uma análise adequada dos contextos das frases no *corpus* de treino.

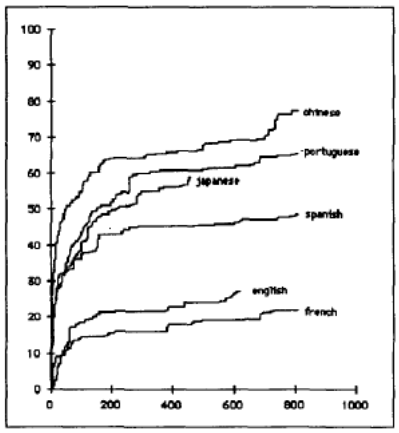

Figura 2.1: Taxa de cobertura (em %) para cada língua em relação ao número de entidades memorizadas no *corpus* de treino.

Verifica-se então que a anotação de entidades mencionadas de acordo com o processo descrito segue a lei de Zipf [2]: um pequeno esforço resulta num grande avanço no reconhecimento, mas para além dessa melhoria inicial, é preciso muito esforço para conseguir aumentar um pouco o desempenho. Este princípio está ilustrado na figura 2.1.

2.2.2 REM usando pistas contextuais e morfológicas

O estudo de Cucerzan e Yarowsky (Cucerzan & Yarowsky, 1999), tal como no caso anterior (Secção 2.2.1), descreve um sistema de reconhecimento de entidades mencionadas independente da língua, mas que segue uma abordagem bastante distinta. Neste caso, os autores recorrem à aprendizagem iterativa, usando informação contextual e informação sobre a morfologia das palavras, com supervisão mínima. É um processo de *bootstrapping*, que começa com a informação extraída dos *corpora* de treino. Relativamente à informação morfológica, o sistema analisa os sufixos e prefixos das entidades candidatas (e.g., nomes terminados em -ia tendem a ser locais: `Áustria`, `Austrália`, `Itália`, `Escócia`, `Suécia`), assim como procura por padrões comuns em entidades multipalavra (e.g., `Associação de Bombeiros Voluntários de Sacavém`, `Associação de Defesa dos Direitos dos Animais`). Os contextos à esquerda e à direita são também bastante importantes e essenciais em situações em que as entidades têm uma estrutura desconhecida, são de origem estrangeira ou são polissémicas. Ao invés de recorrer a bigramas ou trigramas, o sistema processa os contextos da mesma maneira que processa as entidades, permitindo por isso contextos de extensão variável e uma abordagem unificada tanto para a informação interna à palavra (e.g., morfologia, capitalização) como contextual.

[2] A Lei de Zipf, formulada pelo linguista George Kingsley Zipf (1902-1950), afirma que uma palavra com a posição n na tabela de frequências num dado *corpus* tem uma frequência de $\frac{1}{n^k}$ vezes a da palavra mais frequente ($n = 1$), com k entre 1 e 2, consoante a língua.

Em alguns casos a informação morfológica e o contexto envolvente de apenas uma instância da entidade não são suficientes para tomar uma decisão quanto à sua classificação. Contudo, como referido por Katz (Katz, 1996), uma entidade introduzida pela primeira vez será no geral repetida, seja para quebrar a monotonia do uso de pronomes ou para clarificar e enfatizar o seu sentido. Por outro lado, segundo Gale et al. (Gale et al., 1992), se uma palavra polissémica aparece duas ou mais vezes num discurso há uma grande probabilidade de que o seu significado seja o mesmo nesse discurso. Tal não se aplica quando se trabalha com documentos bastante extensos e sem fronteiras definidas, pelo que se torna necessário efectuar uma segmentação prévia do texto, de modo a poder garantir com elevada probabilidade que as ocorrências de cada entidade nesse segmento têm o mesmo significado. No geral, o sistema de Cucerzan e Yarowsky pode ser dividido em oito fases:

1. extracção das entidades e dos contextos do *corpus* de treino;
2. processamento do texto (*corpus de teste*) a ser anotado e extracção de todas as entidades candidatas;
3. extracção de todos os contextos à direita e à esquerda das entidades candidatas;
4. construção de hipóteses usando as palavras individuais, as entidades candidatas e todas as instâncias do contexto à direita e à esquerda;
5. aplicação do processo de *bootstrapping* usando os dados extraídos inicialmente;
6. classificação de cada entidade candidata isoladamente;
7. reclassificação das entidades candidatas tendo em conta as outras entidades envolventes;
8. resolução de conflitos.

O sistema consegue uma medida-F (Medida-F= $\frac{2 \times precisão \times cobertura}{precisão + cobertura}$) de 77% para a língua espanhola e de 72% para o neerlandês. Usando-se, em adição, listas de nomes com os principais países, cidades e nomes de pessoas e de companhias, o sistema apresenta uma melhoria no desempenho de até 2,5% na medida-F.

2.3 REM dependente da língua

Para melhorar o seu desempenho, os sistemas de reconhecimento de entidades mencionadas dependentes da língua utilizam informação sobre a língua em que actuam, seja utilizando listas lexicais, tendo conhecimento sobre as partes do discurso ou recorrendo a regras gramaticais e de contexto específicas para o idioma tratado. Estes sistemas não apresentam resultados efectivos quando aplicados a textos escritos numa língua que não aquela para a qual o sistema foi desenhado. Apresentam-se em seguida dois desses sistemas, para duas línguas distintas: o inglês (Secção 2.3.1) e o japonês (Secção 2.3.2).

2.3.1 Estratégia orientada à língua inglesa

Existem dois tipos de pistas que podem ser usadas em REM para resolver os problemas da ambiguidade: pistas internas localizadas dentro da própria palavra e pistas externas relacionadas com o contexto envolvente.

O estudo de GuoDong e Jian (Zhou & Su, 2002) apresenta um sistema de reconhecimento de entidades mencionadas para a língua inglesa baseado num modelo de Markov não-observável,[3] capaz de integrar e aplicar quatro tipos de pistas internas e externas:

i) traços deterministas simples relacionados com a estrutura interna das palavras, tais como a capitalização;

ii) traços semânticos internos;

iii) traços internos contidos num dicionário de termos;

iv) traços relacionados com o contexto.

Este sistema pretende integrar e aplicar as pistas internas e externas, baseando-se em blocos, em que cada entidade é representada por um bloco. Os traços deterministas simples usados pelo sistema relativamente à estrutura interna das palavras encontram-se representados na tabela 2.2, enquanto que na tabela 2.3 estão reproduzidos os traços semânticos internos associados às categorias que se pretende classificar (datas, valores numéricos, valores temporais, pessoas, locais e organizações). A tabela 2.4, por sua vez, apresenta informação sobre as entidades de cada categoria que já estão presentes no dicionário de termos (lista de palavras).

Usando apenas a informação relativa aos traços da tabela 2.2, o sistema obtém uma medida-F de 77,6%. Por composição, usando os traços da tabela 2.3 e da tabela 2.4 e o contexto, o desempenho aumenta respectivamente para 87,4%, 89,3%, 92,4% e 94,1% (MUC-7). Daqui se pode inferir que a informação semântica é significativa para o reconhecimento das entidades, aumentando o desempenho em mais de 10%. Também a análise do contexto contribui para um aumento de desempenho na ordem dos 5,5%. Por outro lado, em relação ao uso dos traços internos do dicionário de termos, o aumento no desempenho é de apenas 1,2%. Em parte, isto deve-se ao facto de muita da informação contida nas listas ser também capturada noutras fases (como na análise do contexto ou da semântica). No entanto, esta fase tem um contributo mais significativo quando não existe informação explícita na entidade ou no contexto envolvente para a poder classificar (de notar que a informação fornecida pelo dicionário de termos refere-se sempre a entidades que são conhecidas).

[3] Em inglês Hidden Markov Model (HMM).

Traço	Exemplo	Explicação
OneDigitNum	9	número composto de um só dígito
TwoDigitNum	99	número composto de dois dígitos
FourDigitNum	1990	ano composto por quatro dígitos
YearDecade	1990s	década
ContainsDigitAndAlpha	A845-3	código de produto
ContainsDigitAndDash	09-99	data
ContainsDigitAndOneSlash	3/4	data ou fracção
ContainsDigitAndTwoSlashs	19/19/1999	data
ContainsDigitAndComma	19,000	moeda
ContainsDigitAndPeriod	19,000	moeda, percentagem
OtherContainsDigit	123456	outro número
AllCaps	IBM	organização
CapPeriod	M.	inicial de nome próprio
CapOtherPeriod	St.	abreviatura
CapPeriods	N.Y.	abreviatura
FirstWord	First word	sem informação útil sobre a capitalização
InitialCap	Microsoft	palavra com letra maiúscula
LowerCase	dog	palavra com letra minúscula
Other	$	todas as outras palavras

Tabela 2.2: Lista de traços relacionados com a estrutura interna das palavras.

Traço	Exemplo	Explicação
SuffixPERCENT	%	sinal de percentagem (%)
PrefixMONEY	$	prefixo de moeda
SuffixMONEY	dollars	sufixo de moeda
SuffixDATE	day	sufixo de data
WeekDATE	Monday	dia da semana
MonthDATE	July	mês do ano
SeasonDATE	Summer	estação do ano
PeriodDATE1	month	período de tempo
PeriodDATE2	quarter	período
EndDATE	Weekend	fim de data
ModifierDATE	Fiscal	modificador de data
SuffixTIME	a.m.	sufixo de tempo
PeriodTIME	morning	período de tempo
PrefixPERSON1	Mr.	título
PrefixPERSON2	President	cargo
FirstNamePERSON	Michael	nome próprio
SuffixLOC	River	sufixo de localização
SuffixORG	Ltd.	sufixo de organização
Other	six, sixth, etc.	cardinal, ordinal, etc.

Tabela 2.3: Lista de traços semânticos associados às palavras.

Tipo(número de entradas)	Traço	Exemplo
DATE(20)	DATEnGN	Christmas Day:DATE2G2
PERSON(10,000)	PERSOnGN	Bill Gates: PERSON2G2
LOC (5,000)	LOCnGN	Beijing: LOC1G1
ORG(10,000)	ORGnGN	United Nations: ORG2G2

Tabela 2.4: Lista de traços internos do dicionário de termos.

Em relação ao desempenho do sistema à medida que o tamanho do *corpus* de treino aumenta, verifica-se que com um *corpus* inicial de 100KB, a medida-F atinge cerca de 87%, subindo este valor para 90% com um *corpus* de 200KB. Incrementos posteriores de 100KB têm um menor efeito no desempenho: para atingir uma medida-F de 95% é necessário utilizar um *corpus* de 800KB. Nos fóruns de avaliação MUC-6 e MUC-7 (*MUC - Message Understanding Conferences*, n.d.), o sistema obteve medidas-F de 96,6% e 94,1%, respectivamente. Estes valores são superiores aos de sistemas baseados em regras, assim como de outros sistemas baseados em aprendizagem presentes a concurso.

2.3.2 Estratégia orientada à língua japonesa

O estudo de Satoshi Sekine et al. (Sekine & Eriguchi, 2000) apresenta uma análise de vários sistemas de reconhecimento de entidades mencionadas específicos para a língua japonesa no âmbito do projecto IREX (*Information Retrieval and Extraction Exercise*) (*IREX - Information Retrieval and Extraction Exercise*, n.d.). A classificação das entidades efectuou-se segundo oito categorias distintas: ORGANIZATION (organizações), PERSON (pessoas), LOCATION (locais), ARTIFACT (artefactos, e.g., *Pentium II, Prémio Nobel)*, DATE (datas), TIME (expressões temporais), MONEY (expressões referentes a valores monetários) e PERCENT (percentagens).

Não sendo o japonês uma língua indo-europeia nem escrita no alfabeto latino, existem problemas adicionais que não se verificam para o caso do inglês, como por exemplo, o da delimitação das palavras. No sistema de escrita japonês não existem espaços entre diferentes palavras, pelo que a divisão em *tokens* dos textos se torna mais complexa. Por outro lado, uma palavra pode ser composta por vários caracteres, que por si mesmos, também são palavras.

Os quinze sistemas analisados apresentam estratégias diferentes, subdividindo-se nos seguintes grupos:

1. sistemas baseados em regras e padrões criados manualmente;
2. sistemas baseados em regras e padrões parcialmente automáticos criados a partir de um *corpus* de treino;

3. sistemas completamente automatizados, baseados em modelos de máxima entropia, modelos de Markov não-observáveis ou árvores de decisão.

Os três sistemas mais bem classificados vêm um de cada grupo, sendo que o melhor sistema se baseia em regras criadas manualmente (medida-F de 83,86%), o segundo em regras semi-automáticas (medida-F de 80,05%) e o terceiro é um sistema totalmente automatizado (medida-F de 77,37%), tendo todos eles acesso a um extensivo dicionário de termos de dezenas de milhares de nomes de organizações, pessoas e locais. Estes resultados são inferiores àqueles que se obtêm para a língua inglesa, o que se explica em parte pelas características específicas da língua japonesa já mencionadas e também pela introdução da categoria ARTIFACT, em que os sistemas tiveram todos o seu pior desempenho.

2.4 As estratégias ganhadoras

2.4.1 MUC-6

Foi nas conferências do MUC-6 (*MUC - Message Understanding Conferences*, n.d.), em 1995, que foi introduzido pela primeira vez o termo *reconhecimento de entidades mencionadas* e uma avaliação deste tipo foi efectuada, embora a tarefa de REM apareça apenas como um subtarefa no contexto da avaliação dos sistemas, que se focava principalmente em outras actividades da área de extracção de informação. Então, o MUC tinha como foco tarefas de extracção de informação onde informação estruturada relacionada com actividade empresarial e de defesa era extraída de texto não-estruturado, como por exemplo, artigos de jornal. Foi durante o decorrer desta tarefa que se reparou na importância de se reconhecer unidades de informação tais como nomes próprios, organizações, locais, expressões temporais, datas ou unidades monetárias.

O sistema que obteve a melhor classificação para a língua inglesa neste primeiro fórum de avaliação foi o de George R. Krupka (Krupka, 1995), que atingiu uma medida-F de 96,42%.

O sistema baseia-se apenas em regras e padrões manuscritos, contendo também um pequeno dicionário de termos com 530 palavras de nomes de pessoas e organizações.

2.4.2 MUC-7

Na sétima e última edição do fórum de avaliação MUC (*MUC - Message Understanding Conferences*, n.d.), o sistema que obteve a melhor classificação para a língua inglesa foi o de A. Mikheev et al. (Mikheev et al., 1999), que atingiu uma medida-F de 94,51%. Este sistema divide-se em várias fases:

1. divisão do texto em *tokens* de acordo com uma definição pré-estabelecida;

2. marcação de cada palavra no texto com a sua categoria gramatical, usando para isso um modelo de Markov não-observável;

3. atribuição de traços semânticos às palavras (e.g., palavras terminadas em *-an* ou *-ese* geralmente referem-se a nacionalidades: `American`, `Japanese`, `Brazilian`, `Portuguese`);

4. aplicação de regras gramaticais específicas para cada domínio (reconhecimento de organizações, locais, etc.) com ou sem contexto.

O sistema trata as entidades TIMEX e NUMEX de forma diferente das ENAMEX. A razão para esta divisão prende-se com o facto das expressões temporais e númericas serem mais estruturadas e poderem ser capturadas apenas por meio de regras gramaticais. O sistema apresenta gramáticas específicas para anotar as expressões numéricas e temporais, assim como listas de entidades dessas categorias já conhecidas, como por exemplo, nomes de moedas. As expressões ENAMEX, por outro lado, apresentam uma estrutura mais complexa e são mais dependentes do contexto. De acordo com os autores, o contexto é mais importante na determinação da classificação correcta destas entidades do que as regras gramaticais ou as listas de palavras. Somente o contexto pode determinar se `Arthur Andersen` é uma pessoa ou uma companhia, se `Washington` é uma pessoa ou um local ou se `Granada` é um local ou uma organização. Por outro lado, uma vez que uma palavra tenha sido usada com determinado sentido, este não mudará no mesmo texto sem que hajam claras pistas contextuais indicadoras dessa mudança. O sistema usa listas de palavras, mas altera-as dinamicamente, isto é, se durante o processamento do texto se obtiver a partir do contexto informação de que `Granada` é uma organização, a palavra é adicionada à lista respectiva durante o resto do processamento, mas não é usada para a análise de um novo texto, onde a palavra só será adicionada à lista se mais uma vez se obtiver essa informação através do contexto. A identificação das entidades ENAMEX divide-se em cinco fases:

1. uso de regras de sucesso garantido;[5]

2. combinação probabilística parcial (1);

3. uso de regras relaxadas;

4. combinação probabilística parcial (2);

5. tratamento de títulos.

Na primeira fase (uso de regras de sucesso garantido) são usadas regras orientadas ao contexto que só são aplicadas quando a expressão candidata se encontra rodeada por um contexto sugestivo. Por exemplo, `Gerard Klauer` aparenta ser um nome próprio, mas no contexto `analista da Gerard`

[5]Em inglês *sure-fire rules*.

`Klauer`, é um nome de uma organização. Este tipo de regras usa informação sobre elementos que designam companhias (Ltd., Inc., etc.) e títulos (Mr., Dr., Sen.). Nesta fase o sistema usa a informação contida nas listas de entidades como informação provável e não como informação definitiva, verificando sempre se o contexto envolvente é sugestivo e não-contraditório. Um local que se encontra na lista de termos geográficos só é marcado como tal se ocorrer num contexto que sugira uma localização.

Na segunda fase (combinação probabilística parcial) o sistema recolhe todas as entidades já identificadas no documento e gera todas as possíveis ordens parciais das palavras que a compõem (preservando a ordem), marcando-as com a mesma classificação caso ocorram no texto. Por exemplo, se na primeira fase a expressão `Lockheed Martin Production` foi classificada como uma organização por ocorrer num contexto sugestivo de organizações, então todas as instâncias de `Lockheed Martin Production`, `Lockheed Martin`, `Lockheed Production`, `Martin Production`, `Lockheed` e `Martin` serão marcadas como possíveis organizações. Este texto anotado é então utilizado num modelo de máxima entropia pré-treinado que tem em consideração informação contextual tal como a posição na frase e a capitalização. Se o resultado for positivo, a combinação parcial é marcada como entidade ENAMEX.

Na terceira fase (uso de regras relaxadas) aplicam-se novamente regras gramaticais, mas desta vez mais relaxadas no que diz respeito ao contexto e usando extensivamente a informação que já foi descoberta e os dicionários de termos. Por exemplo, se uma palavra com letra inicial maiúscula foi identificada como nome próprio, ocorre seguida de uma ou mais palavras desconhecidas e também com letra inicial maiúscula, o sistema pode assumir que se trata de uma referência a uma pessoa. Nesta fase já não existe a preocupação de que o mesmo nome também possa referir-se a uma organização, já que estas já deveriam ter sido identificadas (na primeira e segunda fases). Os locais e organizações presentes nos dicionários de termos são marcados, sem atender ao contexto.

A quarta fase (repetição da combinação probabilística parcial) processa-se de modo em tudo idêntico à segunda.

Na quinta fase (tratatamento de títulos) classificam-se títulos (frases completamente escritas em letra maiúscula). Esta classificação é realizada tentando combinar as entidades já identificadas nas quatro fases anteriores com as palavras encontradas nos títulos, com verificação num modelo de máxima entropia treinado com títulos de documentos. Por exemplo, no título `MURDOCK SATELLITE EXPLODES ON TAKE-OFF`, `Murdoch` será classificado como pessoa, em concordância com a classificação de `Rupert Murdoch` no texto.

A tabela 2.5 mostra o progresso do desempenho do sistema através das cinco fases descritas anteriormente.

