Extração de Entidades Nomeadas Com Maximização de Entropia (Opennlp)

Richerland Pinto Medeiros [1], Patrícia Bellin Ribeiro.

[1]Curso de Tecnologia em Banco de Dados - Faculdade de Tecnologia de Bauru (FATEC)

Rua Manoel Bento da Cruz, nº 30, Quadra 3 - Centro - 17.015-171 - Bauru, SP – Brasil

[1]richerlandmedeiros@fatec.sp.gov.br, patriciabellin@yahoo.com.br

Abstract. *Through the advent of the technological inclusion, added to the emergence of social networks, the volume of information in the form of text grew dramatically over the last few years and information from these data mass become an interesting strategy tool. This data presents itself on a non-organized form, meaning that the use of such valuable information is hampered by the difficulty to interpret the real information inserted into these masses. The present article aims to present the maximization entropy statistical method for the extraction of named entities enabling tabulation of characteristics referenced by entities in data mass. It has been used for the tests as training base an extraction model for public corpora Amazonia and FlorestaVirgem, both in the Árvores Deitadas format. Is concluded that the statistical machine learning approach for information extraction is efficient when is considered the corpus trained on specific domain text.*

Resumo. *Com o advento do aumento da inclusão tecnológica, somado ao aparecimento das redes sociais, o volume de informação textual cresceu expressivamente nos últimos anos e, com isso, a possibilidade de uso de informações proveniente dessas massas de dados tem se mostrado uma interessante ferramenta estratégica. Tais dados se apresentam de forma desestruturada, ou seja, o uso dessas valiosas informações é dificultado pela complexidade de interpretação da real informação inserido nessas massas de dados. O presente artigo visa apresentar a técnica estatística de maximização de entropia, para a extração de entidades nomeadas, possibilitando a tabulação de características, referenciadas por entidades dentro de massas de dados. Foram utilizados nos testes como base para o treinamento do modelo de extração os corpora públicos Amazônia e FlorestaVirgem ambos no formato Árvores Deitadas. Conclui-se que a abordagem de aprendizado de máquina estatístico, maximização de entropia, para a extração de informações, é eficiente quando levado em consideração o treinamento de um corpus especifico para o domínio de pesquisa.*

1. Introdução

Em um mundo onde a concorrência tem sido cada vez mais acirrada em todos os mercados, investir em inovação tecnológica deixou de ser um plano futuro e se tornou um imperativo para a sobrevivência das empresas. Expressões como *Big Data* e *Data Mining* tem se tornado cada vez mais popular com o crescente volume de dados criados diariamente. Segundo a IBM (2012) graças aos *smartphones*, aos *tablets*, às redes sociais, aos *e-mails* e a outras comunicações

digitais, o mundo cria 2,5 milhões de *terabytes* de novos dados diariamente; 90% dos dados existente hoje foram criados nos últimos dois anos. Em meio a esta imensidão de dados se torna cada vez mais proveitoso a extração automatizada de informações relevantes. Estimativas da Dell Inc. (2010), sugerem que a maior parte destes dados são desestruturados e com um crescimento anual de 60%. Considerando estes fatos, se faz necessário uma metodologia eficiente para o processamento de textos não estruturados a fim de se obter características que possibilitem a classificação, simplificação e identificação de padrões.

A análise de estruturas textuais é extremamente complexa. Na maioria das teorias linguísticas, a linguagens é estruturada em duas partes principais: o léxico, que se refere a um dicionário de palavras dado uma linguagem e a gramatica que representa um sistema de regras para o uso destas.

A proposta do presente trabalho é abordar a extração de entidades nomeadas com aprendizado estatístico através do *framework opensource* Opennlp, criando uma ferramenta de generalização de observações baseado em modelos de extração definidos pela técnica de maximização de entropia; a fim de identificar informações relevantes de bases textuais e estrutura-las, ou seja, transformar informações relevantes encontradas em documentos em linhas e colunas dentro de um sistema de banco de dados relacional.

Os modelos utilizados nos testes foram criados a partir dos corpora anotados públicos Amazônia e FlorestaVirgem do Projeto Floresta Sinta(c)tica [Afonso 2006].

