O Reconhecimento de Entidades Nomeadas por meio de *Conditional Random Fields* para a Língua Portuguesa

Daniela O. F. do Amaral[1], Renata Vieira[1]

[1]Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Caixa Postal 6681 – 90.619-900 – Porto Alegre – RS – Brazil
daniela.amaral@acad.pucrs.br, renata.vieira@pucrs.br

Abstract. Conditional Random Fields (CRF) is a probabilistic method for structured prediction and it has been widely applied in various areas such as Natural Language Processing (NLP), including the Named Entity Recognition (NER), computer vision, and bioinformatics. Therefore, this paper proposes to perform the task of applying the method CRF NER and an evaluation of its performance based on the corpus of HAREM. In summary, the system NERP-CRF achieved the best Precision results when compared to the systems evaluated in the same corpus, proving to be a competitive and effective system.

Resumo. Conditional Random Fields *(CRF) é um método probabilístico de predição estruturada e tem sido amplamente aplicado em diversas áreas, tais como Processamento da Linguagem Natural (PLN), incluindo o Reconhecimento de Entidades Nomeadas (REN), visão computacional e bioinformática. Sendo assim, neste artigo é proposta a realização da tarefa de REN aplicando o método CRF e, sequencialmente, é feita uma avaliação do seu desempenho com base no corpus do HAREM. Conclui-se que, nos testes realizados, o sistema NERP-CRF obteve os melhores resultados de Precisão quando comparado com os sistemas avaliados no mesmo corpus, com plenas condições de ser um sistema competitivo e eficaz.*

1. Introdução

A Extração da Informação (EI) é uma importante tarefa na mineração de texto e tem sido amplamente estudada em vários grupos de pesquisa, incluindo o processamento da linguagem natural, recuperação de informação e mineração na Web. O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa primordial na área de EI, juntamente com a extração de relação entre Entidades Nomeadas (EN) [Jing 2012].

Dentro desse contexto, o REN em textos tem sido amplamente estudado por meio de métodos como aprendizagem supervisionada para classificar entidades do tipo pessoa, lugar e organização em textos ou, ainda, doenças e genes nos resumos das áreas médicas e biológicas [Chinchor et al. 1994]. Esses métodos dependem de recursos caros e extensos para a etiquetagem manual, a qual realiza a identificação das entidades. Os dados etiquetados e o conjunto de *features* extraídas automaticamente são então usados para treinar modelos tais como os Modelos de Markov de Máxima Entropia (MEMMs) [McCallum et al. 2000] ou *Conditional Random Fields* [Lafferty et al. 2001].

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 59–68,
Fortaleza, CE, Brazil, October 21–23, 2013.

Os MEMMs são modelos de uma sequência probabilística condicional, [McCallum et al. 2000], onde cada estado inicial tem um modelo exponencial que captura as características de observação e a distribuição sobre os próximos estados possíveis. Esses modelos exponenciais são treinados por um método apropriado de dimensionamento iterativo no framework de máxima entropia.

O modelo denominado *Conditional Random Fields* (CRF) é um *framework* de modelagem de sequência de dados, que tem todas as vantagens do MEMM e, além disso, resolve o problema do viés dos rótulos. A diferença crítica entre CRF e MMEM é que o MMEM utiliza modelos exponenciais por estados para as probabilidades condicionais dos próximos estados, dado o estado atual. Já o CRF tem um modelo exponencial único para uma probabilidade conjunta de uma sequência de entrada de rótulos, dada uma sequência de observação. Portanto, as influências das diferentes características em estados distintos podem ser tratadas independentemente umas das outras [Lafferty et al. 2001].

Este artigo é estruturado como segue: a Seção 2 elucida o assunto REN e CRF. A Seção 3 expõe uma revisão dos trabalhos relacionados à pesquisa proposta. A Seção 4 descreve o desenvolvimento do sistema NERP-CRF, sua modelagem, implementação e o processo de avaliação. A Seção 5 apresenta os resultados obtidos, bem como a análise de erros. Por fim, a Seção 6 aponta as conclusões e os trabalhos futuros.

2. Reconhecimento de Entidades Nomeadas e *Conditional Random Fields*

O REN consiste na tarefa de identificar as ENs, na sua maioria nomes próprios, a partir de textos de forma livre e classificá-las dentro de um conjunto de tipos de categorias pré-definidas, tais como pessoa, organização e local, as quais remetem a um referente específico [Mota et al. 2007]. Adicionalmente, o REN em textos que abordam os mais variados domínios, além do emprego de extração de relações entre ENs, é uma das tarefas primordiais dentro do trabalho de EI.

