

Universidade Federal do Rio Grande do Sul  
Grupo de Processamento de Linguagens Naturais  
Projeto Expressões Multipalavras

# Verificação automática de substantivos compostos através de reconhecimento de padrões sintáticos

Daniel Emilio Beck (Instituto de Informática UFRGS - [debeck@inf.ufrgs.br](mailto:debeck@inf.ufrgs.br))  
Aline Villavicencio (Instituto de Informática UFRGS - [avillavicencio@inf.ufrgs.br](mailto:avillavicencio@inf.ufrgs.br))  
Marco Idiart (Instituto de Física UFRGS - [marco.idiart@gmail.com](mailto:marco.idiart@gmail.com))

---

---

# *Roteiro*

- Introdução e Motivação
  - Trabalhos Relacionados
  - Detecção de Substantivos Compostos
  - Trabalhos Futuros
  - Referências
- 
-

# Introdução e Motivação

- Expressões Multipalavras (EMs): designam fenômenos lingüísticos sintática (*ad hoc*) semântica (*kick the bucket*) ou estatisticamente (*strong/?powerful/?potent coffee*) idiossincráticos
  - Inclui classes como substantivos compostos (*chocolate factory*), expressões idiomáticas (*spill the beans*), verbos frasais (*rely on*)
  - São um desafio para sistemas de PLN
- 
-

# *Introdução e Motivação*

- Necessita-se de teorias, recursos e métodos lingüísticos e computacionais para o seu tratamento
  - O projeto Expressões Multipalavras tem por objetivo investigar e desenvolver tratamentos e recursos computacionais adequados para PLN
  - Este trabalho visa automatizar a detecção de substantivos compostos (*bus stop, family car, beer glass, orange juice,...*) para a construção de recursos lexicais
- 
-

# *Trabalhos Relacionados*

- Interpretação de Compostos Nominais: Lapata e Keller (2005), Nicholson e Baldwin (2006), Girju (to appear)
  - Outros aspectos de Compostos Nominais: Baldwin e Bond (2003)
  - Interpretação de Verbos de Suporte (light verbs): Fazly e Stevenson (to appear)
  - Detecção de EMs: Zhang et al (2006), Villavicencio et al (2007)
- 
-

# Introdução - Proposta

- Hipótese: “*Um provável nome composto deverá ter uma flexibilidade sintática restrita, ou seja, haverá poucas ou nenhuma ocorrência de formas sintáticas alternativas do candidato*”
  - A partir de uma lista de candidatos, verificar se cada candidato aparece em outras formas sintáticas dentro de um corpus
- 
-

# *Introdução - Exemplo*

- Considerando “bus stop” como um possível candidato
  - Dentro de um corpus, deverão ocorrer muitas ocorrências de “bus stop” em sua forma sintática original
  - Entretanto outras formas sintáticas como “buses stop” ou “stop of bus” deverão aparecer com uma frequência muito menor
  - O que é um indício forte de “bus stop” ser um nome composto
- 
-

# *Detecção de Substantivos Compostos*

- Inicialmente, obter as frequências da forma sintática original e de outras formas a serem selecionadas.
  - Para nomes compostos do tipo N1 N2, são obtidas as frequências da expressão original e de outras expressões montadas variando o número de N1 e/ou N2
  - Repete-se o processo para N2 N1
- 
-



# *Detecção de Substantivos Compostos*

- Por exemplo, para “bus stop”, procuraria-se por “buses stop”, “bus stops” e “buses stops”, além de “stop bus” e variações
- Outras estruturas sintáticas poderão ser analisadas com o decorrer do trabalho

# *Detecção de Substantivos Compostos*

- A partir de uma lista de candidatos e de uma lista de expressões N N que não são nomes compostos, obter as frequências de todas as expressões dentro do corpus
  - De acordo com as frequências obtidas, será verificada a possibilidade de usar uma métrica estatística que permita responder se a expressão é de fato um nome composto
- 
-

# *Detecção de Substantivos Compostos*

- O corpus a ser utilizado será a Web, através da pesquisa automática do Yahoo®
- Serão utilizados também o dicionário CIDE, para obter os plurais das palavras.



# *Trabalhos Futuros*

- Podem ser utilizadas outras estruturas sintáticas, como por exemplo utilizar sinônimos para cada um dos N dentro da expressão
  - Caso o resultado das frequências seja satisfatório, pode-se estender o método a outras formas de nomes compostos, como N P N por exemplo.
- 
-

# Referências

- *Afsaneh Fazly, Suzanne Stevenson and Ryan North. Automatically Learning Semantic Knowledge about Multiword Predicates. To appear in Language Resources and Evaluation.*
- *Jeremy Nicholson and Timothy Baldwin. Interpretation of Compound Nominalisations using Corpus and Web Statistics. Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July 23, 2006, Sydney, Australia*
- Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007), pages 1034-1043, Prague, Czech, 2007.
- Yi Zhang, Aline Villavicencio, Valia Kordoni, and Marco Idiart. Automated multiword expression prediction for grammar engineering. In Proceedings of the ACL 2006 workshop on multiword expressions: identifying and exploiting underlying properties, pages 36-44, Sydney, Australia, 2006.
- Timothy Baldwin and Francis Bond. Learning the countability of English nouns from corpus data, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, p.463-470, July 07-12, 2003, Sapporo, Japan
- Mirella Lapata and Frank Keller: Web-based models for natural language processing. ACM Transactions on Speech and Language Processing (TSLP) 2(1): 1-31 (2005)

# Verificação automática de substantivos compostos através de reconhecimento de padrões sintáticos

Daniel Emilio Beck (Instituto de Informática UFRGS - [debeck@inf.ufrgs.br](mailto:debeck@inf.ufrgs.br))

Aline Villavicencio (Instituto de Informática UFRGS - [avillavicencio@inf.ufrgs.br](mailto:avillavicencio@inf.ufrgs.br))

Marco Idiart (Instituto de Física UFRGS - [marco.idiart@gmail.com](mailto:marco.idiart@gmail.com))

---

---