Extração de Nomes de Pessoas Baseado em Etapas de Etiquetamento

Henrico B. Brum[1], **Sergio L. S. Mergen**[1]

[1]Campus Alegrete - Universidade Federal do Pampa (UNIPAMPA)
CEP – 97.546-550 – Alegrete – RS – Brasil

henrico.brum@gmail.com, sergiomergen@unipampa.edu.br

Abstract. *The identification of person names inside natural language texts can be used for many goals, such as applications related to the information retrieval and data analysis areas. Given this, our paper explores the entity extraction problem by using unsupervised strategies focused on the recognition of person names. The proposed approach employs a classification process divided in three stages, called Location, Filtering and Expansion. Each stage executes tagging rules that satisfies the purpose of the stage. The experiments show the performance of the rules for the classification of textual documents, analyzing the rules individually and collectively.*

Resumo. *A identificação de nomes de pessoas em meio a textos em linguagem natural pode ser usada para diversas finalidades, como por aplicações na área de recuperação de informação e análise de dados. Dessa forma, esse artigo explora a extração de entidades nomeadas por meio de estratégias não supervisionadas com foco no reconhecimento de nomes de pessoas. A abordagem proposta emprega um processo de classificação dividido em três etapas, chamadas de Localização, Filtragem e Expansão. Para cada uma dessas etapas são executadas regras de etiquetamento que atendem os requisitos de cada etapa. Os experimentos demonstram o desempenho das regras utilizadas na classificação de documentos textuais, analisando-as de forma individual e coletiva.*

1. Introdução

Um dos temas de pesquisa que tem recebido bastante atenção atualmente envolve entidades nomeadas (ENs). De acordo com [Krishnan and Manning 2006], entidades nomeadas são termos que caracterizam algum objeto, fornecendo indícios que ajudam a identificá-lo de forma inequívoca. Os exemplos mais comuns de entidades nomeadas incluem nomes de pessoas, nomes de organizações, localidades, entre outros.

Diversas aplicações podem fazer uso de entidades nomeadas, principalmente na área de recuperação de informação e análise de dados. Por exemplo, ao reconhecer que uma entidade nomeada é utilizada em uma consulta, os motores de busca podem se valer dessa informação para aprimorar o resultado da consulta. Outro exemplo envolve a descoberta de conhecimento sobre informações textuais. Algoritmos de processamento de linguagem natural podem se valer das entidades nomeadas contidas em um texto para realizar associações entre os componentes desse texto. Mais adiante, técnicas de aprendizado de máquina poderiam utilizar as associações descobertas em textos históricos para prever comportamentos futuros.

Para que tais aplicações sejam possíveis, é necessário identificar os termos que representem entidades nomeadas. No decorrer dos anos, diversas técnicas de reconhecimento de entidades nomeadas foram propostas. Muitas delas exploram evidências presentes nas sentenças em que os termos aparecem para guiar o reconhecimento. Algumas dessas evidências compreendem a detecção de padrões e a análise da vizinhança (termos que costumam cercar entidades nomeadas) usada em conjunto com a análise da estrutura morfossintática da sentença.

O uso de muitas evidências torna difícil estipular uma equação precisa para identificar se um conjunto de termos equivale a uma entidade nomeada. Por isso, costuma-se recorrer a algoritmos supervisionados de aprendizado de máquina, capazes de atribuir pesos para as palavras de modo a melhorar o cálculo que realiza a classificação. Contudo, tais algoritmos requerem bases de treinamento, contendo rótulos pré-marcados identificando entidades nomeadas, para que novas entidades possam ser reconhecidas no futuro.

Nesse artigo, será analisada a possibilidade de que o reconhecimento de ENs possa ser feito através de uma abordagem que não requer bases de treinamento. Para isso, será apresentado um mecanismo de classificação composto por três etapas: Localização, Filtragem e Expansão. Cada etapa aceita regras com um comportamento específico, cujo objetivo é marcar ou desmarcar palavras como nomes de pessoas. Também são propostas regras que podem ser utilizadas dentro de cada etapa.

Para verificar a qualidade do processo de classificação e das regras proposta, são realizados experimentos utilizando o *benchmark HAREM*, um repositório *XML* de textos demarcados de acordo com o seu tipo. Os experimentos demonstram a importância de cada uma das etapas assim como das regras das quais são compostas.