2. Reconhecimento de Entidades Nomeadas

O reconhecimento de entidades nomeadas é uma importante subtarefa da extração de informação [Amaral 2013] e seu objetivo é realizar a identificação, marcação e extração de elementos pré-determinados em um texto. Esses elementos são definidos no inicio do processo de extração.

Neste trabalho será abordada a técnica de extração de entidades conhecida como Maximização de entropia. Esta técnica tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, a partir de uma série de variáveis explicativas contínuas e/ou binárias e ainda conforme Manning e Schutze (1999) podem ser definidos como uma forma de integralização de informações de diversas fontes heterogêneas para a realização de classificação de fragmentos textuais, ou seja, uma amostra do texto do qual se quer extrair informações é usada como base em um treinamento para que a partir desse ponto o algoritmo de maximização de entropia possa generalizar o raciocínio contido no treinamento. O uso de extração de entidades nomeadas é amplo [Tkachenko e Simanovsky 2012]; pode descrever a importância de fragmentos textuais no contexto da dissertação, auxiliando assim no processo de interpretação e aprendizado de máquina. Aprendizado de máquina é um campo da ciência da computação focado na criação de algoritmos e técnicas que permitam a computadores e equipamentos aprender, isto é alterar seu comportamento em resposta a estímulos externos ou experiências do passado [Alpaydin 2004] e aperfeiçoar seu desempenho em tarefas a partir de padrões, erros ou treinamento prévio.

Há varias situações no qual aprendizado de maquina é necessário para resolver problemas, principalmente em casos onde uma solução em tempo algorítmico não seja viável.

O reconhecimento de entidades nomeadas é um amplo exemplo de uso para técnicas de aprendizado de máquina, pois mesmo com dicionários específicos a extração de fragmentos relevantes seria dificultada, pelo alto índice de ambiguidade a que uma palavra pode ter em diversos contextos [Sasaki et al 2008]. Assim como outras subtarefa, pertence diretamente ao subcampo Processamento de linguagem natural (PLN).

PLN é um importante subcampo de inteligência artificial, somado ao estudo de linguística computacional que estuda os desafios de interpretação, extração e criação textual automática em linguagem humana e produção semântica. O problema de analise semântica é um assunto vastamente estudado, porém ainda sem uma abordagem que o resolva completamente; o campo possui tanto dispositivos simples como N-gramas [Jurafsky e Martin 2008], quanto técnicas mais elaboradas como Modelos Ocultos de Markov [Russel e Norvig 2003].

2.1 Maximização de Entropia

O modelo de maximização de entropia, também conhecido como *Maxent*, trata-se da técnica de estatística indutiva regressão logística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica. Neste caso entropia é definido como uma medida única para incerteza, ou seja, mede a quantidade de informação contida em uma variável aleatória [Carvalho 2012].

Basicamente a técnica promove a construção de um modelo de distribuição de probabilidade p que se aproxime de p', sendo que p' seja uma distribuição de probabilidade obtida através de dados de treinamento [Carvalho 2012]. Os principais termos da maximização de entropia são segundo Filho (2002): **Função característica** é uma função que associa um contexto a uma classificação, retorna 1 quando o par classificação e contexto está correto; **Contexto** é o exemplo de conceito; **Conjunto de treinamento** é o conjunto de pares (classificação, exemplo).

A estimativa de maximização de entropia combina evidencias obtidas no treinamento utilizando log-linear para produzir um modelo em que toda função característica é relacionado. Se B é um conjunto de classes ou neste caso entidades e C é o conjunto de contextos, ou seja o texto das quais as classes são incidentes; o *p(b,c) estimado* deverá ser representado conforme a Equação 1 (Baldridge et al., 2002).

$$H(p) = -\sum_{(b,c) \in BxC} \mathrm{p(b,c)} \log p(b,c) \tag{1}$$

A representação das evidencias são determinadas pela forma *p(b,c)* e as evidencias são codificadas com k características; onde a função característica é descrita conforme a Equação 2.

$$f(a,b) = \begin{cases} 1, se\ a = a'\ e\ b = b' \\ 0, caso\ contrário \end{cases} \tag{2}$$

3. Materiais e Métodos

O Processo de pesquisa, bem como o desenvolvimento da ferramenta contou com uma pesquisa bibliográfica sobre aprendizado de máquina, extração de informações e aprendizado estatístico. Após analise minuciosa de algumas técnicas e ferramentas de extração decidiu-se pelo uso do Framework Opennlp e consequentemente a linguagem de programação Java para a construção da ferramenta, utilizando a técnica estatística de maximização de entropia. A ferramenta de treinamento e consumo de modelos é construída sobre plataforma Web e Banco de dados Mysql.

