Extração de Terminologia Bilíngue Usando Corpora Comparáveis

Kassius Vargas Prestes

Agenda

1 Tarefa

2 Trabalhos Relacionados

3 Objetivos e Motivação

4 Extração Monolíngue

5 Extração Multilíngue

Agenda

1 Tarefa

2 Trabalhos Relacionados

3 Objetivos e Motivação

4 Extração Monolíngue

5 Extração Multilíngue

Tarefa

- Extração de termos automática de copora de domínio específico
  - Um termo é uma representação de um conceito.

    (Ramish, 2012)

    - Ex: neurocognição, robótica, processamento de linguagem natural, aprendizagem de máquina
- Foco em termos multipalavra
  - Ex: engenharia do software, banco de dados relacional, inteligência artificial
- Domínio Específico
  - Engenharia do Software
    - ex: modelo conceitual, rede sem fio
  - Conferências
    - ex: full paper, abstract submission

Agenda

1 Tarefa

2 Trabalhos Relacionados

3 Objetivos e Motivação

4 Extração Monolíngue

5 Extração Multilíngue

Trabalhos Relacionados

- **Extração de Termos**
  - (Ramish, 2012) Propõe um framework aberto e genérico para o tratamento de expressões multipalavra.
  - (Frantzi; Ananiadou; Tsujii, 1998) Introduz o método c-value/nc-value para o reconhecimento automático de termos multipalavra.
  - (Pecina, 2009) Investiga mais de 80 medidas de associação para identificação de expressões multipalavra.
  - (Evert; Krenn, 2009) Verifica a influência do tamanho e do tipo do corpus na identificação de MWEs.

- **Extração Multilíngue**
  - (Tsvetkov; Winter, 2012) Extração de expressões multipalavra de pequenos corpora paralelos.
  - (Caseli et al; 2010) Extração de expressões multipalavra baseada em alinhamento
  - (Daille; Morin, 2005) Extração de terminologia Francês-Inglês de corpora comparáveis

Agenda

1 Tarefa

2 Trabalhos Relacionados

3 Objetivos e Motivação

4 Extração Monolíngue

5 Extração Multilíngue

Objetivos e Motivação

- Objetivo geral:
  - Construir dicionários e ontologias multilíngues para ajudar sistemas de tradução automática
- Objetivos Específicos:
  - 1) Identificar MWEs de corpora de domínio específico
  - 2) Identificar MWEs equivalentes para um par de línguas usando corpora paralelo e comparável
  - 3) Identificar as características sintáticas e semânticas mais adequadas para a extração multilíngue
  - Resultado: Dicionários monolíngues e multilíngues de MWEs de domínio específico

Objetivos e Motivação

- Motivação:
  - Português é uma língua com muito poucos recursos eletrônicos de linguagem
- É difícil de obter um corpus paralelo de um domínio específico
  - O corpora paralelos existentes são mais gerais: Europarl, legendas de filmes
- Como é a performance da extração multilíngue para Português com corpora comparáveis?
  - corpora paralelo = traduções dos mesmos textos
  - corpora comparável = mesmo domínio, textos diferentes

Agenda

1 Tarefa

2 Trabalhos Relacionados

3 Objetivos e Motivação

4 Extração Monolíngue

5 Extração Multilíngue

Extração Monolíngue

- Metodologia:
  - Pré-processamento
  - Filtro Linguístico
  - Filtro Estatístico

Pré-processamento

- Tokenização do texto
  - Expressões Regulares
  - Diferente para cada corpus
    - Termos no GENIA contém caracteres como +-[ ]()
      - ca2+-modulating cyclophilin ligand
      - v-(d)-j recombinase activity

- Extrair todos n-gramas do texto
  - n entre 2 e 4 (parametrizável)
    - exemplos de MWEs: human immunodeficiency virus, retinoic acid
    - exemplo de não-MWEs: human immunodeficiency, expression in, virus type, present study

