NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de *Conditional Random Fields*

A tool for the named entity recognition using conditional random fields

Daniela Oliveira F. do Amaral
Pontifícia Universidade Católica
do Rio Grande do Sul
daniela.amaral@acad.pucrs.br

Renata Vieira
Pontifícia Universidade Católica
do Rio Grande do Sul
renata.vieira@pucrs.br

Resumo

Conditional Random Fields (CRF) é um método probabilístico de predição estruturada que tem sido amplamente aplicado em diversas áreas, tais como a de Processamento da Linguagem Natural (PLN), incluindo o Reconhecimento de Entidades Nomeadas (REN), visão computacional e bioinformática. Nesse sentido, propõe-se 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.

Palavras chave

Reconhecimento de Entidades Nomeadas, Conditional Random Fields, Processamento da Linguagem Natural, Língua Portuguesa.

Abstract

Conditional Random Fields (CRF) is a probabilistic method for structured prediction which 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.

Keywords

Named Entity Recognition, Conditional Random Fields, Natural Language Processing.

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 nos de processamento da linguagem natural, de recuperação de informação e de 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, 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, 2000) ou *Conditional Random Fields* (Lafferty, 2001).

Os MEMMs são modelos de uma sequência probabilística condicional, (McCallum, 2000), em que 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 a partir 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, 2001). Os resultados da Conferência Internacional de Aprendizado de Máquina (*International Conference on Machine Learning* - ICML) no ano de 2001 (LAF01), seguido de outros trabalhos sobre *Conditional Random Fields* (Suakkaphong, 2011), (Lee, 2011), (Lishuang, 2011), indicam que o algoritmo de CRF apresenta um dos melhores desempenhos para o REN.

Sendo assim, o método escolhido foi o CRF e o corpus que receberá a classificação por ele é o do HAREM. O HAREM é um evento de avaliação conjunta da língua portuguesa, organizado pela Linguateca (Santos, 2007). Seu objetivo é o de realizar a avaliação de sistemas reconhecedores de ENs (Santos, 2009). Entre as edições do HAREM temos: o Primeiro HAREM, ocorrido 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. O corpus do HAREM é considerado a principal referência na área de PLN, e caracteriza-se por ter um conjunto de textos anotados e validados por humanos (CD), o que facilita a avaliação do método em estudo.

Com isso, este trabalho teve como motivação o fato de: (i) o REN ter sido pouco explorado utilizando o método de aprendizagem supervisionada CRF para a língua portuguesa; (ii) não existirem propostas de REN aplicando o CRF para identificar as ENs e classificá-las de acordo com as dez categorias do HAREM; e (iii) o método de CRF poder ajudar a identificar um maior número de ENs, o que poderá ser verificado por meio da comparação com outros sistemas.

Portanto, o objetivo geral do presente artigo é utilizar o aprendizado de máquina, ou seja, aplicar CRF para a tarefa de REN em corpora da língua portuguesa e avaliar comparativamente o desempenho desse método com outros sistemas que realizam REN, tendo como base o corpus do HAREM.

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. Sequencialmente, a análise de erros é efetuada na Seção 6. Por fim, a Seção 7 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, 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 (2009), o REN 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 um texto de forma livre, e a saída é um conjunto desses textos anotados, ou seja, uma representação estruturada a partir da entrada de um texto não estruturado.

As três principais abordagens para extração de ENs 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.

Já o CRF são modelos matemáticos probabilísticos, baseados numa abordagem condicional, utilizados com o objetivo de etiquetar e segmentar dados sequenciais (Lafferty, 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 os *Hidden Markov Model* (HMM) (Lafferty, 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

Os trabalhos de Sutton e McCallum (2005), Lafferty (2001) e Chatzis e Demiris (2012), apresentam um *framework* para a construção de modelos probabilísticos para segmentação e etiquetagem de dados sequenciais baseados em CRF. Nesse sentido, temos assistido, durante os últimos anos, a uma explosão de vantagens nos modelos de CRF (Chatzis, 2012), à medida que tais

modelos conseguem alcançar uma previsão de desempenho excelente em uma variedade de cenários. Sendo assim, o processamento de texto por meio da técnica de aprendizado de máquina CRF, é uma das abordagens de maior sucesso para o problema de predição de saída estruturada, com aplicações bem sucedidas em áreas como a bioinformática e o processamento da linguagem natural (PLN). Não obstante, o trabalho de Ratinov e Roth (2009) aponta que o REN pode ser obtido a partir de modelos de classes de palavras, os quais aprendem a partir de rótulos não estruturados. Os autores investigaram a aplicação de REN a partir da necessidade de usar o conhecimento prévio e decisões não locais para a identificação de tais ENs em um texto. Logo, esse modelo que detecta e classifica ENs pode ser uma alternativa para o paradigma de aprendizado supervisionado tradicional como o CRF.