Opennlp é uma ferramenta *open source* de processamento de linguagem natural, baseada em modelos estatísticos e redes neurais. A ferramenta conta com vários recursos de processamento de linguagem natural e possibilita a criação e modelos customizados para cada domínio de pesquisa [Baldridge et al., 2002].

Os modelos dentro da ferramenta são criados baseado no recurso de corpus anotado, segundo definição de Sardinha (2004) corpus é uma coleção de dados linguísticos, como textos ou partes de textos, bem como transcrição de fala em uma determinada língua, escolhidos baseado em uma necessidade, caracterizando-se como uma amostra linguística, os corpora se dividem em Textos puros ou anotados, onde o primeiro não possui nenhuma meta-informação, ou seja, nenhuma marcação de classificação ou identificação, enquanto o segundo possui um conjunto de documentos marcados, ou seja, com o conhecimento que se quer generalizar previamente selecionado dentro do domínio escolhido.

O armazenamento dos corpora de amostra para modelos e os demais dados do sistema são realizadas em banco de dados relacional, que se traduz no relacionamento de tabelas de armazenamento de informações, a fim de organizar os dados e evitar repetições desnecessárias, conforme Date (2004), banco de dados é a persistência de uma coleção de dados, usado em aplicações por uma determinada empresa; o fundamento principal é a capacidade de manter de forma estruturada um repositório de informações que pode ser acessada sempre que necessário. O Software de banco de dados relacional utilizado foi o Mysql que segundo Oracle (2014), é o banco de dados opensource existente mais popular do mundo.

Durante o processo de treinamento foi utilizado como amostra linguística e base dos modelos os corpora, Amazônia com um total de 4.580.000 palavras e 275.000 frases; e FlorestaVirgem com 1.640.000 palavras e 96.000 frases do Projeto Floresta Sintá(c)tica [Afonso 2006], conforme o Figura 1 os recursos possuem as seguintes características.

Corpus Amazonia			Corpus FlorestaVirgem	
ENTIDADE	QUANTIDADE DE ANOTAÇÕES		ENTIDADE	QUANTIDADE DE ANOTAÇÕES
PESSOA	85314		PESSOA	16566
ORGANIZAÇÃO	69654		ORGANIZAÇÃO	17669
LUGAR	52123		LUGAR	11354

Figura 1. Distribuição de Entidades Treinadas

Fonte: Richerland Pinto Medeiros

3.1 Avaliação de Resultados

As avaliações de sistemas de reconhecimento de entidades nomeadas são baseadas na comparação da anotação automática usando o modelo gerado, com a anotação manual realizada em parte do corpus utilizado no treinamento, as medidas utilizadas para na comparação dos modelos criados são a precisão que se refere a fração de candidatos selecionados corretamente, cobertura que se refere a fração de candidatos corretos que foram selecionados e a medida-F que refere a uma relação entre precisão e cobertura [Nadeau e Sekine 2007], conforme a Equação 3.

$$MedidaF = 2 * \frac{\text{Precisão} * \text{Cobertura}}{\text{Precisão} + \text{Cobertura}} \quad (3)$$

Na seção 3.2 de resultado obtidos será apresentado em formato de tabela, os índices de precisão, cobertura e medida-F referente a análise dos copora usados no presente trabalho; bem como será demostrado a tela de estruturação, contendo entidade, informação e a probabilidade da referida informação ser pertencente a entidade.

3.2 Resultados Obtidos

A ferramenta apresentada a seguir é fruto da pesquisa e desenvolvimento da técnica de aprendizado de máquina estatístico maximização de entropia, utilizando o framework de processamento de linguagem natural Opennlp. A Figura 2 descreve o fluxograma de funcionamento, desde a criação do domínio ao consumo do modelo.