Fase	Organizações	Pessoas	Locais
regras de sucesso garantido	C:42 P:98	C:40 P:99	C:36 P:96
combinação prob. parcial (1)	C:75 P:98	C:80 P:99	C:69 P:93
regras relaxadas	C:83 P:96	C:90 P:98	C:86 P:93
combinação prob. parcial (2)	C:85 P:96	C:93 P:97	C:98 P:93
tratamento de títulos	C:91 P:95	C:95 P:97	C:95 P:93

Tabela 2.5: Resultados obtidos pelo sistema de A. Mikheev et al. através das diferentes etapas da análise. C = cobertura, P = precisão.

As regras de sucesso garantido permitem obter uma grande precisão (96-98%), mas têm uma cobertura baixa, isto é, não permitem encontrar um grande número de entidades ENAMEX. Na segunda fase os valores da cobertura sobem consideravelmente (de 33% a 40%) e fases posteriores vão gradualmente anotando mais entidades ENAMEX (aumentando a cobertura), mas ao mesmo tempo introduzindo erros, o que resulta numa ligeira diminuição da precisão (3%-4%).

2.4.3 CoNLL-2002

Na edição de 2002 do fórum de avaliação CoNLL (*CoNLL - Computational Natural Language Learning*, n.d.), o sistema vencedor para a língua espanhola, considerando a medida-F, foi o de Xavier Carreras et al (Carreras et al., 2002). Este sistema consiste em dois módulos separados, sequenciais e independentes entre si, um efectuando o reconhecimento das entidades e outro classificando-as. Ambos os módulos utilizam uma estratégia baseada em aprendizagem, fazendo uso de classificadores binários *AdaBoost*. [4]

No sistema de Xavier Carreras et al. as palavras em redor de uma determinada palavra são codificadas com um conjunto de traços primitivos, juntamente com a sua posição relativa a essa palavra. Os traços primitivos são:

1. o lema da palavra;
2. a parte do discurso a que a palavra pertence;
3. informação relacionada com a ortografia da palavra: *começa com maiúscula, contém dígitos, contém hífens, contém pontuação, é um url*, etc.;
4. o tipo da palavra: *funcional, capitalizada, sinal de pontuação, etc.;*
5. informação sobre se a palavra aparece no dicionário de termos;
6. a previsão da classificação das palavras à esquerda da palavra em questão;

[4] *AdaBoost* (Freund & Schapire, n.d.), abreviatura de *Adaptive Boosting*, é um meta-algoritmo que pode ser usado em conjunção com outros algoritmos de aprendizagem de modo a melhorar o seu desempenho.

7. informação sobre se a palavra é indicadora de um contexto de nome, organização ou local.

A tarefa de REM é efectuada como um combinação de classificadores locais que testam decisões simples em cada palavra do texto. Existem três esquemas de decisão diferentes para o reconhecimento das entidades através da combinação de classificadores: i) BIO ii) Open-Close & I e iii) Open-Close Global.

No esquema BIO, cada palavra é marcada como sendo o início de uma entidade mencionada (marca B), uma palavra dentro de uma entidade mencionada (marca I) ou uma palavra não pertencente a uma entidade mencionada (marca O). Usam-se três classificadores binários para realizar a marcação, cada um correspondendo a uma diferente marca (B, I e O). Quando se realiza a marcação, cada frase é processada da esquerda para a direita, seleccionando-se para cada palavra a marcação com o maior grau de confiança que é coerente com a solução actual.

No esquema Open-Close & I a entidade mencionada é reconhecida através da detecção da palavra que a começa e da palavra que a termina. Uma frase é processada da esquerda para a direita, aplicando três classificadores: o classificador *open* procura o início da entidade e, uma vez detectada, o classificador *close* procura o seu fim. De modo a tornar mais robusta a procura pela palavra que termina a entidade, cada palavra dentro da entidade actual é testada com o classificador I do esquema BIO e, se classificada negativamente, a entidade é forçada a terminar na palavra anterior.

No esquema Open-Close Global procuram-se também os inícios e fins das entidades, mas tomando em consideração a classificação das entidades que aparecem em redor na mesma frase.

A tarefa de classificação de entidades consiste em atribuir um tipo a cada entidade potencial que já foi reconhecida anteriormente. São usadas combinações de dez classificadores binários: os quatro possíveis (não simétricos) *um contra todos* e as três possíveis combinações de *dois contra dois* (PESSOA vs LOCAL, PESSOA vs ORGANIZAÇÃO e LOCAL vs ORGANIZAÇÃO). Além disso, são usados um dicionário de termos e uma lista de palavra sugestivas dos contextos de entidades mencionadas.

O esquema BIO apresenta os melhores resultados para a tarefa de reconhecimento (medida-F de 91,66%). Os resultados relativos às tarefas de identificação e classificação são inferiores aos da tarefa de identificação por si só, tendo a classificação um medida-F de 78,7%, com os melhores resultados obtidos na categoria PESSOA e os piores na categoria MISC (entidades mencionadas que não são pessoas, locais ou organizações). No entanto, o uso de dicionários de termos e outra informação exterior aumenta o desempenho em cerca de 2%.

2.4.4 CoNLL-2003

Na edição de 2003 do fórum de avaliação CoNLL (*CoNLL - Computational Natural Language Learning*, n.d.), o sistema vencedor para a língua inglesa, considerando a medida-F, foi o de Radu Florian et al (Florian et al., 2003). Este sistema é um sistema independente da língua, que utiliza uma combinação de vários métodos estatísticos de classificação (classificação linear robusta, máxima entropia, aprendizagem baseada em transformações e modelo não-observável de Markov) para a detecção e etiquetação das entidades. Cada um dos algoritmos mencionados etiqueta as palavras no texto com uma marca correspondendo à sua posição numa entidade mencionada: (i) começa uma entidade (ii) está dentro de uma entidade, (iii) termina uma entidade ou (iv) não pertence a nenhuma entidade.

Os traços utilizados são de extrema importância para a classificação das entidades. De acordo com T. Zhang et al (Zhang et al., 2002), um espaço de traços rico é a chave para o bom desempenho do sistema. Um sistema de classificação de elevado desempenho que opere num espaço de traços empobrecido é na maior parte dos casos ultrapassado por um sistema de desempenho inferior mas com acesso a um espaço de traços melhorado. De acordo com esta observação, os diferentes métodos de classificação em questão têm acesso a um conjunto diverso de traços, nomeadamente:

1. as palavras e os lemas das cinco palavras circundantes da palavra actual, tanto à esquerda como à direita;
2. informação sobre as partes do discurso da palavra actual e das palavras circundantes;
3. os prefixos e sufixos de dimensão até quatro caracteres da palavra actual e das palavras circundantes;
4. outro tipo de informação sobre a morfologia da palavra: *PalavraEmMaiúsculas, PrimeiraLetraMaiúscula, 2dígitos, 1dígito*, etc.;
5. informação do dicionário de termos, contendo uma lista de 50 000 cidades, 80 000 nomes próprios e 3 500 organizações;
6. informação sobre os blocos de texto;
7. a saída de dois outros sistemas de classificação, usados num sistema de pergunta-resposta da IBM.

O sistema usa ainda um algoritmo de recuperação da capitalização baseado em n-gramas para palavras que aparecem escritas completamente em maiúsculas (geralmente títulos de documentos e cabeçalhos de tabelas).

Na tabela 2.6 encontram-se representados os resultados de cada um dos quatro métodos de classificação para a língua inglesa.

Método	Medida-F (*corpus* 1)	Medida-F (*corpus* 2)
Modelo de Markov não-observável	82,0%	74,6%
Aprendizagem baseada em transformações	88,1%	81,2%
Máxima entropia	90.8%	85.6%
Classificação linear robusta	92,1%	85.5%

Tabela 2.6: Resultados individuais para cada método de classificação usado pelo sistema.

De entre os vários métodos, aqueles que apresentam o melhor desempenho são o método da máxima entropia e o método de classificação linear robusta. Os métodos de classificação linear robusta e modelo de Markov não-observável tendem a obter valores de precisão e cobertura semelhantes, enquanto que os outros dois métodos são mais precisos em sacrifício da cobertura.

Em geral, dados *n* métodos de classificação, pode-se interpretar a combinação desses métodos como uma combinação de distribuições probabilísticas:

$$P(C|w, C_1^n) = f((P_i(C|w, C_1^n))_{i=1...n}) \tag{2.1}$$

em que P_i é a probabilidade de que a classificação seja C segundo a saída do método de classificação i, C_1^n é o conjunto de classificadores usado, f é uma função de combinação, w é uma palavra e C a sua classificação. Um esquema de combinação frequentemente utilizado é o da interpolação linear:

$$P(C|w, C_1^n) = \sum_{i=1}^{n} P(C|w, i, C_i) \cdot P(i|w) = \sum_{i=1}^{n} P_i(C|w, C_i) \cdot \lambda_i(w) \tag{2.2}$$

em que λ_i representa a importância dada ao método de classificação i no contexto da palavra w e $P_i(C|w, C_i)$ é uma estimativa da probabilidade da classificação correcta ser C, dado que a saída do método de classificação i para a palavra w é C_i.

Para a combinação dos métodos referidos foram testadas várias possibilidades, nomeadamente:

1. escolher o resultado do melhor método de classificação;
2. realização de uma votação, em que cada método tem o mesmo peso (caso ocorra um empate, a classificação é escolhida aleatoriamente de entre as duas mais votadas);
3. realização de uma votação, em que cada método tem um peso diferente consoante o seu desempenho individual;
4. realização de uma votação, em que cada método não vota unicamente numa classificação, mas dá votações parciais a cada classificação através da probabilidade $P_i(C|w, C_i)$ na equação 2. Usam-se

Método	Medida-F
Melhor método individual	89,94%
Votação não-pesada	91,23%
Votação pesada	91,56%
Modelo 1	90,4%
Modelo 2	**91,64%**
Combo	**91,63%**

Tabela 2.7: Resultados das combinações de métodos de classificação (sem uso de dicionário de termos).

dois modelos, o modelo 1, em que $P_i(C|w, C_i) = P_i(C|w)$ e o modelo 2, em que $P_i(C|w, C_i) = P_i(C|C_i)$;

5. uso do método de classificação linear robusta para escolher uma função de combinação f a usar na equação 1, baseando-se nos resultados de cada método de classificação (Combo).

Os resultados de cada uma das combinações referidas são apresentados na tabela 2.7. O melhor resultado é obtido pela combinação Combo, que obtém uma medida-F de 91,63%. Ao integrar-se com um dicionário de termos e a saída de dois outros sistemas treinados num *corpus* da IBM de 1,7 milhões de palavras anotadas, o desempenho sobe para 93,9% medida-F. Estas combinações representam uma redução do erro de 17%-20% da medida-F em comparação com o melhor método individual.

2.4.5 HAREM

Na edição de 2005 do fórum de avaliação HAREM (*HAREM - Avaliação de Reconhecimento de Entidades Mencionadas*, n.d.), o sistema vencedor, considerando a medida-F, foi o PALAVRAS-NER de Eckhard Bick (Bick, 2006), que alcançou uma medida-F de 80,61%, considerando apenas as saídas oficiais. Este resultado é inferior aos melhores resultados obtidos nos fóruns de avaliação internacionais, como por exemplo o CoNLL (*CoNLL - Computational Natural Language Learning*, n.d.), em que se obtiveram medidas-F para o inglês e espanhol na ordem dos 90%. Há que ter em consideração, no entanto, que o fórum CoNLL usa diferentes métricas e uma classificação em apenas quatro categorias, enquanto que no HAREM o número de categorias e subcategorias é de 41.

Este sistema é orientado à língua portuguesa e baseia-se em regras manuscritas, tanto ao nível local (reconhecimento de padrões morfológicos) como global (contexto da frase), tendo como base uma gramática constritiva, que trata o reconhecimento de entidades mencionadas como uma tarefa integrante da anotação gramatical. As anotações das categorias candidatas são realizadas em três níveis e desambiguadas através de regras:

1. uso de entradas lexicais conhecidas e dicionários de termos (cerca de 17 000 entradas);

2. predição baseada em padrões morfológicos;

3. predição baseada no contexto para palavras que são desconhecidas.

2.5 Comparação de estratégias

Como descrito nas secções anteriores, existem diferentes tipos de estratégias em reconhecimento de entidades mencionadas, tanto em sistemas independentes da língua como orientados a um idioma em particular, entre as quais se destacam:

i) memorização simples;

ii) uso de pistas morfológicas;

iii) uso do contexto;

iv) uso de modelos estatísticos;

v) aprendizagem (supervisionada ou não-supervisionada);

vi) regras manuscritas.

As várias estratégias podem também ser combinadas entre si, podendo um sistema utilizar, por exemplo, memorização simples na sua base e o contexto para decidir se aplica ou não uma determinada classificação a uma entidade, ou uma combinação de estratégias em que cada uma tem um peso diferente, como apresentado na secção 2.4.4

Como a maior parte dos sistemas de reconhecimento de entidades mencionadas referidos foram criados tendo em conta um determinado fórum de avaliação, as categorias que estes usam na sua classificação diferem entre si, o que torna a comparação mais complexa. Por exemplo, no MUC (*MUC - Message Understanding Conferences*, n.d.) existem apenas três categorias de entidades mencionadas (TIMEX, ENAMEX e NUMEX), enquanto que no HAREM (*HAREM - Avaliação de Reconhecimento de Entidades Mencionadas*, n.d.) existem mais de quarenta. Sabendo que determinado sistema obteve uma medida-F de 90% na categoria ENAMEX no MUC não significa ser possível fazer uma comparação directa com um sistema avaliado no HAREM, em que as entidades ENAMEX estão espalhadas por diversas categorias.

Por outro lado, também não é trivial comparar sistemas com diferentes línguas alvo, já que cada língua apresenta as suas particularidades e para algumas, como o inglês, tem sido realizada mais investigação no domínio das entidades mencionadas do que para outras. Ainda assim, e tendo em conta os resultados da secção 2.2.1, que usa uma estratégia simples e independente da língua, pode-se

definir um limite inferior (*baseline*) que qualquer sistema deve conseguir alcançar para uma determinada língua.

Também os textos usados são distintos entre si: no MUC, por exemplo, são usados domínios restritos, enquanto que noutros fórum de avaliação se usa texto jornalístico, literário ou até mesmo texto de páginas *web*.

De facto, ao olhar para os resultados dos diversos fóruns de avaliação, onde é possível fazer uma comparação mais precisa entre os sistemas, o que se verifica é que não existe nenhuma estratégia que se possa dizer superior a todas as outras. Exceptuando estatégias menos complexas, como a da memorização simples ou o uso exclusivo de um dicionário de termos, as restantes estratégias descritas podem todas apresentar resultados semelhantes, como se pode ver através da descrição dos sistemas vencedores apresentados na secção 2.4, que são todos distintos entre si.

Mais importante do que a estratégia usada é o modo como esta é usada, ou seja, um sistema baseado em regras contendo apenas algumas regras simples tem um desempenho inferior a um sistema baseado em aprendizagem e com análise do contexto. Por outro lado, se um sistema baseado em regras apresenta uma grande variedade de regras específicas à língua e aos domínios, terá um melhor desempenho que um sistema de aprendizagem como o referido, se este, por exemplo, correr sobre um *corpus* de treino de dimensão reduzida.

De acordo com Satoshi Sekine (Sekine, 2004), devido à extensão do número de categorias e ao acesso a *corpora* de dimensões elevadas ($>$ 10GB de texto) a que os sistemas têm hoje acesso, as estratégias de aprendizagem supervisionada em *corpora* anotados tornam-se impraticáveis, seja por causa da inconsistência na classificação manual quando se usam centenas de categorias ou pela própria dimensão dos textos.

Na aprendizagem supervisionada, a dimensão dos dados de treino está directamente relacionada com a precisão do sistema. No entanto, a anotação manual de um texto de grandes dimensões não é uma tarefa trivial. Uma ideia alternativa para contornar este problema é anotar apenas os dados que são marcados com incerteza pelo sistema.

Segundo Satoshi Sekine (Sekine, 2004), o futuro dos sistemas de REM passa pelo uso de estratégias de aprendizagem não-supervisionada ou semi-supervisionada.

A técnica de *bootstrapping*, por exemplo, usa inicialmente apenas um conjunto de sementes. Se se pretender, por exemplo, extrair os nomes de doenças de um texto, pode fornecer-se ao sistema um conjunto inicial de cinco nomes de doenças conhecidos. O sistema, por sua vez, ao encontrar esses nomes no texto, retira daí informação sobre os contextos em que estes ocorrem, que são depois usados de modo a extrair mais entidades. O mesmo método pode ser usado para extrair relações entre, por exemplo, títulos de livros e o seu autor, através de sementes como `Shakespeare` e `Hamlet`.

Na aprendizagem não-supervisionada usa-se a técnica de *clustering*, em que as entidades são agrupadas com base na similaridade entre contextos. Existem também outros métodos, como por exemplo, o uso de conhecimento linguístico de modo a extrair entidades de um *corpus* de grande dimensão, i.e., tentar extrair as entidades usando o conhecimento subjacente ao *corpus* e não olhar para ele como uma mera sequência de caracteres. No entanto, só nos últimos anos se tornou possível utilizar este método em textos com mais de 1GB, pelo que sistemas deste tipo ainda são pouco frequentes.

Obter na tarefa de reconhecimento de entidades mencionadas uma cobertura e precisão de 100% é, como em quase todas as áreas relacionadas com língua natural, algo impossível de alcançar. Hoje em dia, no entanto, conseguem-se obter medidas-F na ordem dos 98% (ou mesmo 99% para alguns domínios mais restritos), sendo que para alguns tipos de entidades, tais como as expressões númericas, é possível atingir uma cobertura e precisão de 100%.

2.6 Sumário

O interesse em sistemas de reconhecimento de entidades mencionadas tem vindo a crescer nos últimos anos, especialmente em conjugação com o desenvolvimento de outras áreas de processamento de língua natural, tais como os sistemas de pergunta-resposta. Também no âmbito da bioinformática o REM se torna um componente tecnológico importante, nomeadamente no reconhecimento de nomes de genes e proteínas. Pode dizer-se que a tarefa de REM está a mudar do âmbito da anotação de nomes próprios para a anotação de uma categoria mais vasta de palavras e expressões que têm interesse para certo tipo de pessoas com necessidades de informação específicas.

Neste capítulo apresentaram-se diversos sistemas de REM, tanto independentes do idioma como orientados a uma língua em particular, e que usam abordagens distintas, como sejam o uso exclusivo de regras manuscritas, de informação morfológica ou contextual, memorização simples, aprendizagem supervisionada e não-supervisionada ou modelos estatísticos.

De entre todas as estratégias mencionadas, e ainda que não seja trivial a comparação de diferentes sistemas para diferentes línguas e orientados a diferentes fóruns de avaliação, nenhuma sobressai como sendo superior a todas as outras, e o uso de qualquer uma delas por determinado sistema permite-lhe apresentar resultados de cobertura e precisão bastante semelhantes, ainda que medidas-F na ordem dos 100% não sejam, como em todo o processamento de língua natural, alcançáveis.

3 Arquitectura e Procedimentos

Neste capítulo aborda-se a arquitectura geral do sistema no qual a função de reconhecimento de entidades mencionadas se insere (secção 3.1), assim como a estrutura das regras e dos léxicos necessários à sua implementação (secção 3.2) e os procedimentos usados na elaboração dessas mesmas regras (secção 3.3). Por fim, na secção 3.4 apresentam-se as directivas utilizadas como guia para efectuar o reconhecimento, tendo em conta tanto a forma de delimitação das entidades mencionadas (identificação) como as categorias e subcategorias em que estas devem ser classificadas (classificação).