Existem diversos trabalhos que também usam CRF e outras abordagens estatísticas para extração de informação textual em PLN e, especificamente, para a tarefa de REN (Finkel, 2005; McCallum e Li 2003).

Sendo assim, a importância de aplicar o CRF para o REN, especialmente, 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. Além disso, o CRF tem sido muito pouco explorado em corpora do nosso idioma, uma vez que trabalhos que visam o processo de identificação e de classificação de EN para o português são raros na literatura. O sistema *Hendrix* (Batista, 2010), por exemplo, 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) gerar um resumo geográfico por meio da criação de uma lista de entidades geográficas descoberta em uma base de conhecimento externa, por exemplo, em uma ontologia.

Tanto quanto o sistema *Hendrix*, os sistemas Priberam ao HAREM, R3M, REMBRANDT, SEI-Geo e CaGE realizam REN para textos da língua portuguesa (Mota, 2008). Com exceção do Hendrix os demais sistemas participaram da trilha do Segundo HAREM e foram comparados com o sistema que desenvolvemos para este trabalho. O Priberam ao HAREM é baseado em um léxico com classificação morfossintática e semântica. Cada entrada do léxico corresponde a uma ligação com um ou mais níveis de uma ontologia multilíngue (Amaral, 2004), podendo corresponder a um ou mais sentidos, os quais possuem diferentes valores morfológicos e semânticos. Para a construção do sistema foram utilizadas regras contextuais. As regras para a tarefa de REN consideram as sequências de nomes próprios, separadas ou não por algumas preposições e o contexto em que as EN são encontradas. Por exemplo, uma EN "João Pedro", classificada como Pessoa, poderá ser classificada como Organização se esta for precedida por uma expressão como "instituto".

Já o R3M aplica aprendizagem supervisionada, utilizando um algoritmo de co-training para inferir regras de classificação (Collins, 1999) no REN. A escolha do algoritmo de *co-training* deve-se ao fato de que este tem grande probabilidade de obter bons resultados de classificação que se aproximam dos 80% de precisão, usando um número muito reduzido de exemplos previamente anotados. As ENs que o R3M classifica compreendem as categorias Pessoa, Organização e Local. A opção por essas três categorias deve-se ao fato de que essas, de uma forma geral, têm sido estudadas mais amplamente dentro da área de extração da informação. Além disso, os desenvolvedores do R3M não tiveram disponibilidade de dedicar mais tempo a esse sistema. Mesmo assim, o R3M foi projetado de modo que permita estender-se ao reconhecimento de outras categorias, assim como incluir o reconhecimento de relações de EN. Esse sistema é uma reimplementação do sistema criado por Mota (Mota, 2009), apresentando várias melhorias.

Além da tarefa de REN realizada pelos sistemas REMBRANDT e SEI-Geo, ambos detectam o Reconhecimento de Relações entre ENs (ReRelEN). O REMBRANDT - Reconhecimento de ENs Baseado em Relações e Análise Detalhada do Texto - por sua vez, utiliza a Wikipédia como base de conhecimento a fim de classificar as ENs, além de um conjunto de regras gramaticais para extrair o seu significado. O REMBRANDT surgiu da necessidade de se criar um sistema de marcação de textos que indique as ENs relacionadas a locais geográficos de forma semântica, como por exemplo, nomes de países, rios, universidades. Seu funcionamento divide-se em três fases primordiais: 1) o reconhecimento de expressões numéricas e geração de candidatas a EN; 2) a classificação de EN e 3) repescagem de ENs sem classificação. Já o SEI-Geo tem o objetivo de fazer REN classificando somente a categoria Local e suas relações. Dentre as características que compõem o SEI-Geo

destacam-se: (i) a incorporação na arquitetura global do sistema *GKBGeographic Knowlegde Base*, o qual estabelece o gerenciamento de conhecimento geográfico; e (ii) a utilização das Geo-ontologias, que exploram as relações entre locais identificados em textos a partir de relações presentes na ontologia. O domínio Organização ajudou significativamente o bom desenvolvimento do SEI-Geo no reconhecimento de relações entre ENs, pois, nos textos, Locais estão situados próximos a Organizações.