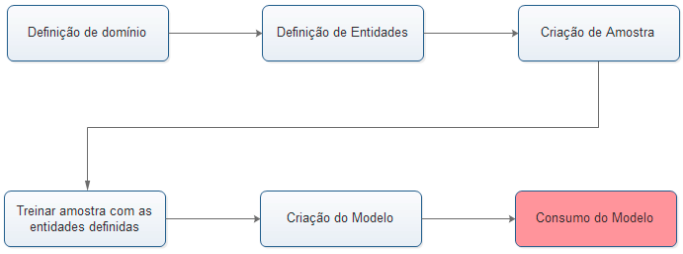

Figura 2. Fluxo de Funcionamento

Fonte: Richerland Pinto Medeiros

Para a definição do domínio foi realizado um cadastro básico, com finalidade de separação dos modelos por domínio de consulta, neste trabalho o domínio, já para a definição de entidades, o observador deverá escolher quais características textuais deseja promover o treinamento, que posteriormente servirá como modelo para a generalização. As entidades escolhidas para este trabalho foram: PESSOA, ORGANIZAÇÃO e LUGAR.

Outro passo importante foi a criação da amostra linguística, pois assim como o cadastro de domínio serve como um separador organizacional, entretanto neste caso separa os arquivos que serão usados para o treinamento da ferramenta, ou seja, é a definição de um repositório de arquivos que posteriormente irá gerar um modelo. Durante os testes foram importados como amostras os corpora Amazônia e Floresta Virgem do projeto Floresta Sintá(c)tica [Afonso 2006].

Cada arquivo dentro da amostra deve ser anotado por um avaliador, buscando os fragmentos que caracterizam a entidade escolhida. A somatória dos arquivos treinados serve como corpus anotado. A Figura 3 descreve o processo de treinamento de um arquivo com as entidades escolhidas.

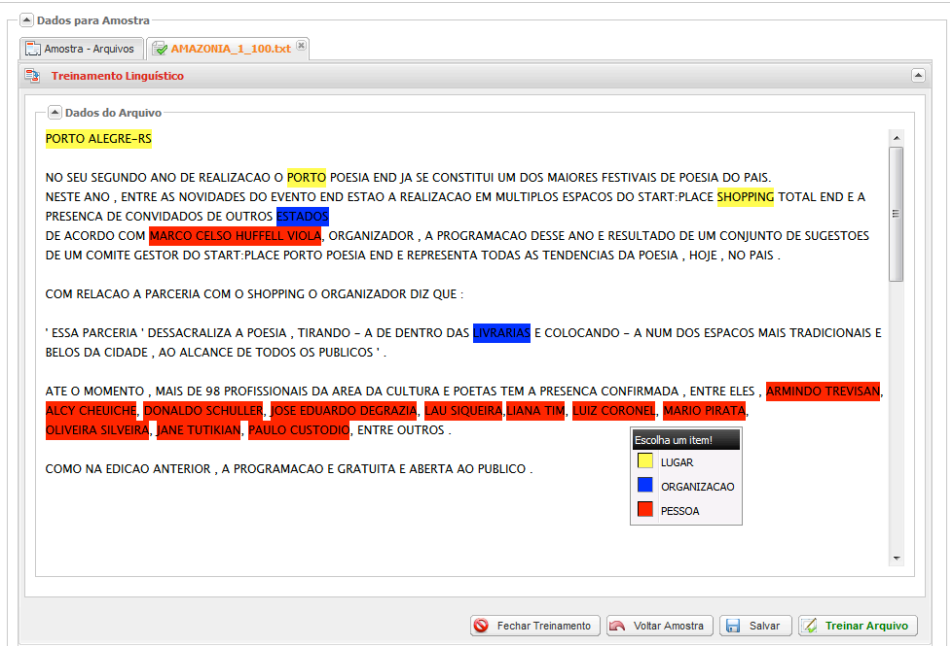

Figura 3. Tela de Treinamento de Arquivos de Amostra

Fonte: Richerland Pinto Medeiros

Após realização o processo de treinamento de todos os arquivos, o corpus anotado já está pronto para ser transformado em modelo. Utilizando a opção "Gerar Modelo" do sistema proposto e consequentemente após a criação do modelo ele pode ser usado para generalizar a extração de entidades nomeadas em outros arquivos, visando extrair palavras e termos candidatos de acordo com as características textuais encontradas no texto correlacionadas ao comportamento treinado no corpus. A Figura 4 apresenta um texto da qual suas entidades foram extraídas automaticamente.