Filtro Linguístico

- Filtra os n-gramas extraídos pelo part-of-speech(POS)
  - N N (Substantivo Substantivo) ex: blood monocytes
  - A N (Adjetivo Substantivo) ex: nuclear factor
  - N N N ex: tumor necrosis factor-alpha
  - A N N ex: peripheral blood lymphocytes
  - N A N
  - N N A
  - ...
- Essa etapa nos permite descartar:
  - of the, in the, ...
  - basicamente expressões com artigos e preposições que são muito frequentes e não interessantes

Filtros Estatísticos

- Medidas para ordenar os candidatos restantes após o filtro linguístico
  - Frequencia
  - c/nc-value (Frantzi; Ananiadou; Tsujii, 1998)
  - Medidas de Associação
    - Pointwise Mutual Information (pmi)
    - Log-Likelihood Ratio
    - Poisson Stirling Measure
    - Mutual Information

Agenda

1 Tarefa

2 Trabalhos Relacionados

3 Objetivos e Motivação

4 Extração Monolíngue

5 Extração Multilíngue

Extração Multilíngue

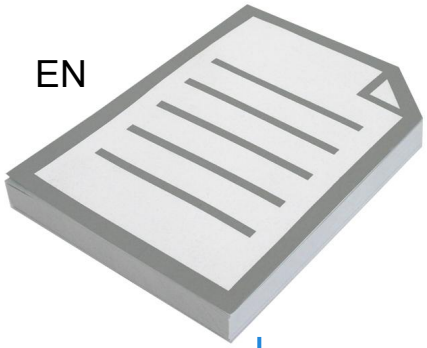
- Extração monolíngue em cada língua;
- Encontrar os termos correspondentes para um dado par de línguas.
  - Por exemplo, queremos encontrar o par:
    - wireless network (EN) = rede sem fio (PT)

Extração Multilíngue

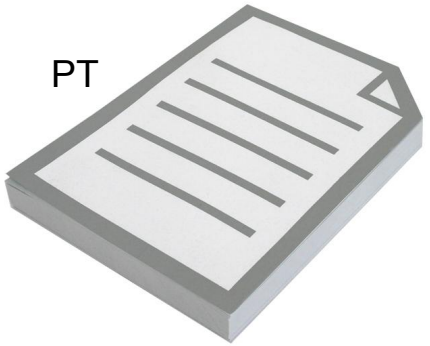
- 3 métodos
  - Método Composicional
    - (Morin; Daille, 2012)
  - Alinhamento de Vetores de Contexto
    - (Déjean; Gaussier; Sadat, 2002)
    - (Morin; Daille, 2005)
  - Método Composicional com Projeção de Contexto
    - (Morin; Daille, 2012)

Extração Multilíngue

- Extrai termos de ambas línguas
  - Etapa comum a todos métodos



EDSON ALVES MENEZES JUNIOR	674535
ERICO FABRICIO MONTEIRO CAVALCANTE	657241
FABIO COSTA DOS SANTOS	692182
FRANCISCO HARISON DA SILVA GOMES	647521
GISELI DA SILVA SIQUEIRA	650544
HELDER COSTA SOUSA	637444
HENDREW GUSTAVO DOS SANTOS	703789
HISSAMY LOPES SARAIVA	672522
IVAN LUIZ SIVA RIBEIRO	607390
JESSICA CLINTIANE CORREA DA SILVA	605123
JHONATA RIBEIRO DA SILVA	641357
JOAO KELVY SILVA DA SILVA	619047
JOAO VICTOR COSTA LIMA	663039
JOSE WILLIAM OLIVEIRA SOARES	702402
JULYANE STEPHANIE PIRES DOS SANTOS	673415
LORENA DA SILVA TRZECIAK	644844
MARIANA DE OLIVEIRA BRAGA	623083
MARLLEN DE SOUZA SANTOS	702290
NIANDRO DE SOUZA MARQUES	640721
PRISCILA DE LIMA MORAIS	664333
RANGEL DE FREITAS ALVES	600233
RODRIGO MONTEIRO DOS SANTOS	673438
ROGERIO DOS SANTOS ALVES	700159
RYLLA BRYANNE MORAES	600383
TAIANE DOS ANJOS SANTOS	623196
TIAGO SOARES MENDES	617547
VITOR AFONSO CARLI ALVES	695756



