Grupo de Processamento de Linguagem Natural do Instituto de Informática (PLN-II)

Aline Villavicencio[1,2]

[1]Instituto de Informática, Universidade Federal do Rio Grande do Sul (Brasil)

[2]Department of Computer Sciences, Bath University (Inglaterra)

Evento de Integracao de PLN
PUC-RS, Janeiro de 2008

Roteiro

1. Integrantes
2. Principais Temas de Pesquisa
3. Common Background

Outline

1 Integrantes

2 Principais Temas de Pesquisa

3 Common Background

Institutos de Informática e Letras da UFRGS

Instituto de Informática

- Aline Villavicencio
- Bruno Menegola
- Carlos Ramisch
- Daniel Beck
- Daniel Germann
- Fernanda Pimenta
- Mario Machado
- Otavio Acosta
- Paulo Schreiner

Institutos de Informática e Letras da UFRGS

Instituto de Letras

- Adriano Zanette (Instituto de Informática)
- Anna Becker Maciel
- Cleci Regina Bevilacqua
- Maria José Finatto
- Leonardo Zilio
- Viviane Possamai

Alguns Colaboradores

- Edson Prestes (Instituto de Informática, UFRGS)
- Jean Paul Sansonnet (LIMSI-CNRS, França)
- Marco Idiart (Instituto de Física, UFRGS)
- Maity Siqueira (Instituto de Letras, UFGRS)
- Maria Alice Pimenta Parente (Instituto de Psicologia, UFRGS)
- Patrícia Jaques (PIPCA, Unisinos)
- Rosa Vicari (Instituto de Informática, UFRGS)
- Sylvie Pesty (IMAG, França)
- Valia Kordoni (Saarland University, Alemanha)
- Viviane Orengo (Instituto de Informática, UFRGS)
- Yi Zhang (Saarland University, Alemanha)

Outline

1 Integrantes

2 Principais Temas de Pesquisa

3 Common Background

Aquisição Automática de Informações Lingüísticas

Modelos para grande escala

- métodos estatísticos
- grandes quantidades de dados
- sem preocupação com plausabilidade cognitiva dos dados, recursos ou algoritmos empregados
- foco em Expressões Multipalavras

Aquisição Automática de Informações Lingüísticas

Modelos para grande escala

- Carlos Ramisch, Marco Idiart, Valia Kordoni, Yi Zhang
- Daniel Beck, Marco Idiart
- Adriano Zanette, Maria José Finatto
- Fernanda Pimenta, Jean Paul Sansonnet, Patrícia Jaques, Rosa Vicari, Sylvie Pesty
- Otávio Acosta, Viviane Orengo
- Paulo Schreiner

Aquisição Automática de Informações Lingüísticas

Modelos Cognitivos-Computacionais

- preocupação com plausabilidade cognitiva dos dados, recursos e algoritmos empregados
  - dados compatíveis com os vistos por uma criança
  - dados processados em ordem, com apenas uma passada pelo corpus
  - aprendizado dinâmico e online

Aquisição Automática de Informações Lingüísticas

Modelos Cognitivos-Computacionais

- Mário Machado, Marco Idiart
- Bruno Menegola, Edson Prestes, Maity Siqueira, Maria Alice Pimenta Parente

Outline

1 Integrantes

2 Principais Temas de Pesquisa

3 Common Background

Background

Multiword Expressions

- Syntactic or semantic properties cannot be derived from their parts [Sag et al., 2002, Villavicencio, 2005]
- phrasal verbs (e.g. *come along*), nominal compounds (e.g. *frying pan*), institutionalised phrases (e.g. *bread and butter*)
- equivalent in number to single words in speakers' lexicon [Jackendoff, 1997]
- fixed (*ad hoc*) vs flexible (*touch/find a nerve*) expressions
- opaque (*kick the bucket*) vs transparent (*eat up*) semantics

Motivation

Challenge for NLP

It is difficult to provide a unified account for the detection of these distinct but related phenomena.

Grammar Engineering

- Lexical coverage is the major barrier to broad-coverage linguistically deep processing
- MWEs comprise a significant part of the missing lexicon

Identification of MWEs

- Given a list of sequences of words to distinguish MWEs (e.g. *in the red*) from random sequences of words (e.g. *of alcohol and*)

Motivation

Challenge for NLP

It is difficult to provide a unified account for the detection of these distinct but related phenomena.

Grammar Engineering

- Lexical coverage is the major barrier to broad-coverage linguistically deep processing
- MWEs comprise a significant part of the missing lexicon

Identification of MWEs

- Given a list of sequences of words to distinguish MWEs (e.g. *in the red*) from random sequences of words (e.g. *of alcohol and*)

Motivation

Challenge for NLP

It is difficult to provide a unified account for the detection of these distinct but related phenomena.

Grammar Engineering

- Lexical coverage is the major barrier to broad-coverage linguistically deep processing
- MWEs comprise a significant part of the missing lexicon

Identification of MWEs

- Given a list of sequences of words to distinguish MWEs (e.g. *in the red*) from random sequences of words (e.g. *of alcohol and*)

Identification of MWEs

[Zhang et al., 2006]

- Error mining based MWE detection
- New MWE entries created with automated lexical acquisition
- Grammar coverage improves significantly

[Villavicencio et al., 2007]

- Validation steps more thoroughly evaluated: for statistical approaches there are two important questions
  - How reliable is the corpus used?
  - How precise is a statistical measure to distinguish the phenomena studied?
- Grammar accuracy investigated

For Further Reading I

Jackendoff, R. (1997).
Twistin' the night away.
Language, 73:534–59.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).
Multiword expressions: A pain in the neck for NLP.
In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Villavicencio, A. (2005).
The availability of verb-particle constructions in lexical resources: How much is enough?
Journal of Computer Speech and Language Processing, 19.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. (2007).
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)*, pages 1034–1043.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006).
Automated multiword expression prediction for grammar engineering.
In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.