3.1 Cadeia de Processamento

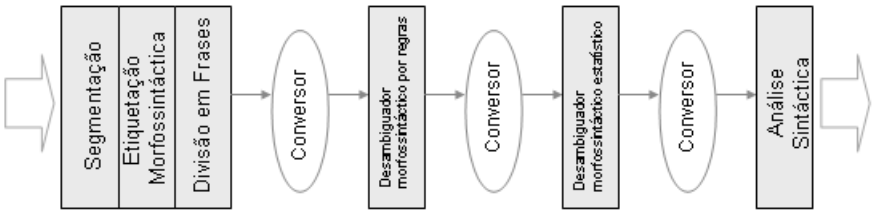

Figura 3.1: Cadeia de Processamento XIP.

O reconhecimento de entidades mencionadas insere-se na cadeia de processamento XIP do *L2F/INESC-ID*[1] em Lisboa. A ferramenta XIP[2] da Xerox é um compilador de regras que permite integrar funcionalidades de *parsing* de texto tanto a nível sintáctico como semântico, e que é usado pelo sistema para marcar e classificar as entidades mencionadas. Contudo, esta análise está inserida na parte final de uma cadeia de processamento (Mamede, 2007), como ilustrado através da figura 3.1.

A primeira tarefa desta cadeia é a segmentação do texto, i.e., a sua divisão em segmentos (ou *tokens*) individuais, efectuando-se também nesta fase a identificação de endereços IP, http e de correio electrónico, abreviaturas, números romanos, números inteiros e decimais, sinais de pontuação e outros símbolos variados como "$" , "%" ou "@" .

[1] Laboratório de Sistemas de Língua Falada do Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento.
[2] *Xerox Incremental Parser.*

Seguidamente é efectuada uma etiquetação morfossintáctica das várias palavras identificadas anteriormente através do sistema Palavroso (Medeiros, 1995), usando para isso um conjunto de etiquetas relativas às partes do discurso (nome, verbo, adjectivo, pronome, advérbio, artigo, preposição, conjunção, numeral, interjeição, marcador da passiva, residual e pontuação), podendo cada uma destas categorias apresentar campos específicos (e.g., género, número, grau, caso, tempo). Contudo, este módulo data de 1992, pelo que apresenta algumas falhas, nomeadamente o facto dos verbos e advérbios não estarem subcategorizados, ser difícil introduzir novas categorias e subcategorias e a lematização não ser adequada à análise sintáctica, visto que, por exemplo, artigos e pronomes apresentam lemas diferentes consoante o seu genéro e número, quando deveriam todos partilhar o lema do masculino singular.

Posteriormente é realizada a divisão do texto em frases, considerando como terminadores de frase os segmentos unicamente constituídos por "." , "!" e "?" , sendo o resultado convertido para o formato XML[3], de modo a poder ser utilizado pelo RuDriCo[4] (Pardal, 2007), o desambiguador morfossintáctico por regras, que efectua algumas correcções à saída do etiquetador morfossintáctico, nomeadamente alterando lemas de pronomes, advérbios e artigos (e.g. "quaisquer" → "qualquer"), realizando a descontracção dos artigos e das preposições (e.g. "no → em + o"), identificando locuções adverbiais (e.g. "à frente de"), agrupando vários segmentos (e.g. "ex" + "aluno" → "ex-aluno") e aplicando regras de desambiguação morfossintáctica.

A saída do RuDriCo é convertida de modo a poder ser utilizada pelo desambiguador morfossintáctico Marv (Ribeiro et al., 2003), que selecciona a etiqueta mais provável para cada palavra utilizando o algoritmo de Viterbi. Este desambiguador guarda as etiquetas preteridas juntamente com aquela escolhida, podendo aceder-se a esta informação posteriormente, caso seja necessário. Contudo, de modo a seleccionar a etiqueta mais provável só é usada informação sobre a categoria, subcategoria e frequência lexical. No caso dos verbos, não é escolhido um lema nem um tempo verbal (e.g. "fui" tanto pode ser uma forma do verbo "ser" como do verbo "ir" , mas o sistema só irá escolher uma das duas leituras). Por outro lado, o *corpus* de treino não tem uma dimensão suficiente (contendo presentemente cerca de 250 000 palavras), pelo que palavras que aparecem nos textos podem não existir no *corpus* de treino, o que afectará a sua frequência lexical e consequente etiquetação. Seguidamente a informação é convertida para o formato de entrada do XIP, onde são aplicadas as gramáticas locais e se introduz informação lexical, sendo também identificadas e classificadas as entidades mencionadas. Por fim, é efectuada uma segmentação em blocos (*chunks*) e são calculadas as dependências entre estes.

A arquitectura da ferramenta XIP está representada graficamente na figura 3.2. Esta ferramenta é um compilador de regras dinâmico que integra funcionalidades de *parsing* ao nível sintáctico e semântico. Uma gramática XIP pode ser usada para extrair diferentes tipos de informação de um texto

[3] *Extensible Markup Language.*
[4] *Rule Driven Converter.*

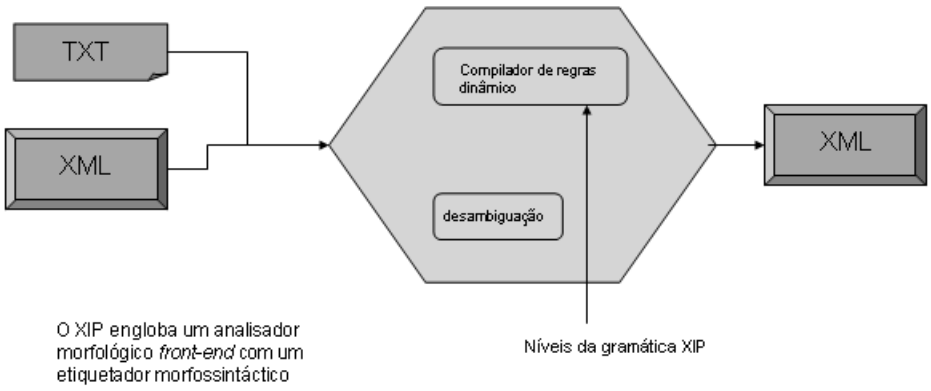

Figura 3.2: Arquitectura XIP.

em língua natural, nomeadamente:

1. Blocos (*chunks*) - e.g., sintagmas nominais, sintagmas verbais, sintagmas preposicionais;
2. Dependências - e.g., sujeito/complemento, passiva/activa;
3. Entidades Mencionadas - e.g., locais, pessoas, organizações;
4. Papéis semânticos - e.g., destino, propósito, duração;
5. Intenções comunicativas - e.g., mudanças de paradigma;
6. Co-referências.

O XIP permite representar e manipular várias características linguísticas, assim como aceder ao contexto circundante. O sistema é independente da língua, sendo que novas regras podem ser criadas sobre as existentes, de modo incremental.

As várias fases de processamento da cadeia da figura 3.1 podem ser parametrizadas, nomeadamente através da utilização de uma lista de abreviaturas no segmentador, um dicionário de palavras no etiquetador morfossintáctico, uma lista de regras de desambiguação e de descontracção no RuDriCo e gramáticas locais e léxicos no XIP.

3.2 Estrutura das Regras e Léxicos

As categorias e subcategorias de cada entidade mencionada são representadas no XIP através de traços (*features*), em que cada um pode tomar uma gama de valores previamente definida. No caso particular das entidades mencionadas, apenas um valor é atribuído aos traços ("+": presença do traço), apresentando-se alguns exemplos desses mesmos traços na tabela 3.1.

Traço	Tipo de Entidades Mencionadas
people	pessoas (e.g. João Silva, Pedro Matos)
location	locais (e.g. Lisboa, Portugal, Serra da Estrela)
event	acontecimentos (e.g. Revolução dos Cravos, Festival da Canção)
org	organizações (e.g. Coca Cola, IBM, Compal)
water	massas de água (e.g. Rio Tejo, Oceano Atlântico, Mar Negro)
title	cargos (e.g. presidente, primeiro-ministro, engenheiro)

Tabela 3.1: Exemplo de traços (*features*) utilizados no reconhecimento de entidades mencionadas.

Uma lista completa de todos os traços utilizados para a classificação de cada tipo de entidade mencionada pode ser encontrada no capítulo 4.

As palavras podem ter mais do que uma leitura, i.e., mais do que um conjunto de traços e de categorias gramaticais. Como exemplo, "olho" tanto pode ser um substantivo como um verbo (1ª pessoa do singular do presente do indicativo do verbo "olhar").

Pode-se introduzir informação lexical no XIP através de ficheiros de léxico, os quais apresentam a estrutura apresentada na figura 3.3.

```
1 Vocabulary:
2
3 Palavra1: +=[traço1=+]
4 Palavra2: +=categoria1[traço1=+]
5 ...
6 PalavraN: +=[traço2=+, lemma="PalavraNAlterada"]
```

Figura 3.3: Estrutura de um ficheiro de léxico.

A linha 1 da figura 3.3 indica o início de um ficheiro de léxico, enquanto que as linhas 3-6 representam adições de vocabulário ou alterações ao vocabulário já existente. A linha 3 tem o significado "adicionar o *traço1* à *Palavra1*" , i.e., colocar o valor "+" no *traço1* associado à palavra. Por outro lado, a linha 4 indica que não só se deverá adicionar o *traço1* à palavra, como se deverá adicionar uma nova leitura com a classe morfológica *categoria1*. Por último, a linha 6 significa que não só deverá ser adicionado o *traço2* à palavra, como também o próprio lema deverá ser alterado para o valor *PalavraNAlterada*.

```
 1 Sequence:
 2
 3 Nível> Regra 1
 4 ...
 5 Nível> Regra N
 6
 7 IDRules:
 8
 9 Nível> Regra 1
10 ...
11 Nível> Regra N
```

Figura 3.4: Estrutura de um ficheiro de regras.

As regras de identificação e classificação de entidades mencionadas, por outro lado, são definidas em gramáticas locais cuja estrutura está representada na figura 3.4.

A linha 1 indica que que as regras definidas nas linhas seguintes são regras de sequência, i.e., a ordem pela qual cada elemento aparece na regra é relevante. Por outro lado, a linha 7 indica que nas linhas seguintes estarão definidas regras de dominância imediata, em que a ordem não é relevante e apenas se procura garantir que todos os elementos presentes na regra aparecem juntos, independentemente da sua posição relativa na frase (e.g. um sintagma nominal poderá será composto de um nome seguido de um adjectivo ou de um adjectivo seguido de um nome, não interessando a ordem pelo qual cada componente aparece, mas apenas que cada elemento esteja presente na frase).

O nível da regra (número positivo $>= 1$) é colocado no início de cada linha de modo a estabelecer prioridades entre as regras i.e., as regras em níveis mais prioritários (mais baixos) serão as primeiras a tentar ser emparelhadas.

As regras utilizadas para identificar e reconhecer as entidades têm a estrutura definida na figura 3.5. As linhas 1-2 apresentam a estrutura das regras de sequência e as linhas 4-5 apresentam a estrutura das regras de dominância imediata.

```
1 CATEGORIA[traço1=+, traço2=+, ... ] = | contexto à esquerda |
2  entidade | contexto à direita | .
3
4 CATEGORIA[traço1=+, traço2=+, ... ] -> | contexto à esquerda |
5  entidade | contexto à direita | .
```

Figura 3.5: Estrutura das regras do XIP.

O significado das regras das figura 3.5 (linhas 1-2 e 4-5) é o seguinte: criar um novo bloco do tipo CATEGORIA contendo a entidade e atribuir-lhe os traços *traço1*, *traço2*, etc., sempre que esta se encontrar numa situação em que o seu contexto à esquerda e à direita correspondam àqueles determinados na regra.

Tanto a atribuição de traços como a presença de qualquer um dos contextos é opcional na estrutura das regras, podendo ter-se uma regra sem contexto, por exemplo, ou uma regra em que é identificada uma entidade, mas não lhe é atribuído nenhum traço.

Visto que se pretende identificar entidades mencionadas, a categoria resultante será necessariamente um nome (NOUN).

A utilização do operador "$->$" é utilizada nas regras de dominância imediata (ID-Rules), enquanto que o operador "=" é utilizado nas regras de sequência. É também possível utilizar o operador "@=" , que indica que se quer obter a maior entidade que emparelhe com a estrutura da regra, ao invés do operador "=" , que emparelha com a entidade mais curta possível que satisfaça os requisitos definidos.

Operador	Exemplo
Concatenação (“,”)	noun, adj
Opção (“( )” ou operadores de Kleene “*” e “+”)	adj*, (adv), noun+
Qualquer categoria (“?”)	det, ?*, noun
Disjunção (“;”)	adv;adj
Exploração de uma sub-árvore (“{ }”)	NP{?*, noun}
Existência de um traço na palavra/bloco	noun[traço1, traço2=” +”]
Ausência de um traço na palavra/bloco	noun[traço1:∼]

Tabela 3.2: Operadores utilizados nas regras do XIP.

Para definir a estrutura das entidades e dos contextos, são utilizadas referências a blocos já existentes e a categorias gramaticais (e.g., *noun, verb, adj*) e a sintaxe apresentada na Tabela 3.2.

3.3 Procedimentos

Nesta secção é descrito o método seguido durante a realização deste trabalho.

A criação de novos léxicos foi feita com base em compilações de palavras previamente recolhidas, que foram transformadas em ficheiros de léxico do XIP através de um *script* de Perl criado para o efeito. Em relação às categorias para as quais não existiam compilações prévias ou às quais foi necessário acrescentar elementos, a recolha de informação foi feita manualmente a partir da Internet, utilizando-se em seguida os procedimentos mencionados anteriormente para os léxicos baseados em compilações de palavras já existentes.

Em relação às regras definidas, e como já foi referido na secção 3.2, tanto podem ser baseadas no contexto como na estrutura da própria entidade. O modo de seleccionar os padrões de reconhecimento e os contextos propícios a um determinado tipo de entidade foi feita através da análise de *corpora* já existentes, tais como aqueles disponíveis no Projecto AC/DC da Linguateca (*Projecto AC/DC*, n.d.), que contém uma colecção de texto jornalístico em português europeu e brasileiro. Escolheu-se usar como *corpus* de referência a colecção CETEMPúblico, que contém edições completas do jornal *Público* da primeira metade da década de 90, num total de cerca de 7 milhões de frases e 191 milhões de palavras.

Uma outra fonte de material para a determinação de contextos e padrões morfológicos foram as tabelas de entidades mencionadas recolhidas por Ana Mendes (Mendes, 2007) para o seu sistema de pergunta-resposta, que utiliza a etiquetação de EMs realizada pelo sistema descrito neste documento. A análise manual das tabelas de entidades mencionadas recolhidas permitiu identificar e corrigir erros nas regras de identificação e classificação.

A avaliação dos resultados obtidos foi feita através da sua comparação com a colecção dourada (anotada) do HAREM, como pode ser consultado no capítulo 5.

3.4 Directivas

Nesta secção apresentam-se as directivas que estão na base do reconhecimento das entidades mencionadas, tanto ao nível da sua delimitação (identificação), como classificação. Começa-se por identificar os critérios gerais de identificação comuns a todos os tipos de entidades, especificando em seguida para cada uma das quatros categorias os seus subtipos e respectivos critérios de classificação, e efectuando uma comparação das directivas com aquelas definidas no fórum de avaliação HAREM (*HAREM - Avaliação de Reconhecimento de Entidades Mencionadas*, n.d.).

3.4.1 Critérios de Identificação Geral

1. Uma entidade mencionada (EM) deve conter pelo menos uma palavra com letra inicial maiúscula e/ou algarismos.
2. Deve ser classificada a EM máxima e não o número máximo de entidades com uma interpretação possível separada, i.e. *reitor da Universidade de Lisboa* deverá ser classificado como um cargo e não separado em três entidades distintas: *reitor* (cargo), *Universidade de Lisboa* (organização), *Lisboa* (local)

3.4.2 Categoria Pessoa

Tipo Individual

1. Os títulos (dr., eng., prof., etc.) usados no tratamento de uma pessoa devem ser incluídos na EM que delimita essa pessoa.
2. Formas de tratamento normalmente usadas para anteceder um nome, tais como presidente, ministro, etc. também devem ser incluídas, assim como graus de parentesco (tia, irmão, avó, etc.) quando fazem parte da forma de tratamento. Outras relações profissionais não devem ser incluídas, assim como profissões que não façam parte da forma de tratamento.
3. Os cargos que estejam separados do nome por uma vírgula não devem ser incluídos no tipo Individual. Se houver vírgula, são incluídos.
4. Diminutivos, alcunhas, iniciais, nomes mitológicos e entidades religiosas são etiquetados nesta categoria. Exemplos: Anocas, Nani, A. Costa, Neptuno, Santo António

Tipo Grupoind

1. O tipo Grupoind representa um grupo de indivíduos (do tipo Individual) que não tem um nome estático como grupo. Exemplos: os Mirandas, o governo de José Sócrates, Vossas Excelências.

Tipo Cargo

1. O tipo Cargo deve ser usado na referência de um posto que é ocupado por uma pessoa, mas que poderá no futuro ser ocupado por outros indivíduos. Ou seja, num dado contexto, pode representar uma pessoa em concreto, mas através da referência ao seu cargo. Exemplos: Papa, Ministro dos Negócios Estrangeiros, Rainha da Inglaterra, Primeiro-Ministro.

Tipo Grupocargo

1. O tipo Grupocargo é análogo ao Grupoind, designando EMs que referem um conjunto de pessoas, através de um cargo. Exemplo: Ministros dos Negócios Estrangeiros da UE.

Tipo Membro

1. O tipo Membro é aplicado quando um indivíduo é mencionado pela organização que representa. Exemplos: um GNR, um Mórmon.

3.4.3 Categoria Organização

Tipo Administração

1. Este tipo pretende etiquetar as organizações relacionadas com a administração e governação de um território, tais como ministérios, municípios, câmaras, autarquias, secretarias de estado, etc. Inclui também as organizações que têm a ver com a governação a nível internacional ou supranacional. Exemplos: Ministério do Ambiente, Câmara Municipal de Lisboa, Secretaria de Estado dos Transportes, ONU, UE.
2. EMs referentes a países, territórios, regiões autónomas ou mesmo territórios ocupados ou ex-colónias, podem ser uma organização, dependendo do contexto.

Tipo Empresa

1. O tipo Empresa abrange organizações com fins lucrativos, como empresas, sociedades, clubes, etc. Exemplos: Xerox, Boavista FC, Círculo de Leitores, Livraria Barata, Microsoft.

Tipo Instituição

1. O tipo Instituição inclui todas as organizações que não possuem fins lucrativos nem um papel directo na governação. Este tipo abrange instituições no sentido estrito, associações e outras organizações de espírito cooperativo, universidades, colectividades, escolas e partidos políticos. Exemplos: Associação de Amizade Portugal-Bulgária, Universidade Técnica de Lisboa, Liceu Maria Amália, Amnistia Internacional, Partido Comunista Português.

Tipo Sub

1. As EMs do tipo Sub referem-se a determinados sectores de uma organização, mas sem autonomia para ser considerada ela própria uma organização, tais como departamentos, secções, assembleias gerais, comissões, comités, secretarias, etc. Exemplos: Comité Geral do PCP, Departamento de Marketing da Xerox, Comissão Winograd, Assembleia Geral do Benfica

2. No caso de sucursais, filiais, empresas em regime de 'franchising', etc, ou seja, onde haja autonomia suficiente para as considerarmos uma organização autónoma, a EM deve ser classificada como uma Empresa, e não uma Sub. Exemplo: Volskwagen Portugal

3.4.4 Categoria Acontecimento

Tipo Efeméride

1. Uma Efeméride é um acontecimento ocorrido no passado e não repetível. Exemplos: o 25 de Abril, o 11 de Setembro, a 2ª Guerra Mundial.

