Geração de features para resolução de correferência: Pessoa, Local e Organização

Evandro B. Fonseca[1], Renata Vieira[1,] Aline A. Vanin[1]

[1]Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brazil

evandro.fonseca@acad.pucrs.br, renata.vieira@pucrs.br, aline.vanin@ymail.com

Abstract. *This work aims at resolving coreference in Portuguese, focusing on categories of named entities Person, Location and Organization. The proposed method uses supervised learning. To this end, the use of features that assist in the correct classification of named entities is critical. The construction and refinement of these features are of great relevance to his task. The performance of many other tasks depends on the correct output of coreference resolution systems, in special the extraction of relationships between named entities.*

Resumo. *Este trabalho tem por objetivo a resolução de correferência em língua portuguesa, tendo como foco entidades nomeadas das categorias Pessoa, Local e Organização. O método proposto utiliza aprendizado supervisionado. Para tal, o uso de features que auxiliem na correta classificação das entidades nomeadas é fundamental. A construção e o refinamento dessas features são de grande relevância para essa tarefa. O desempenho de muitas outras tarefas depende da correta saída de sistemas de resolução de correferência, em especial, a extração de relação entre entidades nomeadas.*

1. Introdução

Este trabalho tem como foco a identificação de entidades nomeadas e suas cadeias de correferência. O objetivo principal é a resolução de correferências em língua portuguesa para os domínios Pessoa, Local e Organização. A resolução de correferências é uma tarefa relevante e também um grande desafio para a área de linguística computacional. Tratando-se da língua portuguesa, esse desafio é ainda maior. Isto é, a quantidade de recursos para a língua portuguesa é limitada se comparada com a quantidade de recursos que temos disponíveis para outras línguas, como o inglês. Collovini et al. (2011) propõem a extração de relação entre entidades nomeadas presentes em textos da língua portuguesa. A extração desse tipo de relação possui um impacto considerável para a área de processamento da linguagem natural, dado o fato que esse tipo de técnica pode melhorar a performance de muitas tarefas. Nesse contexto, a tarefa de reconhecimento de entidades nomeadas tem como objetivo identificar, desambiguar e atribuir uma categoria semântica a essas entidades, como pessoa, organização, entre outras. É nesse ponto que está a contribuição deste trabalho. Tendo textos como entrada, pretende-se gerar cadeias de correferência para categorias específicas de entidades nomeadas e, com isso, pretende-se contribuir para a extração de relações entre entidades, por meio de inferência. Gabbard et al. (2011) mostram que a resolução de correferência pode prover ganhos significativos para a extração de relação

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 158–162,
Fortaleza, CE, Brazil, October 21–23, 2013.

entre entidades nomeadas. Identificando as várias formas de nos referenciarmos à mesma entidade em um determinado texto, é possível tornar mais eficiente o processo de extração de relação entre entidades. Por exemplo, considere a seguinte sentença: "José da Silva reside próximo à Cidade Baixa, em Porto Alegre. O aluno está no primeiro ano de seu mestrado na PUC-RS.". Identificando e criando uma relação de correferência entre as entidades 'José da Silva' e 'aluno', é possível inferir uma relação direta entre as entidades 'José da Silva' e 'PUC-RS' (José da Silva é aluno da PUC-RS). Ao dizermos que José da Silva é um aluno, pode-se classificá-lo como pessoa, assim como dizer que tem relação com PUC-RS.

2.Trabalhos relacionados

Existem muitos trabalhos na literatura voltados à resolução de correferência. Alguns são puramente baseados em regras, outros utilizam uma abordagem mais dinâmica, baseada em aprendizado de máquina. Na CoNLL 2011 (Conference on Computational Natural Language Learning), Lee et al. (2011) apresentaram seu sistema puramente baseado em regras para a resolução de correferências na língua inglesa, ficando em primeiro colocado. O sistema, "*Stanford's Multi-Pass Sieve Coreference Resolution System*", puramente determinístico, atingiu uma eficiência de 57.79%. Essa eficiência foi medida pela média entre três métricas de desempenho (MUC, B-CUBED e CEAFe), descritas em Pradhan et al. (2011). Em 2012, na CoNLL, Fernandes et al. (2012) apresentaram um sistema de aprendizado de máquina baseado em um algoritmo perceptron. O sistema baseou-se em duas técnicas principais: "*latent coreference trees*" e "*entropy guided feature induction*". Para o português, Silva (2011) propôs um sistema de resolução de correferência baseado em domínios específicos (Pessoa, Local e Organização), utilizando um algoritmo de aprendizado não supervisionado. Seu sistema é dividido basicamente em duas fases, sendo elas: identificação das menções (sintagmas nominais) e características e identificação das cadeias de correferência.

