Estud. lingüíst. galega 4 (2012): 13-25
DOI 10.3309/1989-578X-**XX-XX**

Identificação e classificação de entidades mencionadas em galego

Marcos Garcia
Universidade de Santiago de Compostela
marcos.garcia.gonzalez@usc.es
Iria Gayo
Universidade de Santiago de Compostela
iria.delrio@usc.es
Isaac González López
Cilenis Language Technology
isaacjgonzalez@cilenis.com

Recibido o 14/07/2011. Aceptado o 13/10/2011

Resumo

A identificação e a classificação semântica automáticas de entidades mencionadas são tarefas de especial relevância para variadas aplicações do processamento da língua natural, tais como a tradução automática, a extracção de informação ou os sistemas de resposta a perguntas. O presente artigo descreve a adaptação e implementação de diversas ferramentas de código aberto para a identificação e classificação dos seguintes tipos de entidades em galego: (i) datas, (ii) numerais, (iii) quantidades e (iv) nomes próprios. A análise dos três primeiros tipos de entidades realiza-se com o *software* FreeLing através de máquinas de estados finitos. Para a identificação de nomes próprios comparam-se duas estratégias: (i) a utilização de máquinas de estados finitos e (ii) métodos de aprendizagem automática. Finalmente, a classificação semântica dos nomes próprios é realizada com um sistema baseado em regras e recursos obtidos automaticamente. O artigo mostra um conjunto de avaliações para cada um dos módulos apresentados, disponibilizados com licenças livres.

Palabras chave

Processamento da língua natural, reconhecimento de entidades mencionadas, galego

Named Entity Recognition and Classification in Galician

Abstract

Automatic named entity recognition and classification are important tasks for many natural language processing applications, such as machine translation, information extraction or question-answering systems. This paper describes the adaptation and implementation of several open-source systems for the identification and classification of the following named entities in Galician: (i) dates, (ii) numerals, (iii) quantities and (iv) proper nouns. Analysis of the first three types of named entities is performed with the software FreeLing, using finite-state automata. For the proper noun recognition task, two methods were compared: (i) finite-state automata and (ii) machine learning models. Finally, the semantic classification of proper nouns was carried out with a rule-based system that takes advantage of automatically obtained resources. This paper shows some evaluations for each tool, all available under free licenses.

Keywords

Natural language processing, named entity recognition and classification, Galician

Este trabalho foi feito no âmbito de um projecto financiado pelo MCINN com referência FFI2010-14986

1. Introdução

Diversas aplicações do Processamento da Língua Natural (PLN) precisam da execução prévia de sistemas de Reconhecimento de Entidades Mencionadas (REM) para melhorar os seus resultados. O REM pode ser dividido em duas sub-tarefas diferenciadas: a identificação e a classificação das entidades.

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

"**José Francisco** foi ver o **Celta de Vigo** a **Balaídos**"

A segunda (classificação) tem como objectivo etiquetar as entidades reconhecidas com base em classes previamente definidas (datas, quantidades, numerais, pessoas, organizações, etc.). Assim, um resultado possível da aplicação destes sistemas no exemplo anterior poderia ser o seguinte:

"**José Francisco**_*(pessoa)* foi ver o **Celta de Vigo**_*(organização)* a **Balaídos**_*(localização)*"

Esta informação é posteriormente utilizada por sistemas de tradução automática (com o fim de distinguir diferentes tipos de entidades traduzíveis), de extracção de relações (cujos atributos pertencem a uma determinada classe), ou de resposta a perguntas, entre outros.

As tarefas REM referidas são parte de uma disciplina mais abrangente, a Extracção de Informação, cujos sistemas vêm sendo avaliados desde a década de 90 em várias conferências como as MUC[1](*Message Understanding Conference*), as CoNLL[2] (*Conference on Computational Natural Language Learning*) ou as ACE[3] (*Automatic Content Extraction*).

Em galego, as únicas ferramentas que conhecemos que realizam algumas destas tarefas são dois módulos da *suite* de análise FreeLing (Padró *et al.* 2010), dedicados ao reconhecimento de numerais e de quantidades. Neste sentido, revela-se necessária a implementação e adaptação de ferramentas deste tipo para esta língua, bem como uma distribuição e disponibilização que permita o seu uso e melhoria tanto pela comunidade de PLN em geral como pelos investigadores da linguística galega.

O presente artigo tem dois objectivos diferenciados: por um lado, descrever a adaptação e implementação de vários sistemas REM para o galego. Pelo outro, realizar um conjunto de avaliações preliminares para cada um dos sistemas sobre texto real revisto manualmente.