Tipo Organizado

1. Um acontecimento Organizado é um acontecimento multifacetado, que poderá durar vários dias, e geralmente conter vários eventos. Exemplos: o Euro 2004, os Jogos Olímpicos, o Festival de Jazz do Estoril. Quando o acontecimento em questão é um evento periódico, distinguido pelo ano do acontecimento ou pelo seu local, estes (data ou local) devem ser incluídos na etiqueta de acontecimento.

Tipo Evento

1. Um Evento é um acontecimento pontual, organizado ou não. Exemplos: Benfica-Sporting, Britney Spears no Pavilhão Atlântico, Buzinão na Ponte, etc.

2. O Euro 2004, que foi um acontecimento Organizado, incluiu vários Eventos (jogos, festas, conferências, etc).

3.4.5 Categoria Local

Tipo Administrativo

1. O tipo Administrativo Identifica localizações que foram criadas e/ou delimitadas pelo Homem. Inclui países, bairros, regiões geopolíticas, entre outras. Exemplos:Portugal, Rio de Janeiro, Alentejo, América Latina, Alfama

Tipo Correio

1. O tipo Correio abrange todas as referências a locais com indicações completas, tais como moradas, números de salas, salas de cinema. Exemplos: Sala 6, Caixa Postal 2400, Rua da Escola 15B.

Tipo Geográfico

1. O tipo Geográfico indica localizações de geografia física que apenas foram baptizadas (e não construídas) pelo Homem. Exemplos: Serra da Estrela, Mar Negro

Tipo Virtual

1. O tipo Virtual engloba locais como a Internet, e números de telefone ou de fax, desde que contenham ou algarismos ou letras maiúsculas. Também abrange locais de publicação, referidos pelos nomes dos meios de comunicação social. Exemplos: Jornal de Notícias, Telejornal, eBay, 21 555 5555

Tipo Alargado

1. Deve conter referências a locais que não estão nas categorias anteriores, mas que referem um determinado sítio físico, como é o exemplo de pontos de encontro em edifícios, bares, hotéis, praças, centros de congressos, restaurantes, etc. Exemplo: Centro Comercial Amoreiras, Praça da Figueira, Centro de Congressos de Lisboa, Hotel Sheraton.
2. No caso de se referir uma rua, avenida ou praça como um local onde ocorreu ou está localizada qualquer coisa, mas não como se de uma morada ou endereço se tratasse, considera-se um local do tipo Alargado.

3.4.6 Diferenças em relação ao HAREM

As directivas descritas na secção 3.4 são baseadas nas directivas definidas pelo fórum de avaliação HAREM (*HAREM - Avaliação de Reconhecimento de Entidades Mencionadas*, n.d.) para uso na edição de 2005, e cuja versão mais recente é de 18 de Janeiro de 2005.

Uma diferença essencial prende-se com a delimitação das entidades. De acordo com as directivas do HAREM, embora a classificação deva ter em conta o significado da entidade mencionada no texto, a sua delimitação deve restringir-se apenas à parte associada ao nome próprio (em maiúsculas), enquanto que de acordo com as directivas definidas na secção 3.4, toda a entidade deve ser considerada, e não apenas as palavras em maiúscula que dela fazem parte. De modo a exemplificar esta diferença, observem-se as frases 3.1 e 3.2:

Frase 3.1: *"Este fim-de-semana fui à serra da Estrela."*

Frase 3.2: *"O tratado de Tordesilhas foi assinado em 1494."*

De acordo com as directivas do HAREM, são identificadas as entidades mencionadas "Estrela" e "Tordesilhas" .

Não parece ser esta a abordagem mais correcta, já que as palavras que precedem a entidade e que servem para a sua classificação são parte essencial da identidade da mesma, i.e., "serra da Estrela" e "tratado de Tordesilhas" são entidades distintas de "Estrela" (o bairro) e "Tordesilhas" (a cidade), pelo que não tem sentido delimitá-las do mesmo modo, quando existe informação para não o fazer. De acordo com as directivas da secção 3.4, as entidades mencionadas das frases 1 e 2 seriam "serra da Estrela" e "tratado de Tordesilhas" , classificadas respectivamente com as categorias *Local* e *Acontecimento*.

Por outro lado, as indicações nas directivas do HAREM do que diz respeito à classificação morfológica não são consideradas, i.e., as entidades não são marcadas com o seu género ou número.

Além das diferenças já referidas, as categorias e subcategorias (tipos) relativas aos vários tipos de entidades que se pretende classificar são essencialmente iguais àquelas definidas pelo HAREM.

4 Implementação

Neste capítulo são descritas as regras e os léxicos usados na identificação e classificação de cada um dos quatro grupos de entidades mencionadas definidos: locais (secção 4.1), pessoas (secção 4.2), organizações (secção 4.3) e acontecimentos (secção 4.4). Por fim, são mencionadas algumas entidades auxiliares para a classificação, tais como nacionalidades, mas que não são entidades mencionadas (secção 4.5).

4.1 Locais

Pretende-se classificar as entidades mencionadas do tipo local de acordo com as directivas apresentadas na secção 3.4, usando para isso o conjunto de traços da tabela 4.1. Além destes traços, são utilizados também traços auxiliares como *city* (cidade), *country* (país), *continent* (continente), *water* (massas de água), *mountain* (montanhas) ou *cardinal* (ponto cardeal), embora estes não correspondam a nenhum tipo definida nas directivas (e.g. tanto cidades como países são regiões administrativas e tanto massas de água como montanhas são locais geográficos) e sejam apenas usados como auxiliares na classificação das entidades.

Na identificação de locais foi utilizada uma lista de palavras (léxico) contendo 370 cidades e vilas portuguesas, 333 cidades internacionais (não capitais), 224 países, 178 regiões nacionais e internacionais, 159 capitais, 64 ilhas e arquipélagos, 50 estados americanos, 24 estados brasileiros, 17 locais geográficos (cordilheiras e mares) e 6 continentes, além de uma lista dos pontos cardeais e de alguns locais virtuais (e.g. Internet), perfazendo no total 1436 entradas lexicais.

Traço	Subcategoria correspondente
location	local (categoria de topo)
admin_area	tipo administrativo
geographic	tipo geográfico
virtual	tipo virtual
correio	tipo correio
extended	tipo alargado

Tabela 4.1: Traços usados na classificação das entidades do tipo local.

Existem alguns locais que podem ser identificados através da análise da estrutura da entidade, embora não sejam previamente conhecidos. Por exemplo, tendo a informação lexical de que "Europa" é um local, podemos afirmar que "Europa do Norte" ou "norte da Europa" (uma combinação de um local com um ponto cardeal) também deverá ser um local.

```
3> NOUN[location=+,city=+] = ?[lemma:novo];?[lemma:nova], noun[location,city].

3> NOUN[location=+] = ?[lemma:novo];?[lemma:nova], noun[location].
```

Figura 4.1: Identificação de locais do tipo "Nova Iorque" e "Novo México" .

Existem outros padrões que permitem inferir um local através da sua estrutura, nomeadamente as construções representadas nas regras da figura 4.1. Neste caso, tendo conhecimento de que uma cidade é um local (e.g. Iorque, Lisboa), pode afirmar-se que "Nova Iorque" e "Nova Lisboa" também são locais e cidades. Contudo, se não existir a informação de que o local é uma cidade, e se estiver perante uma entidade como "Novo México" ou "Nova Zelândia" , só se pode inferir que se trata de um local, não podendo deduzir-se qualquer informação sobre o seu tipo (estado, país, região, etc.).

De igual modo, construções relativas ao nome oficial de um país (e.g. "República Popular da China" , "República Democrática do Congo" , "República Islâmica da Mauritânia") são indicativos não só de um local, mas mais especificamente de um país.

Existem ainda outros locais que podem ser identificados pela sua estrutura, como por exemplo "Cidade do México" , "Cidade da Guatemala" , "Estados Unidos da América" ou "Emiratos Árabes Unidos" . No caso de locais geográficos como "Mar Negro" , "Rio Tejo" ou "Oceano Atlântico" , o próprio nome contém um indicador que permite classificar a sua categoria e tipo ("mar" , "rio" , "oceano"), sendo também esse o caso de entidades como ilhas, penínsulas ou arquipélagos (e.g., "Ilhas Maurícias" , "Península Ibérica" , "Arquipélago dos Açores").

Apresentam-se na tabela 4.2 outros exemplos de estruturas que permitem identificar locais do tipo administrativo.

distrito de...
região de...
bairro de...
condado de...
estado de ...
vila de...
cidade de ...
lugar de...

Tabela 4.2: Exemplos de indicadores de locais do tipo administrativo.

Também os locais do tipo alargado podem geralmente ser identificados através da sua própria es-

Local	Exemplos
Teatro	Teatro Nacional D. Maria II, Teatro Municipal da Guarda
Estádio	Estádio Alvalade XXI, Estádio da Luz
Hotel	Hotel Altis, Hotel Sheraton
Jardim	Jardim da Estrela, Jardim do Paço
Porto	Porto de Lisboa, Porto de Leixões
Cemitério	Cemitério dos Prazes, Cemitério Novo
Mina	Minas da Panasqueira
Praia	Praia da Rocha, Praia do Meco
Quinta	Quinta do Lago, Quinta da Marinha
Aeroporto	Aeroporto da Portela, Aeroporto de Heathrow
Mosteiro	Mosteiro da Batalha, Mosteiro dos Jerónimos

Tabela 4.3: Exemplos de entidades do tipo alargado que podem ser identificadas a partir da sua estrutura.

trutura, apresentando-se alguns exemplos desse tipo de estruturas na tabela 4.3. Embora a própria entidade contenha um nome que a identifica, há no entanto que considerar os diferentes tipos de constituintes que lhe seguem, já que estes podem ter estrutura e extensão distintas (e.g., "Minas de Jiaohe" e "Minas de Carvão de Jiaohe").

Embora existam, como mencionado anteriormente, alguns locais que podem ser identificados e classificados recorrendo exclusivamente à sua estrutura, a maior parte das entidades mencionadas nesta categoria necessita de ser identificada e classificada recorrendo ao contexto. Um contexto sugestivo de uma localização é um contexto junto ao qual se espera uma entidade mencionada do tipo local. Naturalmente, os contextos não são 100% eficazes, e existem ocasiões em que o uso de determinado contexto resulta na classificação ou identificação errada de uma entidade. Por essa razão, consideram-se apenas aqueles contextos que apresentam um grau de precisão aceitável, precisão esta determinada informalmente através da análise de *corpora* de texto e da frequência com que um entidade ocorre juntamente com determinado contexto.

Verbo
viajar a/para ...
ir a/para ...
vir de/a/para ...
chegar a/de ...
deslocar-se a/para ...
aterrar em ...
regressar a/de ...

Tabela 4.4: Exemplos de verbos de movimento utilizados no reconhecimento de entidades do tipo local.

No caso das entidades mencionadas do tipo local, um dos contextos mais propícios é o dos verbos de movimento que denotam uma direcção ou uma proveniência, como aqueles apresentados na tabela

4.4. Estes verbos são usados como contexto à esquerda na identificação das entidades, não permitindo, no entanto, determinar a subcategoria da mesma. Os contextos não são, no entanto, estáticos, pelo que permitem que existam palavras opcionais entre os contextos e a entidade (e.g., "ir a Lisboa" ou "ir de comboio a Lisboa").

Expressão
ir dar a ...
situar-se em ...
localizar-se em ...
ser em ...
ficar em ...
ficar perto/longe de...
estar em ...
exilar-se em...
andar na escola em ...
jantar/almoçar em...
... ter x habitantes

Tabela 4.5: Exemplos de outros verbos e expressões utilizadas no reconhecimento de entidades do tipo local.

Existem também outros verbos, que embora não denotando movimento, estão ligados à determinação de locais, como por exemplo o verbo "ser" em frases como "Sou de Lisboa" ou o verbo "nascer" em frases como "Nasceu em Belém" . Uma lista de alguns desses verbos e expressões pode ser encontrada na tabela 4.5.

De modo a identificar entidades do tipo geográfico que são rios pode olhar-se a contextos sugestivos como "estuário de..." , "delta de..." , "foz de..." , "...desagua" etc. De modo similar, para identificar entidades relacionadas com massas de água (rios, lagos, mares, oceanos, etc.) recorre-se a contextos como "naufragar em..." , "navegar em..." , "nadar em.." , "velejar em.." , etc.

Os nomes de ruas, avenidas, largos, alamedas, travessas, etc. que não se refiram a uma morada completa são classificados no tipo alargado, ao passo que endereços completos são classificados no tipo correio, sendo que estas últimas regras são colocadas em primeiro lugar. Referências a salas de aula, de cinema, caixas postais, etc. são também incluídas na categoria alargado.

As entidades que são locais do tipo virtual (números de telefone e fax, *urls*, endereços de *e-mail*, etc.) são mais difíceis de identificar num texto, já que muitas vezes aparecem sem qualquer contexto (e.g., seguido de um nome próprio ou entre parênteses). Decidiu-se identificar apenas as entidades que estão explicitamente marcadas como tal no texto, i.e., são precedidas de marcas como "Telefone" , "Telemóvel" ,"Tel." , "Fax:" , "E-mail" , etc. assim como seguindo expressões do tipo"visitar o *site*..." , " o *website* ..." , bem como nomes de publicações através de contextos como "artigo em..." ou " publicado em..." seguidos de uma entidade classificada como organização.

```
4> NOUN[location=+, admin_area=+] @= |noun[location], ?[lemma:e];?[lemma:ou],
(prep), (art)| ?+[maj, toutmaj:~], (prep[lemma:de], (?[lemma:o]), ?+[maj])*.

4> NOUN[location=+, admin_area=+] @= ?+[maj, toutmaj:~], (prep[lemma:de],
(?[lemma:o]), ?+[maj])* |?[lemma:e];?[lemma:ou], (prep), (art), ?[location] |.
```

Figura 4.2: Regras de conjunção e disjunção.

Usam-se as relações de conjunção e disjunção de modo a relacionar entidades do mesmo tipo (ver figura 4.2). No caso de ser ter uma frase como "Pedro fez uma viagem a Tânger e Fez" , a regra de contexto disparada a partir do contexto " uma viagem a/para" só permite classificar como entidade do tipo local a cidade de Tânger. A regra final de conjunção/disjunção permite que uma entidade com letra maiúscula separada de outra entidade mencionada pela conjunção "e" ou "ou" (e opcionalmente uma preposição) seja classificada com a mesma categoria que esta. No entanto, em casos como este não há possibilidade de saber com exactidão o tipo, pelo que se optou por atribuir o tipo mais comum nesta categoria (administrativo) em todos os casos, o que pode resultar numa diminuição na precisão na classificação por tipo, mas que permite uma maior abrangência a nível dessa mesma classificação.

4.2 Pessoas

Traço	Subcategoria correspondente
people	pessoa (categoria de topo)
individual	tipo individual
grupoind	tipo grupoind
postpeople	tipo cargo
postgroup	tipo grupocargo
member	tipo membro

Tabela 4.6: Traços usados na classificação das entidades do tipo pessoa.

Pretende-se classificar as entidades mencionadas do tipo pessoa de acordo com as directivas apresentadas na secção 3.4, usando para isso o conjunto de traços da tabela 4.6. Além destes traços, são utilizados também traços auxiliares como *title* (título) e *relative* (relações familiares), embora estes não correspondam a nenhuma subcategoria definida nas directivas.

Na identificação deste tipo de entidades foi utilizada uma lista de palavras (léxico) contendo 655 nomes próprios portuguesess (masculinos e femininos) aprovados pelo Ministério da Justiça, uma lista de 29 nomes próprios brasileiros, 58 nomes próprios de origem estrangeira (inglesa, francesa e espanhola) e 20 nomes de personagens históricas (e.g. Napoleão, Shakespeare) assim como uma lista de 402 apelidos portugueses. Desta última lista, contudo, foram retirados vários nomes do sistema, visto existirem conflitos com outras entidades mencionadas e palavras comuns (e.g., apelidos como "Braga" ou "Guimarães" referem-se a locais). Existem também apelidos que são palavras comuns (e.g., "Rocha" ,

"Pereira"), mas estes conflitos são resolvidos numa fase anterior da cadeia de processamento. No total, contam-se 1164 entradas lexicais.

Título ou forma de tratamento	Exemplo
Professor	professor Marcelo Rebelo de Sousa
Engenheiro	engenheiro António Guterres
Padre	padre Milícias
General	general Rocha Vieira
Lord	Lord Winston
Papa	papa João Paulo II
Tio	tio João
Senhor	senhor Silva
Dom	D. João II

Tabela 4.7: Alguns exemplos de títulos ou formas de tratamento usados na identificação de entidades do tipo pessoa.

De modo semelhante à identificação e classificação de pessoas descrita na secção 4.1, existem entidades mencionadas do tipo pessoa que podem ser classificadas somente a partir da sua estrutura e outras para as quais é necessário recorrer ao contexto. Incluem-se na primeira categoria aquelas entidades que fazem referência a títulos ou formas de tratamento (e respectivas abreviaturas) como aqueles apresentados na tabela 4.7. Mesmo nos casos em que a profissão não faz parte da forma de tratamento, esta é usada para a identificação, como por exemplo na frase "o serralheiro João das Neves" .

Existem, por outro lado, situações onde o próprio cargo é usado como referência a uma pessoa (e.g., "o primeiro-ministro" , "o senhor padre"), sendo que este tipo de entidades será identificada somente se a palavra referente ao cargo estiver escrita em maiúsculas, e classificada com o tipo cargo. Em geral, cada título pode ainda ter variações associadas, como por exemplo "director" , "director-geral" , "director financeiro" , "professor assistente" , "professor associado" , etc. Incluem-se de igual modo referências a postos religiosos (e.g., "Monsenhor" , "Cardeal" , "Bispo"), cargos de nobreza (e.g., "Rei" , "Conde" , "Xá") e formas de tratamento estrangeiras (e.g. "Miss" , "Madame" , "Monsieur").

Os nomes de santos (e.g. "São João" , "Santo António") são também identificados através da sua estrutura, caso não façam parte de uma entidade previamente identificada como local (e.g. "Vila Real de Santo António" , "Santo Domingo").

Como no caso da identificação de locais (secção 4.1), a maior parte das entidades tem de ser reconhecida através do contexto, apresentando-se na tabela 4.8 alguns exemplos de expressões e verbos usados como contexto à esquerda e na tabela 4.9 alguns exemplos de expressões e verbos usados como contexto à direita para entidades do tipo pessoa.

Tem-se também em conta a voz passiva, pelo que tanto na frase "Al Gore não inventou a Internet" como na frase "A Internet não foi inventada por Al Gore" a entidade "Al Gore" será identificada

Expressões
entrevista a/com...
segundo...
o discurso de...
o sucessor de...
nomear...
falar com...
telefonar a/para...
discutir com...
casar-se com...
divorciar-se de...
ter inveja/ciúmes de...
o golo de...
amigo de...
o assassinato de...

Tabela 4.8: Exemplos de verbos e expressões utilizadas como contexto à esquerda no reconhecimento de entidades do tipo pessoa.

como uma pessoa.

Alguns destes verbos podem ser também usados com locais (e.g., "A França afirmou que não aceitaria intromissões na sua política interna" , "A Noruega lidera a tabela dos países mais ricos"), pelo que se excluem desta categoria todas as entidades que já tenham sido identificadas anteriormente como locais. Numa fase posterior, neste tipo de situações, o local será classificado como uma organização do tipo administrativo.

```
2> NOUN[people=+,individual=+] @= |noun[culture], (punct[comma:+]), ?[lemma:de],
(art)| ?+[maj, location:~, lemma:~secretaria], (prep[lemma:de], (?[lemma:o]), ?+[maj])*.
```

Figura 4.3: Regras utilizadas para identificar pessoas que são autores de obras culturais (e.g., livros, filmes, etc.)