Ainda no âmbito da língua portuguesa, Coreixas (2010) propõe a resolução de correferências com foco nas categorias de entidades nomeadas. Em seu trabalho, para a geração de *features*, a autora faz uso de um analisador sintático proprietário. Esse é o diferencial proposto neste artigo. Um dos objetivos do modelo proposto é a utilização de apenas recursos livres, tendo como foco um produto final aberto a toda comunidade científica. Daí a grande importância para a implementação das *features* selecionadas. Dado o fato de, atualmente, o português possuir poucos recursos, limitar a utilização desses recursos às licenças *open source* pode ser considerado um desafio.

3. Aprendizado de Máquina Voltado à Resolução de Correferências

No contexto de aprendizado de máquina, temos os algoritmos de aprendizado supervisionado. Esses algoritmos aprendem com base em exemplos, previamente classificados. A ideia na utilização desse tipo de algoritmo para o foco deste trabalho é gerar esses exemplos e então codificá-los, de forma que esses algoritmos consigam aprendê-los. É onde entram as *features*. Cada *feature* é responsável por analisar um par de sintagmas nominais, visando identificar se esse par possui determinada propriedade, retornando 'true' ou 'false' para cada uma delas. Por fim, um arquivo é gerado, contendo toda informação proveniente do processamento dessas *features*. Por meio dessa saída, o algoritmo consegue analisar os padrões e aprender com isso, culminando em um modelo de classificação de pares correferentes. Esse modelo é base do sistema de resolução de correferências.

Neste trabalho são utilizados dois corpora: um deles na construção dos pares para treino e outro na avaliação do sistema. Para a construção dos pares, utilizou-se o corpus Summ-it (Collovini (2007)). O Summ-it é um corpus composto por cinquenta textos jornalísticos do caderno de Ciências da Folha de São Paulo. A parte de avaliação será realizada utilizando medidas como abrangência e precisão, sob o corpus do Harem (Freitas et al. (2010)), pelo fato de esse possuir marcação de diversas categorias de entidades e de correferência.

4. Seleção de Features

A seleção de *features* deu-se com base no estudo das features do atual estado da arte. Além disso, Soon et al. (2001) realizaram um experimento com o propósito de verificar o impacto de determinadas features na correta classificação de pares correferentes. Como resultado, os autores constataram que as features String_Match, Alias e Aposto/*Appositive* apresentam um retorno significativo na correta classificação dos pares. Por meio dessas premissas, podemos visualizar as features selecionadas na tabela1.

Tabela 1: Descrição das features

String_Match	Se um SN está contido no outro.
Alias	Se uma das palavras de SN1 é sigla de SN2.
Aposto	Se um SN é aposto do outro.
M_Gênero	Se os sintagmas concordam em gênero (masculino/feminino).
M_Número	Se os sintagmas concordam em número (singular/plural).
IJ_Pronome	Se pelo menos um dos sintagmas possui um pronome.
Cat_Semântica	Se as categorias de entidades (Pessoa, Local, Organização...) são iguais.

A construção dos pares de SNs foi realizada com base nas informações de correferência, contidas no corpus Summ-it. Uma característica levada em consideração na construção desses pares foi a de que todo par deve possuir pelo menos um nome próprio. Como o foco são categorias de entidades, todo SN que se refira a uma pessoa, local ou organização, obrigatoriamente deverá possuir um nome próprio. Tratando-se das *features*, alguns cuidados foram tomados na implementação, de forma a calibrar a precisão e abrangência de cada uma delas. No caso da *feature String_Match*, foram removidos todos os artigos de cada um dos SNs, verificando se SN1 está contido dentro de SN2, ou se SN2 está contido dentro de SN1. Por meio de testes realizados, essa foi a forma que melhor surtiu resultados, minimizando falsos positivos e aumentando sua abrangência e precisão. Para as *features* M_Gênero, M_Número e IJ_Pronome as informações podem ser extraídas do LX-Tagger (2012), um part-of-speech tagger gratuito para o Português.

Para a *feature Alias,* utilizou-se o cálculo de similaridade. Para cada palavra existente em um SN (excluindo stopwords), se a letra inicial for maiúscula, essa é selecionada. Como resultado, tem-se uma string com as iniciais do sintagma com e sem pontos ".", como no exemplo: SN1="Instituto Nacional de Pesquisas Espaciais, Inpe" e SN2="INPE". As siglas geradas pela *feature* serão "I.N.P.E.I." e "INPEI". Após esse

passo, as siglas geradas são comparadas com cada palavra do SN2 pelo cálculo de similaridade. Notemos que 'INPEI' não é exatamente igual ao SN2 'INPE', porém o cálculo de similaridade dará um resultado muito próximo de "1", nos dizendo que as strings são bastante semelhantes. Com isso, conclui-se que SN2 é sigla de SN1.