Primeiramente, são descritos vários dos módulos do pacote FreeLing (que foram melhorados e/ou adaptados) para distintas tarefas de reconhecimento de entidades. Mais concretamente, foram ampliados e melhorados os já existentes (de reconhecimento de numerais e quantidades) e adicionados dois novos módulos: (i) um reconhecedor de datas e horas e (ii) um identificador estatístico de nomes próprios (o qual utiliza aprendizagem automática para detectar as fronteiras de nomes próprios compostos).

De seguida, descreve-se a adaptação para galego de um classificador semântico de nomes próprios já utilizado para a análise do português. O classificador não precisa de *corpora* anotados, uma vez que se baseia em regras e em recursos obtidos de modo automático.

Os resultados das avaliações, apesar de preliminares, indicam que as ferramentas adaptadas para galego têm desempenhos similares (e em alguns casos, superiores) aos que apresentam para outras línguas, bem como outros sistemas com objectivos similares. Além disso, é impor-

[1] http://www-nlpir.nist.gov/related_projects/muc/

[2] http://ifarm.nl/signll/conll/

[3] http://www.itl.nist.gov/iad/mig/tests/ace/

tante salientar que todos os recursos e ferramentas descritos no presente artigo são disponibilizados sob licenças livres.

2. Trabalho relacionado

A escassez de ferramentas de reconhecimento de entidades mencionadas em galego implica também a inexistência de literatura científica a seu respeito. Tendo isto em conta, nesta secção apresentamos brevemente aqueles trabalhos e avaliações conjuntas dedicadas ao inglês (por ser a língua para a qual mais recursos e desenvolvimento existem), bem como ao português e espanhol, pelo facto de serem línguas próximas do galego, cujas ferramentas e estratégias podem ser parcialmente aproveitadas para os objectivos deste trabalho.

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

As *shared task* das conferências CoNLL 2002 e 2003 tiveram avaliações de sistemas de classificação independentes da língua (espanhol e holandês em 2002 e inglês e alemão em 2003), para entidades "enamex". Nestas avaliações, os melhores sistemas obtiveram valores da medida F de 72% (alemão), 88% (inglês), 77% (holandês) e 81% (espanhol). Note-se que este último sistema é a base dos módulos de classificação de entidades mencionadas do FreeLing (não avaliado no presente trabalho por não dispormos de um *corpus* anotado de tamanho suficiente para o seu treino).

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

Detendo-nos nos próprios sistemas de reconhecimento, podemos afirmar que a tendência dominante de desenvolvimento destes recursos é a utilização de regras e de máquinas de estados finitos para a identificação de expressões "timex/numex", e de modelos estatísticos para o tratamento de entidades "enamex".

Existem, contudo, métodos baseados em regras para a classificação de entidades "enamex" (Bick 2006), e modelos híbridos como o apresentado em Ferreira, Balsa e Branco (2007), ambos desenhados para a língua portuguesa.

Os modelos probabilísticos são habitualmente treinados de modo supervisionado, pelo que precisam de *corpora* etiquetados manualmente (Finkel, Grenager e Manning (2005) para o inglês, Carreras *et al*. (2002) para o espanhol ou Ferrández *et al.* (2007) para o português). As ferramentas, por sua vez, utilizam diferentes algoritmos (ou combinações deles) como Conditional Random Fields, AdaBoost, Support Vector Machines ou Hidden Markov Models, entre outros.

A dificuldade de obtenção de recursos de qualidade para a realização do treino dos diferentes modelos inspirou várias estratégias de classificação não-supervisionada (ou semi-supervi-

sionada). Assim, o aumento de fontes semi-estruturadas de fácil acesso (Freebase[4], DBpedia[5]) permite a obtenção de recursos e de *corpora* potencialmente aplicáveis no treino destes modelos. Neste sentido, alguns trabalhos recentes propõem estratégias que tiram proveito de fontes como a Wikipedia[6] para melhorar os sistemas de classificação e extracção (Mika *et al.* 2008). De modo similar, Nothman, Curran e Murphy (2008) utilizam os links internos da Wikipedia para anotar automaticamente entidades em texto não estruturado, empregado posteriormente para treinar modelos estatísticos. Por último, em Gamallo e Garcia (2011) é apresentado um classificador semântico de nomes próprios para o português que utiliza um conjunto de regras e grandes listas de entidades obtidas (semi-)automaticamente. No presente artigo descreve-se a adaptação deste sistema para a análise do galego.

3. Identificação e classificação de entidades mencionadas

Nesta secção são apresentadas brevemente as diferentes tarefas realizadas pelos sistemas disponibilizados, bem como o processo de adaptação e desenvolvimento de cada um deles, avaliados na secção seguinte.