Este artigo está estruturado da seguinte forma: Na seção 2 é apresentado o mecanismo de classificação proposto, as etapas envolvidas e as regras que foram desenvolvidas. Na seção 3 são explicados os experimentos realizados sobre o repositório *HAREM*. Os trabalhos relacionados são descritos na seção 4. Para finalizar, a seção seção 5 traz as conclusões.

2. Regras Propostas

O objetivo geral dos algoritmos de extração de ENs é percorrer blocos de texto e inserir marcações sintáticas nas palavras (etiquetar, ou *tag*) de acordo com o seu tipo. Essa demarcação também ocorre em técnicas do tipo *POS Tagger* (*Part of speech tagger*), onde o objetivo é etiquetar as palavras de acordo com a sua classe gramatical. Por esse motivo, algoritmos desse tipo são chamados de etiquetadores.

Como o foco deste trabalho compreende o reconhecimento de nomes de pessoas, consideramos apenas dois tipos de etiqueta: 'N' (nome) e 'O' (outro). Dessa forma, dada uma sentença qualquer, o objetivo do etiquetamento é rotular as palavras como 'N' ou 'O', conforme ilustrado na Figura 1. Observa-se que a etiqueta 'O' caracteriza qualquer coisa que não seja nome de pessoa, como numeral, título, verbo, adverbio ou até mesmo de uma entidades nomeada que não represente um indivíduo.

Para atingir esse objetivo é proposto um etiquetador baseado em etapas de execução, conforme apresentado na Figura 2. O etiquetamento é dividido em três etapas: Localização, Filtragem e Expansão. Cada etapa é composta por um conjunto de

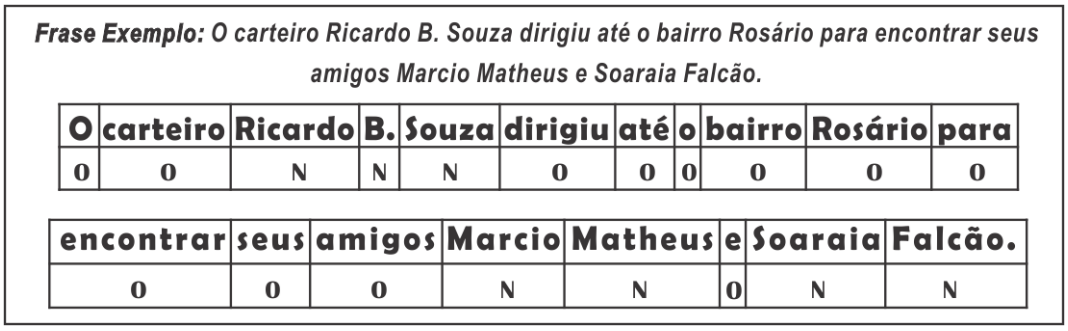

Figura 1. Exemplo de marcação de palavras em uma sentença

regras que atendem ao propósito da etapa. De modo geral, a Etapa de Localização tem o propósito de encontrar candidatos a nomes de pessoas, a Etapa de Filtragem tem o propósito de eliminar candidatos erroneamente identificados, e a etapa de expansão tem o objetivo de encontrar novos candidatos a partir dos que sobreviveram à etapa de filtragem.

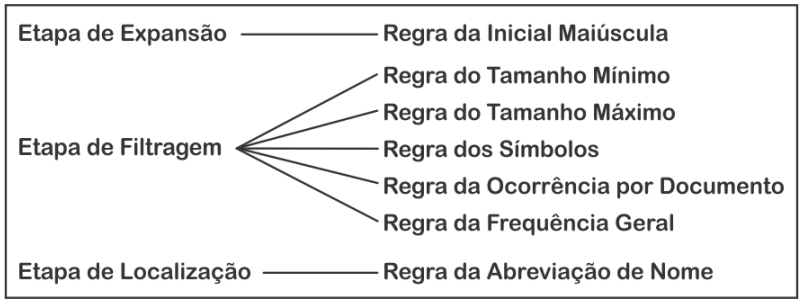

Figura 2. Categorias de Regras Propostas

Para demonstrar as características e problemas enfrentados na implementação das regras de cada uma das etapas, utilizaremos um exemplo (Figura 3) composto por três parágrafos retirados da internet[1]. Os termos sublinhados nos textos referem-se a entidades nomeadas que deveriam ser detectadas como nomes de pessoas.