EDSON ALVES MENEZES JUNIOR	674535
ERICO FABRICIO MONTEIRO CAVALCANTE	657241
FABIO COSTA DOS SANTOS	692182
FRANCISCO HARISON DA SILVA GOMES	647521
GISELI DA SILVA SIQUEIRA	650544
HELDER COSTA SOUSA	637444
HENDREW GUSTAVO DOS SANTOS	703789
HISSAMY LOPES SARAIVA	672522
IVAN LUIZ SIVA RIBEIRO	607390
JESSICA CLINTIANE CORREA DA SILVA	605123
JHONATA RIBEIRO DA SILVA	641357
JOAO KELVY SILVA DA SILVA	619047
JOAO VICTOR COSTA LIMA	663039
JOSE WILLIAM OLIVEIRA SOARES	702402
JULYANE STEPHANIE PIRES DOS SANTOS	673415
LORENA DA SILVA TRZECIAK	644844
MARIANA DE OLIVEIRA BRAGA	623083
MARLLEN DE SOUZA SANTOS	702290
NIANDRO DE SOUZA MARQUES	640721
PRISCILA DE LIMA MORAIS	664333
RANGEL DE FREITAS ALVES	600233
RODRIGO MONTEIRO DOS SANTOS	673438
ROGERIO DOS SANTOS ALVES	700159
RYLLA BRYANNE MORAES	600383
TAIANE DOS ANJOS SANTOS	623196
TIAGO SOARES MENDES	617547
VITOR AFONSO CARLI ALVES	695756

Método Composicional

- Tradução
  - Usa um **dicionário bilíngue** para traduzir cada palavra de cada termo da língua de origem
- Geração dos candidatos
  - Gera todas as $n! \prod_{i=1}^{n} t_i$ combinações possíveis de candidatos a tradução do termo
- Seleção das traduções
  - Seleciona os candidatos que estão nos termos encontrados na língua alvo de acordo com a frequência

Alinhamento de Vetores de Contexto

- Construção dos Vetores de Contexto
  - palavras que co-ocorrem com o termo extraído
  - junto com uma medida de associação
    - pmi
    - log-likelihood

```
17395  Performing cost-benefit analyses to determine whether requirements are best met
17396  Developing partitioning algorithms (and other processes) to allocate all preser
17397  Partitioning large systems into (successive layers of) subsystems and component
17398  Interfacing with the design and implementation engineers and architects, so tha
17399  Ensuring that a maximally robust design is developed.
17400  Generating a set of acceptance test requirements, together with the designers,
17401  Generating products such as sketches, models, an early user guide, and prototyp
17402  Ensuring that all architectural products and products with architectural input
17403  [edit] Systems architect: topics
17404  Large systems architecture was developed as a way to handle systems too large f
```

Alinhamento de Vetores de Contexto

- acceptance test
  - generating 0.98
  - set 0.5
  - requirements 1.2
  - together 0.32
  - designers -0.25
  - ...

Alinhamento de Vetores de Contexto

- Traduz os Vetores de Contexto de uma língua
  - Precisa de um **dicionário bilíngue**

- acceptance test
  - generating - gerando 0.98
  - set - conjunto 0.5
  - requirements - requisitos 1.2
  - together - juntos 0.32
  - designers - designers -0.25
  - ...

Alinhamento de Vetores de Contexto

- Compara os Vetores de Contexto
  - Similaridade de Cossenos
- Os vetores mais próximos são as traduções mais prováveis.

acceptance test

gerando - generating	0.98
conjunto - set	0.5
requisitos - requirements	1.2
juntos - together	0.32
designers - designers	-0.25

...

teste de aceitação

conjunto	0.8
requisitos	0.2
designers	0.15
engenharia	0.45
software	-0.39

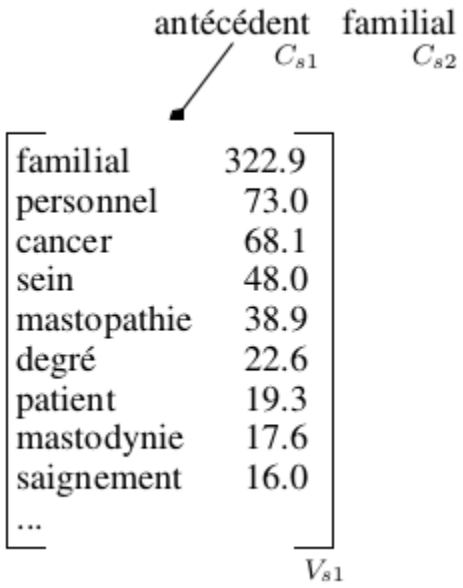
...