3.1. Numerais

O primeiro dos processos descritos é o reconhecimento de expressões numerais, realizado através de um módulo específico do pacote FreeLing. O reconhecedor de expressões numerais aplica-se depois do *tokenizador*, pelo que a sua entrada se encontra já dividida em *tokens* (elementos individuais como palavras ou sinais de pontuação).

Este módulo é composto por um conjunto de máquinas de estados finitos que detectam expressões numerais em vários formatos: numérico ("7,4", "325.275") e extenso ("trescentos vinte cinco mil douscentos setenta e cinco", "un millón e medio"), assim como outras formas lexicais tais como "decenas", "milleiros", "cuartos", etc. Além da identificação, o módulo normaliza as entidades, atribuindo um lema numérico a cada uma das expressões reconhecidas ("24", "vinte e catro", "dúas ducias" → *24*).

O módulo de reconhecimento de expressões numerais já estava disponível em versões anteriores do FreeLing (2.2), pelo que para o presente trabalho só foram realizadas pequenas correcções de erros de identificação.

3.2. Datas

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

Este módulo reconhece formas como horas, dias da semana (e as suas partes: "mediodía", "mañá", "madrugada"...), meses, séculos, anos, etc., que podem aparecer de modo individual ("xullo", "12:24h.") ou em diferentes combinações ("sete da mañá", "luns, vinte e sete de xullo de mil novecentos oitenta", "xaneiro do 1968", etc.). As máquinas de estados finitos identificam também outro tipo de expressões comuns como "o pasado mes de xullo" ou "as sete e cuarto da tarde".

[4] http://www.freebase.com

[5] http://www.dbpedia.org

[6] http://www.wikipedia.org

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

Este módulo não se encontrava disponível em anteriores versões do FreeLing para galego, pelo que foi preciso desenvolvê-lo por completo.

3.3. Quantidades

O reconhecedor de quantidades (também incluído no FreeLing) é dependente do reconhecedor de expressões numerais. Assim, consiste num conjunto de máquinas de estados finitos, ao qual há que acrescentar um ficheiro externo com etiquetas e expressões regulares relativas a quantidades, unidades monetárias, longitudes, etc.

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

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

Versões anteriores do FreeLing (2.2) já disponibilizavam este módulo para a análise do galego. Contudo, tanto o número de máquinas de estados finitos como de unidades e quantidades foram aumentados consideravelmente.

3.4. Nomes próprios

O tratamento dos nomes próprios apresentado no presente artigo divide-se em duas tarefas diferentes: (i) a identificação e (ii) a classificação semântica. A identificação consiste na detecção correcta das fronteiras de um nome próprio ("*Museo_do_Pobo_Galego*" —identificado como um único nome próprio— *vs* *"*Museo* do *Pobo_Galego*", onde erroneamente se detectam dois nomes próprios). A classificação consiste na atribuição ao nome próprio de uma etiqueta que denote uma classe semântica previamente definida (no nosso caso: "pessoa", "organização", "localização" e "outro").

O primeiro dos processos foi realizado através de dois módulos do FreeLing (com o fim de avaliar idependentemente o desempenho de cada um deles), enquanto que a classificação foi realizada com um sistema de código aberto baseado em regras e recursos, não precisando portanto de *corpus* de treino (Gamallo e Garcia 2011).

3.4.1. Identificação

Utilizaram-se dois módulos diferentes para a identificação de nomes próprios: *basic* e *BIO*. O primeiro (*basic*) consiste numa máquina de estados finitos que detecta sequências de palavras que começam por maiúsculas, e numa lista de palavras funcionais ("de", "por", etc.) que podem ocupar uma posição intermédia em nomes próprios compostos. Esta estratégia identifica expressões como "John Lennon" ou "Universidade de Vigo", e em conjunto com um desambiguador morfossintáctico (já incluído no FreeLing, (Garcia e Gamallo 2010)), detecta com alta precisão se um *token* em posição inicial de oração é ou não é um nome próprio ("*Café* con leite" *vs* "*Café* Starbucks de Barcelona").

Este método não precisa de *corpus* de treino, sendo a sua adaptação e execução rápidas. Contudo, existem casos em que o identificador falha sistematicamente, uma vez que não são atribuídos valores de probabilidade para cada um dos elementos que podem formar o nome próprio: assim, tanto a expressão "*Consellería_de_Educación*" como *"[a esa hora chegaba] *Sarkozy_de_Roma*" são analisadas como um único nome próprio.

O segundo dos módulos de identificação de nomes próprios do FreeLing tenta corrigir estes e outro tipo de erros implementando o método estatístico *BIO*. Esta estratégia de aprendizagem automática precisa de um *corpus* de treino anotado, cujos nomes próprios sejam divididos em B (*begin*) e I (*inside*), para além dos elementos que não formam parte dos nomes próprios (O, *outside*). O *corpus* de treino, bem como um conjunto de atributos lexicais (que incluem listas de nomes próprios frequentes, palavras funcionais, etc.), permitem criar um classificador que detecte as fronteiras dos nomes próprios, em função das probabilidades de cada *token* ser B, I ou O.