Segundo Sureka et al. [Sureka et al. 2009], o REN e a posterior classificação de tais entidades é uma técnica amplamente utilizada no PLN e consiste na identificação de nomes de entidades-chave presentes na forma livre de dados textuais. A entrada para o sistema de extração de entidade nomeada é o texto de forma livre, e a saída é um conjunto das chamadas anotações, ou seja, grupo de caracteres extraídos de trechos do texto de entrada. A saída do sistema de extração de entidades nomeadas é, basicamente, uma representação estruturada a partir da entrada de um texto não estruturado.

As três principais abordagens para extração de entidades nomeadas são: sistemas baseados em regras, sistemas baseados em aprendizado de máquina e abordagens híbridas. Sistemas baseados em regras ou sistemas baseados no conhecimento consistem em definir heurísticas na forma de expressões regulares ou de padrões linguísticos. Sistemas baseados em aprendizado de máquina utilizam algoritmos e técnicas que permitam ao computador aprender.

O objetivo deste trabalho é utilizar o aprendizado de máquina, ou seja, aplicar CRF para REN em textos da Língua Portuguesa e, em sequência, avaliar o desempenho do método com base no corpus do HAREM.

CRF são modelos matemáticos probabilísticos, baseados numa abordagem condicional, utilizados com o objetivo de etiquetar e segmentar dados sequenciais

[Lafferty et al. 2001]. O CRF é uma forma de modelo grafo não direcionado que define uma única distribuição logaritmicamente linear sobre sequências de rótulos, dada uma sequência de observação particular. A vantagem primária dos modelos de CRF sobre outros formalismos, como por exemplo, os *Hidden Markov Model* (HMM) [Lafferty et al. 2001], é a sua natureza condicional, pois resulta no abrandamento de pressupostos sobre a independência dos estados, necessários para os modelos HMM, a fim de assegurar uma inferência tratável.

3. Trabalhos Relacionados

Conforme Chatzis e Demiris [Chatzis e Demiris 2012], durante os últimos anos temos assistido a uma explosão de vantagens nos modelos de CRF, à medida que tais modelos conseguem alcançar uma previsão de desempenho excelente em uma variedade de cenários. Sendo assim, uma das abordagens de maior sucesso para o problema de predição de saída estruturada, com aplicações bem sucedidas, inclui o processamento de texto e áreas como da bioinformática e do processamento da linguagem natural.

A importância de aplicar o CRF para o REN em textos da língua portuguesa deve-se ao fato de que essa técnica de aprendizado de máquina possibilita a extração automática de EN a partir de um grande conjunto de dados com uma capacidade de resposta mais rápida do que outras técnicas já utilizadas, como a implantação de heurísticas ou de sistemas baseados em regras [Mota e Santos 2008]. Além disso, o CRF tem sido muito pouco explorado em corpora do nosso idioma, uma vez que trabalhos que visem o processo de identificação e classificação de EN para o português são raros na literatura.

Dentre outros trabalhos relacionados, destacam-se os de Sutton e McCallum [Sutton e McCallum 2005], Lafferty et al. [Lafferty et al. 2001] e Chatzis e Demiris [Chatzis e Demiris 2012], os quais apresentam um *framework* para a construção de modelos probabilísticos para segmentação e etiquetagem de dados sequenciais baseados em CRF.

O trabalho de Ratinov e Roth [Ratinov e Roth 2009] investigou a aplicação do Reconhecimento de Entidades Nomeadas a partir da necessidade de usar o conhecimento prévio e decisões não locais para a identificação de tais entidades nomeadas em um texto.

O sistema *Hendrix* [Batista et al. 2010] foi elaborado com o propósito de extrair entidades geográficas de documentos em português e produzir o seu resumo geográfico. O processo dividiu-se em três partes: (i) Reconhecer Entidades Geográficas em um documento, ou seja, nomes de ruas, rios, serras, utilizando CRF; (ii) Desambiguar significados geográficos a fim de eliminar nomes idênticos aos extraídos dos textos; (iii) Geração de um resumo geográfico: criar uma lista de entidades geográficas descoberta em uma base de conhecimento externa, por exemplo, em uma ontologia.

