Uso de Informações Semânticas na Identificação de Anáforas Indiretas e Associativas

Luiz Carlos Ribeiro Jr.[1], **Sandra Collovini**[1], **Patrícia N. Gonçalves**[1], **Vinicius Muller**[1]
Renata Vieira[1]

[1]Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Unisinos 950 – 93022-000 – São Leopoldo – RS – Brasil

{lucarijr,scollovini,pnunesg,vmuller}@turing.unisinos.br

renatav@unisinos.br

Abstract. *This paper presents the evaluation of semantic features for anaphora classification, showing the usefulness of such features to this process.*

Resumo. *Este trabalho avalia o uso de características semânticas na classificação de expressões anafóricas, mostrando que elas são úteis nesse processo.*

1. Introdução

A Resolução de Anáforas é uma tarefa de grande importância para diversas áreas de Processamento de Linguagem Natural (PLN), entre as quais estão Extração de Informação, Sumarização de Textos, Tradução Automática etc. [Collovini and Vieira 2006a] propõe a classificação de expressões anafóricas em 4 classes: *Novas-no-Discurso*, *Associativas*, *Diretas* e *Indiretas* (seção 2), com base em um conjunto de 16 características morfossintáticas. Porém, devido ao fato das expressões *Novas-no-Discurso* existirem em número muito maior, e isso ser pernicioso para o modelo gerado na fase de treinamento, esse trabalho propôs um balanceamento das classes. Os resultados reportados em [Collovini and Vieira 2006a] foram ainda insatisfatórios para as *Indiretas* e *Associativas*. Em [Coelho et al. 2006], foi proposta uma abordagem baseada no uso de informações semânticas para a resolução de anáforas nominais, através da qual foram obtidos bons resultados. Este artigo propõe utilizar esse tipo de informação na classificação. Sendo assim, o objetivo deste trabalho é propor o uso de características semânticas para auxiliar na classificação de anáforas, utilizando as informações semântica gerada pelo parser PALAVRAS [Bick 2000]. Cabe salientar que todos os experimentos analisam somente as descrições definidas (sintagmas nominais iniciados por artigos definidos).

2. Classes, Corpus e Características

Neste trabalho, as classes das descrições definidas consideradas foram derivadas de [Prince 1981, Vieira and Poesio 2000], sendo apresentadas a seguir:

- **Novas-no-discurso**: descrições definidas que introduzem entidades que são novas no discurso, ou seja, que não foram mencionadas anteriormente no texto.
- **Associativas**: descrições definidas que introduzem novas entidades no discurso, entretanto, possuem uma dependência semântica com uma expressão antecedente.

- **Diretas**: descrições definidas que possuem relação de identidade com seus antecedentes e apresentam o mesmo nome-núcleo.
- **Indiretas**: descrições definidas que possuem uma relação de identidade com seus antecedentes, porém, possuem diferentes nomes-núcleo.

Os experimentos apresentados neste artigo foram realizados num corpus composto por 24 textos jornalísticos da Folha de São Paulo (FSP), escritos em Português do Brasil, correspondendo a uma parte do corpus do NILC[1], que foi utilizado no aprendizado e um outro corpus constituído por 25 textos do jornal Público (Publico), escritos em Português Europeu, pertencentes ao corpus CETEMPublico[2], usado na validação do classificador. Os textos foram anotados automaticamente com informações sintáticas utilizando o parser PALAVRAS e manualmente com informações de correferência utilizando a ferramenta MMAX[3]. Os resultados da anotação manual do corpus FSP e Público são apresentados na tabela 1.