Foram treinados cinco modelos diferentes do método estatístico, utilizando o algoritmo AdaBoost (Carreras *et al*. 2002), em função da frequência dos atributos no *corpus* de treino (desde os que apresentam uma frequência superior a 1%, até todos os extraídos). Para tarefas de treino e teste dos identificadores utilizou-se o *corpus* do projecto GariCoter (Barcala *et al*. 2007), de aproximadamente 240.000 *tokens* (com cerca de 12.000 nomes próprios).

3.4.2. Classificação

A classificação de nomes próprios é uma tarefa que consiste na atribuição, depois de identificadas as fronteiras do nome próprio, de uma etiqueta semântica previamente estabelecida a cada um deles. Apesar de existirem tarefas que requerem uma classificação mais detalhada, as etiquetas utilizadas no nosso sistema são as "enamex" (ver secção 2), amplamente utilizadas no reconhecimento de entidades desde a avaliação MUC-6. Estas etiquetas diferenciam três classes principais: PER (pessoa), ORG (organização) e LOC (localização), às quais desde a conferência CoNLL 2002 se acrescenta MISC (outro), para classificar as entidades que não pertencem a nenhum dos tipos anteriores.

A classificação de determinados nomes próprios de acordo com as etiquetas estabelecidas provocou algumas diferenças tanto nas várias edições das avaliações referidas como em outras (Segundo HAREM, por exemplo). Dois dos principais problemas que surgem na classificação de nomes próprios são a polissemia e a metonímia. Neste sentido, determinados nomes de países, cidades, etc. podem ser classificados (para além de LOC), como ORG ("Bélxica asinou o Tratado de Roma"), como PER ("Vigo opúxose á reforma do sector naval"), etc., utilizando critérios de etiquetagem diferentes.

Nos nossos testes, tanto na implementação do sistema de classificação, como na anotação manual do *corpus* de teste, só foi considerada a homonímia, ignorando-se portanto as interpretações metonímicas das entidades mencionadas (que podem ser identificadas em processos posteriores de análise) (veja-se Gamallo e Garcia (2011) para uma discussão mais pormenorizada).

A propósito das características dos sistemas de classificação de entidades mencionadas, cabe referir que existem diversas estratégias para o seu desenvolvimento. A par dos modelos estatísticos (que precisam de um *corpus* de treino com anotação semântica de alta qualidade), existem sistemas baseados em regras, os quais analisando o contexto linguístico mais próximo das entidades, as classificam de acordo com as etiquetas predefinidas.

O sistema avaliado no presente trabalho é uma adaptação do classificador que Gamallo e Garcia (2011) propõem para o português. Enquadra-se nos sistemas de regras, pelo que não precisa de um *corpus* de treino anotado. Para além de um conjunto de regras, o classificador utiliza

várias listas de entidades de cada classe (*gazetteers*) e de *trigger words*, palavras que sugerem a classificação de entidades num determinado tipo ("amigo": PER, "empresa": ORG, etc.).

Com o fim de obtermos de modo (semi-)automático os recursos (*gazetteers* e *trigger words*), utilizámos a seguinte estratégia:

Para extrair as *trigger words*, procura-se na árvore de categorias da Wikipedia (em galego) um conjunto de categorias que sejam subclasses de pessoas, organizações e localizações. Este processo realiza-se seleccionando categorias que contenham as palavras "persoa", "organización" ou "lugar" (e sinónimos) como núcleo da categoria (p. ex. "Organizacións internacionais").

Depois, escolhem-se os hipónimos de cada categoria e seleccionam-se os núcleos ("Asociación Europea de Libre Comercio" → "asociación"; "Comisión Central para a Navegación no Rin" → "comisión", etc.). As listas obtidas por este processo são depois utilizadas como *trigger words* e como palavras semente para a obtenção dos *gazetteers*.

