Universidade Federal do Rio Grande do Sul
Instituto de Informática
Programa de Pós-Graduação em Computação

Identificação e Tratamento de Expressões Multipalavras aplicado à Recuperação de Informação

Otávio Costa Acosta

{ocacosta@inf.ufrgs.br}

Orientadora: Profa. Dra. Aline Villavicencio

Co-orientadora: Profa. Dra. Viviane Orengo

Roteiro

- Motivação
- Introdução
- Expressões Multipalavras
- Objetivo
- Atividades

Motivação

- Refinar sistemas de RI utilizando informações de PLN
  - Incorporando informações sobre Expressões Multipalavras (EMs)
    - Indexação de termos atômicos pode causar perda semântica
    - Ex.: Banco de Dados
  - RI-Multilingüe

Introdução

PLN para RI

• Sistemas de RI utilizam métodos estatísticos tradicionais que não utilizam PLN

  • Termo composto identificado a partir de uma determinada freqüência.

• PLN é conveniente para a RI [Sparck-Jones, 1997]

  • Melhor gerenciamento para identificação de conteúdo

Introdução

EMs para RI

• Seleção de termos adequados é importante para a melhoria de sistemas de RI

• Indexação de termos atômicos pode causar perda semântica

• Existe a hipótese que as EMs gerem uma melhoria significativa na precisão de sistemas de RI

Expressões Multipalavras

O que são Expressões Multipalavras (EMs)?

• Seqüencias de palavras que têm características sintáticas, semânticas ou estatísticas idiossincráticas.

- Substantivos Compostos – Ex.: *bus stop*
- Verbos Frasais – Ex.: *break down*

• Significado diferente caso sejam analisadas separadamente

Ex.: *banco de dados*, *kick the bucket*

Expressões Multipalavras

Exemplo de Expressão Multipalavra

- “Bode expiatório”
  - Bode – animal
  - Expiatório – expiação, castigar ou sofrer uma pena
  - Bode expiatório – alguém escolhido para levar a culpa

Expressões Multipalavras

- Natureza variada
- Classes com características específicas
  - *Institutionalized phrases, lexicalized phrases*
- Dificuldade de reconhecimento

Expressões Multipalavras

Dificuldades no Reconhecimento de EMs

- Determinar os limites de uma EM
  - Ex.: Banco de Dados Relacional x Banco de Dados x Banco x Dados Relacional x Dados x etc.
- Divergência entre línguas
  - EMs em Português x EMs em Inglês

Técnicas para o descobrimento de EMs

- Simbólicos
- Estatísticos
    - Lapata e Keller, 2005
    - Baldwin e Bond, 2003
    - Nicholson e Baldwin, 2006
    - Zhang et al., 2006
    - Villavicencio et al., 2007

Objetivo

Hipótese

• Integração de informações sobre EMs em sistemas de Recuperação de Informação Multilingües (RI-ML) supõe-se trazer melhorias nos resultados obtidos;

Atividades

- Investigar a aplicação de métodos para a identificação e tratamento automático de Expressões Multipalavras (EMs) encontradas em corpora;

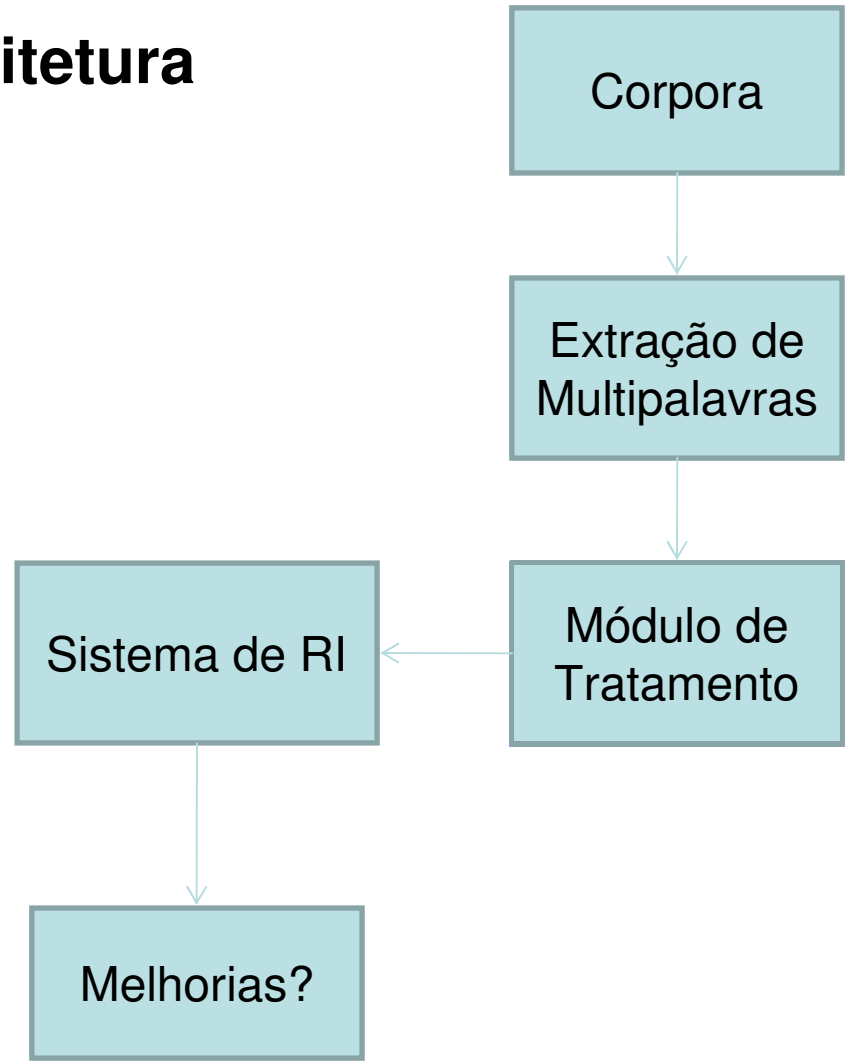
- Criação de um dicionário de EMs para RI monolíngüe;