Por fim, o sistema CaGE trata do problema do reconhecimento e desambiguação de nomes de locais. Essa é uma tarefa muito importante na geo-codificação de documentos textuais (Martins, 2009). O objetivo principal do sistema CaGE é atribuir a área geográfica e o âmbito temporal aos documentos de modo geral, combinando a informação diferente extraída do texto. As categorias que tal sistema classifica são: Pessoa, Local, Organização e Tempo. O CaGe caracteriza-se por ser um método híbrido, o qual utiliza dicionários e regras de desambiguação. Quatro etapas resumem uma sequência de operações de processamento que compõem o algoritmo do sistema: 1) identificação inicial das ENs; 2) classificação das entidades mencionadas e tratamento da ambiguidade; 3) desambiguação completa de entidades geográficas e temporais; e 4) atribuição de âmbitos geográficos e temporais aos documentos.

Dessa forma, o nosso trabalho difere dos demais na aplicabilidade do modelo de CRF, o qual vem demonstrando bons resultados frente a outros métodos que utilizam aprendizado de máquina para a tarefa de REN. Além disso, a literatura apresenta muitos poucos trabalhos que identificam e classificam ENs, utilizando as dez categorias do HAREM, em corpus da língua portuguesa por meio de modelos probabilísticos.

4 NERP-CRF

Esta seção descreve o desenvolvimento do sistema NERP-CRF (Amaral, 2013) 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 que é dividido em um conjunto de textos para treino e um conjunto de textos para teste. A CD do HAREM foi o corpus utilizado 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.

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) (Bick, 2000) 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, 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 as notações, demonstraram que a notação BILOU se equivale à 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, 2009) também fizeram testes com as duas notações, concluindo também 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.

Features	Descrição das features
1) tag	Etiqueta POS de cada palavra de acordo com a sua classe gramatical. Ex.: artigo, adjetivo, verbo.
2) word	A própria palavra do texto, ignorando letras maiúsculas e minúsculas;
3) prevW	A palavra anterior a que está sendo analisada no texto, ignorando letras maiúsculas e minúsculas.
4) prevT	Classe gramatical da palavra anterior. Ex.: artigo, adjetivo, verbo.
5) prevCap	A palavra anterior totalmente formada por letras minúsculas, formada por letras minúsculas e maiúsculas ou por letras maiúsculas. Cada uma dessas palavras pode receber um dos atributos: 'min', 'maxmin' ou 'max'.
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. Essas palavras podem receber um dos atributos: 'max', 'min' ou 'sim'.
15) simb	Caso a palavra seja composta por símbolos, dígitos ou letras. Tais palavras recebem o atributo 'alfa'.

Tabela 1: Features implementadas no NERP-CRF.

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 (a) cria o vetor de POS e o vetor de *features*; (b) envia esses vetores para o modelo de CRF gerado que, por sua vez, (c) treina e (d) 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.

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. A avaliação do desempenho do modelo treinado para o 'teste 1' utilizou a técnica de *Cross Validation* (Arlot, 2010), com cinco repetições (*5 – fold cross validation)*. Trabalhou-se com 5 *folds* porque foi empregada uma pequena quantidade de textos, 129, para os testes iniciais, incluindo 670.610 palavras. Esse procedimento resultou em 7.610 ENs identificadas pelo NERP-CRF num valor máximo de 17.767 ENs identificadas por humanos nessa mesma CD. 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. Os dois conjuntos somam 258 textos e aproximadamente 804.179 palavras. O novo corpus recebe a classificação do CRF abordando as dez categorias do HAREM, citadas no "Teste 1". Essa estrutura foi arquitetada com o objetivo de verificar o desempenho do NERP-CRF em um maior número de textos e avaliá-lo perante os resultados obtidos por ele com os outros sistemas participantes do Segundo HAREM.

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, 2008), o qual determinou as métricas Precisão, Abrangência e Medida-F a cada um deles nas tarefas de REN.

