Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem

Autores

Cid Ivan da Costa Carvalho

Davis Macedo Vasconcelos

Orientador: Leonel Figueiredo de Alencar

Sumário

Proposta

Para Kveton e Oliva (2002):
Os **erros de etiquetagem** constituem desvios das regularidades que se espera que o sistema aprenda, resultando num modelo falso da língua.

Um **método de correção** desses erros num processo de pós-etiquetagem, levando em conta a detecção de *n-gramas* impossíveis na língua a ser modelada.

Metodologia

Como ponto de partida deste trabalho, compilamos um pequeno *corpus* de textos de comunicação mediada por computador (comentários de blog sobre educação).

sistematizamos os erros cometidos pelo etiquetador morfossintático **Aelius**.

Erros de *unigramas*

unigramas impossíveis são erros os quais estão absolutamente e localmente detectados, ou seja, fora do contexto.

Palavra	Etiqueta do Aelius	Etiqueta correta	Categoria da etiqueta correta
professor	NPR	N	Nome
Oi	N	INTJ	Interjeição
Escola	NPR	N	Nome
escolar	VB	ADJ-G	Adjetivo

Erros de *bigramas*

Bigramas impossíveis, num corpus etiquetado, um podem ser resultados de mal formação do texto fonte e de **marcação incorreta** dos dados de treinamento.

Palavra	Etiqueta do Aelius	Etiqueta correta	Categoria da etiqueta correta
Concordo	N	VB-P	Verbo no presente
Envio	VB-P	N	Nome
disponibilize	N	VB-SP	Presente do subjuntivo

O contextualização dos erros de bigramas

Concordo/N totalmente/ADV contigo/P+PRO ./.

Envio/VB-P de/P verbas/N-P para/P este/D fim/N …

... que/C o/D município/N disponibilize/N uma/D-UM-F verba/N …

Considerações

Após essa classificação, vamos elaborar um primeiro conjunto de regras para **correção automática** da etiquetagem que corrija os unigramas e os bigramas impossíveis.

Um segundo conjunto de regras procuraremos detectar e corrigir os *n-gramas* impossíveis.