Extração de nomes de pessoas em textos em português: uma abordagem usando gramáticas locais

Juliana Pinheiro Campos Pirovani, Elias Silva de Oliveira

Extração de nomes de pessoas em textos em português: uma abordagem usando gramáticas locais

Juliana P. C. Pirovani[1], **Elias de Oliveira**[1]

[1]Programa de Pós-Graduação em Informática - Universidade Federal do Espírito Santo (UFES)
Caixa Postal 01-9011, 29060-970 - Vitória - ES - Brasil

juliana.campos@ufes.br, elias@lcad.inf.ufes.br

Abstract. *The automatic identification of person names in texts is an important task in natural language processing. This paper describes an approach for identifying people's names in Portuguese texts, based on a local grammar built after a linguistic study of these texts. A local grammar to the book Senhora, a classic of José de Alencar, was built and subsequently applied to articles in a local newspaper of the Espírito Santo. The goal was to observe the appropriation of a grammar built from a context into another. It was observed that it is possible, but some adaptations are necessary to improve performance. For the book Senhora was obtained 99% of Recall and 100% of Precision. The results are promising.*

Resumo. *A identificação automática de nomes de pessoas em textos é uma tarefa importante no processamento de linguagem natural. Esse artigo descreve uma abordagem para identificar nomes de pessoas em textos em português, com base em uma gramática local construída após um estudo linguístico desses textos. Uma gramática para o livro Senhora, um clássico de José de Alencar, foi construída e posteriormente aplicada a artigos de um jornal local do Espírito Santo. O objetivo foi observar a apropriação da gramática construída para um contexto em outro. Foi observado que isso é possível, mas algumas adaptações são necessárias para melhorar o desempenho. Para o livro Senhora foi obtido 99% de Recall e 100% de Precision. Os resultados são promissores.*

1. Introdução

A pesquisa na área de identificação de nomes de pessoas em textos tem crescido significativamente no meio acadêmico. Em 1995, a Message Understanding Conference - 6 (MUC-6) promoveu o desenvolvimento dessa área ao adicionar como tarefa a identificação e classificação de entidades nomeadas (Named Entity Task) [Grishman and Sundheim 1996]. Um dos três tipos de entidades nomeadas, a ENAMEX, engloba, dentre vários nomes próprios, os nomes de pessoas.

A extração de nomes de pessoas em textos é uma tarefa importante no processamento de linguagem natural e na recuperação de informação (RI). Os nomes de pessoas aparecem com frequência em textos e podem ser considerados uma fonte de informação essencial. No contexto de redes sociais, eles podem ser úteis para identificar a quem o texto se refere e compreender melhor o assunto do texto, possibilitando sua classificação. Além disso, nomes identificados podem ser usados para melhorar a classificação e clusterização de documentos. Em [Friburger et al. 2002], por exemplo,

nomes próprios foram utilizados no cálculo de similaridade, apresentando melhores resultados no processo de clusterização. Em textos jornalísticos analisados por eles, nomes de pessoas representam 39.8% dos nomes próprios.

São muitos os desafios envolvidos na extração automática de nomes de pessoas em textos [Traboulsi 2004]. Dentre eles, o fato de grande parte dos nomes que aparecem nos textos serem palavras desconhecidas e, por isso, nem sempre um dicionário pode auxiliar a busca. Outro problema é que não existe um padrão para escrita de nomes. Algumas vezes o nome completo é apresentado, outras vezes apenas partes do nome aparecem (primeiro nome ou sobrenome) e alguns nomes podem aparecer abreviados. A identificação de nomes depende do idioma, da base de dados e do domínio considerado. Além disso, nomes de pessoas, lugares e organizações são escritos de forma semelhante.

As principais abordagens utilizadas pelos sistemas que buscam identificar nomes automaticamente em textos são: a linguística (baseada em regras) [Black et al. 1998], a probabilística (baseada em aprendizagem de máquina) [Bikel et al. 1997] e a híbrida (combinação das duas anteriores) [Lin 1998]. De acordo com [Friburger and Maurel 2004], a abordagem mais frequente é a linguística. As regras dessa abordagem possibilitam a construção de uma gramática local. "Gramáticas locais são regras que governam a escolha simultânea de um conjunto de palavras usadas em um contexto especialista" [Traboulsi et al. 2004, p.1].