O NERP-CRF, no 'Teste 1', apresentou os melhores resultados para as medidas de Precisão e de Medida-F em relação aos outros sistemas, respectivamente, 83,48% e 57,92% (Tabela 2) (Figura 1). Esse resultado é baseado em um único corpus para treino e teste, apesar de validá-lo com *Cross-validation*.

Sistemas	Precisão	Abrangência	Medida-F
NERP-CRF	83,48%	44,35%	57,92%
Priberam	64,17%	51,46%	57,11%
Rembrandt	64,97%	50,36%	56,74%
R3M	76,44%	25,20%	37,90%
CaGE	44,99%	27,57%	34,19%
SEI-Geo	74,85%	11,66%	20,17%

Tabela 2: Resultados do NERP-CRF comparado com os sistemas apresentados para o 'Teste 1'.

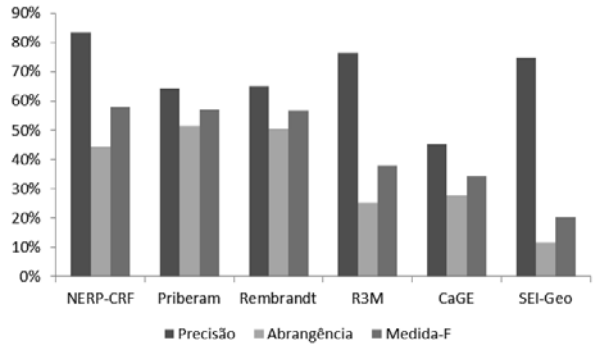

Figura 1: Desempenho do NERP-CRF comparado graficamente com os sistemas no 'Teste 1'.

Com a finalidade de ver o comportamento do aprendizado em outro corpus, realizamos o 'Teste 2', o qual apresentou 80,77% de Precisão como o melhor resultado do NERP-CRF (Tabela 3). 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% (Figura 2).

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 e Abstração. Isso fez com que o NERP-CRF treinasse menos com essas categorias e gerasse um modelo menos abrangente. Nesse cenário, consideram-se os nossos resultados muito positivos, principalmente no que tange ao valor de Precisão alcançado pelo NERP-CRF.

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

Tabela 3: Resultados do NERP-CRF comparado com os sistemas apresentados para o 'Teste 2'.

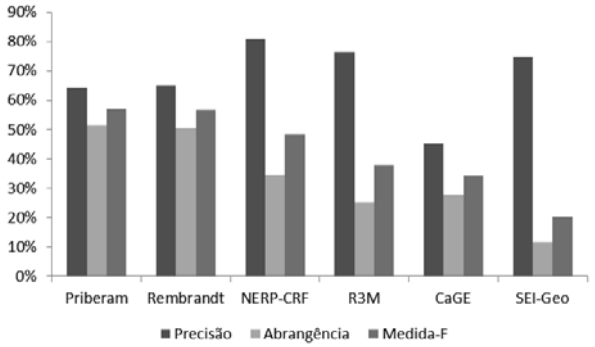

Figura 2: Desempenho do NERP-CRF comparado graficamente com os sistemas no 'Teste 2'.

A Tabela 4 reporta os resultados das métricas de Precisão, Abrangência e Medida-F de acordo com as dez categorias estabelecidas pelo evento do HAREM. Logo, pode-se observar de acordo com os valores da Medida-F, que o NERP-CRF classificou melhor as categorias Tempo, Pessoa, Valor e Local. Em contra partida, houve um maior número de erros com as categorias Abstração, Outra, Coisa e Acontecimento, devido a poucos exemplos no corpus de treino.

CATEGORIAS	Abrangência	Precisão	Medida-F
TEMPO	68,05%	83,99%	75,18%
PESSOA	71,89%	61,57%	66,33%
VALOR	54,42%	78,23%	64,19%
LOCAL	57,22%	52,06%	54,51%
ORGANIZACAO	43,44%	44,75%	44,08%
OBRA	28,48%	40,71%	33,52%
ACONTECIMENTO	22,76%	50,83%	31,44%
COISA	7,36%	26,80%	11,55%
OUTRA	4,74%	43,49%	8,55%
ABSTRAÇÃO	3,65%	16,45%	5,97%

Tabela 4: Resultados com as dez categorias.