Tendo sido classificadas as entidades do tipo obra (e.g. títulos de livros e filmes), é possível identificar uma pessoa através da estrutura da regra da figura 4.3, do tipo *«A Insustentável Leveza do Ser», de Milan Kundera*.

A relação de aposto permite também identificar entidades do tipo pessoa, quer quando a entidade em si é o aposto, quer quando este é uma profissão, um título ou uma descrição que permite classificar o sintagma nominal a que este se liga como uma pessoa (e.g., "Cavaco Silva, o presidente da república..." , "o primeiro-ministro, José Sócrates...").

Depois de efectuada a classificação das entidades que são organizações (secção 4.3), é possível também identificar entidades do tipo cargo através de contextos como "Presidente da Microsoft" , em que "Microsoft" é uma organização.

Um outro contexto particular tem que ver com o uso de iniciais para designar um nome (geralmente

Expressões
... dizer
... afirmar
... declarar
... referir
... aceitar
... mencionar
... liderar
... marcar um golo
... casar-se
... nascer
... morrer
... trabalhar
... fundar
... escrever
... vencer
... suicidar-se
... ter x anos

Tabela 4.9: Exemplos de verbos e expressões utilizadas como contexto à direita no reconhecimento de entidades do tipo pessoa.

em texto de entrevista). Optou-se por considerar que se uma entidade do tipo pessoa está seguida de uma expressão abreviada entre parênteses, então essa expressão também será uma pessoa e referir-se-á à mesma entidade (e.g., "Maria Silva dos Reis (M.S.R.)")

Também no texto de entrevista (e nas peças teatrais) é comum encontrar o nome dos diferentes intervenientes antes de cada fala, na maior parte dos casos abreviado, pelo que se considera uma abreviatura a seguir a um parágrafo seguida do sinal de pontuação ":" como um sinal da existência de uma entidade do tipo pessoa (de acordo com as directivas definidas também os nomes de jornais ou revistas são considerados pessoas quando representam o entrevistador). Contudo, esta regra acaba por identificar também outras expressões que não são necessariamente pessoas, tais como "Secretaria do I.S.T.: 21 0000 999 " .

Recorrendo à informação sobre palavras que são nacionalidades (secção 4.5) é possível também observar um padrão comum, em que o nome da pessoa é precedido ou antecido da sua nacionalidade (e.g."o britânico James Smith" , " os espanhóis Pablo e Juan Dominguez" , "Paris Hilton, a americana mais falada do momento").

Através das relações familiares, é possível também identificar entidades do tipo pessoa, em expressões como "Henry Fonda é o pai de Jane Fonda" ou "Bashar Al-Assad, o filho de Hafez Al-Assad" .

De maneira semelhante ao caso dos locais, é também feito uso das relações de conjunção e disjunção de modo a relacionar entidades do mesmo tipo. No caso de ser ter uma frase como "Maria telefonou a Pedro e a Joana" , a regra de contexto disparada a partir do contexto " telefonar a" só permite classi-

ficar como entidade do tipo pessoa “Pedro” . A regra final de conjunção/disjunção permite que uma entidade com letra maiúscula separada de outra entidade mencionada pela conjunção “e” ou “ou” (e opcionalmente uma preposição) seja classificada com a mesma categoria que esta. No entanto, em casos como este não há possibilidade de saber com exactidão o tipo, pelo que se optou por atribuir o tipo mais comum nesta categoria (individual), o que pode resultar numa diminuição na precisão na classificação por tipo, mas que permite uma maior abrangência a nível dessa mesma classificação.

São classificados com o tipo grupomembro as referências a grupos ou organizações (e.g., “Polícia” , “GNR” , “Testemunhas de Jeová”) quando precedidos do artigo indefinido singular. Outras entidades do tipo pessoa classificada neste tipo incluem referências a clubes de futebol em junção com verbos como “jogar” , “derrotar” ou “vencer” , em que a equipa ou selecção é vista não como uma organização, mas como um conjunto de pessoas. Também os nomes de povos antigos ou modernos, cuja referência venha precedida do artigo definido no plural são classificadas nesta categoria (e.g., “os Romanos” , “os Incas”).

São classificados com o tipo grupo individual as referências a cargos quando o elemento principal se encontra no plural (e.g. “Ministros dos Negócios Estrangeiros”), assim como referências a, por exemplo, famílias, em que a presença da palavra família ou o artigo definido masculino plural servem de contexto à classificação (e.g., “os Mirandas” , “a família Braga”).

Por último, é efectuada a junção de nomes, sempre que se tiver um sintagma nominal já marcado como pessoa seguido de outro sintagma nominal cuja palavra inicial começa por maiúscula. Nesse caso, as duas entidades são agrupadas e marcadas como uma só com a classificação pessoa e tipo individual.

4.3 Organizações

Traço	Subcategoria correspondente
org	organização (categoria de topo)
administration	tipo administrativo
institution	tipo instituição
suborg	tipo sub
company	tipo empresa

Tabela 4.10: Traços usados na classificação das entidades do tipo organização.

Pretende-se classificar as entidades mencionadas do tipo organização de acordo com as directivas apresentadas na secção 3.4, usando para isso o conjunto de traços da tabela 4.10.

Na identificação deste tipo de entidades foi utilizada uma lista de palavras (léxico) contendo 128 nomes de empresas (portuguesas e estrangeiras), organizações internacionais, siglas de partidos políticos e outras organizações estatais.

Estrutura	Exemplo
Município/Autarquia/Concelho	Concelho de Lisboa
União	União Europeia
Universidade/Faculdade/Instituto	Universidade Técnica de Lisboa
Associação/Liga/Grupo/Conselho/Federação	Associação dos Amigos dos Animais
Bombeiros	Bombeiros Voluntários de Massamá
Biblioteca/Arquivo	Biblioteca Municipal de Bragança
Comunidade	Comunidade de Países de Língua Portuguesa
Ministério/Secretaria de Estado	Ministério das Obras Públicas
Embaixada	Embaixada de Portugal em Madrid
Banco	Banco Millenium BCP
Assembleia	Assembleia Nacional Francesa
Partido	Partido Comunista Português
Fábrica	Fábrica da Coca-Cola
Igreja	Igreja Universal do Reino de Deus
Polícia	Polícia Federal
Exército/Armada	Exército Popular de Libertação
Hospital	Hospital de Santa Maria

Tabela 4.11: Exemplos de estruturas utilizadas no reconhecimento de entidades do tipo organização.

Uma estrutura indicativa de organizações, mais precisamente do tipo empresa, é aquela em que o nome contém no final “S.A.” , “Lda.” , ou em inglês “Ltd.” . Estas entidades contêm ainda outras características identificadoras, como o uso de expressões como “& Filhos”

De modo idêntico, duas expressões em maiúscula separadas pelo símbolo “&” são consideradas também como organizaçãoes (e.g. “AT&T” , “Barnes & Noble”).

Na tabela 4.11 encontram-se alguns exemplos de estruturas de entidades do tipo organização que podem ser identificadas e classificadas olhando apenas para os seus constituintes.

Contexto
...invadir/capturar/conquistar/derrotar...
...separar-se (de)...
...tornar-se independente (de)...
...afirmar/declarar/dizer
...assinar/ratificar/acordar
...condenar/apoiar/pressionar
...aceder/torna-se membro (de)...
liderar/comandar/dirigir...
a ameaça de....
a garantia de...
relação com...
autonomia em relação a...

Tabela 4.12: Exemplos de contextos utilizados no reconhecimento de entidades do tipo organização que também são locais.

Entidades que se referem a publicações como revistas e jornais são classificadas com o tipo empresa

e identificadas a partir de estruturas como “Diário...” , “Jornal (de)...” , “Rádio (de)...” , “TV...” ou em inglês “...Times” .

São também usadas estruturas da língua inglesa para entidades que são comuns, tais como “University of...” ou “... University” .

Em determinadas situações, uma referência a um país ou a uma capital é classificada como organização, quando este se refere a um governo ou a outra entidade política (e.g. “Bruxelas avisa Portugal de que não pode ultrapassar o défice de 3%” , “A Alemanha invadiu a Polónia em 1939”). Apresentam-se na tabela 4.12 alguns exemplos de contextos, tanto à esquerda como à direita, que permitem reclassificar uma entidade do tipo local como uma organização. De notar que alguns destes contextos permitem classificar simultaneamente duas entidades, como é o caso de frases como “Espanha assinou um acordo de paz com Portugal” , em que o contexto “assinar um acordo” permite classificar tanto a entidade à esquerda (Espanha) como à direita (Portugal) como organizações.

Entidades como hóteis, pensões, centros de congressos, pousadas, etc. são ambíguas, no sentido que tanto podem representar um local como referir-se à organização em si (e.g. “A conferência realiza-se no Hotel Sheraton” vs “o Hotel Sheraton emprega 200 pessoas”). Opta-se neste caso por considerar que uma referência a uma organização deste tipo é um local sempre que vier precedida de uma preposição como “em” ou ” a” , e que representa uma organização caso contrário.

Entidades como “Estado” , “Governo” , “Procuradoria” ou derivados como “governo da China” , “Estado Maior” ou “Procuradoria Geral da República” são sempre classificadas como organizações do tipo administrativo quando ocorrem em letra maiúscula, embora em alguns casos isto leve a uma classificação errada (e.g. ” Estado Novo”).

Uma sequência de letras totalmente em maiúscula ou uma entidade que já foi classificada como organização seguida de um nome de país é também classificada como organização (e.g. “Volskwagen Portugal” , “Epson do Brasil”).

Contexto
publicidade/anúncio a/de...
a sede de...
o site de ...
licencidado/graduado/doutorado por...
empregado/funcionário/trabalhador/técnico de...
administração de...
editado/distribuído/comercializado por ...

Tabela 4.13: Exemplos de contextos à esquerda utilizadas no reconhecimento de entidades do tipo organização.

De modo idêntico à identificação de pessoas (secção 4.2), considera-se que se uma entidade do tipo

organização está seguida de uma expressão abreviada entre parênteses, então essa expressão também será uma organização e referir-se-á à mesma entidade (e.g., "Instituto Superior Técnico (I.S.T.)").

Organizações terroristas são identificadas através de contextos como "atentados de..." ou " atentados reinvidicados por..." . Isto permite identificar entidades como "ETA" , "IRA" , "Al-Qaeda" ou "Hamas" .

Novamente, de maneira semelhante ao caso dos locais e pessoas, é também feito uso das relações de conjunção e disjunção de modo a relacionar entidades do mesmo tipo. No caso de ser ter uma frase como "Os EUA avisaram a Rússia e China" , a regra de contexto disparada a partir do contexto " avisar" relacionada com locais só permite classificar como entidade do tipo organização "Rússia" . A regra final de conjunção/disjunção permite que uma entidade com letra maiúscula separada de outra entidade mencionada pela conjunção "e" ou "ou" (e opcionalmente uma preposição) seja classificada com a mesma categoria que esta. No entanto, em casos como este não há possibilidade de saber com exactidão o tipo, pelo que se optou por não atribui-lo tipo nestes casos, o que resulta numa diminuição na precisão na classificação por tipo, mas que permite uma maior abrangência a nível dessa mesma classificação.

4.4 Acontecimentos

Traço	Subcategoria correspondente
event	acontecimento (categoria de topo)
ephem	tipo efeméride
organized	tipo organizado
eventac	tipo evento

Tabela 4.14: Traços usados na classificação das entidades do tipo acontecimento.

Pretende-se classificar as entidades mencionadas do tipo acontecimento de acordo com as directivas apresentadas na secção 3.4, usando para isso o conjunto de traços da tabela 4.14.

São também utilizadas algumas estruturas não gerais que identificam entidades específicas como "Queima das Fitas" , "Recepção ao Caloiro" ou "Volta a Portugal" .

Alguns exemplos de estruturas que permitem identificar entidades do tipo acontecimento estão apresentadas na tabela 4.15.

Entidades do tipo efeméride são também identificadas através da sua estrutura, apresentando-se alguns exemplos na tabela 4.16

A maior parte das entidades desta categoria podem também ser precedidas por um número ordinal

Estrutura	Exemplo
Feira	Feira Internacional de Lisboa
Simpósio	Simpósio Internacional sobre as Alterações Climáticas
Semana/Mês	Semana da Música
Campeonato/Torneio/Copa/Taça	Campeonato Nacional de Futebol
Grande Prémio	Grande Prémio do Estoril
Cerimónia	Cerimónia de Encerramento da Expo 98
Exposição	Exposição Internacional de Pintura Abstracta de Barcelona
Cimeira	Cimeira do Rio de Janeiro
Jogos	Jogos Olímpicos de 2000
Seminário	Seminário de Gestão em Saúde
Jornada	Jornadas da Juventude

Tabela 4.15: Exemplos de estrutura utilizadas no reconhecimento de entidades do tipo acontecimento.

Estrutura	Exemplo
Batalha	Batalha de Aljubarrota
Revolução	Revolução Francesa
Guerra	Guerra dos Cem Anos
Dia	Dia da Mãe

Tabela 4.16: Exemplos de estrutura utilizadas no reconhecimento de entidades do tipo acontecimento.

(geralmente em numeração romana), como por exemplo “3º Simpósio Sobre o Tabagismo” , “II Guerra Mundial” , “12º Jornada da Liga de Honra” .

A estrutura “País x País” ou “Equipa x Equipa” é comum na denominação de jogos de futebol ou outras modalidades desportivas, pelo que é classificada através de uma regra e classificada como acontecimento.

Entidades que se referem a partidas desportivas com o formato “Equipa-Equipa” não são, no entanto, identificadas, já que a ferramenta XIP considera toda a expressão como uma única palavra.

São também identificadas algumas entidades contendo palavras em inglês que são comuns, tais como “Show” , “Party” ou “Rave” .

De modo idêntico à identificação de pessoas e organizações, usa-se o contexto particular do uso de iniciais para designar uma acontecimento, pelo que se optou por considerar que se uma entidade do tipo acontecimento está seguida de uma expressão abreviada entre parênteses, então essa expressão também será uma organização e referir-se-á à mesma entidade (e.g., “Feira Internacional do Móvel (F.I.M.)”).

4.5 Outros

Além dos léxicos referidos nas secções anteriores (4.1 a 4.4), foi também usada como auxiliar no reconhecimento uma lista de 2440 nacionalidades e gentílicos (e.g. “americano” , “saudita” , “lisboeta”) e uma lista de 639 profissões, num total de 3079 entradas lexicais.

5 Avaliação e Resultados

Neste capítulo são descritos os procedimentos utilizados na avaliação do sistema de reconhecimento de entidades mencionadas descrito neste documento (secção 5.1) e comparados os resultados obtidos nessa avaliação com o desempenho de outros sistemas de reconhecimento de entidades mencionadas para a língua portuguesa relativamente às categorias e subcategorias consideradas (secção 5.2).

5.1 Procedimentos

De modo a avaliar as tarefas de identificação e classificação de entidades mencionadas, recorreu-se ao sistema de avaliação disponibilizado pelo fórum de avaliação HAREM, que permite avaliar a correcção dos resultados através do uso de uma colecção dourada previamente anotada manualmente.

O texto original, de acordo com as regras de etiquetagem do HAREM, deve conter cada EM rotulada por uma etiqueta de abertura e uma etiqueta de fecho, semelhante às etiquetas usadas em XML. Na etiqueta de abertura estão contidas a categoria e tipo atribuídos. A etiqueta de fecho contém apenas a categoria.

Os tipos (ou subcategorias) são colocadas entre aspas e tanto estes como a categoria de topo devem estar em maiúsculas, não conter acentos, cedilhas ou espaçamento. Não devem de igual modo existir espaços entre a entidade mencionada e as etiquetas que a rodeiam, e caracteres como aspas ou parênteses não devem ser incluídos na parte rotulada. Apresenta-se um exemplo de uma rotulação segundo o formato do HAREM na figura 5.1.

Utilizou-se a saída *-indent* do XIP, que apresenta a estrutura da figura 5.2 e que contém informação sobre os traços atribuídos a cada uma das palavras no texto, de modo a realizar a etiquetação definida anteriormente.

O <PESSOA TIPO="INDIVIDUAL">João Silva</PESSOA> é médico.

Mora em <LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>.

Figura 5.1: Exemplos de etiquetação de EMs de acordo com o HAREM.

```
1  TOP(1-5)+[cat:0]
2   NP(1-2)-[np:+,noun:+,!start:+,first:+]
3    ART(1-1)+[toutmaj:+,maj:+,def:+,masc:+,sg:+,art:+,hmmselection:+,!start:+,first:+]
4      O(1-1)+[toutmaj:+,maj:+,def:+,masc:+,sg:+,art:+,hmmselection:+,!start:+,first:+]
5    NOUN(2-2)+[maj:+,people:+,proper:+,masc:+,fem:+,sg:+,noun:+,hmmselection:+,last:+,first:+]
6      João(2-2)+[maj:+,people:+,proper:+,masc:+,fem:+,sg:+,noun:+,hmmselection:+,last:+,first:+]
7   VF(3-3)-[fin:+,verb:+]
8    VERB(3-3)+[cop:+,pres:+,ind:+,3p:+,sg:+,verb:+,hmmselection:+,last:+,first:+]
9      vive(3-3)+[cop:+,pres:+,ind:+,3p:+,sg:+,verb:+,hmmselection:+,last:+,first:+]
10  PP(4-5)-[pp:+,noun:+,!end:+,last:+]
11   PREP(4-4)+[sg:+,prep:+,hmmselection:+,first:+]
12     em(4-4)+[sg:+,prep:+,hmmselection:+,first:+]
13   NOUN(5-5)+[maj:+,proper:+,capital:+,city:+,location:+,masc:+,fem:+,sg:+,noun:+,hmmselection:+,!end:+,last:+]
14     Lisboa(5-5)+[maj:+,proper:+,capital:+,city:+,location:+,masc:+,fem:+,sg:+,noun:+,hmmselection:+,!end:+,last:+]
```

Figura 5.2: Ficheiro -indent do XIP após processamento da frase ” O João vive em Lisboa” .

```
1  <DOC>
2  <DOCID>HAREM-XXX-00000</DOCID>
3  <GENERO>Jornalístico</GENERO>
4  <ORIGEM>PT</ORIGEM>
5  <TEXTO>
6
7  ...
8
9
10 </TEXTO>
11 </DOC>
```

Figura 5.3: Exemplo da estrutura de um documento da colecção do HAREM.

Foi criado um programa Perl que recebe a saída *-indent* do XIP e um ficheiro contendo a lista de traços a classificar e que devolve o texto etiquetado segundo o formato do HAREM.

Contudo, durante o processamento do texto são feitas alterações à sua estrutura, nomeadamente no tratamento de contracções (e.g. “no” , “da”) e de clíticos (e.g. “deu-lhe” , “viu-a”), que são separados nos seus componentes individuais (e.g. “no” →“em + o”). Como tal, o resultado obtido é passado a um outro programa em Perl que realiza de novo a contracção das preposições dentro dos sintagmas nominais que compõem as entidades mencionadas. Fora das entidades, o texto resultante pode ser diferente do original, algo que não afecta a avaliação da tarefa de identificação e classificação.