Projeção de Contexto

1. Extração de Contexto
2. Transferência para a língua destino
3. Geração das traduções candidatas
4. Ordenamento das traduções candidatas

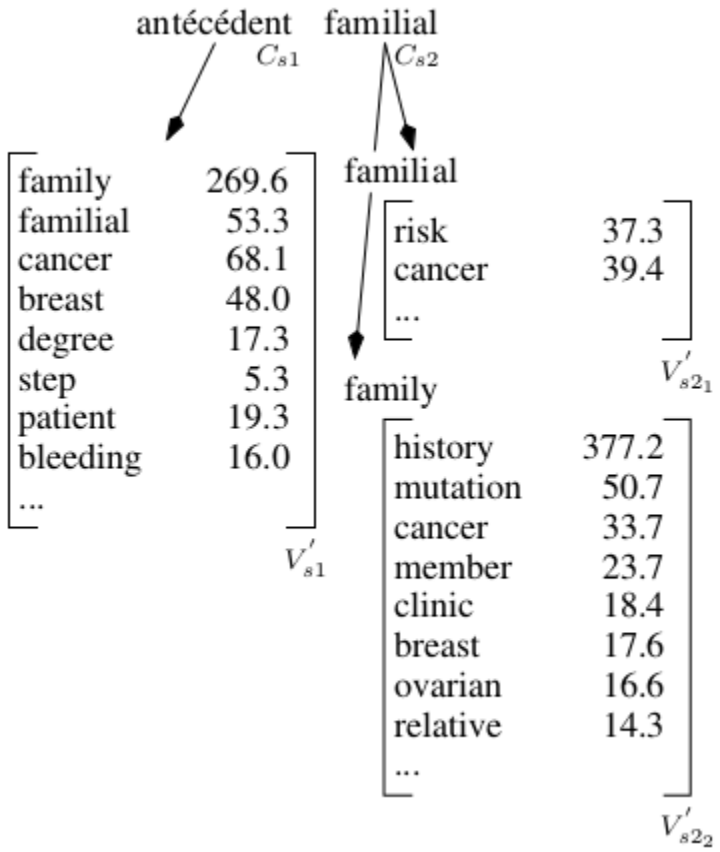
Projeção de Contexto

1. Extração de Contexto
   - Busca em um dicionário bilíngue por cada palavra do termo, se não encontra substituí por informação de contexto