2.1. Etapa de Localização

A etapa de Localização tem o objetivo de encontrar o máximo possível de nomes de pessoas no texto. Para isso, devem ser projetadas regras de localização que tenham uma alta cobertura, mesmo que sejam associadas a uma precisão baixa. Nesse momento não existe a preocupação com excesso de informação irrelevante ou inválida, uma vez que a etapa de filtragem se encarregará de eliminar os falsos positivos encontrados.

Para fins de etiquetamento, considera-se que inicialmente todas palavras são marcadas com o rótulo 'O'. A medida que as regras descubram candidatos, o rótulo das palavras vai sendo adaptado.

Neste artigo, consideramos uma única regra de localização, baseada na observação de que nomes de pessoas, em geral, são escritos com a primeira letra maiúscula. A definição da regra encontra-se abaixo:

[1]Parágrafos retirados dos artigos referentes a *Hebe Camargo*, *Quentin Tarantino* e *Jodie Foster* presentes no domínio http://pt.wikipedia.org.br

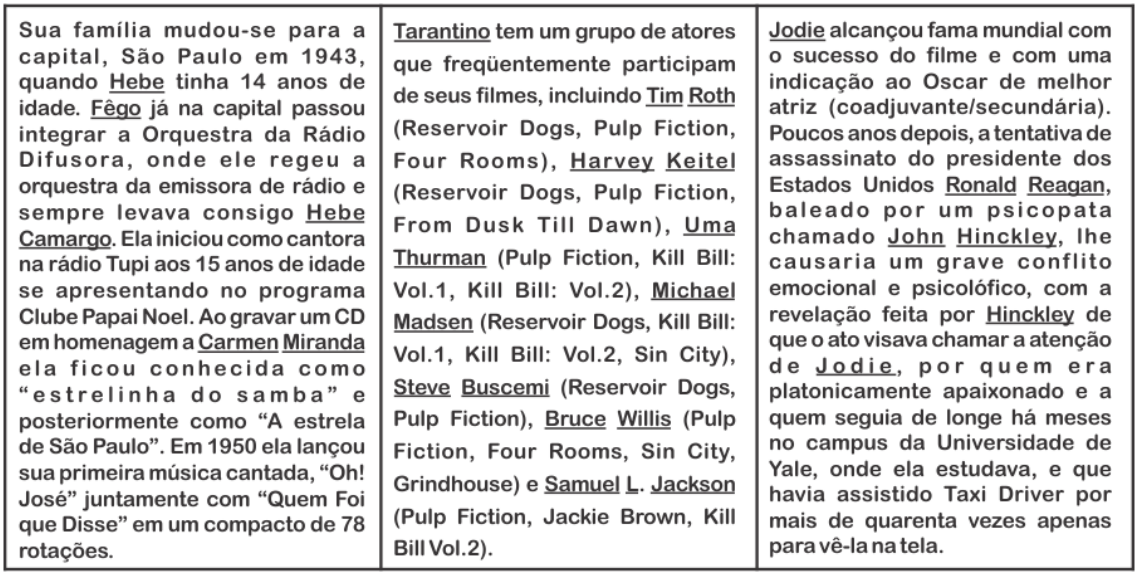

Figura 3. Exemplo de parágrafos de textos distintos

Regra da Inicial Maiúscula: Todas as palavras que possuam a primeira letra em maiúscula recebem a marcação 'N'.

Para exemplificar a regra, considere a Figura 4. Como pode ser observado, todos os nomes de pessoas foram corretamente marcados. Porém, muitas outras palavras que não se referiam a pessoas também foram marcadas como tal, como siglas ou substantivos que iniciavam sentenças. Esse comportamento está em consonância com o propósito da etapa de localização, uma vez que a regra conseguiu uma alta taxa de cobertura.