Table 1. Anotação manual do corpora

Corpora (%)	Diretas	Indiretas	Associativas	Novas-no-discurso	**Total**
FSP	285 (27%)	116 (11%)	94 (9%)	550 (53%)	1045
Publico	326 (22%)	135 (9%)	75 (5%)	961 (64%)	1497

Neste trabalho, além de 16 características morfossintáticas, baseadas num estudo da literatura e uma análise do corpus FSP detalhados em [Collovini and Vieira 2006b], utilizamos mais 2 de natureza semântica. As novas caracteristicas foram propostas a partir de uma análise das informações semânticas fornecidas pelo parser PALAVRAS, visando usá-las como um auxiliar na distinção entre as classes apresentadas, principalmente as *Associativas* e *Indiretas*. Segundo [Bick 2006] existem cerca de 160 etiquetas semânticas organizadas de tal forma que, dependendo do contexto textual, um substantivo pode ter uma ou mais delas a ele associadas. Para tirar proveito dessas informações foram propostas as características *SEM_NOT_DIR* e *SEMANTIC_WINDOW*, ambas levando em consideração somente o núcleo do sintagma nominal. A característica *SEM_NOT_DIR* é representada através de valores booleanos (*true* ou *false*). Durante seu processamento, o texto é percorrido do sintagma nominal que está sendo analisado até sua primeira sentença, esta assumindo o valor *true* se for encontrado algum sintagma nominal de nome-núcleo diferente e, no mínimo, uma etiqueta semântica igual. A característica *SEMANTIC_WINDOW* funciona de forma semelhante, porém, esta é representada por valores numéricos, ao invés dos booleanos. Seu valor é definido contando-se o número de sintagmas nominais que validam a condição de teste (a mesma da característica anterior). Outra particularidade desta característica é que as comparações são feitas de acordo com uma ”janela” (de tamanho 8) que indica o número máximo de sentenças anteriores no texto que podem ser consideradas.

3. Experimentos de Classificação

Nesta seção são descritos os experimentos realizados utilizando características semânticas na tarefa de classificação das expressões anafóricas. Para demonstrar as vantagens de utilizarmos tais características, os resultados obtidos são comparados com

[1]http://www.nilc.icmp.usp.br/nilc
[2]http://www.linguateca.pt/CETEMPublico
[3]http://www.eml-research.de/english/research/nlp/download/mmax.php

[Collovini and Vieira 2006a], baseados somente em características morfossintáticas. Para um melhor entendimento, convencionamos chamar os experimentos utilizando as características semânticas de *InfSem* e os realizados em [Collovini and Vieira 2006a] de *Baseline*. É importante salientar que para a classificação foi utilizado o algoritmo *J48* do pacote Weka[4] com *10-fold cross validation*. Para avaliação dos resultados foram utilizadas as medidas de precisão (P), abrangência (A), f-measure (F) e taxa de acertos (C). Vale lembrar também que todas as bases utilizadas nos testes foram balanceadas através da duplicação das suas instâncias, esse método foi utilizado em [Collovini and Vieira 2006a], onde é melhor descrito. Os resultados obtidos no experimento de classificação das expressões anafóricas estão na tabela 2.

Table 2. Resultados da classificação com 4 classes

	Classes	P	A	F	C
Baseline	novas-no-discurso	49%	47%	48%	51%
	associativas	43%	77%	55%	
	diretas	68%	78%	73%	
	indiretas	46%	4%	7%	
InfSem	novas-no-discurso	60%	40%	48%	59%
	associativas	57%	70%	63%	
	diretas	73%	78%	76%	
	indiretas	47%	50%	49%	