Os documentos disponibilizados na colecção do HAREM têm a estrutura da figura 5.3. As etiquetas $< DOC >$ e $< /DOC >$ (linhas 1 e 11) delimitam cada documento individual na colecção, as etiquetas $< DOCID >$ e $< /DOCID >$ (linha 2) delimitam o código único de identificação de cada documento, as etiquetas $< GENERO >$ e $< /GENERO >$ (linha 3) delimitam a definição do género do texto em

Categoria	Percentagem de textos
Jornalístico	33,4%
Web	33,3%
CorreioElectrónico	12,1%
Literário	5,6%
Entrevista	5,2%
Expositivo	5%
Político	5%
Técnico	1,2%

Tabela 5.1: Distribuição dos vários géneros de texto na colecção do HAREM.

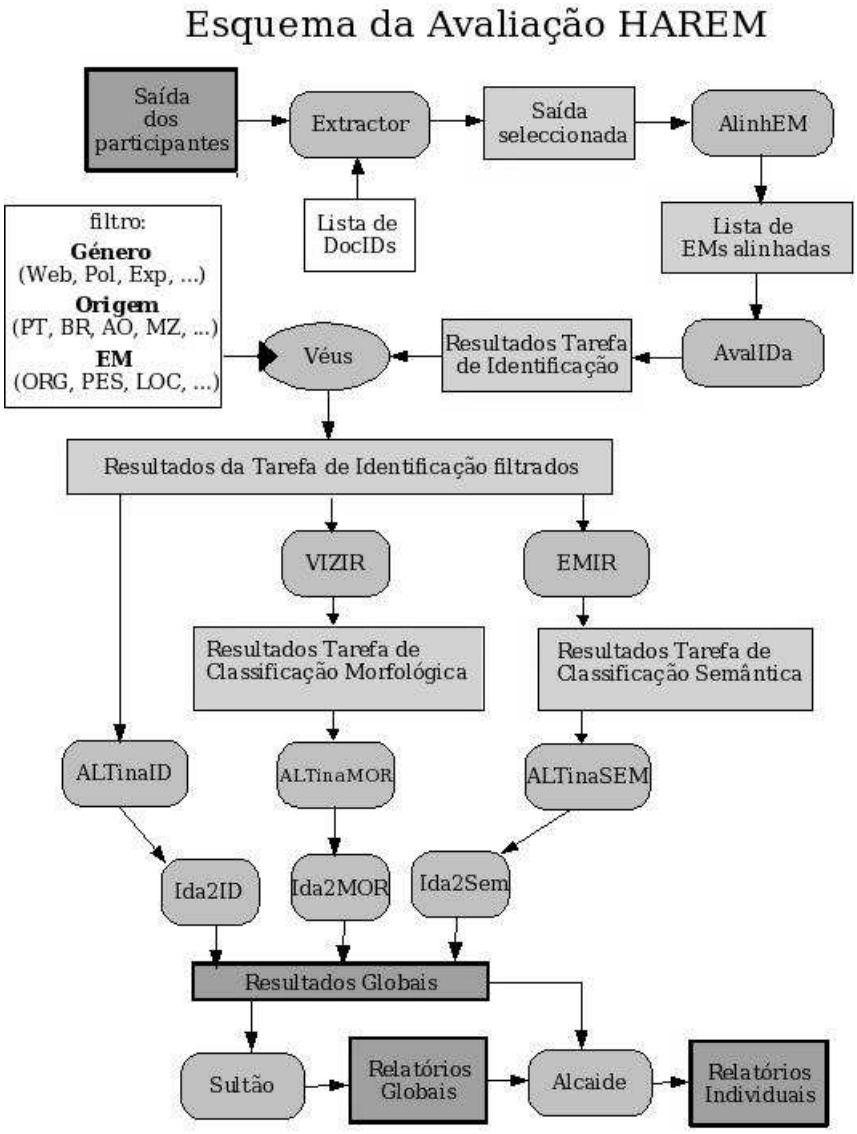

Figura 5.4: Diagrama de avaliação do HAREM.

questão (texto jornalístico, entrevista, literatura, etc.), as etiquetas $< ORIGEM >$ e $< /ORIGEM >$ (linha 4) definem a origem do texto (Portugal, Brasil, etc.) e por fim as etiquetas $< TEXTO >$ e $< /TEXTO >$ (linhas 5 e 10) delimitam o texto que deve ser marcado no que diz respeito às entidades mencionadas nele presentes.

A colecção do HAREM contém 1202 textos em duas variantes (português europeu e português brasileiro) e oito géneros, cuja distribuição se pode observar na tabela 5.1. A colecção dourada i.e., a colecção de textos anotados manualmente, corresponde a cerca de 11% dos textos presentes na colecção do HAREM e as proporções de textos de cada género não são equivalentes àquelas da colecção tomada no seu todo.

O esquema de avaliação do HAREM está apresentado no diagrama da figura 5.4 (*HAREM - Ava-*

liação de Reconhecimento de Entidades Mencionadas, n.d.). A saída etiquetada é passada através de um programa em Perl que selecciona dos textos processados aqueles que fazem parte da colecção dourada, através da análise dos identificadores dos documentos.

Esta saída seleccionada é então passada a um programa em Java (AlinhEM) que faz o alinhamento das entidades mencionadas, i.e., compara o texto da colecção dourada com o da saída e devolve um ficheiro contendo pares de entidades mencionadas. Quando não existe correspondência é usado o campo "null" . Além disso, o alinhador também etiqueta cada aparição de uma entidade mencionada com um número, o que permite fazer a distinção entre entidades iguais que aparecem mais do que uma vez num texto.

Em seguida esta saída é utilizada por outro programa (AvalIDa), que avalia a correcta identificação das entidades mencionadas. O programa verifica se as fronteiras estão correctas e se todas as palavras que foram etiquetadas correspondem realmente a uma entidade, sem olhar à sua classificação semântica (categorias e tipos). O programa marca os pares de entidades alinhadas relativamente à identificação com os valores: em falta, correcta, espúria, parcialmente correcta por defeito ou parcialmente correcta por excesso.

Posteriormente a saída do AvalIDa passa por um conjunto de filtros (Véus), onde é possível escolher as categorias e subcategorias que se quer considerar, assim como ignorar géneros ou origens de texto específicas, sendo que todas as outras entidades mencionadas são ignoradas e não consideradas na avaliação posterior. Neste trabalho, só são consideradas as entidades do tipo local, pessoa, organização e acontecimento, e respectivas subcategorias.

Seguidamente é efectuada a tarefa de avaliação semântica (Emir). A saída é utilizada para verificar a classicação em relação às categorias e subcategorias (tipos), apenas das entidades mencionadas que foram delimitadas correctamente.

A saída resultante dos filtros pode também ser utilizada para realizar a avaliação morfológica (Vizir), algo que não é considerado no âmbito deste trabalho.

Como existem casos na colecção dourada em que a etiquetação manual não foi unânime ou é ambígua e existem alternativas na delimitação, a saída do AvalIDA (avaliação da identificação) e do Emir (avaliação semântica) passa por dois programas de escolha de alternativas (AltinaID e AltinaSEM), que escolhem das alternativas (se estas existirem) aquelas que permitem obter um melhor resultado para o sistema avaliado.

As saídas dos programas de escolha de alternativas são em seguida passadas aos programas de cálculo de resultados individuais de modo a calcular os resultados finais da avaliação de identificação e classificação (Ida2ID e Ida2Sem).

Os resultados individuais de cada uma das tarefas são então combinados entre si de modo a se obter um relatório HTML de resultados globais (ferramentas Sultão e Alcaide).

5.1.1 Medidas

Nesta subsecção são apresentadas as medidas usadas na avaliação da tarefa de identificação e classificação de entidades mencionadas.

No que diz respeito à tarefa de identificação, esta tem como objectivo medir a eficiência do sistema em delimitar as entidades de forma correcta, em comparação com as entidades previamente anotadas existentes na colecção dourada.

Para esta avaliação é importante a noção de átomo, que se define como sendo qualquer sequência de letras ou dígitos individuais.

O avaliador da tarefa de identificação (AvalIDa) atribui as seguintes classificações:

1. Correcto - quando o átomo inicial e o átomo final da entidade mencionada são iguais na saída do sistema e na colecção dourada e o número total de átomos é igual entre si;
2. Parcialmente correcto por defeito - quando pelo menos um átomo da saída do sistema corresponde a um átomo de uma entidade mencionada na colecção dourada e o número total de átomos da entidade mencionada na saída do sistema é menor do que o número de átomos respectivos na colecção dourada;
3. Parcialmente correcto por excesso - quando pelo menos um átomo da saída do sistema corresponde a um átomo de uma entidade mencionada na colecção dourada e o número de átomos na entidade mencionada na saída do sistema é maior ou igual ao número de átomos respectivos na colecção dourada;
4. Em falta - quando o sistema falha a detecção correcta de qualquer átomo de uma certa entidade mencionada presente na colecção dourada;
5. Espúrio - quando foi delimitada uma alegada entidade mencionada que não consta na colecção dourada, quer parcial ou totalmente.

Enquanto que às entidades mencionadas correctamente identificadas é atribuída a pontuação 1 e aos espúrios e entidades em falta a pontuação 0, as entidades mencionadas identificadas como parcialmente correctas são pontuadas segundo a fórmula 5.1.

$$0,5 \times (\frac{nc}{nd}) \tag{5.1}$$

onde nc representa a cardinalidade da intersecção dos átomos das duas entidades e nd a cardinalidade da reunião dos átomos das duas entidades.

A avaliação da classificação semântica tem como objectivo medir a capacidade do sistema em conseguir classificar uma entidade mencionada tendo em conta a hierarquia de categorias e tipos definidos pelo HAREM. A classificação semântica pode ser avaliada em quatro modalidades:

1. Classificação semântica por categorias - apenas é considerada a categoria na etiqueta;
2. Classificação semântica por tipo - apenas são avaliadas as entidades cuja categoria foi classificada correctamente, e apenas em relação ao seu tipo;
3. Classificação semântica combinada - é avaliada tanto a correcção das categorias como dos tipos da entidade mencionada, através de uma pontuação que combina as duas;
4. Classificação semântica plana - os pares categoria-tipo são avaliados atomicamente, considerando-se apenas como correctas as entidades que tenham categoria e tipo correctos.

No caso da classificação semântica combinada, a pontuação a atribuir é:

i) 0, se a categoria não estiver correcta;

ii) 1, se a categoria estiver correcta, mas o tipo estiver errado;

iii) $1+(1-\frac{nc}{n})-\frac{ne}{n}$, se a categoria e pelo menos um dos tipos estiver correcto, em que nc é o número de tipos correctamente identificados, ne o número de tipos espúrios e n o número de tipos possíveis nessa categoria. No caso de locais, pessoas, organizações e acontecimentos, o valor de n é 5, 6, 4 e 3, respectivamente, pelo que os valores máximos da pontuação são 1.8, 1.833, 1.75 e 1.666.

A precisão é uma medida da qualidade da resposta do sistema que mede a proporção de respostas correctas em relação a todas as respostas dadas pelo sistema. Na tarefa de identificação a precisão mede a relação entre as entidades correctas e parcialmente correctas de todas as entidades identificadas pelo sistema, e é calculada de acordo com a fórmula 5.2.

$$Precisão = \frac{(Num\ EMs\ Correctas + x)}{Num\ EMs\ Identificadas} \tag{5.2}$$

em que x é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada.

Em relação à classificação semântica, há que ter em conta as quatro modalidades descritas anteriomente: classicação por categorias, classificação por tipo, classificação semântica combinada e classificação semântica plana.

No que diz respeito à classificação por categorias, o cálculo da precisão está definido na fórmula 5.3.

$$Precisão = \frac{(Num\ EMs\ com\ Identificacão\ e\ Categoria\ Correcta + y)}{Num\ EMs\ Classificadas} \tag{5.3}$$

em que y é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com a categoria correcta.

Em relação à modalidade de classificação por tipos, a classificação é, por definição, sempre relativa, e dada pela fórmula 5.4.

$$Precisão = \frac{(Num\ EMs\ com\ Identificacão,\ Categoria\ e\ Tipo\ Correctos + z)}{Num\ EMs\ Total\ ou\ Parcialmente\ Identif.\ e\ Classif.\ na\ Categoria\ Correcta} \tag{5.4}$$

em que z é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com categoria e tipo correctos.

Para a classificação semântica combinada, a precisão mede o grau de sucesso de acordo com a classificação máxima (calculada assumindo que todas as categorias e tipos propostos pelo sistema estão correctos) e é dada pela fórmula 5.5.

$$Precisão = \frac{ValorMedidaSemânticaSistema}{ValorMáximo\ Medida\ Semântica\ p/\ Saida\ do\ Sistema} \tag{5.5}$$

No caso da classificação plana, a precisão é dada pela fórmula 5.6.

$$Precisão = \frac{((Num\ EMs\ com\ Identificacão,\ Categoria\ e\ Tipo\ Correctos + z)}{Num\ EMs\ Classificadas} \tag{5.6}$$

em que z é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com categoria e tipo correctos.

A abrangência (ou cobertura) mede a percentagem de respostas correctas que o sistema conseguiu recuperar. Na tarefa de identificação a abrangência mede a quantidade de entidades mencionadas da colecção dourada que foram identificadas e é dada pela fórmula 5.7.

$$Abrangência = \frac{(Num\ EMs\ Correctas + x)}{Num\ EMs\ ColeccãoDourada} \tag{5.7}$$

onde x é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada

De modo similar ao cálculo da precisão, a abrangência para a classificação semântica é definida diferentemente para cada uma das modalidades de avaliação. O cálculo da abrangência no caso da avaliação por categorias é dado pelas fórmula 5.8.

$$Abrangência = \frac{(Num\ EMs\ com\ Identificacão\ e\ Categoria\ Correctas + y)}{Num\ EMs\ na\ ColeccãoDourada} \tag{5.8}$$

em que y é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com categoria correcta.

No caso da classificação por tipos a abrangência é dada pela fórmula 5.9.

$$Abrangência = \frac{(Num\ EMs\ com\ Identificacão,\ Categoria\ e\ Tipo\ Correctos + z)}{Num\ EMs\ Correctamente\ Classificadas\ na\ Categoria} \tag{5.9}$$

em que z é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com categoria e tipo correctos.

Na avaliação da classificação semântica combinada, a abrangência mede o nível de cobertura de acordo com a classificação máxima (se tanto as categorias como os tipos enviados estiverem correctos) e é dada pela fórmulas 5.10.

$$Abrangência = \frac{ValorMedidaSemânticaSistema}{ValorMáximo\ Medida\ na\ ColeccãoDourada} \tag{5.10}$$

Por fim, relativamente à classificação plana, o valor da abrangência é dado pela fórmula 5.11.

$$Abrangência = \frac{(Num\ EMs\ com\ Identificacão,\ Categoria\ e\ Tipo\ Correctos + z)}{Num\ EMs\ na\ ColeccãoDourada} \tag{5.11}$$

em que z é o somatório dos valores obtidos para cada entidade mencionada parcialmente identificada e com categoria e tipo correctos.

A medida-F combina as medidas de precisão e de abrangência para cada tarefa, de acordo com a fórmula 5.12.

$$Medida - F = \frac{(2 \times Precisão \times Abrangência)}{(Precisão + Abrangência)} \tag{5.12}$$

A sobregeração mede o excesso de resultados espúrios que um sistema produz, ou seja, quantas vezes produz resultados errados. Relativamente à tarefa de identificação, a sobregeração mede quantas

entidades mencionada identificadas pelo sistema não existem na colecção dourada e é calculada através da fórmula 5.13.

$$Sobregeracão = \frac{Num\ EMs\ Espúrias}{Num\ EMs\ Identificadas} \tag{5.13}$$

A sobregeração na classificação semântica mede o número de entidades mencionadas com uma classificação semântica espúria, em comparação com a colecção dourada. No caso da avaliação por categorias, a sobregeração é dada pela fórmulas 5.14.

$$Sobregeracão = \frac{Num\ EMs\ Espúrias\ na\ Categoria}{Num\ EMs\ Classificadas\ na\ Categoria} \tag{5.14}$$

Em relação à avaliação por tipos, a sobregeração é dada pela fórmula 5.15.

$$Sobregeracão = \frac{Num\ EMs\ Espúrias\ no\ Tipo}{Num\ EMs\ Identificadas\ e\ Classif.\ na\ Categoria\ e\ Tipo} \tag{5.15}$$

No caso da classificação plana, a sobregeração é calculada segundo as fórmula 5.16.

$$Sobregeracão = \frac{Num\ EMs\ Espúrias\ na\ Categoria\ ou\ Tipo}{Num\ EMs\ Classificadas\ na\ Categoria\ e\ Tipo} \tag{5.16}$$

A subgeração é uma medida de quanto faltou ao sistema analisar, dada a solução conhecida, i.e., a colecção dourada.

A subgeração, relativamente à tarefa de identificação, mede a quantidade de entidades mencionadas que existem na colecção dourada que não foram identificadas pelo sistema e é calculada através da fórmula 5.17.

$$Subgeracão = \frac{Num\ EMs\ em\ Falta}{Num\ EMs\ ColeccãoDourada} \tag{5.17}$$

A subgeração na classificação semântica mede as classificações semânticas em falta. No caso da avaliação por categorias, a subgeração é calculada de acordo com as fórmula 5.18.

$$Subgeracão = \frac{Num\ EMs\ em\ Falta\ na\ Categoria}{Num\ EMs\ Classificadas\ na\ Categoria} \tag{5.18}$$

No caso da avaliação por tipos, a subgeração é dada pela fórmula 5.19.

$$Subgeracão = \frac{Num\ EMs\ em\ Falta\ no\ Tipo}{Num\ EMs\ Parcial\ ou\ Totalmente\ Identif.\ com\ Tipo\ na\ ColeccãoDourada} \quad (5.19)$$

Por último, no que diz respeito à avaliação plana, a subgeração é calculada de acordo com a fórmula 5.20.

$$Subgeracão = \frac{Num\ EMs\ em\ Falta\ no\ Tipo}{Num\ EMs\ Classificadas\ na\ Categoria\ na\ ColeccãoDourada} \quad (5.20)$$

5.2 Resultados

O reconhecimento de entidades mencionadas subdivide-se em duas tarefas distintas: a identificação (ou delimitação) das entidades e a classificação das mesmas. Tendo isto em consideração, serão apresentados os resultados de cada uma destas duas tarefas independentemente.

É importante salientar que, no âmbito do HAREM, as saídas de cada sistema são anónimas, pelo que serão referidas através do *alias* que lhes é atribuído na avaliação automática. O sistema descrito neste documento, por outro lado, será referido como *l2f*.

Embora os resultados do sistema descrito neste documento e o dos restantes participantes no HAREM sejam apresentados conjuntamente para efeitos de comparação, é de salientar que a avaliação não foi realizada simultaneamente, sendo que os resultados dos sistemas participantes no HAREM são os valores oficiais da edição de 2005 deste fórum, enquanto que os valores obtidos para o sistem l2f são de uma avaliação posterior em Agosto de 2007, como descrito na secção 5.1.

Relativamente à tarefa de identificação, os resultados para a categoria *local* encontram-se discriminados na tabela 5.2.

Em termos de medida-f, o sistema classifica-se em quarto lugar (0.6754), sendo o segundo sistema com a melhor precisão (79.82%) e o quinto no que diz respeito à abrangência (58.54%). De notar, no entanto, que o sistema com a melhor precisão (92%) tem uma cobertura de apenas 2.15%.

Em relação à categoria *pessoa*, os resultados encontram-se discriminados na tabela 5.3. Em termos de medida-f, o sistema classifica-se em segundo lugar (0.6118), sendo que de todos os sistemas é aquele que apresenta a melhor precisão (74.91%) e a segunda melhor abrangência (51.69%), embora com uma diferença de mais de 20% em relação ao primeiro classificado.