Projeção de Contexto

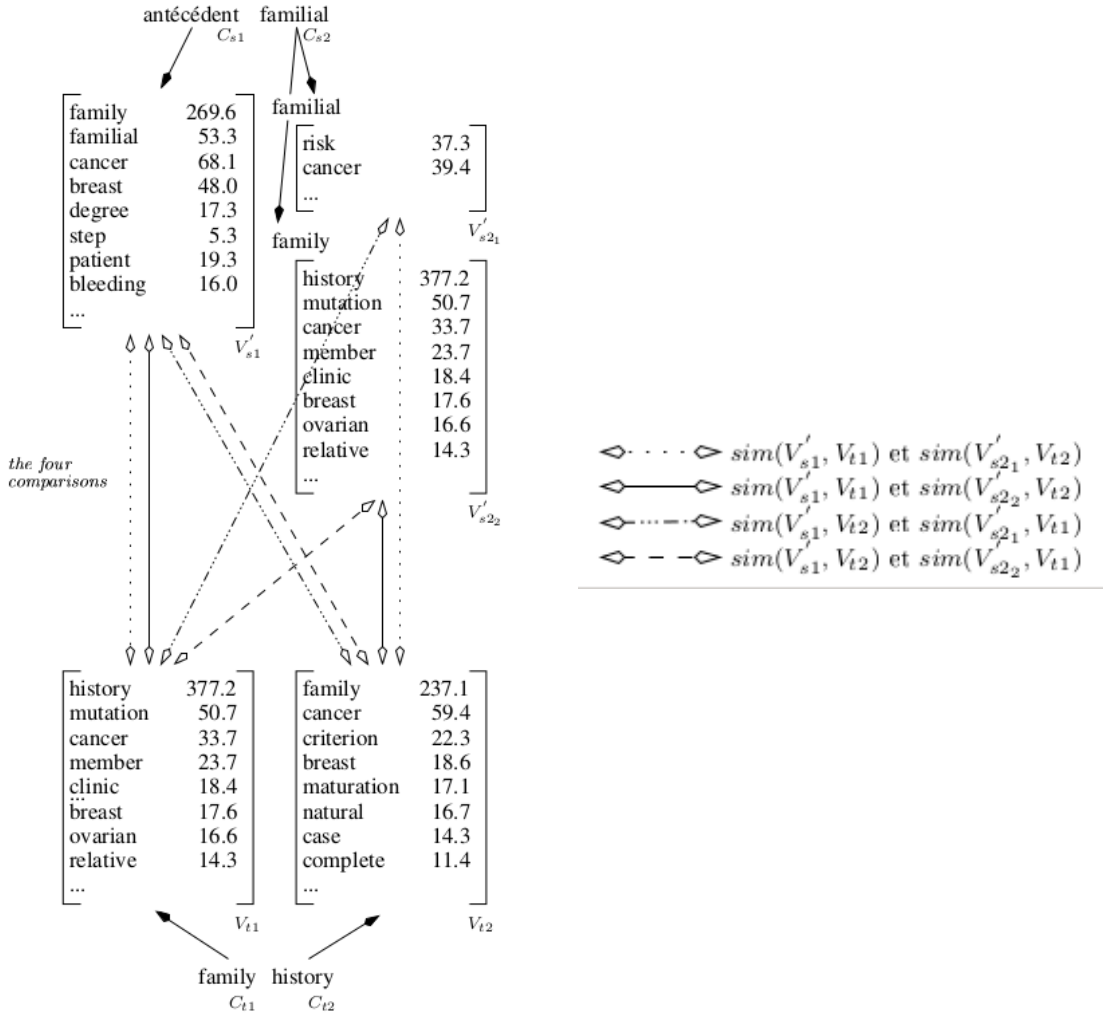
2. Transferência para a língua destino
   - palavra no dicionário:
     - Busca contexto na língua alvo
   - palavra fora do dicionário:
     - Traduz contexto da língua origem para língua alvo usando dicionário.



Projeção de Contexto

3. Geração das traduções candidatas
    - Todas as possibilidades de alinhamento são consideradas no cálculo das similaridades entre os vetores de contexto do termo na língua origem na língua alvo.