4. NERP-CRF

Esta seção descreve o desenvolvimento do sistema denominado NERP-CRF: desde o pré-processamento dos textos, o modelo gerado pelo CRF para o REN até a avaliação empregada.

4.1 Modelagem do Sistema

A elaboração do modelo consiste em duas etapas: treino e teste. Dessa forma, adotamos um corpus, para este trabalho, que é dividido em um conjunto de textos para treino e um conjunto de textos para teste. O corpus adotado foi criado pelo HAREM, evento de avaliação conjunta da língua portuguesa, organizado pela Linguateca [Santos e Cardoso 2008]. Seu objetivo é o de realizar a avaliação de sistemas reconhecedores de EN [Santos 2009]. Entre as edições do HAREM temos: o Primeiro HAREM, decorrido no ano de 2004, e o Segundo HAREM, em 2008. A Coleção Dourada (CD) é um subconjunto da coleção do HAREM, sendo utilizada para tarefa de avaliação dos sistemas que tratam REN. As ENs foram identificadas e classificadas por todos os sistemas participantes do evento, sendo que a sua classificação foi dividida em categorias, tipos e subtipos. Destacam-se para essa pesquisa dez categorias: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro.

Optou-se, especificamente, por trabalhar com os corpora do HAREM, por serem, primeiramente, a principal referência na área, utilizados pela maioria dos trabalhos relacionados ao REN, e, segundo, devido ao fato de eles disponibilizarem um conjunto de textos anotados e validados por humanos (CD), o que facilita a avaliação do método em estudo [Mota e Santos 2008].

Os textos, utilizados como entrada para o NERP-CRF, estão no formato XML com a marcação das entidades e sofreram dois procedimentos, os quais pertencem ao pré-processamento do sistema: primeiro, a etiquetagem de cada palavra por meio do *Part-of-Speech* (POS) *tagging* [Schmid 1994] e segundo, a segmentação em sentenças a fim de que a complexidade seja menor ao aplicar o algoritmo de CRF nos textos de entrada.

Após a conclusão da etiquetagem POS e da segmentação das sentenças, determinou-se como as ENs seriam identificadas. Para tal, foi feito um estudo de duas notações citadas na literatura: BIO e BILOU [Ratinov e Roth 2009]. A primeira possui o seguinte significado: B (*Begin*) significa a primeira palavra da EN; I (*Inside*) uma ou mais palavras que se localizam entre as entidades; e O (*Outside*) a palavra não é uma EN. Já a segunda notação tem a mesma descrição do BIO, acrescentando-se as seguintes particularidades: L (*Last*) a última palavra reconhecida como EN e U (*Unit*) quando a EN for uma única palavra.

Para o presente trabalho, utilizou-se a notação BILOU por dois motivos: (i) Testes aplicados sob a CD do Segundo HAREM, empregando ambas notações, demonstraram que a notação BILOU se equivale a BIO, conforme os resultados apresentados. Isso porque o BILOU facilita o processo de classificação feito pelo sistema desenvolvido por possuir mais duas identificações: L(*Last*) e U(*Unit*); e (ii) os autores [Ratinov e Roth 2009] também fizeram testes com as duas notações, concluindo também com os seus resultados obtidos que, apesar do formalismo BIO ser amplamente adotado, o BILOU o supera significativamente.

Depois da identificação das EN por meio do BILOU, foi gerado o vetor de *features.* Tal vetor corresponde aos dados de entrada que serão aplicados ao sistema de aprendizado do CRF. As *features* têm o objetivo de caracterizar todas as palavras do corpus escolhido para esse processo, direcionando o CRF na identificação e na classificação das ENs. A Tabela 1 apresenta a lista de *features* criadas.

Dois vetores são considerados como entrada para o CRF na etapa de treino: primeiro, o vetor contendo a etiquetagem POS, as categorias estabelecidas pela Conferência do HAREM e a notação BILOU, e segundo, o vetor de *features*.

Na etapa de teste um conjunto de textos é enviado ao NERP-CRF. O referido sistema cria o vetor de POS e o vetor de *features*; envia esses vetores para o modelo de CRF gerado que, por sua vez, treina e classifica as ENs do corpus trabalhado. Por fim, são apresentados aos usuários do sistema as ENs extraídas e as métricas precisão e abrangência. O sistema é concluído com o vetor de saída, o qual classifica o texto com a notação BILOU e com as dez categorias conforme o Segundo HAREM.

Tabela 1. : *Features* implantadas no NERP-CRF.