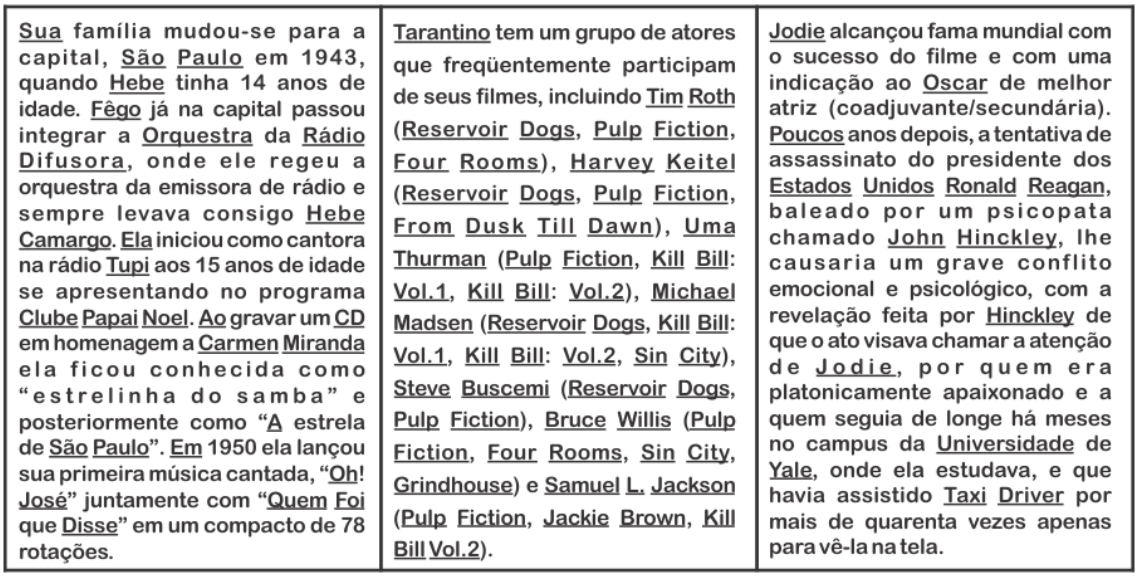

Figura 4. Marcação de Palavras na Etapa de Localização

Apesar de considerar-se apenas uma regra na Etapa de Localização, outras regras poderiam ser concebidas, caso elas contenham indícios fortes de que uma palavra corresponde a nome de pessoa, mesmo sem iniciar em maiúscula. A complementação das

regras de Localização pode ser abordada em trabalhos futuros.

2.2. Etapa de Filtragem

Apesar da boa cobertura oferecida pela Etapa de Localização, novas regras são necessárias para eliminar as palavras indevidas. As regras que possuem esse intuito são encontradas na Etapa de Filtragem. Nesta etapa, apenas as palavras previamente etiquetadas como 'N' são processadas. Caso qualquer uma das regras de filtragem for satisfeita, a palavra volta a ser marcada como 'O'.

As regras de filtragem adotadas analisam as palavras e as descartam em relação à condições como tamanho mínimo, máximo, presença de símbolos e numerais na palavra e frequência geral da palavra. A seguir, cada uma das regras é apresentada com mais detalhes:

Regra do Tamanho Mínimo Palavras com menos de três caracteres são marcadas com 'O'.

Essa regra parte da observação de que nomes de pessoas normalmente possuem mais do que dois caracteres. Para exemplificar, essa regra poderia ser usada para remover palavras erroneamente marcadas como 'N' pela etapa anterior, como artigos (ex. "A"), conjunções (ex. "Em") e algumas siglas (ex. "CD").

Regra do Tamanho Máximo Palavras com mais de 10 caracteres são marcadas com 'O'.

Essa regra parte da observação de que nomes de pessoas normalmente possuem menos do que onze caracteres. Para exemplificar, essa regra poderia ser usada para remover palavras erroneamente marcadas como 'N' pela etapa anterior, como nomes de filmes (ex. *Grindhouse*) e títulos de instituições (ex. *Universidade*).

Regra dos Símbolos Palavras que possuam em sua composição símbolos e numerais são marcadas com 'O'.