A coluna Entidade, se refere a categoria da entidade ao qual a informação foi encontrada, enquanto a coluna Informação se refere ao fragmento de texto encontrando no contexto que foi relacionado a entidade; por fim a coluna Probabilidade menciona, o índice de certeza que a ferramenta tem, baseado no modelo de que a informação mencionada na coluna Informação é relativa ao tipo da coluna Entidade; este índice é baseado diretamente na quantidade de

informações contidas no modelo treinado e sua eficácia esperada; não possuindo referência com os índices precisão, cobertura e medida-F de avaliação geral dos modelos.

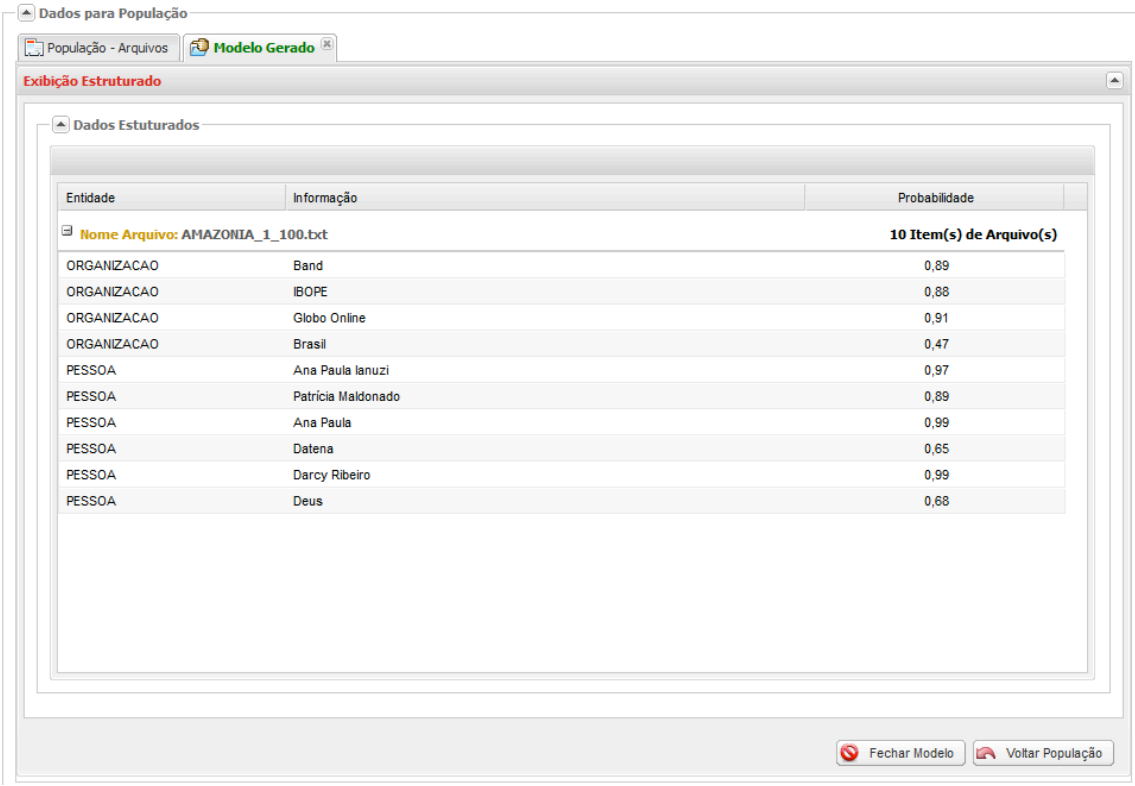

Figura 4. Tela de Estruturação de Entidades

Fonte: Richerland Pinto Medeiros

Considerando o funcionamento da ferramenta para o modelo proposta em cima dos corpora Amazônia e FlorestaVirgem [Afonso 2006], um fragmento de 10% foi extraída de cada corpus anotado para avaliação dos resultados, analisando as medidas de precisão, cobertura e medida-F, conforme mencionadas na seção 3.1 de avaliação de resultados. A Tabela 1 mostra os números da análise do consumo do modelo de cada corpus.