6 Análise de Erros

Com base em uma análise dos textos utilizados como entrada para testar o NERP-CRF, constata-se que o sistema, tanto para o 'Teste 1' quanto para o 'Teste 2', não identificou determinadas ENs ou não as classificou corretamente. A Tabela 4 apresenta

alguns erros encontrados após a execução do NERP-CRF. A notação apresentada (Tabela 5) pela saída desse sistema refere-se ao POS tagger de cada EN, seguido da notação BILOU e da classificação dessas entidades. Por exemplo, substantivo, etiquetado como <n>; preposição <prp>; nome próprio <prop>; verbo finito <v-fin>; numeral <num>; artigo <art>. Quanto à identificação e a classificação das ENs o NERP-CRF apresentou a notação conforme alguns exemplos: <I-Obra> EN identificada como Inside e classificada como Obra; <L-Obra> Last e Obra; < B-Pessoa> Begin e classificação Pessoa; <U - Org> Unit e Organização. As ENs que não foram identificadas e não receberam classificação, foram marcadas pelo sistema como: <O-OUT>.

Percebeu-se que a má formatação de alguns textos, como por exemplo, a falta de pontuação e a anotação incorreta pelo POS tagger afetaram os resultados. A delimitação errônea de ENs, como em "Diário de Notícias", marcado pelo NERP-CRF como O I L, mas identificado pelo corpus de referência como B I L, prejudicou também o resultado do sistema. Outro erro em destaque foi a não identificação da preposição 'de' e de suas combinações com artigos, como I (Inside), no caso de ENs compostas, como "Fenando de Bulhões" e "Igreja dos Mártires". Esses erros podem ser sanados com a aplicação de algoritmos de classificação como o de Viterbi (Finkel, 2005), abordagem utilizada em ferramentas com propósitos similares (FreeLing User Manual, 2013). Outra alternativa seria o AdaBoosting (Carreras, 2003).

NERP-CRF	CD do HAREM
Diário <n, O-OUT> **de** <prp, I-Obra> **Notícias** <n, L-Obra>	**Diário**<n, B-Org> **de** <prp, I-Org> **Notícias** <n, L-Org>
Fernando<prop,B-Pessoa> **de** <prp, O-Out **Bulhões** <prop, L-Local>	**Fernando**<prop,B-Pessoa> **de** <prp, I-Pessoa> **Bulhões** <prop, L-Pessoa>
Igreja <v-fin, O -OUT> **dos** <n, O - OUT> **Mártires**<prop,U-Pessoa>	**Igreja** <v-fin, B -Local> **dos** <n, I - Local> **Mártires** <prop, L -Local>
RF <prop, U- Org>	**RF** <prop, U-Coisa>
IFF <prop,U- Org>	**IFF** <prop, U-Coisa>
Friendly <prop, U-Local>	**Friendly** <prop,U- Abstração>
em <prp, O-OUT> **1973** <num, U-Tempo>	**em** <prp, B-Tempo **1973** <num, L-Tempo>
desde <prp, O-OUT> **os** <art, I-Tempo> **anos** <n, I-Tempo> **1990** <num, L-Tempo>	**desde** <prp, B-Tempo> **os** <art, I-Tempo> **anos** <n, I-TempoO> **1990** <num, L-Tempo>

Tabela 5: Alguns erros apresentados pelo NERP-CRF.

Outro ponto relevante foram os erros de classificação das ENs. Podemos citar as siglas "RF" e "IFF", consideradas como ENs, as quais deveriam ter sido classificadas como "Coisa", porém o sistema considerou-as como "Organização". As palavras estrangeiras sofreram o mesmo tipo de erro, como 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 ocorreu com a categoria "Abstração", a qual tem pouca exemplificação no corpus de referência. 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. Por exemplo, na EN "em 1973", a preposição "em" não foi identificada como EN. O correto seria que o NERP-CRF a tivesse classificado como B-Tempo. Situação semelhante também ocorreu com outra EN de Tempo, "desde os anos 1990". O sistema não reconheceu a preposição "desde" como EN e, consequentemente, não a classificou.

7 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, 2001) (Ratinov, 2009).

O NERP-CRF foi o sistema desenvolvido para executar 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%. 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.

A análise de erros mostrou que o NERP-CRF precisa melhorar a identificação e a classificação das EN. Dentre os erros que ocorreram, aqueles mais frequentes foram: marcação pela notação BILOU, erros de classificação entre as categorias Local e Pessoa, classificação de sigla e de palavras estrangeiras, identificação e classificação de EN de Tempo.