Essa regra parte da observação de que nomes de pessoas não possuem símbolos especiais. A presença desses símbolos está normalmente associada a termos referentes a e-mails e endereços. A implementação dessa regra ignora hifens, acentos e apóstrofos, pois estes podem ser encontrados em nomes de pessoas. Para exemplificar, essa regra poderia ser usada para remover palavras erroneamente marcadas como 'N' pela etapa anterior, como em *"Kill Bill Vol.2"*.

Regra da Frequência por Documento Palavras que ocorram com frequência maior do que α em todos os documentos são marcada como 'O', sendo α um valor percentual.

Essa regra parte da observação de que o mesmo nome de pessoa não costuma aparecer com frequência elevada no conjunto total de documentos. Sendo assim, as palavras comuns (de frequência elevada) são descartadas, enquanto as palavras raras (de frequência baixa) são conservadas. Para exemplificar, considere a Figura 4, e suponha que α seja igual a 50%. Nesse caso, a regra removeria a palavra 'Ela' como nome de pessoa, visto que esta palavra tem uma frequência igual a 66,6% (a palavra aparece no primeiro documento e no terceiro documento, sendo que temos 3 documentos no total).

O ponto de corte α pode ser definido de forma empírica, conforme demonstrado

na seção de experimentos.

2.3. Etapa de Expansão

A Etapa de Expansão prevê o uso de marcações prévias para encontrar novos nomes de pessoas. Assim como na Etapa de Localização, as regras nesta etapa são aditivas. Ou seja, elas tem o objetivo de mudar as marcações de 'O' para 'N'.

Neste artigo será considerada uma única regra de expansão, a regra de Abreviação de Nome:

Regra de Abreviação de Nome: Uma palavra é marcada como 'N', se possuir um único caractere maiúsculo (ignorando caracteres de pontuação), e a sucessora possuir a etiqueta 'N'.

Essa regra parte da observação de que caracteres isolados sucedidos por um nome de pessoa provavelmente se referem a uma abreviação de um nome composto. Para exemplificar, através da aplicação desta regra, a palavra 'L.' de 'Samuel L. Jackson', que seria marcada como 'O' pela regra de tamanho mínimo executada durante a etapa anterior, voltaria a ser marcada como 'N' durante a expansão.

3. Experimentos e Resultados

A qualidade do processo de classificação de nomes de pessoas proposto neste trabalho é avaliada através do *HAREM*, um repositório de textos marcados usado em experimentos no campo de Processamento de Linguagem Natural[2]. O *HAREM* é composto por um documento *XML* com 129 documentos com marcações de entidades nomeadas distribuídas em categorias. Ao todo são 74.298 palavras, sendo 2.622 representando nomes de pessoas. No documento *XML* os nomes de pessoas são marcados com a categoria "PESSOA"e tipo "INDIVIDUAL".

Para fins de avaliação usou-se as métricas de Precisão e Cobertura. A Precisão é dada pelo número de palavras classificadas como nome de pessoas que realmente são nomes de pessoas no *HAREM*. Já a Cobertura é dada pelo número de palavras classificadas como nome de pessoas em função do total de nomes de pessoas existentes no *HAREM*. Também usou-se a medida F, que calcula uma média harmônica entre os valores de precisão e cobertura para chegar a um valor que exprima o impacto conjunto dessas métricas.

Para a Regra da Frequência por Documento, o valor de α foi determinado com base em uma análise manual de uma amostragem de documentos contidos no *HAREM*. A partir dessa amostragem, procurou-se pelo nome da pessoa (entidade nomeada da categoria "PESSOA"e tipo "INDIVIDUAL") que aparece com maior frequência. O valor de frequência encontrado foi usado como ponto de corte α.

A Figura 5 apresenta os resultados obtidos com a aplicação das regras de Localização e Filtragem. Em um dos casos avaliados, é avaliada apenas a Etapa de Localização, ou seja, a regra da Inicial Maiúscula. Nos demais casos, a regra da Inicial Maiúscula é aplicada juntamente com alguma técnica da etapa de filtragem.