Esse trabalho apresenta uma abordagem baseada em gramática local para identificação de nomes de pessoas em textos escritos na língua portuguesa. Essa abordagem utiliza o contexto em que o nome próprio aparece no texto, mas também faz uso de dicionários.

Esse artigo está estruturado em 6 seções. Na Seção 2 são apresentados trabalhos correlatos que buscam identificar nomes de pessoas em outros idiomas. A Seção 3 apresenta a metodologia utilizada no desenvolvimento do trabalho. A construção de uma gramática local para o texto Senhora de José de Alencar é descrita passo a passo na Seção 4. Na Seção 5 são apresentados os resultados dos experimentos realizados com essa gramática e a Seção 6 apresenta as conclusões e trabalhos futuros.

2. Trabalhos correlatos

Uma abordagem para identificação de nomes em textos jornalísticos escritos em francês é descrita em [Friburger and Maurel 2004]. Para isso, foi realizada uma análise linguística de textos com o objetivo de construir uma série de transdutores de estados finitos que transformam o texto. Essa análise buscou identificar o contexto à direita e à esquerda de nomes de pessoas. O sistema CasSys que implementa a abordagem utilizou ferramentas do sistema Intex. São usados dois transdutores para extração de nomes, sendo que o primeiro usa uma lista de regras descrevendo a gramática local e o segundo usa os nomes identificados pelo primeiro para encontrar os nomes restantes.

Traboulsi [Traboulsi 2004] apresenta uma gramática local para identificação de nomes de pessoas e organizações em textos jornalísticos de notícias financeiras em inglês. Ele propõe um método para identificar automaticamente padrões de nomes próprios que utiliza colocação e análise de concordância para gerar a gramática local. Sua análise mostrou que a palavra "disse" ocorreu em 47% dos contextos contendo nomes próprios. Ele

também apresenta uma ferramenta chamada NExtract para identificar nomes automaticamente. Sua arquitetura possui três componentes: um analisador léxico, um analisador de sublinguagem que usa uma série de transdutores de estados finitos para gerar uma gramática local e um classificador de nomes próprios.

Uma abordagem baseada em gramática local foi avaliada em [Bayraktar and Temizel 2008] com o objetivo de extrair nomes de pessoas em textos de notícias financeiras turcas. Foram realizados os mesmos passos propostos em [Traboulsi 2004]. Apesar de relatar sucesso na extração de nomes, a construção das regras foi muito difícil, devido provavelmente à formação das palavras turcas.

[Traboulsi 2009] também seguiu a abordagem proposta em [Traboulsi 2004], apresentando uma gramática local para extração de nomes árabes. Transdutores de estados finitos e um pequeno vocabulário contendo verbos que indicam ações humanas, preposições e símbolos de pontuação foram construídos com essa finalidade. A avaliação foi realizada para bases de dados pequenas.

Em [Baptista 1998] são apresentadas algumas propriedades linguísticas de nomes próprios considerando o português de Portugal. O objetivo foi auxiliar o processamento automático de textos nessa língua. São apresentadas algumas propriedades das variações formais (flexão de número e gênero) como o fato de alguns nomes próprios aceitarem plural (Ex: Os Antónios são espertos). São apresentadas também algumas restrições combinatórias como a existência de proposições entre nomes (Ex: Maria de Lurdes) e a existência de poucos nomes compostos conectados por hífen (Ex: José-Maria). Essas informações são representadas por autômatos de estados finitos e também é apresentada uma proposta de formalização de nomes próprios em dicionários.

Esse trabalho busca apresentar uma gramática local como foi realizado em [Friburger and Maurel 2004], [Traboulsi 2004],[Bayraktar and Temizel 2008] e [Traboulsi 2009], considerando as particularidades da língua Portuguesa falada no Brasil, que difere também da proposta apresentada em [Baptista 1998].

3. Metodologia

Para alcançar os propósitos desse trabalho foram utilizadas duas bases de dados. Foi necessário identificar manualmente os nomes de pessoas que ocorrem nessas bases para comparar posteriormente com aqueles identificados automaticamente.