Features	Descrição das *features*
1) 'tag'	Etiqueta POS de cada palavra de acordo com a sua classe gramatical;
2) 'word'	A própria palavra, ignorando letras maiúsculas e minúsculas;
3) 'prevW'	A palavra anterior, ignorando letras maiúsculas e minúsculas;
4) 'prevT'	Classe gramatical da palavra anterior;
5) 'prevCap'	A palavra anterior totalmente formada por letras minúsculas, formada por letras minúsculas e maiúsculas ou por letras maiúsculas;
6) 'prev2W'	Igual a *feature* 3, porém considerando a palavra que está na posição p-2;
7) 'prev2T'	O mesmo que a *feature* 4, considerando a palavra que está na posição p-2;
8) 'prev2Cap'	Igual a *feature* 5, porém considerando a palavra que está na posição p-2;
9) 'nextW'	A palavra subsequente àquela que está sendo analisada, ignorando maiúsculas e minúsculas;
10) 'nextT'	A classe gramatical da palavra subsequente à que está sendo analisada;
11) 'nextCap'	o mesmo que a *feature* 5, levando em consideração a palavra subsequente àquela que está sendo analisada;
12)'next2W', 'next2T', 'next2Cap'	Semelhante as *features* 3, 4 e 5, mas para a palavra na posição p + 2;
13) 'cap'	O mesmo que a *feature* 5, mas para palavra atual que está sendo analisada;
14) 'ini'	Se a palavra iniciar com letra maiúscula, minúscula ou símbolos;
15) 'simb'	Caso a palavra seja composta por símbolos, dígitos ou letras.

4.2 Descrição dos Testes Realizados

Dois testes foram realizados utilizando o sistema NERP-CRF, com as seguintes características:

‘Teste 1’: empregou a CD do Segundo HAREM para treinar e testar o modelo de CRF, o qual faz a classificação de dez categorias: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro. A avaliação do desempenho do modelo treinado para o “teste 1 utilizou a técnica de *Cross Validation* [ARL10], com cinco repetições (*5 – fold cross validation).* Trabalhou-se com 5 *folds* porque foi empregado uma pequena quantidade de textos, 129, para os testes iniciais. Dado o conjunto de textos da CD do Segundo HAREM, utilizou-se a cada *fold,* 80% do conjunto de textos para treino e 20% para teste, de modo que a cada repetição do *Cross Validation,* não se empregasse o mesmo conjunto de teste das *folds* anteriores e assim, não reduzisse, significativamente, o número de casos para teste. A finalidade de executar esse experimento foi para verificar o desempenho do NERP-CRF utilizando apenas o corpus citado.

‘Teste 2’: caracteriza-se por trabalhar com a CD do Primeiro HAREM para treino, a qual abrange 129 textos e a CD do Segundo HAREM para teste formada por mais 129 textos. O novo corpus recebe a classificação do CRF abordando as dez categorias do HAREM, citadas no “Teste 2”. Essa estrutura foi arquitetada com o objetivo de verificar o desempenho do CRF em um maior número de textos e avaliá-lo perante os resultados obtidos por ele com os outros sistemas participantes do Segundo HAREM (Tabela 2).

5. Resultados

A comparação dos resultados do NERP-CRF com os sistemas que participaram da Conferência do Segundo HAREM foram obtidos por meio do SAHARA [Mota e Santos 2008], o qual determinou as métricas Precisão, Abrangência e Medida-F a cada um deles nas tarefas de reconhecimento e classificação de EN. O NERP-CRF, no ‘Teste 1’, apresentou o melhor resultado para Medida-F (57,92%) em relação aos outros sistemas. Esse resultado é tendencioso uma vez que utilizamos um único corpus para treino e teste, apesar de validá-lo com *Cross-validation.*

Com a finalidade de resolver esse problema, realizamos o ‘Teste 2’, o qual apresentou 80,77% de Precisão como o melhor resultado do NERP-CRF (Tabela 2). A Medida-F ocupou a terceira posição em relação aos sistemas em comparação, 48,43%. Essa última métrica não alcançou a melhor posição como no ‘Teste 1’ devido a uma baixa Abrangência de classificação, 34,59%.