Com relação as classes *Novas-no-discurso* e *Diretas*, ocorreram algumas alterações nos valores da precisão e abrangência. Entretanto, para ambas as classes a f-measure manteve-se praticamente igual. Para as classes *Associativas* e *Indiretas* foram obtidos ganhos relevantes em suas taxas, principalmente nas *Indiretas*. Quanto as *Associativas*, apesar de ter ocorrido uma pequena perda na abrangência, ocorreram ganhos na precisão, o que acarretou em uma melhor F-measure. Já com relação as *Indiretas*, apesar da precisão ter praticamente se mantido, a taxa de abrangência subiu de 4% para 50%. Essa diferença fez com que a f-measure do *InfSem* fosse para 49%, contra 7% do *Baseline*. Analisando a taxa de acertos do *Baseline* e do *InfSem*, podemos nos enganar sobre as vantagens obtidas através do uso das características semânticas. Isto por que apresentam resultados próximos. Entretanto, no *Baseline* não foram classificadas praticamente nenhuma *Indiretas*, o mesmo não ocorrendo no *InfSem*. Além disso, experimentos de validação do classificador com uma base de dados nova e sem balanceamento foram realizadas. Essa base foi construída a partir do corpus Público e os resultados são apresentados na tabela 3. Analisando os resultados é possível observar uma perda nas taxas de acertos comparado ao corpus de aprendizado. Entretanto, isto é natural, já que a Árvore de Decisão busca se especializar nos exemplos de treino. Já a classe *Diretas* apresentou perda na precisão, em compensação, ocorreu ganho na abrangência. Para as classses *Associativas* e *Indiretas* podemos concluir que foram obtidos bons resultados.

4. Considerações Finais e Trabalhos Futuros

O uso de informações semânticas para a classificação de expressões anafóricas mostrou-se promissor. Apesar de ficar claro os melhores resultados na classificação das *Indiretas* e *Associativas* do *InfSem* para o *Baseline*, algumas limitações foram constatadas. A dificuldade de diferenciar as classes *Indiretas* e *Associativas* da classe *Novas-no-discurso*

[4]http://www.cs.waikato.ac.nz/ml/weka/

Table 3. Resultados da validação com 4 classes

	Classes	P	A	F	C
Baseline	novas-no-discurso	85%	47%	60%	52%
	associativas	1%	59%	17%	
	diretas	61%	88%	72%	
	indiretas	11%	4%	6%	
InfSem	novas-no-discurso	85%	42%	56%	50%
	associativas	13%	49%	20%	
	diretas	61%	87%	72%	
	indiretas	12%	26%	17%	

mostrou-se um dos principais problemas, influenciando até mesmo a validação. Entretanto, as informações semânticas podem ainda ser melhor exploradas, através da elaboração das características propostas. Acreditamos que a grande contribuição do trabalho é demonstrar que o uso das informações semânticas melhora o desempenho do classificador. Sendo assim, como trabalho futuro serão realizados estudos mais aprofundados sobre as informações semânticas presentes no PALAVRAS. A partir desse estudo podem ser propostas características mais robustas e complexas, por exemplo, levando em consideração a hierarquia das informações semânticas.

References

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Arhus University, Arhus.

Bick, E. (2006). Noun sense tagging: Semantic prototype annotation of a portuguese treebank. In Hajič, J. and Nivre, J., editors, *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006)*, Prague, Czech Republic.

Coelho, J. C. B., Muller, V., Collovini, S., Vieira, R., and Rino, L. (2006). Resolving portuguese nominal anaphora. In Renata Vieira, Paulo Quaresma, M. d. G. V. N. N. M. C. O. . M. C. D., editor, *7th Workshop on Computational Processing of Written and Spoken Language (PROPOR'2006)*, Itatiaia, RJ. Springer.

Collovini, S. and Vieira, R. (2006a). Anáforas nominais definidas: balanceamento de corpus e classificação. In *IV Workshop de Tecnologia da Informação e Linguagem Humana TIL*, Ribeirão Preto, SP. Proceeding of the Brazilian Symposium on Artificial Intelligence.

Collovini, S. and Vieira, R. (2006b). Learning discourse new references in portuguese texts. In *IFIP Conference on Artificial Intelligence - IFIP AI 2006. IFIP World Computer Congress (WCC2006)*, Santiago, Chile.

Prince, E. F. (1981). Toaward taxonomy of given-new information. In *P. Cole, editor Radical Gramatics*, pages 223–256, New York. Academic Press.

Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579.