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

Uma vez obtidos os recursos necessários para a aplicação do classificador, este aplica a seguinte estratégia de etiquetagem:

- Se a entidade aparece só numa das listas de *gazetteers*, é classificada de acordo com a classe da lista.
- Se for uma forma homónima (aparece em várias listas), ou desconhecida, procura-se nos três *tokens* anteriores e posteriores à entidade para verificar se correspondem a alguma *trigger word* conhecida (p.ex., "o meu *amigo* Anselmo"). Se houver várias *trigger words*, é preferida a mais próxima (ou a que se encontre antes da entidade). O algoritmo contém um conjunto de regras que evitam a classificação, entre outros fenómenos, de complementos preposicionais (em função das preposições que possam existir entre uma *trigger word* e uma entidade). Assim, a entidade "Banco de Portugal" no contexto "fundador do *Banco de Portugal*" não será classificada como pessoa apesar de ter a *trigger word* (da classe PER) "fundador".
- Se a entidade não é conhecida nem tem *trigger words* próximas, analisa-se a sua forma para verificar se contém *trigger words* internas ("*Museo* do Pobo Galego"), se é acrónimo, etc.
- Finalmente, se as regras anteriores não conseguem atribuir uma etiqueta, um último conjunto de regras decide entre a classificação como MISC ou como ORG, em função tanto da forma do nome próprio como do contexto morfossintáctico mais próximo.

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

4. Testes e avaliação

Nesta secção são apresentados os diversos testes realizados para avaliar cada um dos módulos descritos. Primeiro, mostram-se os testes de avaliação dos reconhecedores de base numérica (numerais, datas e quantidades). A seguir, descrevem-se as experiências de identificação de nomes próprios e, finalmente, mostra-se a avaliação do sistema de classificação de nomes próprios.

O primeiro conjunto de testes tem como objectivo a realização de uma avaliação preliminar dos reconhecedores de expressões numerais, datas e quantidades sobre texto real. Com este fim, foram seleccionadas aleatoriamente 10 notícias do jornal em galego Galicia Hoxe (de todas as secções), criando um *corpus* de aproximadamente 10.000 *tokens*, com quase 300 entidades de base numérica etiquetadas manualmente.

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

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

A Tabela 1 mostra os resultados das avaliações referidas, tendo em conta a etiquetagem de cada tipo de entidades, bem como o desempenho geral dos reconhecedores. A precisão é o número de entidades correctamente reconhecidas dividido pelo número total de entidades reconhecidas; o *recall* refere-se ao número de entidades correctamente reconhecidas dividido pelo número total de entidades do *corpus* de teste. Finalmente, a medida F é a média harmónica entre a precisão e o *recall*.

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

Tabela 1. Número de entidades, precisão, *recall* e medida F dos reconhecedores de expressões numéricas, datas e quantidades

Para além das entidades *Números* e *Datas* (reconhecidas pelos módulos do mesmo nome), as *Percentagens*, *Moedas* e *Unidades* foram analisadas pelo reconhecedor de quantidades. As principais diferenças entre as avaliações *Dura* e *Branda* têm a ver com a classificação de expressões numéricas em contextos ambíguos, às quais o sistema atribui a etiqueta *Número*. Assim,

entre as duas avaliações, a anotação de numerais passa de 63% a 93%, a de moedas de 63% a 100%, e a de datas de 73% a 95%. Tendo em conta as propriedades dos módulos adaptados, assim como a ambiguidade de expressões como as referidas nos parágrafos anteriores, a avaliação *Branda* dos reconhecedores de numerais, quantidades e datas mostra que o desempenho destes módulos se situa em redor dos 95% (note-se, contudo, que o número de entidades analisadas é pequeno).

O conjunto seguinte de testes avaliou o identificador de nomes próprios em texto livre. O primeiro método testou o desempenho do módulo *basic* do FreeLing, que utiliza uma máquina de estados finitos. O *corpus* de teste utilizado foi uma selecção aleatória de 40.000 *tokens* do *corpus* do projecto GariCoter (Barcala *et al*. 2007), também empregue para testar o modelo estatístico *BIO*. Para treinar os classificadores *BIO*, utilizou-se a parte restante do mesmo *corpus*, de uns 200.000 *tokens*.

A avaliação foi realizada de acordo com as directrizes do CoNLL, tendo em conta a anotação tanto dos *tokens* B como dos I. A Tabela 2 mostra os resultados da avaliação do módulo *basic*, bem como o melhor modelo estatístico (que utiliza o total dos atributos extraídos do *corpus* de treino; cabe referir, contudo, que as diferenças máximas entre os diferentes modelos *BIO* foram de 0,5% na medida F).

Modelo	**Prec.**	**Rec.**	**Med. F**
basic	87,72%	92,31%	89,96%
BIO	94,86%	94,77%	94,81%

Tabela 2. Precisão, *recall* e medida F dos reconhecedores de nomes próprios

Os resultados da avaliação dos identificadores de nomes próprios mostram que o método probabilístico, treinado com um *corpus* anotado manualmente de 200.000 *tokens*, obtém resultados notoriamente superiores (quase 5 pontos acima) ao modelo *basic*. Contudo, note-se que os resultados deste último modelo, que não precisa de *corpus* de treino e que tem uma execução mais rápida, estão próximos dos 90%.