A primeira base é o livro Senhora, um clássico de José de Alencar que pode ser obtido no Portal Domínio Público [Domínio Público 2014]. A versão utilizada nesse trabalho é a da Fundação Biblioteca Nacional. A identificação dos nomes nesse texto foi realizada por alunos dos Cursos de Ciência da Computação e Sistemas de Informação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo (CCA/UFES). Foram identificadas 1712 ocorrências de nomes de pessoas nas 124 páginas do livro.

A segunda é uma classe de artigos da base A Tribuna [Tribuna 2014]. Essa é uma coleção de 45907 textos jornalísticos escritos em português publicados pelo jornal local A Tribuna do Espírito Santo. Esses textos estão distribuídos em 21 classes, representando as várias seções do jornal como Cidades, Ciência e Tecnologia, Concursos, Economia, Família, Política, Tudo a Ver, TV Tudo, etc. Para o experimento apresentado nesse trabalho, 30 artigos da classe "Tudo a ver" (TAV) foram utilizados. Essa é a menor classe da

coleção e aborda assuntos diversos como moda, decoração, celebridades e filmes. Nela, foram identificadas 116 ocorrências de nomes de pessoas manualmente, sendo que em 4 artigos não apareciam nomes.

Inicialmente foi realizado um estudo linguístico do livro Senhora. Foi observado como os nomes de pessoas aparecem nesse livro e qual o contexto à direita e à esquerda desses nomes. O objetivo era identificar palavras que pudessem indicar de alguma forma que o que aparece antes ou depois é um nome, ou seja, verificar a existência de regras onde nomes aparecem para posteriormente construir uma gramática local para essa base.

A ferramenta UNITEX [Unitex 2014] foi utilizada para realizar o pré-processamento da base, a construção da gramática local e sua utilização para identificação de nomes. UNITEX é uma versão open-source da ferramenta Intex. É um conjunto de softwares livre para processamento de texto em linguagem natural, contendo ferramentas para pré-processamento de textos, aplicação de dicionários e casamento de padrões. Foi desenvolvida na universidade Marne-La- Vallée (França) tendo como principal desenvolvedor Sébastien Paumier. O artigo [Muniz et al. 2005] apresenta o desenvolvimento dos recursos linguísticos para o Português Brasileiro (UNITEX-PB) nessa ferramenta.

Como a ferramenta UNITEX manipula textos Unicode, os textos analisados foram convertidos para essa codificação antes do pré-processamento. Para representar as regras de gramáticas locais, essa ferramenta permite criar redes de transição recursivas (*recursive transition networks* - RTN), um formalismo semelhante aos autômatos de estados finitos [Sipser and Queiroz 2007]. Essas redes são representadas como grafos.

Após a construção da gramática local e identificação automática de nomes para o livro Senhora, a gramática construída também foi aplicada à classe TAV da base A Tribuna. O objetivo foi observar a apropriação da gramática construída para um contexto em outro contexto. A idéia foi comparar se a gramática construída especificamente para um texto convencional como o livro Senhora, também se aplica a outros textos não tão convencionais como os artigos da base A Tribuna.

A avaliação de desempenho foi realizada analisando as métricas *Recall* e *Precision* apresentadas nas Equações 1 e 2 respectivamente.

$$Recall = \frac{\text{Total de nomes de pessoas identificados corretamente}}{\text{Total de nomes de pessoas realmente existentes no texto}} \quad (1)$$

$$Precision = \frac{\text{Total de nomes de pessoas identificados corretamente}}{\text{Total de nomes de pessoas identificados}} \quad (2)$$

A métrica *Recall* representa a quantidade de acertos no total de nomes existentes no texto. Quanto maior o *Recall*, menor a quantidade de falso-negativos, que corresponde a quantidade de nomes que não foram identificados. Já a métrica *Precision* reflete a quantidade de acertos no total de nomes identificados. Uma alta taxa de *Precision* indica menos falso-positivos, que são palavras que foram identificadas como nomes erroneamente. 100% de *Precision* indica, portanto, a não existência de falso-positivos.

4. Gramática local para extração de nomes de pessoas no livro Senhora

Analisando o livro Senhora, foi observado que todos os nomes se iniciam com letras maiúsculas e que 93% do total de nomes de pessoas são escritos apresentando um único nome. Esse nome único pode ser o primeiro nome ou algum sobrenome que aparece isolado após a primeira aparição do nome completo. Por exemplo, o nome "Aurélia" aparece 586 vezes no texto após ser apresentado inicialmente como "Aurélia Camargo". A única exceção ocorre para nomes de celebridades como "Shakespeare", onde apenas o primeiro nome ou nome artístico é apresentado. Para o restante, mais de dois nomes é apresentado, sendo que preposições podem aparecer entre nomes.