[2]Disponível em 'http://www.linguateca.pt/harem/'

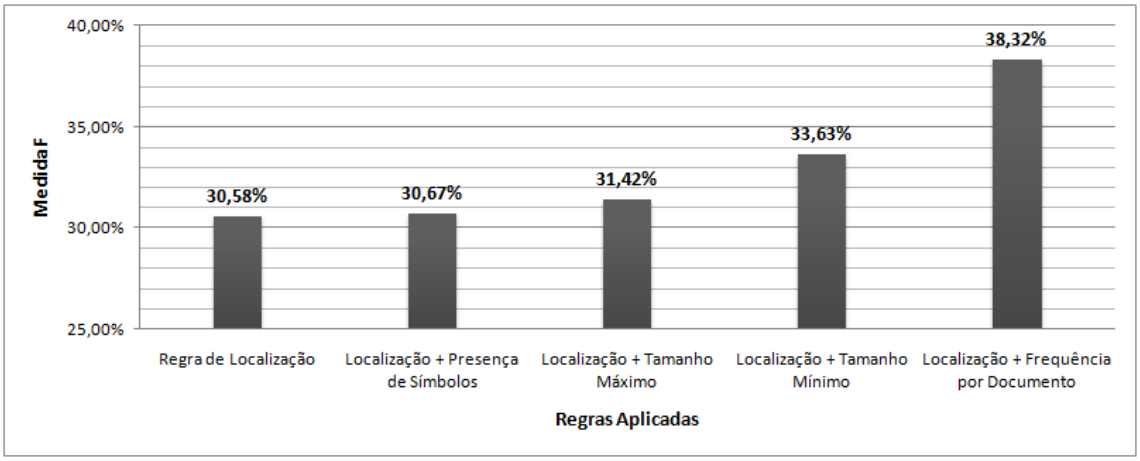

Figura 5. Regras de FiltragemAplicadas Individualmente

Com a regra de Localização isolada foi obtida uma cobertura de 100% na análise dos 129 documentos, o que indica que todos os nomes de pessoas foram corretamente marcados como tal. Porém, a precisão foi de 18%, o que significa que muitas palavras foram erroneamente marcadas como nomes de pessoas. A medida F resultante atingiu 30,58%.

Como a Figura demonstra, as demais regras avaliadas procuram filtrar o resultado obtido pela Etapa de Localização de modo a aumentar a precisão. De modo geral, em todas as regras testadas houve aumento da precisão, tendo como contrapartida uma redução na cobertura. Observa-se que a regra de Frequência por Documento foi a que obteve um melhor desempenho.

A Tabela 1 exibe alguns exemplos de palavras que perderam a marcação 'N' acertadamente após o uso de regras específicas de Filtragem. Como pode-se ver, as regras são complementares, o que indica que elas possam ser usadas em conjunto para obter uma melhor filtragem.

Regra de Filtragem	**Exemplos de Nomes Filtrados**
Regra de Tamanho Mínimo	Siglas (como *RS* e *SP*) Exclamações (como *Ah*) Títulos (como *Sr* e *Dr*)
Regra do Tamanho Máximo	Cidades (cidades (como *Teresópolis*) Países (como *Grã-Bretanha* e *Afeganistão*
Regra dos Símbolos	Valores (como *R$2,00*) Fórmulas químicas (como *H5N1*)

Tabela 1. Exemplos de Palavras Filtradas

Visando verificar essa possibilidade, realizou-se outro experimento em que regras foram sucessivamente adicionadas ao processo de classificação, seguindo o fluxo estabelecido pelo processo proposto, em que executam-se em ordem as regras de Localização,

Filtragem e Expansão. A Figura 6 ilustra os resultados obtidos.