Os resultados da tarefa de identificação para a categoria *organização* estão descritos na tabela 5.4. O sistema apresenta a melhor medida-f (0.5979), a terceira melhor precisão (71.3%) e a segunda melhor abrangência (51.48%) de todos os sistemas avaliados.

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
qatar	68.03	73.91	0.7085	0.2859	0.2290
kuwait	67.78	71.37	0.6953	0.2922	0.2612
meca	74.11	63.87	0.6861	0.2097	0.3293
l2f	**79.82**	**58.54**	**0.6754**	**0.1689**	**0.3882**
bagdad	66.48	65.42	0.6595	0.2754	0.2942
abudhabi	71.79	44.29	0.5478	0.2552	0.5422
rabat	71.25	43.95	0.5437	0.255	0.5422
oman	74.55	42.73	0.5432	0.2034	0.5450
amã	36.89	35.03	0.3594	0.5013	0.5408
cairo	38.53	32.28	0.3512	0.5676	0.6321
casablanca	34.27	30.94	0.3252	0.6059	0.6434
nicósia	39.45	25.28	0.3082	0.4432	0.6472
doha	39.30	25.12	0.3065	0.4443	0.6488
damasco	92.07	2.15	0.04195	0	0.9775

Tabela 5.2: Resultados da tarefa de identificação de locais (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
túnis	58.75	72.72	0.6499	0.3433	0.1988
l2f	**74.91**	**51.69**	**0.6118**	**0.1984**	**0.4490**
jerusalém	66.65	53.78	0.5953	0.2757	0.4137
cairo	64.97	53.69	0.5879	0.2834	0.4049
ancara	64.27	51.35	0.5709	0.2503	0.3971
kuwait	59.23	35.80	0.4463	0.2876	0.5664
abudhabi	70.08	29.84	0.4186	0.2087	0.6621
bahrein	70.08	29.84	0.4186	0.2087	0.6621
teerão	24.62	26.52	0.2553	0.6945	0.6719
qatar	18.82	25.45	0.2164	0.7617	0.6650
argel	0	0	0	1.000	1.000

Tabela 5.3: Resultados da tarefa de identificação de pessoas (ordenados por medida-f).

Os resultados da tarefa de identificação para a categoria *acontecimento* estão descritos na tabela 5.5. O sistema apresenta a segunda melhor medida-f (0.4780), a segunda melhor precisão (61.02%) e a terceira melhor abrangência. Em termos globais, esta é a categoria que apresenta os piores resultados, com uma medida-f abaixo dos 0.5, embora seja o segundo melhor classificado em relação aos outros sistemas participantes.

Visto que cada sistema concorrente ao HAREM não precisa necessariamente de concorrer a todas as categorias ou a todos os tipos dentro de cada categoria, torna-se difícil efectuar uma comparação global dos resultados. No entanto, pode-se tomar em consideração os resultados selectivos globais, i.e., as medidas alcançadas por cada sistema na totalidade de categorias que se propôs identificar. Isto poderá significar, consoante o sistema, desde apenas uma categoria ao total das categorias do HAREM. Neste cenário, o sistema classifica-se em quinto lugar, com uma medida-f de 0.6978, apresentando a melhor

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
l2f	**71.30**	**51.48**	**0.5979**	**0.2210**	**0.4347**
mascate	51.01	62.72	0.5626	0.4160	0.2923
riad	62.38	44.50	0.5195	0.2933	0.5084
marraquexe	60.89	42.48	0.5005	0.3118	0.5293
abudhabi	67.38	30.98	0.4245	0.2091	0.6259
oman	71.46	29.33	0.4159	0.2030	0.6708
gaza	71.46	29.33	0.4159	0.2030	0.6708
eritreia	76.03	19.27	0.3074	0.1488	0.7812
asmara	28.06	30.32	0.2915	0.6122	0.5967
qatar	25.18	34.19	0.2900	0.6456	0.5377
túnis	67.72	4.751	0.08879	0.2687	0.9476

Tabela 5.4: Resultados da tarefa de identificação de organizações (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
rabat	50.76	46.61	0.4859	0.4356	0.4818
l2f	**61.02**	**39.28**	**0.4780**	**0.3191**	**0.5479**
eritreia	47.03	39.70	0.4305	0.4674	0.5505
ancara	50.12	28.05	0.3597	0.4262	0.6789
marraquexe	26.04	43.82	0.3267	0.6473	0.4309
argel	87.50	6.422	0.1197	0.1250	0.9358
meca	0	0	0	1.000	1.000

Tabela 5.5: Resultados da tarefa de identificação de acontecimentos (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
eritreia	78.50	82.84	0.8061	0.07913	0.07329
asmara	77.15	84.35	0.8059	0.09134	0.03575
bahrein	76.85	83.56	0.8006	0.08966	0.04035
damasco	77.43	69.57	0.7329	0.09524	0.2079
l2f	**83.03**	**60.12**	**0.6978**	**0.0844**	**0.3384**
riad	76.31	58.40	0.6616	0.09725	0.3157
ancara	59.45	64.39	0.6182	0.2018	0.1607
jerusalém	56.95	64.39	0.6044	0.2353	0.1607
doha	57.21	63.51	0.6020	0.2315	0.1707
oman	58.57	52.12	0.5516	0.3408	0.4240
dakar	58.44	51.93	0.5499	0.3413	0.4240
tripoli	47.32	54.50	0.5066	0.1119	0.1687
rabat	36.89	35.03	0.3594	0.5013	0.5408
kuwait	39.45	25.28	0.3082	0.4432	0.6472
qatar	57.40	17.72	0.2708	0.2330	0.7866
iémen	47.12	10.98	0.1781	0.1596	0.8101

Tabela 5.6: Resultados da tarefa de identificação relativamente ao conjunto de categorias que cada sistema se propôs identificar (ordenados por medida-f).

precisão (83.03%) e a nona melhor abrangência (60.12%). Pode-se afirmar que, no geral, o sistema não identifica tantas entidades quanto vários dos outros sistemas, mas aquelas que identifica estão mais frequentemente correctas.

É importante referir o facto dos valores da precisão e da abrangência na tarefa de avaliação combinada serem superiores à soma dos valores obtidos nas tarefas individuais em cada categoria. Isto deve-se ao facto de se considerar como correcta a entidade mencionada de maior extensão, pelo que uma expressão como *presidente do Brasil* só é correctamente identificada como entidade mencionada num cenário que inclua a categoria *pessoa*. Num cenário contendo apenas a categoria *local*, somente a entidade mencionada *Brasil* é identificada e classificada, o que é considerado um espúrio. Assim, o facto de reconhecer a categoria *pessoa* e *local* simultaneamente aumenta a precisão da categoria local e consequentemente dos resultados globais. Por outro lado, algumas entidades na colecção dourada têm reconhecimentos alternativos. Por exemplo, "Benfica-Sporting" poderá ser classificado como um acontecimento ou alternativamente como duas organizações. Se a opção tomada for a de reconhecer a entidade como um todo, então ao correr a avaliação somente sobre a categoria organização ter-se-ão dois espúrios. Correndo a avaliação global, mais uma vez este problema é eliminado, aumentando a precisão.

Cerca de metade dos textos da colecção dourada HAREM são textos de origem brasileira, o que tem influência nos resultados do sistema a vários níveis. Por um lado, os léxicos de locais, siglas, marcas ou nomes brasileiros que foram incluídos são bastante reduzidos em comparação com aqueles de origem portuguesa. Por outro lado, a ortografia, sintaxe, colocação pronomial e vocabulário do português brasileiro podem impedir que um determinado contexto seja detectado. Por exemplo, embora só a forma *fui ao Brasil* seja aceite em português europeu, a variação com a preposição "em" também ocorre no português do Brasil (e.g. *fui no Brasil*). Realizando a avaliação apenas sobre textos cuja origem é Portugal, os resultados globais apresentam uma melhoria de cerca de 0.04 medida-f, 5% na abrangência e 3% na precisão.

Em relação à tarefa de classificação, i.e., a atribuição de categoria e tipo a cada entidade previamente identificada, a avaliação pode efectuar-se segundo quatro cenários distintos, como descrito na secção 5.1.1.

Os resultados obtidos para a tarefa de classificação semântica por categorias em relação à categoria local estão discriminados na tabela 5.7.

O sistema posicionou-se em quarto lugar em termos de medida-f (0.6873), apresentando a segunda melhor precisão (81%) e a quinta melhor abrangência (54.81%).

Os resultados obtidos para a tarefa de classificação semântica combinada em relação à mesma categoria estão discriminados na tabela 5.8.

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
cairo	69.00	74.84	0.7180	0.2852	0.2286
riad	68.59	72.10	0.7030	0.2915	0.2608
abudhabi	75.15	64.77	0.6957	0.2097	0.3293
l2f	**81.00**	**59.69**	**0.6873**	**0.1664**	**0.3835**
damasco	68.22	67.13	0.6767	0.2754	0.2942
sana	76.02	43.57	0.5540	0.2034	0.5450
qatar	72.49	44.72	0.5531	0.2552	0.5422
dakar	72.06	44.45	0.5498	0.2552	0.5422
kuwait	40.17	38.08	0.3910	0.5004	0.5399
bahrein	39.95	33.47	0.3642	0.5676	0.6321
jerusalém	35.77	32.26	0.3393	0.6050	0.6429
marraquexe	43.45	27.80	0.3391	0.4419	0.6462
iémen	43.31	27.64	0.3375	0.4431	0.6478
asmara	94.48	2.203	0.04305	0	0.9775

Tabela 5.7: Resultados da tarefa de classificação semântica por categorias para a categoria local (ordenados por medida-f).

Saída	Precisão Máxima do Sistema (%)	Abrangência Máxima na CD(%)	Medida-F
cairo	64.09	69.83	0.6683
riad	63.45	66.85	0.6511
damasco	65.52	64.58	0.6505
abudhabi	69.37	59.90	0.6429
l2f	**77.31**	**54.81**	**0.6414**
qatar	70.65	43.64	0.5395
dakar	69.14	42.71	0.5280
sana	71.87	36.70	0.4859
jerusalém	32.81	29.78	0.3122
kuwait	36.37	26.81	0.3087
bahrein	40.17	19.27	0.2604
marraquexe	39.04	19.07	0.2563
iémen	38.96	18.98	0.2553
asmara	94.48	2.203	0.04305

Tabela 5.8: Resultados da tarefa de classificação semântica combinada para a categoria local (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
damasco	62.39	61.40	0.6189	0.3591	0.3693
cairo	58.59	63.55	0.6097	0.4031	0.3493
riad	57.36	60.29	0.5879	0.4178	0.3864
abudhabi	62.44	53.81	0.5780	0.3579	0.4466
l2f	**66.10**	**48.71**	**0.5609**	**0.2217**	**0.4666**
qatar	68.56	42.29	0.5232	0.3099	0.5735
dakar	65.70	40.53	0.5013	0.3385	0.5912
sana	49.04	28.11	0.3573	0.3310	0.7122
jerusalém	29.57	26.67	0.2804	0.6957	0.7207
kuwait	13.41	12.72	0.1306	0.5762	0.8466
iémen	12.79	8.162	0.09965	0.5282	0.8954
marraquexe	12.76	8.162	0.09956	0.5281	0.8954
asmara	94.48	2.203	0.04305	0	0.9775
bahrein	1.809	1.516	0.01649	0.5676	0.9807

Tabela 5.9: Resultados da tarefa de classificação semântica plana para a categoria local (ordenados por medida-f).

O sistema posicionou-se em quarto lugar em termos de medida-f (0.6414), apresentando a segunda melhor precisão (77.71%) e a quinta melhor abrangência (48.71%).

Os resultados obtidos para a tarefa de classificação semântica plana em relação à mesma categoria estão representados na tabela 5.9.

O sistema posicionou-se em quinto lugar em termos de medida-f (0.5609), apresentando a segunda melhor precisão (66.10%) e a quinta melhor abrangência (59.69%).

Os resultados obtidos para a tarefa de classificação semântica combinada por tipo, i.e., considerando apenas as entidades cuja categoria foi classificada correctamente em relação à categoria local, estão discriminados na tabela 5.10.

O sistema posicionou-se em quinto lugar em termos de medida-f (0.8301), apresentando a quarta melhor precisão (87.5%) e a oitava melhor abrangência (79.01%). Isto indica que embora o sistema tenha um bom desempenho na classificação da categoria local, erra mais frequentemente no tipo da mesma que a maioria dos sistemas concorrentes.

Em relação à categoria pessoa, os resultados obtidos para a tarefa de classificação semântica por categorias estão apresentados na tabela 5.11.

O sistema posicionou-se em segundo lugar em termos de medida-f (0.6461), apresentando a melhor precisão (76.76%) e a terceira melhor abrangência (55.78%).

Os resultados obtidos para a tarefa de classificação semântica combinada em relação à mesma categoria estão representados na tabela 5.12.

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
nicósia	94.48	97.86	0.9614	0	0
teerão	92.06	92.38	0.9222	0.07343	0.06842
cairo	88.21	88.52	0.8836	0.1119	0.1070
manama	86.12	87.00	0.8656	0.1153	0.1063
l2f	**87.5**	**79.01**	**0.8301**	**0.1102**	**0.1934**
dakar	81.97	82.48	0.8222	0.1649	0.1566
gaza	80.95	81.65	0.8130	0.1783	0.1701
marraquexe	79.00	80.23	0.7961	0.1875	0.1749
meca	75.01	74.84	0.7493	0.2280	0.2162
bahrein	77.87	61.77	0.6889	0.2027	0.3675
riad	57.69	27.74	0.3747	0.3188	0.6655
damasco	53.23	23.28	0.3239	0.3438	0.7016
bagdad	52.95	23.17	0.3224	0.3472	0.7029
bengazi	85.77	4.120	0.07862	0	0.9476

Tabela 5.10: Resultados da tarefa de classificação semântica por tipo para a categoria local (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
luxor	61.07	75.23	0.6742	0.3409	0.1979
l2f	**76.76**	**55.78**	**0.6461**	**0.1970**	**0.42**
oman	69.46	55.61	0.6176	0.2709	0.4105
manama	68.07	55.82	0.6134	0.2786	0.4017
marraquexe	68.47	54.71	0.6082	0.2503	0.3971
iémen	63.69	38.50	0.4799	0.2876	0.5664
teerão	73.64	31.36	0.4398	0.2087	0.6621
amã	73.64	31.36	0.4398	0.2087	0.6621
abudhabi	27.09	28.98	0.2801	0.6902	0.6667
bengazi	20.78	28.09	0.2389	0.7599	0.6637
meca	0	0	0	1.000	1.000

Tabela 5.11: Resultados da tarefa de classificação semântica por categorias para a categoria pessoa (ordenados por medida-f).

Saída	Precisão Máxima do Sistema (%)	Abrangência Máxima na CD(%)	Medida-F
luxor	59.33	73.43	0.6563
l2f	**77.59**	**53.93**	**0.6362**
marraquexe	66.31	53.56	0.5926
manama	64.81	53.64	0.5870
oman	65.29	52.20	0.5801
teerão	72.96	31.21	0.4372
amã	72.96	31.21	0.4372
iémen	57.69	22.10	0.3195
bengazi	19.53	26.67	0.2255
abudhabi	26.98	16.44	0.2043
meca	0	0	0

Tabela 5.12: Resultados da tarefa de classificação semântica combinada para a categoria pessoa (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
luxor	57.85	71.27	0.6386	0.3843	0.2493
l2f	**71.16**	**51.71**	**0.5990**	**0.2217**	**0.4666**
marraquexe	65.32	52.19	0.5802	0.2918	0.4312
manama	62.24	51.03	0.5608	0.3495	0.4618
oman	60.09	48.11	0.5344	0.3543	0.4976
amã	72.88	31.03	0.4353	0.2179	0.6660
teerão	72.88	31.03	0.4353	0.2179	0.6660
bengazi	18.47	24.96	0.2123	0.7967	0.7164
iémen	3.998	2.417	0.03013	0.2956	0.9678
abudhabi	1.307	1.398	0.01351	0.6975	0.9855
meca	0	0	0	1.000	1.000

Tabela 5.13: Resultados da tarefa de classificação semântica plana para a categoria pessoa (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
tripoli	92.10	91.84	0.9197	0.01159	0.01156
iémen	92.10	91.84	0.9197	0.01159	0.01156
l2f	**76.76**	**55.78**	**0.9124**	**0.1970**	**0.4200**
dakar	87.98	89.26	0.8862	0.06235	0.06083
sana	87.12	86.56	0.8684	0.05537	0.05663
eritreia	86.68	86.25	0.8647	0.08896	0.09180
meca	85.51	82.52	0.8399	0.1086	0.1398
cairo	77.62	74.67	0.7612	0.1485	0.1516
luxor	65.13	5.574	0.1027	0.1316	0.9257
marraquexe	62.73	4.294	0.08038	0.3478	0.9554

Tabela 5.14: Resultados da tarefa de classificação semântica por tipo para a categoria pessoa (ordenados por medida-f).

O sistema posicionou-se em segundo lugar em termos de medida-f (0.6362), apresentando a melhor precisão (77.59%) e a segunda melhor abrangência (53.93%).

Os resultados obtidos para a tarefa de classificação semântica plana em relação à mesma categoria estão representados na tabela 5.13.

O sistema posicionou-se em segundo lugar em termos de medida-f (0.5990), apresentando a terceira precisão (71.16%) e a terceira melhor abrangência (51.71%).

Os resultados obtidos para a tarefa de classificação semântica por tipos em relação à mesma categoria estão representados na tabela 5.14.

O sistema posicionou-se em segundo lugar em termos de medida-f (0.9124), apresentando a oitava melhor precisão (76.76%) e a oitava melhor abrangência (55.78%). De modo semelhante à categoria local, o sistema apresenta um desempenho pior do que a maior dos sistemas na classificação do tipo deste grupo de entidades, embora a classificação da própria categoria em si tenha um desempenho

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
l2f	**74.79**	**53.37**	**0.6229**	**0.2124**	**0.4349**
jerusalém	53.30	65.40	0.5873	0.4143	0.2917
eritreia	64.94	46.33	0.5408	0.2933	0.5084
oman	63.30	44.17	0.5203	0.3118	0.5293
bahrein	72.04	33.12	0.4538	0.2091	0.6259
asmara	73.88	30.32	0.4300	0.2030	0.6708
teerão	73.88	30.32	0.4300	0.2030	0.6708
dakar	79.33	20.10	0.3208	0.1488	0.7812
amã	27.63	37.51	0.3182	0.6456	0.5377
sana	30.50	32.95	0.3168	0.6122	0.5967
ancara	69.77	4.895	0.09148	0.2687	0.9476

Tabela 5.15: Resultados da tarefa de classificação semântica por categorias para a categoria organização (ordenados por medida-f).

Saída	Precisão Máxima do Sistema (%)	Abrangência Máxima na CD(%)	Medida-F
l2f	**69.35**	**42.64**	**0.5282**
jerusalém	45.26	55.92	0.5003
eritreia	57.63	41.17	0.4803
oman	55.88	39.04	0.4596
bahrein	66.44	31.59	0.4282
asmara	59.64	24.85	0.3508
teerão	59.64	24.85	0.3508
dakar	71.11	17.32	0.2785
sana	25.50	27.61	0.2651
amã	27.46	21.44	0.2408
ancara	68.75	2.797	0.05376

Tabela 5.16: Resultados da tarefa de classificação semântica combinada para a categoria organização (ordenados por medida-f).

superior à média.

Em relação à categoria organização, resultados obtidos para a tarefa de classificação semântica por categoria estão representados na tabela 5.15.

O sistema posicionou-se em primeiro lugar em termos de medida-f (0.6229), apresentando a melhor precisão (74.79%) e a segunda melhor abrangência (53.37%).