Observou-se que 11,7% dos nomes podem ser reconhecidos através do seu contexto à esquerda. Esses nomes são precedidos por palavras que indicam sua presença. Essas palavras podem ser abreviações de pronomes de tratamento como Sr., Sra. e D., ou verbos que se referem a ações humanas como aconselha, disse, afirmou, dentre outros. Os subgrafos responsáveis por reconhecer regras desse tipo são apresentados nas Figuras 1 e 2. O código PRE é utilizado nos dicionários do UNITEX para representar qualquer palavra iniciada com letra maiúscula.

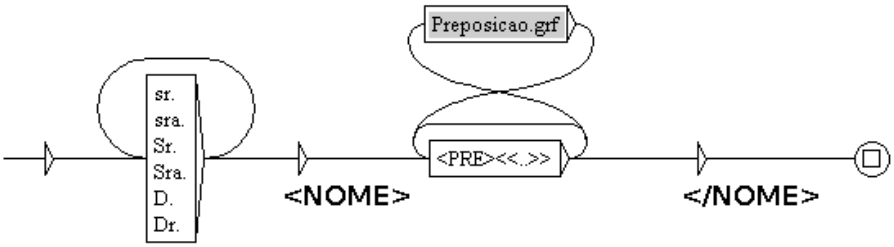

Figura 1. Subgrafo ReconhecePronomesDeTratamento.grf criado no UNITEX

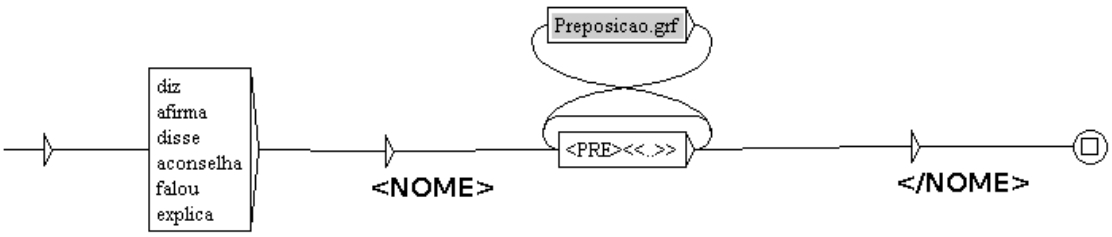

Figura 2. Subgrafo ReconheceAcoesHum.grf criado no UNITEX

Observe que esses subgrafos reconhecem qualquer sequência de palavras iniciando com letras maiúsculas após os pronomes de tratamento ou verbos. As preposições "de", "da", "do", "das" e "dos" que podem aparecer entre nomes são reconhecidas pelo subgrafo Preposicao.grf. Além disso, observe na Figura 1 que os pronomes de tratamento podem se repetir. Isso foi feito para reconhecer nomes que aparecem no texto como "Sra. D. Emília Camargo" e "Sr. Dr. Torquato da Costa Ribeiro". Outros exemplos de nomes reconhecidos pelos subgrafos das Figuras 1 e 2 são: "Sr. Lemos", "D. Firmina" e "disse Aurélia".

Como o contexto não permite identificar uma quantidade considerável de nomes, também foram criadas regras considerando a classificação gramatical e semântica das

palavras após a aplicação de dicionários. Além do dicionário da língua portuguesa, o da língua inglesa também foi aplicado. Isso porque alguns nomes identificados manualmente no livro eram nomes estrangeiros como Shakespeare e Stolz.

O grafo da Figura 3 reconhece duas palavras iniciando com letra maiúscula através dos subgrafos Primeiro_Nome.grf e Ultimo_Nome.grf. Observe que a regra descrita em Ultimo_Nome.grf pode ser repetida diversas vezes.