Tomando por base a análise de erros, sugere-se um trabalho futuro com experimentos que utilizem algoritmos de meta aprendizagem, como combinação de classificadores, para aumentar a efetividade do NERP-CRF. Resultados interessantes baseados em anotações BIO foram obtidos com o uso de AdaBoosting (Carreras et al. 2003). A atual versão do NERP-CRF já utiliza anotações BILOU, logo, acredita-se que tanto a abrangência como a precisão do processo proposto possa ser melhorada com esse tipo de abordagem. Especificamente, busca-se melhorar a qualidade da anotação BILOU, induzir *features* e classificar ENs consideradas ambíguas. Adicionalmente, sugere-se também experiências com outros *parsers* e eventual comparações com o desempenho obtido para outras línguas.

Acredita-se que o teste com outros *parsers* possibilitará um melhor resultado de Abrangência pelo NERP-CRF. O FreeLing (Padró, 2010) e o PALAVRAS (Bick, 2000) são os *parsers* que serão utilizados para etiquetar o mesmo corpus empregado na fase de pré-processamento.

O CRF pode implementar, eficientemente, a seleção de *features* e de algoritmos de indução de *features*. Isso quer dizer que, em vez 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, 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 ENs que possuem como acrônimo, 'Pontifícia Universidade Católica do Rio Grande do Sul' e 'PUCRS' devem ser identificadas como a mesma entidade. Portanto, essas devem receber a mesma classificação. As soluções para a correta categorização de ENs, nesse caso, pode ser a aplicabilidade, como da Correferência (Black, 1998) (Lee, 2011) e de recursos externos, como o emprego de *Gazetters* (Ratinov, 2009).

Referência

Amaral, C.; Figueira, H.; Mendes, A.; Mendes, P.; Pinto, C. 2004. A workbench for developing natural language processing tools. In: *1st Workshop on International Proofing Tools and Language Technologies*, Patras, Greece, July 1-2.

Amaral, D.O.F. 2012. O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa. M.Sc. dissertation, PUCRS.

Arlot, S.; Celisse, A. 2010. A survey of cross-validation procedures for model selection. Statistics Surveys, vol. 04, p. 4, 40.

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.

Bick, E. 2000. The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. *Aarhus University Press.*

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 Understanding Conference (MUC-7)*.

Carreras, X.; Màrquez, L.; Padró, L. 2003. A simple named entity extractor using adaboost. In *Proceedings of CoNLL-2003 Shared Task Edmonton, Canada*.

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.*

Collins, M.; Singer, Y. 1999. Unsupervised models for named entity classification. In: *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, p.100–110.

Finkel, Jenny R.; Grenager, T.; Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL*, p. 363–370.

FreeLing User Manual, October 2013. In *http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf*

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.

Lishuang L.; Degen H.; Dan L. 2011. Recognizing Chinese Person Names based on Hybrid Models. *Advanced Intelligence*, vol. 3: 219-228.

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.

Martins, B. 2009. Geographically aware Web text mining. Tese de Doutorado, Faculdade de Ciências, Universidade de Lisboa, p. 155-157.

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

McCallum, A.; Li, W. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of 7th conference on natural language learning, CoNLL.*

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.

Mota, C. 2009. How to keep up with language dynamics: A case study on named entity recognition. Tese de doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Padró, L.; Collado, M.; Reese, S.; Lloberes, M.; Castellon, I. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference, LREC, ELRA, La Valletta, Malta.*

Ratinov, L.; Roth, D. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In: *13th Conference on Computational Natural Language Learning*, CONLL, p. 147-155.

Santos, D.; Cardoso, N. 2007. Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, capítulo 1, p. 1–16.

Santos, D.; Cabral, L. M. 2009. GikiCLEF: Crosscultural issues in an international setting: asking non-english-centered questions to wikipedia. *Cross Language Evaluation Forum: Working notes for CLEF*.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In: *International Conference on New Methods in Language Processing*, 1994, p. 44-49.

Suakkaphong, N.; Zhang, Z.; Chen, H. 2011. Disease Named Entity Recognition Using Semisupervised Learning and Conditional Random Fields. *Journal of the American Society for Information Science and Technology*, vol. 62, p. 727-737.

Sureka, Ashish S.; Pranav, P. M.; Kishore, I. V. 2009. Polarity Classification of Subjective Words Using Common-Sense Knowledge-Base. In: *12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, p. 486-493.