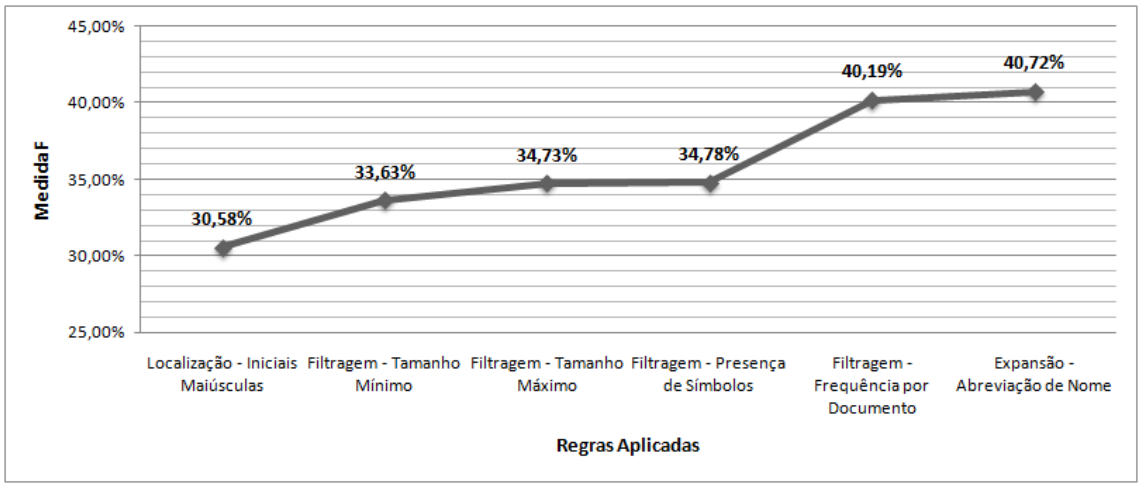

Figura 6. Evolução de Eficiência ao Longo das Regras

Conforme pode-se observar, o desempenho da classificação aumenta conforme novas regras são adicionadas. É interessante destacar a importância da existência de Regras de Expansão. No caso avaliado, a Regra de Abreviação de Nome conseguiu aumentar a cobertura sem perda na precisão, o que resultou em um valor melhor da Medida F. Ao final do processo, atingiu-se uma Medida F equivalente a 40,72%, o que é um resultado razoavelmente expressivo, considerando o uso de técnicas de mineração não-assistidas.

4. Trabalhos Relacionados

O problema de extração de entidades (ou reconhecimento de entidades) nomeadas contempla tanto a localização das entidades nomeadas quanto uma identificação mais precisa dos tipos específicos dessas entidades as quais esse conjunto se refere. As duas tarefas podem ser realizadas por algoritmos de classificação, sendo que na localização os termos são classificados em duas categorias (positivo ou negativo) enquanto na identificação o número de categorias é maior, e depende da aplicação [Nadeau and Sekine 2007].

Os algoritmos de classificação podem ser supervisionados, não supervisionados e semi-supervisionados. Os algoritmos supervisionados partem de *córpus* textuais pré-marcados e um conjunto de dimensões (*features*), e descobrem regras sobre as dimensões que determinam em que categoria os termos se enquadram [McCallum and Li 2003]. Esse tipo de abordagem é útil para tarefas de classificação em que as regras são complexas demais para serem definidas manualmente, como por exemplo, para identificar inequivocamente o tipo de entidade localizada. No entanto, elas dependem da presença entidades nomeadas pré-rotuladas(também chamadas de *gazetters*) para funcionar.

Os algoritmos semi-supervisionados não dependem propriamente de *gazetters*, mas partem de algumas sementes (*seeds*) referentes ao tipo de entidade que se deseja buscar. As sementes são padrões textuais que costumam aparecer junto ao tipo de entidade de interesse. Sentenças que contém essas sementes são recuperadas, e a partir da análise dessas sentenças se busca identificar novas entidades e novos padrões de texto relevantes. Os padrões encontrados se transformam em novas sementes, de modo que o processo de busca e análise possa ocorrer sucessivamente, até que um limiar seja atingido.

[Pasca et al. 2006] Esse tipo de abordagem é útil quando se pretende descobrir entidades nomeadas existentes em um conjunto de documentos indexados.

Assim como os algoritmos supervisionados, os não supervisionados também usam regras sobre dimensões para realizar o processo de descoberta. Uma abordagem típica adotada envolve a clusterização, que agrupa termos que possivelmente se referem ao mesmo tipo de entidade nomeada [Cucchiarelli and Velardi 2001]. Além disso, alguns trabalhos mais específicos podem se valer de outras técnicas que não recorram ao aprendizado de máquina. Por exemplo, [Alfonseca and Manandhar 2002] descreve uma técnica que atribui um tipo de entidade nomeada a uma palavra com base no análise do contexto da palavra (as palavras que estão a sua volta). Se o contexto costuma aparecer mais seguidamente associado a um tipo específico de EN, a palavra é considerada como sendo do mesmo tipo.