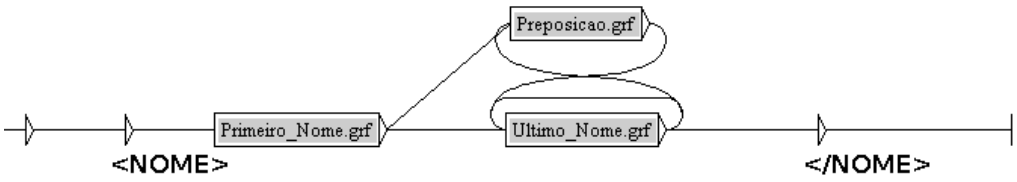

Figura 3. Subgrafo PrimeiroUltimoNome.grf criado no UNITEX

A Figura 4 descreve o reconhecimento do primeiro nome. Um primeiro nome é reconhecido se a palavra possui apenas a primeira letra maiúscula, se não é a palavra "Rua" e é um nome próprio (<N+Pr>). A exceção para a palavra "Rua" foi inserida, para evitar o reconhecimento de nomes de Ruas como "Rua do Ouvidor" e "Rua de Santa Teresa".

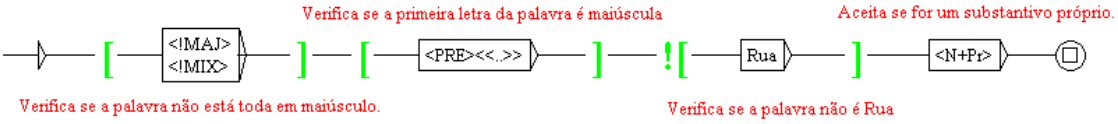

Figura 4. Subgrafo Primeiro_Nome.grf criado no UNITEX

O grafo Ultimo_Nome.grf é semelhante ao da Figura 4, porém aceita palavras que não estão no dicionário. Isso significa que o primeiro nome sempre deve ser um nome de pessoa que consta no dicionário, mas os sobrenomes podem ser palavras desconhecidas. Isso é necessário para reconhecer nomes como "Otávio Feuillet".

A gramática foi aprimorada diversas vezes analisando os falso-positivos e falso-negativos até obter a gramática da Figura 5, composta dos subgrafos apresentados anteriormente.

5. Resultados

Após aplicar a gramática local apresentada na seção anterior para identificação de nomes, foi realizada uma segunda análise do texto buscando identificar os nomes remanescentes, especialmente o primeiro nome ou último nome que aparecem isolados no texto após o nome completo.

Usando essa estratégia para identificação automática de nomes no texto Senhora, foi obtido 99% de *Recall* e 100% de *Precision*, identificando corretamente 1699 nomes no texto. Alguns falso-negativos foram: Amaralzinha, Bernardina, Cretten. 100% de

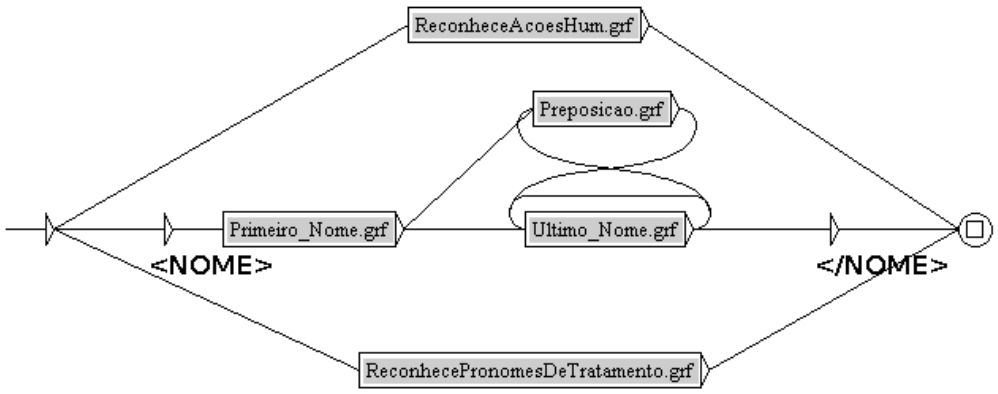

Figura 5. Grafo ReconheceNomes.grf criado no UNITEX

precisão indica que todas as palavras identificadas realmente eram nomes de pessoas, ou seja, não foram encontrados falso-positivos.