Corpus Amazonia			Corpus FlorestaVirgem	
Precisão	72,30%		Precisão	73%
Cobertura	72,80%		Cobertura	64%
Medida-F	75,30%		Medida-F	68,60%

Tabela 1 – Distribuição de Entidades Treinadas

Fonte: Baldridge et al., 2002

Onde o uso do recurso *TokenNameFinderEvaluator* do *framework* Opennlp representa uma avaliação completa trazendo todos índices necessários para a avaliação. [Baldridge et al., 2002].

O sistema foi modelado para ser abstrato em relação a entidades, ou seja, caso o observador deseje criar outras entidades, para domínios específicos, como esportes, jornais e etc., será possível, desde que realize o treinamento de um modelo especifico. A Figura 5 apresenta o Diagrama Entidade e Relacionamento do projeto.

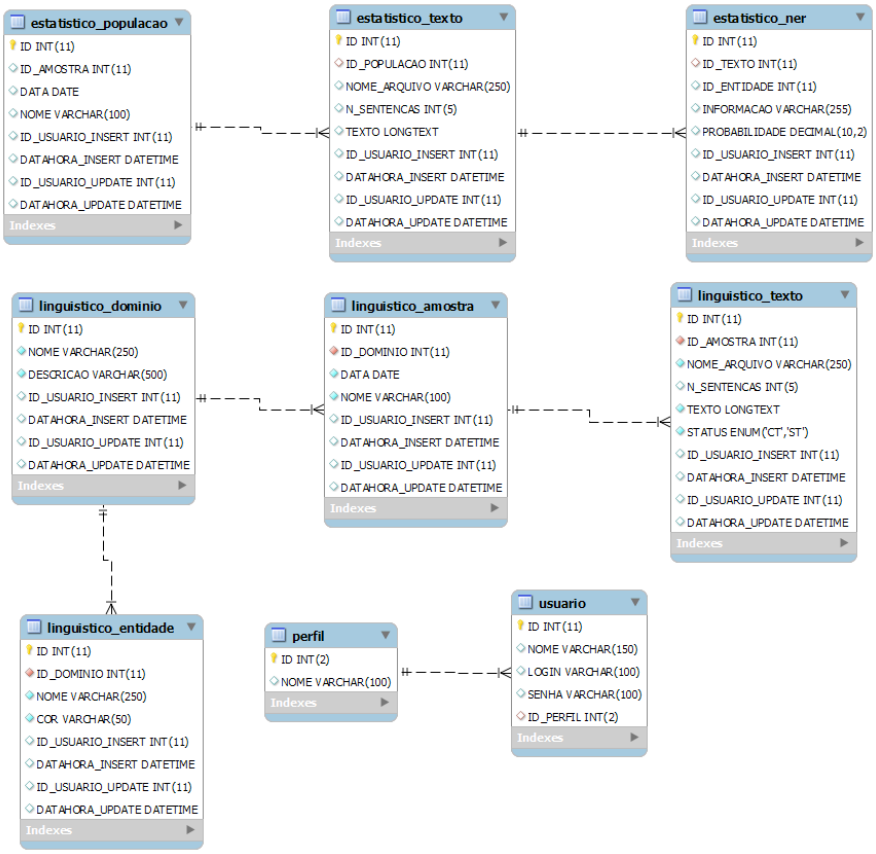

Figura 5. Diagrama Entidade e Relacionamento

Fonte: Richerland Pinto Medeiros

4. Conclusão

A ferramenta proposta atingiu o resultado esperado extraindo as entidades previstas de arquivos que não eram contemplados no treinamento. A ferramenta Opennlp mostrou-se uma ferramenta eficiente na extração de entidades nomeadas ao usar a técnica estatística de maximização de entropia; entretanto o custo em tempo envolvido no processo de treinamento pode dificultar o uso desta técnica. Considerando o teste com os corpora Amazônia e FlorestaVirgem, em que obtiveram 75,30% e 68,60% respectivamente em sua análise da medida-F, ficou evidente que a quantidade de texto não marcado do corpus influencia negativamente na criação do modelo, pois o corpus FlorestaVirgem com mais frases e menos treinamento trouxe uma cobertura 64% e

precisão 73%, contra 72,80% de cobertura e 72,30% de precisão do corpus Amazônia, que possuía uma maior abrangência de treinamento apesar de possuir menos frases e palavras.