A desigualdade dos resultados entre os dois testes ocorreu, principalmente, por dois motivos: a mudança do corpus de treino e de validação além do número reduzido de exemplos para determinadas categorias, por exemplo, Coisa, Abstração. Isso faz com que o CRF treine menos com essas categorias e gere um modelo menos abrangente para elas. Nesse cenário, consideram-se os nossos resultados muito positivos, principalmente no que tange ao valor de Precisão alcançado pelo NERP-CRF.

Tabela 2. NERP-CRF comparado com os sistemas apresentados para o 'Teste 2'.

Sistemas	Precisão	Abrangência	Medida-F
NERP-CRF	80,77%	34,59%	48,43%
Priberam	64,17%	51,46%	57,11%
R3M	76,44%	25,20%	37,90%
Rembrandt	64,97%	50,36%	56,74%
SEI-Geo	74,85%	11,66%	20,17%
CaGE	44,99%	27,57%	34,19%

5.1. Análise de Erros

Com base em uma análise dos textos utilizados como entrada para testar o CRF, constata-se que o sistema, tanto para o 'Teste 1' quanto para o 'Teste 2', não identificou determinadas ENs ou não as identificou corretamente. Percebeu-se que a má formatação de alguns textos, como por exemplo, a falta de pontuação e a anotação incorreta pelo *POS tagger* afetou significativamente os resultados. A delimitação errônea de ENs, como em "Ministério da Cultura", marcado pelo CRF como BIU, mas identificado pela CD como B I L, prejudicou também o resultado do sistema. Outro erro em destaque foi a não identificação da preposição 'de' e suas combinações com artigos, como I (*Inside*), no caso de ENs compostas, como "Fenando de Bulhões" e "Igreja dos Mártires".

Outro ponto relevante foram os erros de classificação das ENs. Podemos citar as siglas "RF" e "IFF", consideradas como ENs, deveriam ter sido classificadas como "Coisa", porém o sistema considerou-as como "Organização". As palavras estrangeiras sofreram o mesmo tipo de erro, como por exemplo, a EN "*Friendly*" que foi classificada como "Local", ao passo que deveria ter recebido "Abstração" como classificação correta. Percebeu-se também que houve pouco contexto para classificar corretamente certas ENs, como por exemplo, a categoria "Abstração", a qual tem pouca exemplificação na CD anotada. Além disso, são ENs que não seguem padrão algum de escrita, ou seja, não há uma sintaxe própria para essa categoria que faça com que o sistema aprenda corretamente a identificá-la. Já a categoria "Tempo" apresenta-se num formato que a identifica com mais clareza, isto é, possui um padrão bem rígido de sintaxe como <um número> de <outro número>, indicando data, ou até mesmo outras palavras indicativas de tempo como "desde", "enquanto" e "quando". Mesmo assim, o sistema teve dificuldade de classificá-la, pois esse tipo de EN pode não iniciar com letra maiúscula, o que prejudicou o aprendizado feito pelo NERP-CRF.

6. Conclusões e Trabalhos Futuros

CRF oferece uma combinação única de propriedades: modelos treinados para etiquetar e segmentar sequências de dados; combinação de arbitrariedade, *features* de observação aglomeradas, decodificação e treinamento eficiente baseado em programação dinâmica e

estimativa de parâmetro garantida para encontrar o ótimo global [Lafferty et al. 2001] [Ratinov e Roth 2009].

O NERP-CRF foi o sistema desenvolvido para realizar duas funções: a identificação de ENs e a classificação dessas com base nas dez categorias do HAREM: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro.

Dois testes foram realizados. Um deles utilizou a CD do Segundo HAREM para treino e teste, obtendo Medida-F de 57,92%. Um outro teste empregou a CD do Primeiro HAREM para treinar o modelo de CRF e a CD do Segundo HAREM para testar o mesmo modelo gerado. Nesse caso, as métricas obtidas foram: 80,77% de Precisão, 34,59% de Abrangência e 48,43% de Medida-F. A Precisão foi o melhor resultado quando comparado com os outros sistemas. Já a Medida-F apresentou o terceiro melhor resultado, ficando abaixo dos sistemas Priberam e Rembrandt, que apresentaram maior abrangência. O modelo proposto, baseado em CRF e no conjunto de *features* estabelecidas, gerou um sistema eficaz, competitivo, sendo ainda passível de fácil adaptação e modificação.