É interessante observar o bom desempenho obtido com o livro Senhora. Esse é um texto popular que apresenta um padrão para escrita de nomes e não possui uma variedade de nomes muito grande. Além disso, muitas repetições de nomes e sobrenomes aparecem no texto. Apenas 304 nomes foram identificados pela gramática. Todos os outros foram identificados na segunda etapa que reconhece parte dos nomes encontrados na etapa anterior.

Também é importante destacar que o total de nomes de pessoas realmente existentes no texto, utilizado no cálculo do *Recall*, foi identificado através de uma análise humana que é passível de erro. Foi realizada uma análise dos nomes identificados pelo programa e foram observados nomes identificados por ele que os alunos não identificaram ao analisar o texto em busca de nomes de pessoas. Ou seja, o número de nomes realmente existentes no texto é maior do que 1712, que foram encontrados pelos alunos.

A mesma estratégia foi adotada na classe TAV da base A Tribuna. Nela, foram identificados corretamente 63 dos 116 nomes existentes e 9 falso-positivos, obtendo assim 54% de *Recall* e 88% de *Precision*. Alguns falso-positivos observados foram: Rosa Chá, Brastemp Prêt e Warner Home Vídeo; e alguns falso-negativos foram: Betty Feliz, Adélia Borges e Jô Soares.

Algumas alterações foram feitas na gramática criada para o livro Senhora buscando melhorar o desempenho para classe TAV. Analisando os artigos dessa classe, foi observado que 40% dos nomes podem ser identificados usando o contexto à esquerda. Desses, 93% contém aposto especificativo e são escritos de uma das seguintes formas:

<artigo definido><adjetivo><NOME>

ou

<artigo definido><título ou profissão><NOME>

Exemplos: o carioca Maxime Perelmuter, a bela Gabriela Furlan, a empresária Gisela Maia, etc.

Para reconhecer nomes seguindo essa regra, o subgrafo da Figura 6 foi adicionado à gramática apresentada na Figura 5 . Os códigos DET, A e Hum são códigos utilizados nos dicionários do UNITEX indicando respectivamente: determinante , adjetivo e código semântico usado para se referir a humano. O código Hum é atribuído a palavras como latino, médico, atriz, etc.

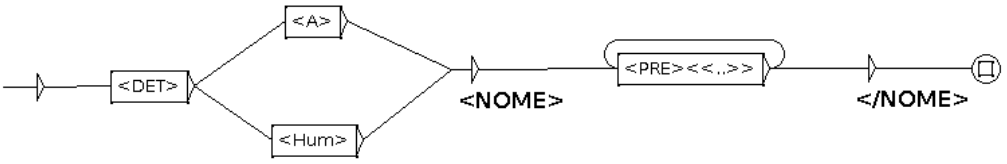

Figura 6. Subgrafo ReconheceAposto.grf criado no UNITEX

Outra alteração foi feita no subgrafo da Figura 4. O primeiro nome é reconhecido se a palavra possui apenas a primeira letra maiúscula e é um nome próprio (<N+Pr>) ou um substantivo relacionado a humano (<N+Hum>). Essa última regra possibilita reconhecer alguns apelidos, por exemplo Mari, no reconhecimento do nome "Mari Nicácio".

Também foi criado um subgrafo chamado ReconheceNomesComp.grf que reconhece nomes próprios compostos identificados pelo dicionário (<N+Pr+Hum>).

Adicionando essas regras, os novos valores de *Recall* e *Precision* foram 71% e 87% respectivamente. A métrica *Precision* diminuiu devido a inserção das novas regras. Inserindo a regra para reconhecer aposto, por exemplo, foi reconhecida a palavra Universia pois ela aparece no texto como "o portal Universia".

Mesmo adaptando a gramática para classe TAV, os valores de *Recall* e *Precision* foram baixos comparados a outros da literatura para sistemas que identificam nomes de forma automática. Segundo [Friburger and Maurel 2004], a maioria dos sistemas obtém uma taxa em torno de 90% em *Recall* e *Precision*. Porém, nesses trabalhos, o contexto à direita e à esquerda dos nomes de pessoas possibilitaram identificar grande parte deles no texto, o que não aconteceu nos textos da classe TAV. Conforme já mencionado, apenas 40% dos nomes foram identificados pelo contexto. Além disso, outro fator que dificulta a identificação é a variedade de nomes apresentados: nomes estrangeiros, nomes artísticos, apelidos, etc. Também aparecem muitos nomes de marcas e empresas, que são escritos de forma semelhante aos nomes de pessoas.

