Aquisição léxica de verbos-partícula utilizando corpora paralelos

Paulo Schreiner
<paulo.schreiner@gmail.com>

PLN-II
Instituto de Informática - UFRGS

9 de janeiro

Introdução

- Aplicações de PLN muitas vezes dependem da qualidade e abrangência de recursos lingüísticos
- Criação manual destes recursos cara, trabalhosa
- Obter correspondências entre expressões multi-palavra em inglês e português
- Utilizando recursos disponíveis:
  - Corpus paralelo alinhado por sentença (Europarl)
  - Etiquetador morfossintático (Treetagger)
  - Ferramente para gerar léxico bilíngüe (NATools)
- Alinhador léxico que utiliza informações de bigramas
- Extração de EMPs a partir de alinhamentos gerados pelo alinhador.

Léxicos bilíngües

- Estabelece correspondência entre palavras entre 2 línguas
- Geração utilizando corpora paralelos
  1. Escolha de pontuação a ser usada
  2. Cálculo das pontuações
  3. Ordenação em ordem decrescente
  4. Seleção das melhores entradas
- Associações indiretas

Sell house.
Vender casa.

- Multi-palavras
- Exemplo: NATools

Alinhadores léxicos

- Estabelecem correspondência entre palavras de 2 frases paralelas
- Amenizam o problema das associações indiretas
- *sell* $\Leftrightarrow$ *vender* e *house* $\Leftrightarrow$ *casa*

But this programme must go on.
Mas este programa dever viver.

but	mas
this	este
programme	programa
must	dever
go on	viver
.	.

Tabela: Exemplo de alinhamento léxico

LIHLA

É um alinhador léxico

- Parte de léxico bilíngüe
  1. Correspondência exata
  2. Melhor candidato de acordo com o léxico bilíngüe
     - Heurística para EMPs
  3. Cognatos
- LIHLA trabalha com palavras simples
- Gera erros de alinhamento para algumas EMPs

. . . should not tempt us into give up old ingredient. . .
. . . mas não dever cair em erro de desistir de antigo conteúdo. . .

- Não consegue alinhar *give up* com *desistir*

Greedy Aligner

- Utiliza léxico bilíngüe com unigramas e bigramas

This can not go on.
This This-can can can-not not not-go go go-on on on-. .

- Ordena os alinhamentos possíveis em ordem descrescente
- Primeiro alinhamento da lista é considerado válido
- Outros alinhamentos que contenham tokens presentes no primeiro são retirados da lista
- Repete até a lista ficar vazia

Greedy Aligner - exemplo

Frases

this cannot go on .
isto não poder continuar .

Associações

. $\Leftrightarrow$. $\rightarrow$ 0.7972
cannot $\Leftrightarrow$ não-poder $\rightarrow$ 0.4175
cannot $\Leftrightarrow$ não $\rightarrow$ 0.2145
cannot $\Leftrightarrow$ poder $\rightarrow$ 0.1627
this $\Leftrightarrow$ isto $\rightarrow$ 0.1608
this-cannot $\Leftrightarrow$ isto-não $\rightarrow$ 0.0834
this-cannot $\Leftrightarrow$ não-poder $\rightarrow$ 0.0442
go-on $\Leftrightarrow$ continuar $\rightarrow$ 0.0380
cannot-go $\Leftrightarrow$ poder-continuar $\rightarrow$ 0.0378
outras entradas $\rightarrow$ 0.0

Extração de CVPs

- Boa precisão do Greedy Aligner para CVPs
- Geração de léxico bilíngüe, com corpus de treinamento
- Busca por verbos seguidos de partículas no corpus de teste
- Alinhamento utilizando o Greedy Aligner
- Extração de correspondências interessantes

Exemplo de saída

take out $\Leftrightarrow$ suprimir - 1
carry over $\Leftrightarrow$ transitar - 3
pass on $\Leftrightarrow$ transmitir - 9
watch over $\Leftrightarrow$ vigiar - 1

Testes para avaliar o desempenho

- Etiquetamento na origem
  - Utiliza dados do etiquetador morfossintático
  - Extrai somente verbos seguidos por partícula
  - 223 entradas, 94% atestadas
- Etiquetamento na origem e destino
  - Extrai somente verbos seguidos por partícula, associados com uma expressão que contenha um verbo em português
  - Exclui entradas incorretas como *wash up* $\Leftrightarrow$ *cadáver*
  - Mas também corretas como *phase out* $\Leftrightarrow$ *eliminação gradual*

