Identificação de Sintagmas Nominais na Língua Portuguesa

Aprendizado de Máquina II

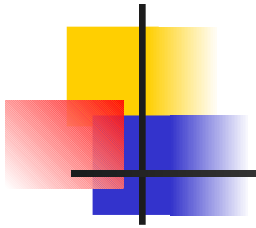
Eduardo Cardoso
Iam Jabour
Prof.: Ruy Luiz Milidiú

O que é Sintagma Nominal?

- Cícero dos Santos (2005): “sintagma consiste num conjunto de elementos que constituem uma unidade significativa dentro da sentença e que mantêm entre si relações de dependência e de ordem”.

- Perini (2003): “É o sintagma que pode ser sujeito em alguma oração”.

- Exemplo: Esse professor é um neurótico.
  - Esse professor é SN
  - Um neurótico é SN



Motivação

- Apresentar resultados com a técnica SVM na língua portuguesa
  - (Kudo & Matsumoto, 2001) precisão de 94.15% e abrangência de 94.29% em inglês, utilizando comitê
- Utilizar comitês na tentativa de melhorar os resultados obtidos na disciplina AM I.
- Utilizar mais técnicas apresentadas na literatura
- Comparar os resultados com ETL e TBL

Corpora

- Corpora SRN-CLIC
  - 5268 (sentenças) registros já rotulados e
  - Utilizados no trabalho (Santos, 2009)

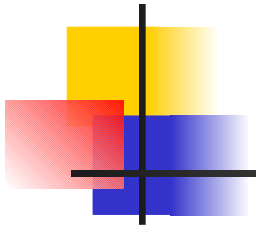
- Utilizado o conjunto de etiquetas IOB1 (Sang e Veenstra, 1999):
  - I (*In*) – pertence a um SN
  - O (*Out*) – não pertence a um SN
  - B (*Begin*) - primeira palavra de um SN que apareceu imediatamente depois de um outro SN.

Resultados Existentes

Corpora SRN-CLIC

	Precision %	Recall %	F1 %
TBL[1]	87.17	88.26	87.71
ETL[1]	89.66	89.51	89.58
SVM[2]	80.48	82.34	81.40
J48[2]	83.26	85.43	84.33

1 – Resultados apresentados por [Santos, 2009]
2 – Resultados obtidos na disciplina AM I



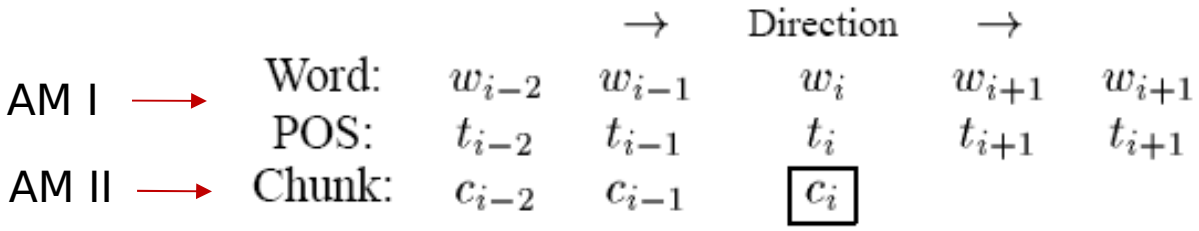
Abordagem

- Framework de data-mining WEKA.
  - Framework de data-mining amplamente difundido e com diversos recursos avançados para a tarefa.
- Utilização da predição anterior
- Utilização da técnica comitê
  - Geração de um conjunto de modelos;
  - Utilização de modelos SVM junto aos de árvore de decisão.

Abordagem

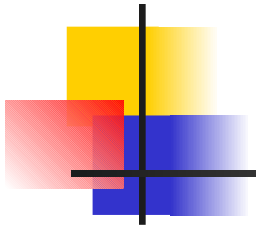
- Features
  - Exemplo:

O/ART/I rato/N/I comeu/V/O o/ART/I queijo/N/I ././O



W_{-2}	W_{-1}	W_0	W_{+1}	W_{+2}	T_{-2}	T_{-1}	T_0	T_{+1}	T_{+2}	C_0
O	rato	comeu	o	queijo	ART	N	V	ART	N	O

W_{-2}	W_{-1}	W_0	W_{+1}	W_{+2}	T_{-2}	T_{-1}	T_0	T_{+1}	T_{+2}	C_{-1}	C_0
O	rato	comeu	o	queijo	ART	N	V	ART	N	I	O



Experimento

- Envolve o contexto ao redor da palavra
  - i-ésima palavra: janela [-k, +j]:
    - k palavras e classes gramaticais atrás da palavra (i)
    - j palavras e classes gramaticais para frente da palavra (i)
- Menor contexto: aprende pouco
- Maior contexto: muito específico
- Utilização da janela aprendida em AM I [-3,+2]

Resultados

Corpora SRN-CLIC

	Precision %	Recall %	F1 %
SVM^{f2}	83.95	80.20	82.03
SVM^{b2}	80.29	74.22	77.14
J48^{f2}	85.91	86.76	86.33
J48^{b3}	87.32	87.32	87.32

f – forward
b – backward

Resultados

Corpora SRN-CLIC

	Precision %	Recall %	F1 %
TBL[1]	87.17	88.26	87.71
ETL[1]	89.66	89.51	89.58
SVM[2]	83.95	80.20	82.03
J48[2]	87.32	87.32	87.32

1 – Resultados apresentados por [Santos, 2009]
2 – Melhores resultados obtidos na disciplina AM II

Referências

- Weka Machine Learning Project. Disponível em http://www.cs.waikato.ac.nz/~ml/index.html

- Taku Kudo , Yuji Matsumoto, Chunking with support vector machines, Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001, p.1-8, June 01-07, 2001, Pittsburgh, Pennsylvania.

- PERINI, M. A. Gramática Descritiva do Português. Editora Ática, São Paulo, 4 edição, 2003. ISBN 85-08-05550-1.

- Santos, Cícero Nogueira dos; Milidiú, Ruy Luiz. Aprendizado de Transformações Guiado por Entropia. Rio de Janeiro, 2009. 85p. Tese de Doutorado —Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.