Como trabalho futuro pode se buscar uma abordagem mista com o método de maximização de entropia, usando dicionários de palavras, ou ontologias, para que entidades mais especificas possam serem extraídas diminuindo a ambiguidade e alcançando uma medida-F maior.

5. Referências Bibliográficas

Afonso, S. (2006) "A floresta sintá(c)tica como recurso". Linguateca, http://www.linguateca.pt/floresta/corpus.html.

Aires, R. (2000) "V X Implementação, Adaptação, Combinação e Avaliação de Etiquetadores para o Português do Brasil". São Paulo, Universidade de São Paulo, Instituto de Ciências Matemáticas de São Carlos.

Alpaydin, E. (2004) "Introduction To Machine Learning". MIT Press.

Amaral, D. O. F. (2013) "O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa". Rio Grande do Sul , Pontifícia Universidade Católica, Mestrado em Ciências da Computação.

Baldridge et al. (2002). "The Opennlp Maximum Entropy Package". SourceForge, Apache.

Banko, M., Etzione, O. (2008) "The Tradeoff Between Open and Traditional Relation Extraction". Columbia University: The Association for Computer Linguistics.

Candido JR, A. (2008) "Criação de um Ambiente para o Processamento de Corpus de Português Histórico". São Paulo, Universidade de São Paulo, Mestrado em Ciências da Computação.

Carvalho, W. S. (2012) "Reconhecimento de Entidades Mencionadas em Português utilizando Aprendizado de Maquina". São Paulo, Universidade de São Paulo, Mestrado em Ciências da Computação.

Crawford, R. (1994) "Na era do Capital Humano". São Paulo, Atlas.

Nadeau, D., Sekine, S. "A Survey of named entity recognition and classification". Linguisticae Investigationes, Páginas 3-26.

Dell, "Object Storage a Fresh Approach to Long-Term File Storage", http://i.dell.com/sites/doccontent/business/solutions/whitepapers/pt/Documents/object-storage-overview_br.pdf.

Filho. A. A. A. (2002) "Maximização de Entropia em Linguística Computacional para a Língua Portuguesa", São Paulo, Universidade de São Paulo, Mestrado em Ciências da Computação.

Ibm, "What is Big Data?", http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

Oracle, "What is Mysql", http://dev.mysql.com/doc/refman/4.1/en/what-is-mysql.html.

Jiang, J. (2012) "Information Extraction From Text", In *Mining Text Data*. Research Collection School Of Information Systems.

Jurafsky, D., Martin, J. H. (2008) “Speech and Language Processing”. Englewood Cliffs, NJ, USA: Prentice Hall.

Date, C. J. (2003) “Introdução a Sistemas de BANCOS de Dados”, Elsevier.

Manning, C. D., Schutze, H. (1999) “Foundations of Statistical Natural Language Processing”. The MIT Press.

Norvig, P., Russell, S. J. (2003) “Artificial Intelligence – A Modern Approach”, Prentice Hall, 2. ed. Englewood Cliffs.

Sardinha, T. B. (2004) “Linguística de Corpus”, São Paulo: Manole, 2004. v. 1. 410p.

Sasaki et al (2008) “How to Make the Most of NE Dictionaries in Statistical NER”, ACL Workshop, Columbus, OH, USA.

Simanovsky, A. Tkachenko, M. (2012) “Proceedings of KONVENS”.

Zacara, R. C. C. (2012) “Anotação e Classificação Automática de Entidades Nomeadas em Notícias Esportivas em Português Brasileiro”, São Paulo, Universidade de São Paulo, Mestrado em Ciências da Computação.