Os trabalhos futuros, os quais podem dar melhoria aos resultados apresentados, determinam-se em duas abordagens de pesquisa: algoritmos de indução de *features* e classificação de EN consideradas ambíguas. O CRF pode implementar, eficientemente, a seleção de *features* e de algoritmos de indução de *features.* Isso quer dizer que, ao invés de especificar antecipadamente quais *features* serão utilizadas, pode-se iniciar a partir de regras que geram *features* e avaliam o benefício dessas geradas automaticamente sobre os dados [Lafferty et al. 2001].

Outra abordagem de pesquisa futura é a classificação correta de uma mesma EN apresentada de formas diferentes, por exemplo: a EN ‘Pontifícia Universidade Católica do Rio Grande do Sul’ pode receber a mesma classificação ou ser categorizada como Organização e Local, dependendo do contexto no qual essas entidades estão inseridas. Outra situação que pode ocorrer é que quando as ENs ‘Pontifícia Universidade Católica do Rio Grande do Sul’ e ‘PUCRS’ são a mesma entidade e, portanto, devem receber a mesma classificação. As soluções para a correta categorização de EN nesse caso pode ser a aplicabilidade, como por exemplo, da Correferência [Black er al. 1998] [Lee et al. 2011] e de recursos externos, como o emprego de *Gazetters* [Ratinov e Roth 2009].

Referências

Batista, S.; Silva, J.; Couto, F. e Behera, B. (2010) "Geographic Signatures for Semantic Retrieval", In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, ACM, p.18-19.

Black, W. J., Rinaldi, F. e Mowatt, D. (1998) "Facile: Description of the NE system used for MUC-7", In *Proceedings of the 7th Message Under-standing Conference (MUC-7)*.

Chinchor, N.; Hirschman, L. e Lewis, D. (1994) "Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3)", In *Computational Linguistics*, p. 409-449.

Chatzis, Sotirio P. e Demiris, Yiannis. (2012) "The echo state conditional random field model for sequential data modeling", In *International Journal of Expert Systems with Applications.*

Jing, J. (2012) "Information extraction from text", In *Mining Text Data*, p. 11-41.

Lafferty, J.; McCallum, A. e Pereira, F. (2001) "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In *Proceedings of the 18th International Conference on Machine Learning.*

Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M. e Jurafsky, D. (2011) "Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task", In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task,* p. 28-34.

Mansouri, A.; Affendey, Lilly S. e Mamat, A. (2008) "Named Entity Recognition Approache", In *International Journal of Computer Science and Network Security*, Vol. 8 Nº.2.

McCallum, A.; Freitag, D. e Pereira, F. (2000) "Maximum entropy Markov models for information extraction and segmentation", In *International Conference on Machine Learning.*

Mota, C.; Santos, D. e Ranchhod, E. (2007) "Avaliação de reconhecimento de entidades mencionadas: Princípio de Harem", In *Diana Santos, editor, Avaliação Conjunta: Um novo paradigma no processamento computacional da língua portuguesa*, capítulo 14, IST Press, p. 161–176.

Mota, C. e Santos, D. (2008) "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM", http://www.linguateca.pt/LivroSegundoHAREM/, Dezembro.

Nadeau, D. e Sekine, S. (2007) "A survey of named entity recognition and classification", In *Journal Linguistica e Investigationes, National Research Council Canada,* Vol. 30, p. 3-26.

Pinto, D.; McCallum, A.; Wei, X. e Croft, W. B. (2003) "Table extraction using conditional random fields". In *Proceedings of the ACM SIGIR.*

Ratinov, L. e Roth, D. (2009) "Design Challenges and Misconceptions in Named Entity Recognition", In *Proceedings of the 13th Conference on Computational Natural Language Learning.*

Santos, D. e Cardoso, N. (2008) "Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área", http://www.linguateca.pt/aval_conjunta/LivroHAREM/, Dezembro.

Santos, D. (2009) " Caminhos percorridos no mapa da portuguesificação: A linguateca em perspectiva", http://www.linguateca.pt/Diana/download/Santos2009Linguamatica.pdf, Dezembro.

Schmid, H. (1994) "Probabilistic part-of-speech tagging using decision tree", In *Proceedings of the International Conference on New Methods in Language Processing*.

Sureka, A.; Mirajkar, P. P. e Varma, K. I. (2009) "Polarity Classification of Subjective Words Using Common-Sense Knowledge-Base", In *Proceedings of the 2nd Bangalore Annual Compute Conference, ACM.*

Sutton, C. e Mccallum, A. (2005) " Piecewise training for structured prediction", In *Conference on Uncertainty in Artificial Intelligence*, p. 165-194.