Finalmente, os últimos testes foram dedicados a conhecer o funcionamento do sistema de classificação semântica de nomes próprios baseado em regras e recursos. Para isso foi anotada manualmente uma selecção aleatória de 20.000 *tokens* (com perto de 1.000 nomes próprios) do *corpus* de teste utilizado para a avaliação do sistema de identificação de nomes próprios. Tendo em conta que o sistema é dependente de listas externas, e que o tamanho da Wikipedia em galego é notoriamente menor que o de outras línguas (além de que, muitos dos nomes próprios podem ser independentes da língua), foram realizados quatro testes diferentes:

Para o primeiro teste utilizaram-se unicamente os *gazetteers* extraídos da versão galega da Wikipedia; no segundo e no terceiro juntaram-se a estes os extraídos da versão portuguesa e espanhola, respectivamente; finalmente, levou-se a cabo uma avaliação com as listas de entidades de três versões da Wikipedia (galega, portuguesa e espanhola). Em todas as avaliações foi utilizado o mesmo conjunto de *trigger words*, obtidas da Wikipedia em galego.

A Tabela 3 contém o número de entidades de cada uma das listas de *gazetteers* (onde *gl*=galego, *pt*=português e *es*=espanhol), bem como o número de *trigger words* utilizadas. Na Tabela 4 podemos ver os resultados dos diferentes testes de classificação de nomes próprios. Estes testes foram realizados utilizando o modelo *BIO* para o reconhecimento dos nomes próprios. Testes preliminares em que foi usado o método *basic* tiveram resultados com valores da medida F ≈3% mais baixos.

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

Tabela 3. Número de *trigger words* e *gazetteers* extraídos da Wikipedia

Gazetteers	Prec.	Rec.	Med. F
gl	62,15%	62,84%	62,49%
gl+pt	74,11%	74,94%	**74,52%**
gl+es	59,80%	60,47%	60,14%
gl+pt+es	68,31%	69,08%	68,69%

Tabela 4. Precisão, *recall* e medida F do classificador de nomes próprios em función dos *gazetteers*

O primeiro conjunto de resultados mostra que a utilización de *gazetteers* extraídos unicamente da Wikipedia em galego não é suficiente para conseguir um bom desempenho de um sistema baseado em regras e recursos, obtendo valores da medida F de 62,49%. A utilização de listas de entidades em espanhol e português melhora, portanto, a qualidade do sistema (68,69%). Note-se, contudo, que o aumento da medida F é superior com as listas *gl+pt* (menor que o conjunto *gl+pt+es*), pelo que se infere que os *gazetteers* extraídos da Wikipedia em espanhol podem ter algum tipo de ruído (o que explicaria os valores do teste *gl+es*). Assim, os resultados do sistema apresentado com os *gazetteers* do galego e do português ultrapassam 74% de medida F. A Tabela 5 mostra os resultados individuais de cada classe "enamex". A análise destes resultados indica-nos que os tipos LOC e ORG (e, em menor medida, PER) têm bons resultados tanto em termos de precisão como de *recall*. Contudo, os valores de *recall* da classe MISC são notoriamente mais baixos. Note-se que tanto o tipo de entidades quanto a sua contextualização são mais heterogéneas do que as restantes. Neste sentido, e tendo em conta que o sistema apresentado se baseia em regras e recursos, cabe analisar pormenorizadamente a classificação da classe MISC, com o fim de melhorar a sua etiquetagem em posteriores versões do classificador.

Classe	Núm.	Prec.	Rec.	Med. F
LOC	280	81,89%	83,20%	82,54%
PER	121	61,31%	77,06%	68,29%
ORG	438	73,79%	80%	76,77%
MISC	85	62,5%	7,94%	14,08%

Tabela 5. Número, precisão, *recall* e medida F vs classe de entidades "enamex". *Gazetteers* gl+pt

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

É preciso ter em conta que as listas de *gazetteers* utilizadas não tiveram nenhum tipo de revisão nem filtro. Neste sentido, a adaptação das listas portuguesas para galego pode ser uma boa estratégia de melhoramento do sistema (Malvar *et al.* 2010). Além disso, a aplicação de al-

gum tipo de filtro e/ou revisão sobre os *gazetteers*, assim como a utilização de listas com maior número de entidades (como as do inglês, por exemplo), podem contribuir para o aumento da precisão do classificador semântico aqui proposto.

Os resultados obtidos pelos distintos sistemas apresentados não são facilmente comparáveis com os de sistemas concebidos para outras línguas, devido às características dos *corpora* de teste, bem como aos próprios objectivos de cada um dos módulos de reconhecimento. Assim, os resultados dos sistemas de reconhecimento de numerais, datas e quantidades só podem entender-se como preliminares, uma vez que o *corpus* de teste não tem um tamanho suficiente para considerá-los definitivos.