Os resultados obtidos para a tarefa de classificação semântica combinada estão discriminados na tabela 5.16.

O sistema posicionou-se em primeiro lugar em termos de medida-f (0.5282), apresentando a segunda melhor precisão (69.37%) e a segunda melhor abrangência (42.64%).

Os resultados obtidos para a tarefa de classificação semântica plana estão discriminados na tabela 5.17.

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
bahrein	64.26	29.54	0.4048	0.3136	0.6719
eritreia	48.06	34.28	0.4002	0.4751	0.6381
jerusalém	35.26	43.27	0.3886	0.6222	0.5396
oman	46.14	32.19	0.3793	0.4978	0.6579
l2f	**39.72**	**28.35**	**0.3308**	**0.2875**	**0.6921**
teerão	42.78	17.56	0.2489	0.5457	0.8104
asmara	42.78	17.56	0.2489	0.5457	0.8104
dakar	53.69	13.61	0.2171	0.3264	0.8513
sana	18.95	20.48	0.1969	0.7766	0.7618
ancara	0	0	0	0.2687	1.000
amã	0	0	0	0.6456	1.000

Tabela 5.17: Resultados da tarefa de classificação semântica plana para a categoria organização (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
damasco	81.25	78.98	0.8010	0.1322	0.1229
iémen	68.71	69.74	0.6922	0.2600	0.2638
bengazi	67.79	68.39	0.6809	0.2731	0.2733
bahrein	71.00	62.17	0.6629	0.2350	0.3206
l2f	**79.45**	**50.17**	**0.6150**	**0.1502**	**0.4552**
cairo	60.29	61.18	0.6073	0.3547	0.3496
qatar	53.67	53.33	0.5350	0.4299	0.4241
luxor	53.67	53.33	0.5350	0.4299	0.4241
doha	48.87	50.77	0.4980	0.4239	0.4093
rabat	0	0	0	0	1.000
oman	0	0	0	0	1.000

Tabela 5.18: Resultados da tarefa de classificação semântica por tipo para a categoria organização (ordenados por medida-f).

O sistema posicionou-se em quinto lugar em termos de medida-f (0.3308), apresentando a sétima melhor precisão (39.72%) e a quinta melhor abrangência (28.35%).

Os resultados obtidos para a tarefa de classificação semântica por tipo estão discriminados na tabela 5.16.

O sistema posicionou-se em quinto lugar em termos de medida-f (0.6150), apresentando a segunda melhor precisão (79.45%) e a pior abrangência (50.17%) (exceptuando os sistemas que têm medida-f 0). Neste caso, como para as categorias local e pessoa referidas anteriormente, parece ser a classificação de tipos dentro da categoria topo o ponto franco do sistema. A falta de abrangência deve-se em parte, no entanto, à decisão de não classificar a entidade com um tipo quando a sua classificação é incerta.

Finalmente, em relação à categoria acontecimento, os resultados obtidos para a tarefa de classificação semântica por categoria estão apresentados na tabela 5.19.

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
l2f	**74.79**	**53.37**	**0.6229**	**0.2124**	**0.4349**
bagdad	53.01	48.67	0.5074	0.4356	0.4818
sana	49.49	41.78	0.4531	0.4674	0.5505
dakar	52.69	29.49	0.3782	0.4262	0.6789
amã	28.89	48.62	0.3624	0.6473	0.4309
bengazi	87.50	6.422	0.1197	0.1250	0.9358
mascate	0	0	0	1.000	1.000

Tabela 5.19: Resultados da tarefa de classificação semântica por categorias para a categoria acontecimento (ordenados por medida-f).

Saída	Precisão Máxima do Sistema (%)	Abrangência Máxima na CD(%)	Medida-F
l2f	**62.94**	**54.63**	**0.5850**
bagdad	47.26	43.05	0.4506
sana	45.35	37.94	0.4132
dakar	46.74	26.16	0.3354
amã	24.56	41.33	0.3081
bengazi	85.29	5.321	0.1002
mascate	0	0	0

Tabela 5.20: Resultados da tarefa de classificação semântica combinada para a categoria acontecimento (ordenados por medida-f).

O sistema posicionou-se em primeiro lugar em termos de medida-f (0.6229), apresentando a segunda melhor precisão (74.79%) e a melhor abrangência (53.37%)

Os resultados obtidos para a tarefa de classificação semântica combinada estão apresentados na tabela 5.20.

O sistema posicionou-se em primeiro lugar em termos de medida-f (0.5850), apresentando a melhor precisão (62.94%) e a melhor abrangência (54.63%)

Os resultados obtidos para a tarefa de classificação semântica plana estão apresentados na tabela 5.21.

O sistema posicionou-se em primeiro lugar em termos de medida-f (0.5664), apresentando a melhor precisão (61.68%) e a melhor abrangência (52.37%)

Por último, os resultados obtidos para a tarefa de classificação semântica por tipo estão apresentados na tabela 5.22.

O sistema posicionou-se em primeiro lugar em termos de medida-f (0.8811), apresentando a segunda melhor precisão (89.53%) e a melhor abrangência (86.74%). Em oposição às três categorias anteriores (local, pessoa e organização), a classificação de tipos na categoria acontecimento apresenta os melhores resultados de todos os sistemas. Esta diferença poder-se-á explicar em parte devido ao redu-

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
l2f	**61.68**	**52.37**	**0.5664**	**0.3555**	**0.4340**
bagdad	37.71	34.63	0.3610	0.5743	0.6273
sana	38.14	32.19	0.3492	0.5652	0.6514
dakar	37.80	21.16	0.2713	0.6066	0.7798
amã	18.05	30.39	0.2265	0.7778	0.6341
bengazi	50.00	3.670	0.06838	0.1250	0.9633
mascate	0	0	0	1.000	1.000

Tabela 5.21: Resultados da tarefa de classificação semântica plana para a categoria acontecimento (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
l2f	**89.53**	**86.74**	**0.8811**	**0.064**	**0.062**
bagdad	74.66	71.61	0.7311	0.1915	0.2245
riad	100.0	57.14	0.7273	0	0.4286
abudhabi	69.26	66.83	0.6802	0.2545	0.2807
sana	65.89	65.89	0.6589	0.3143	0.3143
bengazi	51.20	53.39	0.5227	0.3699	0.3571

Tabela 5.22: Resultados da tarefa de classificação semântica por tipo para a categoria acontecimento (ordenados por medida-f).

zido número de tipos presentes nesta categoria (apenas 3) e ao menor número de sistemas concorrentes.

De modo idêntico ao utilizado para a comparação global da tarefa de identificação, apresentam-se os resultados selectivos globais para todos os sistemas participantes, sendo que os resultados em relação à classificação semântica por categoria estão apresentados na tabela 5.23.

O sistema obteve o primeiro lugar em termos de medida-f (0.6561), apresentando a segunda melhor precisão (77.42%) e a segunda melhor abrangência (56.93%).

Os resultados em relação à classificação semântica combinada estão apresentados na tabela 5.24.

O sistema obteve o primeiro lugar em termos de medida-f (0.6111), apresentando a segunda melhor precisão (74.80%) e a terceira melhor abrangência (51.65%).

Os resultados em relação à classificação semântica plana estão apresentados na tabela 5.25.

O sistema obteve o terceiro lugar em termos de medida-f (0.5145), apresentando a segunda melhor precisão (60.72%) e a quarta melhor abrangência (44.65%).

Os resultados em relação à classificação semântica por tipos a estão apresentados na tabela 5.26.

O sistema obteve o terceiro lugar em termos de medida-f (0.8126), apresentando a segunda melhor precisão (88.35%) e a oitava melhor abrangência (75.22%).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
l2f	**77.42**	**56.93**	**0.6561**	**0.1945**	**0.4064**
tripoli	61.30	64.81	0.6301	0.3416	0.3131
túnis	68.67	52.78	0.5968	0.2584	0.4286
damasco	62.31	53.27	0.5744	0.3383	0.4363
manama	61.70	51.97	0.5642	0.3422	0.4475
qatar	50.54	44.99	0.4760	0.4579	0.5237
rabat	50.41	44.82	0.4745	0.4582	0.5244
dakar	40.17	38.08	0.3910	0.5004	0.5399
sana	43.45	27.80	0.3391	0.4419	0.6462
nicósia	32.35	35.19	0.3371	0.6313	0.5989
oman	31.26	34.88	0.3297	0.6418	0.5979
luxor	28.02	25.35	0.2662	0.6987	0.7264
meca	50.11	15.47	0.2364	0.4694	0.8406
bahrein	43.42	9.743	0.1591	0.4417	0.8758

Tabela 5.23: Resultados da tarefa de classificação semântica por categorias relativamente ao conjunto de categorias que cada sistema se propôs identificar (ordenados por medida-f).

Saída	Precisão Máxima do Sistema (%)	Abrangência Máxima na CD(%)	Medida-F
l2f	**74.80**	**51.65**	**0.6111**
tripoli	56.30	60.42	0.5829
túnis	65.10	51.13	0.5728
damasco	57.28	49.85	0.5330
manama	56.79	48.73	0.5245
qatar	47.02	42.65	0.4473
rabat	46.57	42.25	0.4430
dakar	36.37	26.81	0.3087
oman	27.06	31.66	0.2918
nicósia	32.20	24.64	0.2792
sana	39.04	19.07	0.2563
luxor	31.66	19.66	0.2426
meca	49.57	13.49	0.2121
bahrein	38.76	7.025	0.1189

Tabela 5.24: Resultados da tarefa de classificação semântica combinada relativamente ao conjunto de categorias que cada sistema se propôs identificar (ordenados por medida-f).

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
túnis	63.13	48.52	0.5487	0.3349	0.4843
tripoli	51.50	54.45	0.5293	0.4569	0.4295
l2f	**60.72**	**44.65**	**0.5145**	**0.2540**	**0.5366**
damasco	52.39	44.79	0.4830	0.4432	0.5292
manama	52.16	43.94	0.4770	0.4489	0.5364
qatar	43.48	38.70	0.4095	0.5406	0.5944
rabat	42.86	38.11	0.4034	0.5467	0.6004
oman	25.03	27.94	0.2641	0.7290	0.6910
meca	44.81	13.84	0.2114	0.4855	0.8585
dakar	13.41	12.72	0.1306	0.5762	0.8466
luxor	13.46	12.18	0.1279	0.7319	0.8724
nicósia	11.00	11.97	0.1147	0.6551	0.8675
sana	12.76	8.162	0.09956	0.5281	0.8954
bahrein	12.82	2.877	0.04700	0.5270	0.9631

Tabela 5.25: Resultados da tarefa de classificação semântica plana relativamente ao conjunto de categorias que cada sistema se propôs identificar (ordenados por medida-f) .

Saída	Precisão (%)	Abrangência(%)	Medida-F	Sobregeração	Subgeração
abudhabi	91.65	86.82	0.8917	0.03304	0.1120
argel	85.12	84.90	0.8501	0.1032	0.09763
l2f	**88.35**	**75.22**	**0.8126**	**0.086**	**0.219**
luxor	80.20	81.24	0.8072	0.1525	0.1486
casablanca	80.21	79.53	0.7987	0.1598	0.1641
cairo	79.80	79.60	0.7970	0.1624	0.1602
tripoli	79.10	80.13	0.7961	0.1635	0.1596
sana	78.39	79.42	0.7890	0.1733	0.1676
nicósia	69.89	69.47	0.6968	0.2436	0.2316
teerão	77.85	44.51	0.5664	0.1918	0.5335
riad	76.82	29.84	0.4299	0.1665	0.6697
bagdad	57.69	27.74	0.3747	0.3188	0.6655
ancara	53.23	23.28	0.3239	0.3438	0.7016
iémen	52.95	23.17	0.3224	0.3472	0.7029

Tabela 5.26: Resultados da tarefa de classificação semântica por tipo relativamente ao conjunto de categorias que cada sistema se propôs identificar (ordenados por medida-f).

Em termos globais, confirma-se o que foi referido anteriormente: o sistema obtém um desempenho bastante melhor na classificação da categoria de topo do que na classificação dos tipos em que esta se subdivide, sendo que obtém o pior desempenho de identificação na categoria acontecimento, mas melhores resultados em termos da classificação nesta mesma categoria.

6 Conclusão e Trabalho Futuro

Neste documento apresentaram-se e compararam-se várias estratégias utilizadas na tarefa de reconhecimento de entidades mencionadas e descreveu-se um sistema de REM para a língua portuguesa cujo objectivo é identificar e classificar entidades do tipo local, pessoa, organização e acontecimento, utilizando uma abordagem manual, orientada à língua portuguesa e baseada em regras e léxicos.

O sistema foi avaliado de acordo com os critérios de avaliação do HAREM, fórum de avaliação de entidades mencionadas para a língua portuguesa, e os resultados comparados com aqueles obtidos pelos sistemas participantes na edição de 2005.

Tendo em conta esses resultados, pode afirmar-se que, no geral, o sistema classificou-se acima da média em todas as categorias em que participou, tendo um desempenho particularmente bom no reconhecimento de entidades do tipo organização, onde foi o sistema mais bem classificado em termos de medida-f.

Contudo, existem ainda bastantes entidades que não são identificadas (a abrangência total ronda os 60%) e a precisão da classificação em relação aos tipos (abaixo da categoria de topo) apresentam, à excepção da categoria acontecimento, resultados abaixo da média. Tendo isto em consideração, existe ainda trabalho a fazer no sentido de não só diminuir o erro, a subgeração e a sobregeração no reconhecimento, mas também aumentar a abrangência.

De realçar, no entanto, que existem certas situações para o qual o sistema não está correntemente preparado para lidar, algumas delas por razões relacionadas com a própria cadeia de processamento e ferramentas usadas e outras mesmas intrísecas aos textos, mas que podem ser integradas num trabalho futuro, nomeadamente:

1. Aumento da informação lexical disponível. O sistema tem um número de entradas lexicais para as categorias que classifica que não excede as 3000, o que é inferior à maioria dos sistemas analisados que fazem uso de apenas regras e léxicos para o reconhecimento. Um léxico extensivo é particularmente importante no reconhecimento de entidades que contêm nomes estrangeiros, por exemplo;

2. Resolução de ambiguidades entre as diferentes categorias, olhando para a estrutura completa da entidade e para o significado semântico dos seus constituintes, ao invés de considerar apenas as

partes essenciais da mesma. Por exemplo, a expressão " Organização Estrutural da Membrana" , do ramo da Biologia, é classificada como organização, visto que o sistema só toma em consideração o constituinte inicial da entidade, sugestivo de uma organização;

3. Integração e classificação de mais categorias no sistema. De momento, o facto de não se classificar certas categorias com determinado tipo dá origem a erros na classificação de entidades de outra categoria. Quanto mais entidades estão identificadas e classificadas, mais fácil é proceder à identificação e classificação de novas entidades;

4. Possibilidade de reconhecimento de entidades mencionadas sobre texto sem acentuação. Neste momento o sistema não identifica entidades como "Suiça" ou "Sao Paulo" ou cujos contextos não estejam correctamente acentuados, embora alguns textos da Colecção Dourada HAREM, particularmente textos *web*, não tenham qualquer acentuação;

5. Possibilidade de reconhecimento de entidades com ortografia errada. É comum encontrar entidades na Colecção Dourada com ortografia errada (e.g. um letra trocada ou em falta). As situações mais comuns poderão eventualmente ser consideradas, de modo a poder classificar uma entidade conhecida com um erro de ortografia menor;

6. Melhoramento do reconhecimento sobre texto de português brasileiro, tanto ao nível do léxico como da estrutura sintáctica e gramatical, e tendo em conta as diferenças ortográficas entre os dois países;

7. Tratamento de anáforas, particularmente quando a mesma entidade é mencionada de diversos modos (e.g. " O Liceu Maria Amália" , mas posteriormente " o Maria Amália" , o "Instituto Superior Técnico" , mas posteriormente "o Técnico" ;)

8. Possibilidade de recorrer ao contexto extra-frase. Neste momento o sistema só é capaz de processar e analisar uma frase de cada vez, perdendo-se qualquer contexto relevante que esteja incluido em frases quer antes, quer depois daquela em que a entidade se encontra inserida.

Bibliografia

ACE - Automatic Content Extraction. (n.d.). http://www.nist.gov/speech/tests/ace/.

Bick, E. (2006). *Functional Aspects in Portuguese NER.* http://poloxldb.linguateca.pt/harem/publicacoes/HAREM2006Bick.pdf.

Carreras, X., Márques, L., & Padró, L. (2002). Named Entity Extraction using AdaBoost . In *Proceedings of the CoNLL-2002.* Taipei, Taiwan.

CLEF - Cross-Language Evaluation Forum. (n.d.). http://www.clef-campaign.org/.

CoNLL - Computational Natural Language Learning. (n.d.). http://www.cnts.ua.ac.be/conll/.

Cucerzan, S., & Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of 1999 joint sigdat conference on emnlp and vlc.* University of Maryland, MD.

Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings of conll-2003* (p. 168-171). Edmonton, Canada.

Freund, Y., & Schapire, R. (n.d.). *AdaBoost.* http://www.site.uottawa.ca/~stan/csi5387/boost-tut-ppr.pdf.

Gale, W., Church, K., & Yarowsky, D. (1992). One sense per discourse. In *Proceeding of the 4th darpa speech and natural language workshop* (p. 233-237).

HAREM - Avaliação de Reconhecimento de Entidades Mencionadas. (n.d.). http://poloxldb.linguateca.pt/harem.php.

IREX - Information Retrieval and Extraction Exercise. (n.d.). http://cs.nyu.edu/projects/proteus/irex.

Katz, S. M. (1996). Distribution of context words and phrases in text and language modeling. *Natural Language Engineering*, 15-59.

Krupka, G. R. (1995). SRA:Description of the SRA System as Used for MUC-6. In *Proceedings of the 1995 MUC-6.* Maryland, USA.

Mamede, N. (2007). *A Cadeia de Processamento XIP em Maio de 2007.*

Medeiros, J. C. (1995). *Processamento morfológico e correcção ortográfica do português.* Portugal.

Mendes, A. (2007). Clefomania, QA@L2F: Primeiros Passos.

Mikheev, A., Grover, C., & Moens, M. (1999). Description of the LTG system used for MUC-7. In *Proceedings of 1999 muc-7.* University of Edinburgh.

MUC - Message Understanding Conferences. (n.d.). `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/`.

Palmer, D. D., & Day, D. S. (1997). A statistical profile of the named entity task. In *Proceedings of fifth acl conference for applied natural language processing (anlp-97).* Washington, DC.

Pardal, J. P. (2007, Maio). *Manual do Utilizador do RuDriCo.* L^2F – Laboratório de Sistemas de Língua Falada.

Projecto AC/DC. (n.d.). `http://acdc.linguateca.pt`.

Ribeiro, R., Mamede, N. J., , & Trancoso, I. (2003). Using morphossyntactic information in tts systems: comparing strategies for european portuguese. In *Computational processing of the portuguese language: 6th international workshop, propor 2003, faro, portugal, june 26-27, 2003. proceedings* (Vol. 2721). Springer.

Sekine, S. (2004). *Named entity: History and future.* (New York University)

Sekine, S., & Eriguchi, Y. (2000). Japanese named entity extraction evaluation: analysis of results. In *Proceedings of the 18th conference on computational linguistics.* Saarbrücken, Germany.

Zhang, T., Damerau, F., & Johnson, D. (2002, March). Text chunking based on a generalization of a winnow. *Journal of Machine Learning Research*, *2*, 615-637.

Zhou, G., & Su, J. (2002). Named entity recognition using an HMM based chunk tagger. In *Proceedings of the 40th annual meeting of the acl* (p. 473-480). Philadelphia, PA.