De modo geral, existem uma série de dimensões (*features*) que podem ser usadas para o reconhecimento e identificação de ENs, tanto em abordagens supervisionadas quanto não supervisionadas. Os exemplos mais comuns envolvem análise de maiúsculas/minúsculas, dígitos para a identificação dos componentes de uma data, análise dos radicais e afixos de palavras [Bick 2004] e padrões sumarizadores que atribuem um tipo de dados a palavra [Collins 2002].

Também existem dimensões estatísticas, relacionadas a fatos presentes no *corpi textual* estudado. Por exemplo, palavras que aparecem tanto na forma maiúscula quanto na forma minúscula são consideradas como sendo o mesmo tipo de substantivo que em alguns casos aparece no início de frases [Mikheev 1999]. O conceito de raridade também é utilizado para o reconhecimento de entidades nomeadas. Por exemplo, em [da Silva et al. 2004], palavras compostas não são consideradas entidades nomeadas caso elas possuam um termo longo e raro. Já [Shinyama and Sekine 2004] parte da observação de que textos de notícia de uma mesma época costumam conter a mesma entidade. Isso permite descobrir entidades nomeadas raras, mas que foram destaque em algum período específico no tempo. Curiosamennte, os dois últimos trabalhos citados utilizam a frequência de modo inverso ao adotado neste artigo. Ou seja, quanto menos frequente o termo, menores as chances de ele ser uma entidade nomeada. Contudo, é importante destacar que a regra de Frequência usada neste trabalho tem um foco ortogonal, uma vez que o intuito é de filtragem, e não a de descoberta de candidatos.

5. Conclusão

Este artigo apresentou um processo para reconhecimento de Entidades Nomeadas do tipo 'nome de pessoa' baseada na aplicação de regras em três etapas distintas: localização, filtragem e expansão. Também foram propostos exemplos de regras que poderiam ser executadas dentro de cada uma das etapas.

Os resultados na classificação dos nomes de pessoas usando o repositório marcado *HAREM* mostraram que as regras conseguem atingir uma cobertura elevada e uma precisão baixa, o que gerou um escore de Medida F intermediário. Isso sugere que novas Regras de Filtragem e Expansão precisam ser concebidas, as primeiras para remover palavras da lista de nomes e a segunda para corrigir eventuais enganos cometidos durante a filtragem.

Além da criação de novas regras, deseja-se também parametrizar algumas das

regras propostas, como as baseadas em tamanhos mínimos e máximos das palavras. Para isso, será verificado através de experimentos a existência de um valor adequado a ser usado como ponto de corte por número de caracteres.

Como análise final, concluímos que as regras testadas atingiram um desempenho satisfatório, considerando a inexistência de uma etapa prévia de treinamento. Para fins de comparação, pretende-se implementar um algoritmo supervisionado de aprendizado de máquina baseada nessas mesmas regras, e verificar em que aspectos as abordagens supervisionada e a não supervisionada se diferenciam.

Referências

Alfonseca, E. and Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. In *In: Proceedings of the 1 st International Conference on General WordNet*.

Bick, E. (2004). A named entity recognizer for danish. In *LREC*. European Language Resources Association.

Collins, M. (2002). Ranking algorithms for named-entity extraction: boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 489–496, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cucchiarelli, A. and Velardi, P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Comput. Linguist.*, 27(1):123–131.

da Silva, J. F., Kozareva, Z., and Lopes, J. G. P. (2004). Cluster analysis and classification of named entities. In *LREC*. European Language Resources Association.

Krishnan, V. and Manning, C. D. (2006). An effective two-stage model for exploiting non-local dependencies in named entity recognition. In Calzolari, N., Cardie, C., and Isabelle, P., editors, *ACL*. The Association for Computer Linguistics.

McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikheev, A. (1999). A knowledge-free method for capitalized word disambiguation. In Dale, R. and Church, K. W., editors, *ACL*. ACL.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher: John Benjamins Publishing Company.

Pasca, M., Lin, D., Bigham, J., Lifchits, A., and Jain, A. (2006). Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pages 1400–1405. AAAI Press.

Shinyama, Y. and Sekine, S. (2004). Named entity discovery using comparable news articles. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.