Assim como essa estratégia foi aplicada à classe TAV, ela poderia ser aplicada a outros textos e a gramática seria adaptada para identificar nomes que aparecem com uma estrutura diferente nesse novo contexto. Em uma outra classe da base A Tribuna, chamada Mulher, existem nomes como PAULA Shalders, com o primeiro nome todo em letras maiúsculas, algo não reconhecido pela gramática construída. Outra estrutura diferente apresentada nessa mesma classe é "As jornalistas e editoras de moda Márcia Disitzer e Sílvia Vieira", com nomes separados pela conjunção "e".

6. Conclusão

Nesse trabalho foi construída uma gramática local com o objetivo de identificar nomes de pessoas automaticamente em textos escritos em português. Essa gramática foi construída

para uma base de dados específica, o livro Senhora de José de Alencar, mas foi apresentada a possibilidade de utilizá-la também para identificar nomes de outras bases, como a classe TAV da base A Tribuna.

O desempenho obtido para a classe TAV foi menor do que o obtido para o livro Senhora já que a gramática não foi construída especificamente para ela e devido às particularidades na apresentação de nomes em cada base, confirmando que a identificação automática dos nomes é dependente da base de dados. A classe TAV se mostrou uma base atípica comparada a outras da literatura, devido a diversidade de assuntos abordados e de nomes apresentados.

A principal dificuldade encontrada durante o trabalho foi a criação manual das regras. Esse é um processo lento e as regras precisam ser analisadas e adaptadas de acordo com a base em estudo.

Como trabalho futuro, será estudada a possibilidade de transformar uma gramática local em outra, inserindo adaptações de forma automática para que ela seja capaz de reconhecer, de forma eficiente, nomes em uma outra base. A princípio isso pode ser realizado com o auxílio de um humano informando os nomes que não foram identificados pela gramática original. Além disso, pretende-se: utilizar a gramática construída para identificar nomes em toda a base A Tribuna com o objetivo de encontrar novas regras para reconhecer nomes; e também construir um modelo estatístico com o objetivo de realizar uma auditoria no trabalho humano de identificação de entidades nomeadas em textos, evitando erros no cálculo das métricas que não são considerados atualmente.

Referências

(2014). Jornal a tribuna. [Acesso em: 14/10/2014].

(2014). Portal domínio público. [Acesso em: 15/10/14].

(2014). Unitex. [Acesso em: 15/10/2014].

Baptista, J. (1998). A local grammar of proper nouns. *Seminários de Linguística*, 2:21–37.

Bayraktar, O. and Temizel, T. T. (2008). Person name extraction from turkish financial news text using local grammar-based approach. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*, pages 1–4. IEEE.

Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R. (1997). Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics.

Black, W. J., Rinaldi, F., and Mowatt, D. (1998). Facile: Description of the ne system used for muc-7. In *Proceedings of the 7th Message Understanding Conference*.

Friburger, N. and Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93–104.

Friburger, N., Maurel, D., and Giacometti, A. (2002). Textual similarity based on proper names. In *Proceedings of the workshop Mathematical/Formal methods in Information Retrieval (MFIR â2002) at the 25th ACM SIGIR Conference*, pages 155–167.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING*, volume 96, pages 466–471.

Lin, D. (1998). Using collocation statistics in information extraction. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, volume 66. Citeseer.

Muniz, M. C., Nunes, M. D. G. V., Laporte, E., et al. (2005). Unitex-pb, a set of flexible language resources for brazilian portuguese. In *Proceedings of the Workshop on Technology on Information and Human Language (TIL)*, pages 2059–2068.

Sipser, M. and Queiroz, R. J. G. B. (2007). *Introdução à teoria da computação*. Thomson Learning.

Traboulsi, H. (2004). *A local grammar for proper names*. PhD thesis, MPhil Thesis. Surrey University.

Traboulsi, H. (2009). Arabic named entity extraction: A local grammar-based approach. In *IMCSIT*, pages 139–143. Citeseer.

Traboulsi, H., Cheng, D., and Ahmad, K. (2004). Text corpora, local grammars and prediction. In *LREC*.