A tarefa de identificação de nomes próprios foi avaliada sobre um *corpus* mais diversificado e de maior extensão, obtendo-se valores na medida F similares aos resultantes da aplicação destes sistemas para outras línguas (Carreras *et al.* 2002).

A comparação entre os resultados do último dos módulos, o classificador semântico de nomes próprios, é mais complexa: por um lado, os objectivos de diferentes classificadores costumam ser diferentes, em função do número de tipos e subtipos de entidades que pretendam classificar. Por outro lado, o tamanho e tipologia do *corpus* de teste é também muito variável, bem como a anotação de entidades potencialmente ambíguas. Tendo isto em conta, e observando que, por exemplo, os melhores sistemas das avaliações CoNLL (2002 e 2003) diferem em mais de 16 pontos percentuais (72,41% para o alemão e 88,76% para o inglês), os resultados obtidos com diferentes métricas e *corpora* não podem ser directamente comparáveis. O mesmo acontece se observarmos os resultados das avaliações do Segundo HAREM, cujos valores foram obtidos utilizando tipos e subtipos diferentes na classificação de entidades. Nesta avaliação, a métrica mais próxima da realizada no presente artigo é o "Cenário Selectivo 2", que inclui as categorias "local" (com dois subtipos: "humano" e "físico"), "organização", "pessoa" e "tempo", na qual o sistema XIP-L2F/Xerox_3 obteve valores da medida F de 63,26%. Por último, o sistema de classificação semântica avaliado neste artigo teve quase os mesmos resultados (medida F de ≈75%) do que na análise do Português (Gamallo e Garcia 2011)".

Em suma, esta secção apresentou os resultados de um conjunto de avaliações sobre os diferentes sistemas de reconhecimento de entidades mencionadas adaptadas e implementadas para a análise do galego, mostrando que é possível aproveitar para o galego sistemas já desenvolvidos para outras línguas e obter resultados similares.

5. Conclusões e trabalho futuro

O presente trabalho descreve a adaptação e implementação de sistemas de reconhecimento de entidades mencionadas em galego.

Foram melhorados os módulos já existentes de reconhecimento de expressões numéricas e de quantidades da *suite* Freeling, e foi adicionado um novo módulo de reconhecimento de datas e horas para o galego. Adicionalmente, foi treinado um identificador estatístico de nomes próprios sobre um *corpus* de mais de 200.000 *tokens*.

Para além dos módulos referidos, foi adaptado para o galego um sistema de classificação semântica de nomes próprios. O sistema baseia-se em regras e recursos obtidos de modo semi-automático e foi utilizado com resultados similares na análise do português.

As avaliações dos diferentes sistemas descritos são ainda preliminares, nomeadamente devido à pequena dimensão dos *corpora* utilizados. Contudo, os resultados sugerem que o desempenho de cada um deles se aproxima dos valores obtidos pelos reconhecedores de entidades dos tipos "timex", "numex" e "enamex" em avaliações com métricas similares, tais como as *shared tasks* das CoNLL (Tjong Kim Sang e de Meulder 2003).

Como trabalho futuro, para além de realizar diferentes testes em textos de diversas tipologias e extensões, torna-se necessário implementar sistemas de desambiguação de entidades de base numérica (datas, números, etc.), bem como o treino de classificadores estatísticos de nomes próprios, com o fim de comparar o seu desempenho com o modelo proposto neste artigo. Além disso, é preciso realizar uma análise pormenorizada dos erros produzidos pelo classificador de nomes próprios, com vista a melhorar a sua precisão na etiquetagem de entidades potencialmente ambíguas ou mal identificadas por módulos anteriores (como "rúa 5 de Outubro", etc.).

Finalmente, cabe referir que todos os recursos apresentados e utilizados no presente trabalho são disponibilizados com licenças livres (alguns deles na versão 3.0 do FreeLing), o que permite que a comunidade científica possa utilizá-los livremente, bem como melhorá-los e/ou modificá-los em função dos seus objectivos.

Agradecimentos

Agradecemos aos revisores anónimos pelas sugestões que contribuíram para o aperfeiçoamento deste artigo.

Referências bibliográficas

Barcala, Francisco Mario *et al.* (2007): "A corpus and lexical resources for multi-word terminology extraction in the field of economy in a minority language", em Zygmunt Vetulani (ed.), *Human Language Technology as a Challenge for Computer Science and Linguistics. Proceedings of the 3rd Language and Technology Conference*. Poznand: Wydawnictwo Poznańskie, 359-363 (http://www.grupocole.org/cole/library/ps/BarDomGamLopMosRojSanSot2007b.pdf).