Projeção de Contexto



Projeção de Contexto

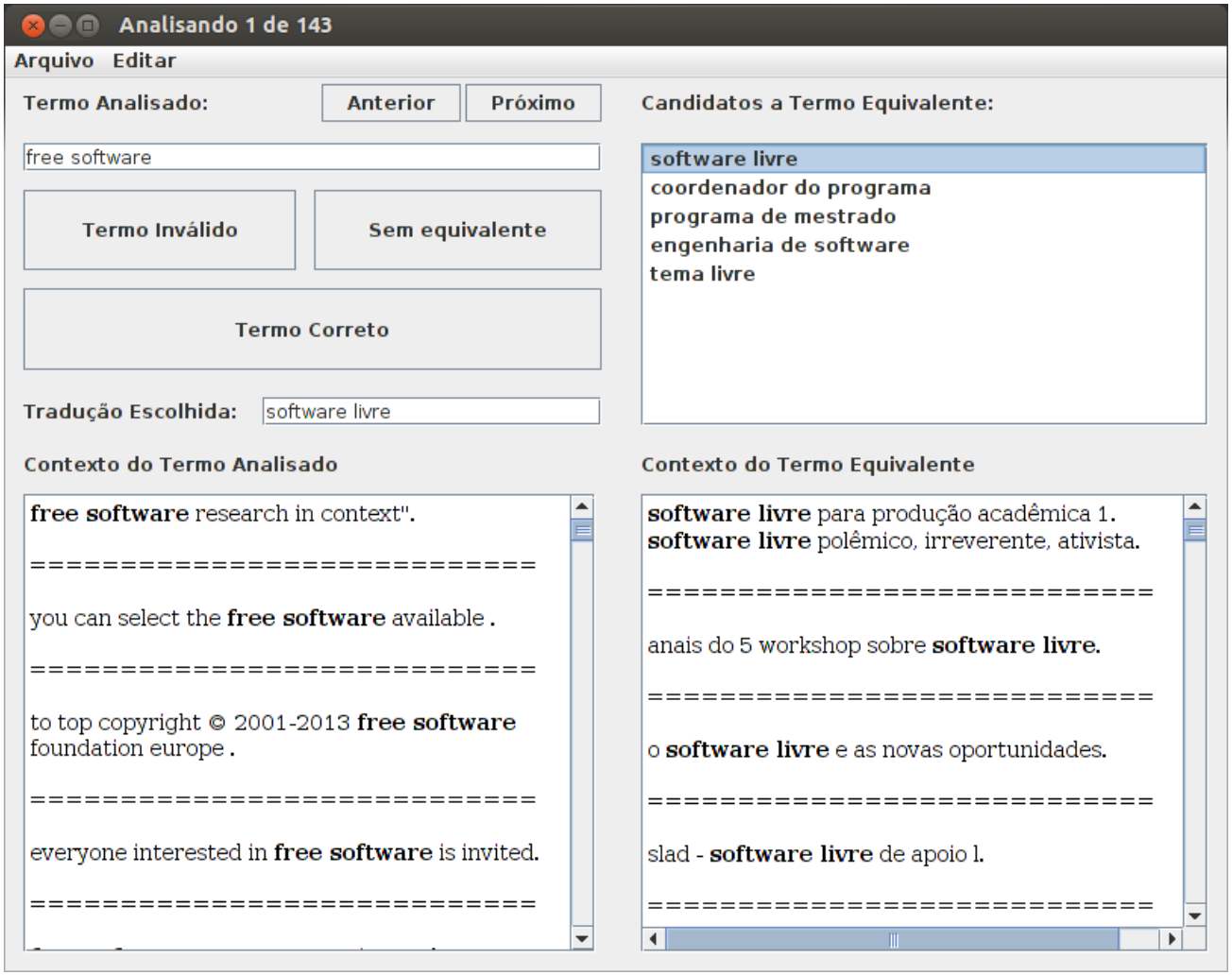
4. Ordenamento das traduções candidatas
    - As traduções candidatas são ordenadas pela ordem decrescente dos scores de similaridade

Exemplo de Resultados

- Top n traduções candidatas

phone features	números de telefone	0.2834
	estabelecimentos comerciais	0.2784
	suas pesquisas	0.2754
	nome do contato	0.2733
	recursos do telefone	***0.2694***
light sensor	***sensor de luz***	***0.2832***
	luz de fundo	0.2475
	alto-falante do dock	0.2244
	brilho da tela	0.2011
	suas configurações	0.1982

Avaliação

- 
Analisando 1 de 143
Arquivo Editar
Termo Analisado:
Anterior
Próximo
free software
Termo Inválido
Sem equivalente
Termo Correto
Tradução Escolhida:
software livre
Candidatos a Termo Equivalente:
software livre
coordenador do programa
programa de mestrado
engenharia de software
tema livre
Contexto do Termo Analisado
free software research in context".
==========
you can select the free software available .
==========
to top copyright © 2001-2013 free software foundation europe .
==========
everyone interested in free software is invited.
==========
Contexto do Termo Equivalente
software livre para produção acadêmica 1.
software livre polêmico, irreverente, ativista.
==========
anais do 5 workshop sobre software livre.
==========
o software livre e as novas oportunidades.
==========
slad - software livre de apoio l.
==========

Obrigado

Kassius Vargas Prestes
kassiusvargasprestes@gmail.com

Cosine Similarity

- A, B are vectors
- Each "dimension" of the vector is a context word
- Each A_i is the velue of the association of the context word with the term.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$