Substantivos compostos

- Verificar a aplicabilidade da técnica para outros tipos de EMPs
- Substantivos compostos são menos flexíveis sintanticamente e normalmente composicionais
- Mesmo assim, tradução palavra-a-palavra nem sempre é aceitável
- Foram encontrados 3 tipos principais de traduções para substantivos compostos do inglês ao português
  1. oil tanker $\Leftrightarrow$ petroleiro
  2. world market $\Leftrightarrow$ mercado mundial
  3. fishing agreement $\Leftrightarrow$ acordo de pesca
- Tipo 3 não pode ser modelado através de bigramas

Teste para substantivos compostos

- Análogo aos testes com CVPs
- Utiliza o mesmo léxico
- Busca por substantivos seguidos por substantivos em inglês no resto do corpus
- Alinhamento léxico utilizando o Greedy Aligner
- Busca por alinhamentos de substantivos compostos no corpus alinhado
- 100 primeiros SCs distintos avaliados
- Saída é um SC em inglês, seguido de sua tradução candidata

Exemplo de saída

farm budget $\Leftrightarrow$ orçamento agrícola - 1
monitoring centre $\Leftrightarrow$ observatório - 2
Middle Ages $\Leftrightarrow$ idade Média - 8

Avaliação dos alinhadores

- 200 frases contendo CVP selecionadas aleatoriamente
- Alinhamento da CVP com a tradução realizado manualmente
- Traduções como *back down* $\Leftrightarrow$ *recuo* corretas
- 161 destas frases continham uma tradução da CVP

	LIHLA	Greedy
Precisão	1.00	1.00
Recall	0.11 (18)	0.19 (30)

Tabela: Desempenho dos alinhadores com CVP

Precisão: Alinhamentos **corretos** de CVP encontrados / Alinhamentos de CVP encontrados
Recall: Alinhamentos (corretos) de CVP / Total de CVPs

Avaliação dos alinhadores

- Expressão muito complexa
  - Modelo de bigramas é insuficiente para o fenômeno lingüístico
  - tighten up $\Leftrightarrow$ tornar mais rigoroso
- Erro do lematizador
  - *Como já foi observado*, classifica *observado* como adjetivo
  - Lema fica *observado* em vez de *observar*
- Ruído nas partículas
  - Alinhamento (errado) da partícula com uma palavra na língua alvo com escore mais alto que o da CVP
  - Baixa freqüência de CVP e alta freqüência da partícula
- Construções onde V+P é traduzido como V (*eat up*)

Resultados para extração de CVPs

- Correspondências encontradas no corpus de teste classificadas manualmente

Threshold	1	2	3	5
Total	223	141	117	82
Correto	0.74 (164)	0.87 (122)	0.88 (103)	0.88 (72)
Quase	0.09 (20)	0.07 (10)	0.07 (8)	0.06 (5)
Incorreto	0.17 (39)	0.06 (9)	0.05 (6)	0.06 (5)

Tabela: CVP x qualquer token

Threshold	1	2	3	5
Total	173	121	100	70
Correto	0.88 (152)	0.93 (113)	0.94 (94)	0.94 (66)
Quase	0.08 (14)	0.05 (6)	0.05 (5)	0.043 (3)
Incorreto	0.04 (7)	0.02 (2)	0.01 (1)	0.014 (1)

Tabela: CVP x entrada contendo verbo

Substantivos compostos

- 100 primeiros tipos distintos
- Classificados manualmente

Total	100
Correto	0.57 (57)
Trigrama (tipo 3)	0.18 (18)
Incorreto	0.25 (25)

Tabela: Extração de substantivos compostos

- Uso de trigramas deve melhorar a performance

Conclusão

- Este trabalho teve como objetivo investigar a aquisição de conhecimento léxico a partir de corpora paralelo, expandindo métodos existentes para o uso de bigramas.
- Uso de bigramas no modelo da língua ajuda no processamento de EMPs.
- Identificação de traduções de CVPs muito boa ao utilizar dados de um etiquetador morfossintático.
- Identificação de traduções para substantivos compostos mais modesta.
- Método de baixo custo, utiliza ferramentas e recursos amplamente disponíveis.
- Alinhador léxico que utiliza conhecimentos de bigramas
- Pares de EMP e tradução candidatos a serem incluídos em algum dicionário
- Independente de linguagem
- Ajustes podem ser feitos.

Fim

Obrigado

Paulo Schreiner
paulo.schreiner@gmail.com