Na *feature* Cat_Semântica foram utilizados dois recursos que possuem/geram informações referentes à categorias de entidades. O Repentino (REPositório para reconhecimento de ENtidades Nomeadas) Repentino (2013) e um Sistema de Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa (NERP-CRF) (Amaral (2013)). Existem limitações em ambos os recursos utilizados, dentro do contexto de busca utilizado pela *feature* implementada. No caso do Repentino, o problema são as ambiguidades. Por exemplo, ao buscarmos pela entidade 'Amazônia', o Repentino pode conter tanto a categoria 'Organização' quanto 'Local' – respectivamente, Banco da Amazônia e Amazônia. Já o sistema NERP-CRF possui uma taxa de acerto de 83,99%. Isto é, não classifica todas as entidades corretamente. Pensando nessas duas limitações, a ideia foi buscar pelas entidades presentes nos SNs em ambos os recursos, alinhando seus resultados. Dessa forma, a saída torna-se mais confiável e precisa. Para os casos em que o Repentino retorna mais de uma categoria, verifica-se se uma delas combina com a saída do sistema NERP-CRF e, se for o caso, a categoria fornecida por NERP-CRF é atribuída à entidade. Caso não seja encontrada a categoria semântica da entidade no Repentino, são usadas heurísticas para a resolução da categoria, como listas, contendo sinônimos como: Instituto, Empresa, Organização. Essas listas serão construídas por meio de um dicionário. A *feature* 'Aposto' ainda não está implementada, porém sua arquitetura será baseada em regras heurísticas, como uso de vírgulas ou travessões entre dois SNs, como em "Porto Alegre, antiga Porto dos Casais".

5. Conclusão e Trabalhos Futuros

Neste trabalho, foram apresentadas as *features* propostas para gerar um sistema de resolução de correferências, com foco em categorias de entidades nomeadas. Apesar de haver muitos trabalhos relacionados, poucos focam em categorias específicas de entidades. O uso de categorias específicas de entidades nomeadas tem um impacto positivo na tarefa de resolução de correferência, já que cada categoria apresenta características distintas e bem definidas (Coreixas (2010)). Já o motivo pela escolha das categorias selecionadas deu-se objetivando auxiliar sistemas de extração de relação entre entidades nomeadas, como o de Collovini et al. (2011). Como trabalho futuro, visa-se a construção do modelo e posteriormente do sistema de resolução de correferências.

6. Referências

Collovini, S., Grando, F., Souza, M., Freitas L. And Vieira, R., **Semantic Relations Extraction in the Organization Domain**, IADIS International Conference Applied Computing, 2011

Gabbard, R., Freedman, M., Weischedel, R. M. "**Coreference for Learning to Extract Relations: Yes, Virginia,Coreference Matters**". In: Proceedings 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pages 288–293, Portland, Oregon, 2011

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M. and Jurafsky, D. **Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task**, Conference on computational natural language learning, 2011.

CoNLL2011, **Conference on computational natural language learning**, Disponivel em: http://conll.cemantix.org/2011/ Acessado em 05/08/2012.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R. and N. Xue **Modeling Unrestricted Coreference in OntoNotes**, CoNLL Shared Task, 2011.

Fernandes, E., Santos, C. and Milidiú, R. **Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution**, Conference on computational natural language learning, 2012.

Silva, J., **Resolução de Correferência em Múltiplos Documentos Utilizando Aprendizado Não Supervisionado**, Dissertação de Mestrado, USP, 2011.

Coreixas, T., **Resolução de Correferência e Categorias de Entidades Nomeadas**, Dissertação de Mestrado, Pontifícia Universidade Católica Do Rio Grande Do Sul, 2010.

Soon, W. M., Ng, H. T., Lim, D. C. Y. **A Machine Learning Approach to Coreference Resolution of Noun Phrases**". In: Computational Linguistics, pages 521-544, 2001

Collovini, S., Carbonel, T., Fuchs, J., Coelho, J., Rino, L. e Vieira, R., **"Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática"**. In: V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXVII Congresso da SBC, Rio de Janeiro, 2007.

Amaral, D., **Reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa**, Dissertação de mestrado, Pontifícia Universidade Católica do Rio Grande do Sul,2013

Repentino, **REPositório para reconhecimento de ENtidades Nomeadas**, Disponivel em: http://labclup.letras.up.pt/repentino/faq.html, Acessado em 10/03/2013.

Freitas, C., Mota, C., Santos, D., Oliveira, H. and Carvalho P., **Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese**, Linguateca, FCCN, 2010.

LX-Tagger, **Language Resources and Technology for Portuguese University of Lisbon**, NLX-Natural Language and Speech Group Disponível em: http://lxcenter.di.fc.ul.pt/tools/en/LXTaggerEN.html, Acesso em: 05/12/2012