- Criação de um dicionário de tradução de EMs para RI multilíngüe;

- Integração de módulo de tratamento de EMs em sistemas de RI; (Arquitetura)

- Avaliação empírica de potenciais benefícios dos resultados obtidos. (CLEF)

Arquitetura



Atividades

- Investigar a aplicação de métodos para a identificação e tratamento automático de Expressões Multipalavras (EMs) encontradas em corpora;

- Criação de um dicionário de EMs para RI monolíngüe;

- Criação de um dicionário de tradução de EMs para RI multilíngüe;

- Integração de módulo de tratamento de EMs em sistemas de RI; (Arquitetura)

- Avaliação empírica de potenciais benefícios dos resultados obtidos. (CLEF)

CLEF

- CLEF - http://www.clef-campaign.org/

  - Avaliação imparcial
  - Campanha de avaliação que ocorre em várias etapas
  - Participantes recebem uma coleção de testes
  - Posteriormente recebem os tópicos de consulta
  - Resultados são enviados à coordenação do CLEF
  - É dado o resultado da avaliação
  - Finaliza com um workshop

Bibliografia

- Baldwin, T., Bannard, C., Tanaka, T., Widdows, D. - ***An Empirical Model of Multiword Expression Decomposability*** - 2003
- Baldwin, T., Villavicencio, A. - ***Extracting the Unextractable: A Case Study on Verb-particles*** - 2002
- Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A. - ***Towards Best Practice for Multiword Expressions in Computational Lexicons*** - 2002
- Coelho, A. - ***Stemming para a Língua Portuguesa: estudo, análise e melhoria do algoritmo RSLP*** - 2007
- Cross-Language Evaluation Forum (CLEF) - Disponível em: ***http://www.clef-campaign.org -*** Acesso em: outubro 2007
- Evans, D. A. and Zhai, C. - ***Noun-phrase analysis in unrestricted text for information retrieval*** – 1996
- Evert S., Krenn B. - ***Using small random samples for the manual evaluation of statistical association measures*** - 2005
- Manning, C., Shütze, H. – ***Foundations of Statistical Natural Language Processing***
- Orengo, V.M. ***A study on the use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval*** - 2006
- Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickiger, D. - ***Multiword Expressions: A Pain in the Neck for NLP*** – 2002
- Salton, G. and McGill, M. J. - ***Introduction to Modern Information Retrieval*** - 1983
- Sparck Jones, K. – ***What's the role of NLP in Text Retrieval*** – 1997
- Villada Moirón, B., Tiedemann, J. - ***Identifying idiomatic expressions using automatic word-alignment(.pdf)*** - 2006
- Villavicencio, A., Baldwin, T., Waldron, B. - ***A Multilingual Database of Idioms*** - 2004
- Villavicencio, A., Bond, F., Korhonen, A., McCarthy, D. – ***Introduction to the Special Issue on Multiword Expressions: Having a crack at a hard nut*** – 2005
- Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., Ramisch, C. - ***Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering*** – 2007.
- Zhang, Y., Kordoni, V., Villavicencio, A., Idiart, M. - ***Automated Multiword Expression Prediction for Grammar Engineering*** - 2006

Dúvidas? Sugestões?

{ocacosta@inf.ufrgs.br}

Universidade Federal do Rio Grande do Sul
Instituto de Informática
Programa de Pós-Graduação em Computação

Identificação e Tratamento de Expressões Multipalavras aplicado à Recuperação de Informação

Otávio Costa Acosta

{ocacosta@inf.ufrgs.br}

Orientadora: Profa. Dra. Aline Villavicencio

Co-orientadora: Profa. Dra. Viviane Orengo