Bick, Eckhard (2006): "Functional aspects on Portuguese NER", em Renata Vieira *et al.* (eds.), *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Language (PROPOR 2006). Lecture Notes in Computer Science*, vol. 3960. Berlin / Heidelberg: Springer-Verlag, 260-263 (http://193.136.2.105/aval_conjunta/LivroHAREM/Cap12-SantosCardoso2007-Bick.pdf).

Carreras, Xavier *et al.* (2002): "Named entity extraction using AdaBoost", em *Proceedings of the 6th Conference on Computational Natural Language Learning (CoNLL 2002)*. Taipei: Association for Computational Linguistics (ACL), 167-170 (http://acl.ldc.upenn.edu/W/W02/W02-2004.pdf).

Ferrández, Óscar *et al.* (2007): "Tackling HAREM's portuguese named entity recognition task with spanish resources", em Diana Santos / Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca (http://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap11-SantosCardoso2007-Ferrandezetal.pdf).

Ferreira, Eduardo / João Balsa / António Branco (2007): "Combining rule-based and statistical methods for named entity recognition in Portuguese", em *V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL 2007). Anais do XXVII Congresso da Sociedade Brasileira de Computação.* Salvador: Sociedade Brasileira de Computação (SBC), 1615-1624 (http://www.di.fc.ul.pt/%7Eahb/FerreiraBalsaBranco2007.pdf).

Finkel, Jenny Rose / Trond Grenager / Christopher Manning (2005): "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", em *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. Ann Arbor: Association for Computational Linguistics (ACL), 363-370 (http://acl.ldc.upenn.edu/P/P05/P05-1045.pdf).

Gamallo, Pablo / Marcos Garcia (2011): "A Resource-Based Method for Named Entity Extraction and Classification", em Luís Antunes / H. Sofia Pinto (eds.), *Proceedings of the XV Portuguese Conference on Artificial Intelligence (EPIA 2011). Progress in Artificial Intelligence. Lecture Notes in Computer Science (LNAI)*, vol. 7026. Berlin / Heidelberg: Springer-Verlag, 610-623.

Garcia, Marcos / Pablo Gamallo (2010): "Análise morfossintáctica para Português Europeu e Galego. Problemas, Soluções e Avaliação", *Linguamática. Revista para o Processamento Automático das Línguas Ibéricas* 2(2), 59-67 (http://linguamatica.com/index.php/linguamatica/article/download/56/87).

Leach, Geoffrey / Andrew Wilson (1996): "Recommendations for the Morphosyntactic Annotation of Corpora". Relatório Técnico. Expert Advisory Group on Language Engineering Standards (EAGLES) (http://tagmatica.fr/doc/EaglesAnnotate.pdf).

Malvar, Paulo *et al.* (2010): "Vencendo a escassez de recursos computacionais. Carvalho: Tradutor

Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português", *Linguamática. Revista para o Processamento Automático das Línguas Ibéricas* 2(2), 31-38 (http://linguamatica.com/index.php/linguamatica/article/download/57/81).

Mika, Peter *et al.* (2008): "Learning to tag and tagging to learn: A case study on Wikipedia", *IEEE Inteligent Systems* 23(5), 26-33 (http://research.yahoo.com/files/wikipedia-ieee.pdf).

Mikheev, Andrei / Claire Grover / Marc Moens (1998): "Description of the LTG system used for MUC-7", em *Proceedings of the 7th Message Understanding Conference*. Morgan Kaufman (http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/ltg_muc7.pdf).

Mota, Cristina / Diana Santos (eds.) (2008): *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo* HAREM. Linguateca (http://www.linguateca.pt/LivroSegundoHAREM/).

Nothman, Joel / James R. Curran / Tara Murphy (2008): "Transforming Wikipedia into Named Entity Training Data", em Nicola Stokes / David Powers (eds.), *Proceedings of the Australasian Language Technology Workshop*, vol. 6. Hobart: Australasian Language Technology Association, 124-132 (http://aclweb.org/anthology/U/U08/U08-1016.pdf).

Padró, Lluís *et al.* (2010): "FreeLing 2.1: Five Years of Open-Source Language Processing Tools", em Nicoletta Calzolari *et al.* (eds.), *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. Valletta: European Language Resources Association (ELRA) (http://www.lsi.upc.edu/~nlp/papers/padro10b.pdf).

Santos, Diana / Nuno Cardoso (eds.) (2007): *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca (http://www.linguateca.pt/LivroHAREM/).

Tjong Kim Sang, Erik F. / Fien de Meulder (2003). "*Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*", em Walter Daelemans / Miles Osborne (eds.), *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*. Edmonton: Association for Computational Linguistics (ACL), 142-147 *(http://acl.ldc.upenn.edu/W/W03/W03-0419.pdf)*.