Lingüística Computacional

Parte 02: Alguns Problemas Frequentes

Marcelo Finger

Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

2º Semestre 2019

Preâmbulo

O que são "palavras"?

- É preciso separar dois conceitos diferentes: **ocorrências** e **tipos** de palavras

O que são "palavras"?

- É preciso separar dois conceitos diferentes: **ocorrências** e **tipos** de palavras
- Ocorrências de palavras (*word tokens*):
  E.g. { `"O"`, `"menino"`, `"viu"`, `"o"`, `"vizinho"` }

O que são "palavras"?

- É preciso separar dois conceitos diferentes: **ocorrências** e **tipos** de palavras
- Ocorrências de palavras (*word tokens*):
  E.g. { `"O"`, `"menino"`, `"viu"`, `"o"`, `"vizinho"` }
- Tipos de palavras (*word types*):
  E.g. { `o`, `menino`, `viu`, `vizinho` }

O que são "palavras"?

- É preciso separar dois conceitos diferentes: **ocorrências** e **tipos** de palavras
- Ocorrências de palavras (*word tokens*):
  E.g. { `"O"`, `"menino"`, `"viu"`, `"o"`, `"vizinho"` }
- Tipos de palavras (*word types*):
  E.g. { `o`, `menino`, `viu`, `vizinho` }
- *Ocorrências* são **instanciações** de *tipos* de palavras

O que são "problemas"?

Dependem de:

- Linguagem
- Aplicação
- Modelo matemático
- Algoritmo a ser implementado, etc

Os problemas podem ser compostos

Pré-processamento

Tokenização

- **Entrada**: "Mais vale um asno que me carregue que um cavalo que me derrube"
- **Saída**: ["Mais", "vale", "um", "asno", "que", "me", "carregue", "que", "um", "cavalo", "que", "me", "derrube"]

Filtragem (e.g. de etiquetas)

Ex: Formato XML de etiquetas morfossintáticas

- **Entrada**: "<ADV>Mais</ADV> vale um asno <REL>que</REL> me carregue <CONJS> que </CONJS> um cavalo <REL>que</REL> me derrube"
- **Saída**: "Mais vale um asno que me carregue que um cavalo que me derrube"

Filtragem (e.g. de etiquetas)

Ex: Formato XML de etiquetas morfossintáticas

- **Entrada**: "<ADV>Mais</ADV> vale um asno <REL>que</REL> me carregue <CONJS> que </CONJS> um cavalo <REL>que</REL> me derrube"
- **Saída**: "Mais vale um asno que me carregue que um cavalo que me derrube"
- **Saída**: ["Mais", "vale", "um", "asno", "que", "me", "carregue", "que", "um", "cavalo", "que", "me", "derrube"] + entradas em um BD *lembrando* das etiquetas

Identificação de Ponto Final

Nem todos os pontos (“.”) finalizam um período. Por exemplo

1. Esse logo depois do 1
2. Pontos após abreviações: Dr., Jr., Ling. Comp.
3. Em catalão existem palavras como: intel.ligencia
4. . Em textos . antigos os pontos aparecem em lugares inesperados

É um problema contextual e dependente de língua.
Influencia o processo de tokenização de textos

Expansão de Abreviações

- Ling. Comp. $\Longrightarrow$ Linguísitica Computacional
- IME $\Longrightarrow$ Instituto de Matemática e Estatísica
- bj em v té+ $\Longrightarrow$ Beijo em você. Até mais

Dependente de língua, de contexto, de mídia.

Separação em Sentenças

Separar em sentenças:

```
Existem coisas que deixam, louco um prof.  de Ling.
Comp.  Mas as três mais irritantes:  1.  frases sem
verbo; 2.  coisas incompletas.
```

Quantas sentenças há no texto acima?

Lematização

- **Entrada**: `Extrair totalmente os lemas (radicais) das palavras.`
- **Saída**: {`Extrai,total,o,lema,radic,de,a,palavr`}

Requer conhecimento linguístico, dicionário de lemas

Stemmer

- **Entrada**: `Extrair aproximadamente os lemas (radicais) das palavras.`
- **Saída**: {`Extra, aproximada, o, lema, radic, d, palavr`}

Não usa conhecimento linguístico

Stop-words (palavras muito comuns)

- **Entrada**: `E extrair as palavras que não são informação relevante`
- **Saída**: {`extrair, palavras, não, informação, relevante`}

Dependente de língua e de aplicação

Sintaxe

Separação de junções

- naquelas $\Longrightarrow$ em aquelas
- àquilo $\Longrightarrow$ a aquilo
- pela $\Longrightarrow$ por a (ambíguo)

Processamento de Expressões Multipalavras

Unidades lexicais complexas, com significado pré-teórico não decomponível em suas partes

- A lei entrou imediatamente em vigor $\Longrightarrow$ A lei entrou_em_vigor imediatamente
- Ele é uma mão na roda $\Longrightarrow$ Ele é uma mão_na_roda

A definição de o que é uma expressão multipalavra pode depender da aplicação

Etiquetagem Morfosintática

Associar a palavras em contexto uma etiqueta morfossintática (*Part-of-speech, PoS tagging*). Ex:
Entrada: A primeira coisa que ignoramos, é quando ha-de ser o dia do Juiso:
Saída: A/D-F primeira/ADJ-F coisa/N que/WPRO ignoramos/VB-P ,/, é/SR-P quando/CONJS ha-de/HV-P+P ser/SR o/D dia/N do/P+D Juiso/NPR :/.

Segmentação (*Chunking*)

Identificar as principais unidades sintáticas constituíntes

Entrada: A primeira coisa que ignoramos, é quando ha-de ser o dia do Juiso:

Saída: A primeira coisa que ignoramos, é quando ha-de ser o dia do Juiso:

Majoritariamente, sintagmas verbais e nominais

Parseamento Raso (Shallow Parsing)

Identificas os sintagnas vervais e nominais, apenas
Entrada: A primeira coisa que ignoramos, é quando ha-de ser o dia do Juiso:
Saída: <SN>A primeira coisa</SN> <SV>que ignoramos</SV>, é quando <SV>ha-de ser <SN>o dia do Juiso</SN><SV>:

Parseamento Profundo (Deep Parsing)

```
( IP-MAT (IP-MAT-3
        (NP-SBJ (D-F A)
            (ADJ-F primeira)
            (N coisa)
            (CP-REL (WNP-1 (WPRO que))
                (IP-SUB (NP-ACC *T*-1)
                    (NP-SBJ *pro*)
                    (VB-P ignoramos)))) (PUNC ,)
        (SR-P é)
        (CP-QUE (WADVP-2 (WADV quando))
            (IP-SUB (ADVP *T*-2)
                (HV-P ha-)
                (PP (P -de)
                    (IP-INF (SR ser)))
                (NP-SBJ (D o)
                    (N dia)
                    (PP (P d@)
                        (NP (D @o)
                            (NPR Juiso)))))))
    (PUNC :) )
```

Reconhecimento de Entidades Mencionada (NER)

Não menos relevante foi a influência da pressão social exercida pela
<EM CATEG="ABSTR|ACONTECIMENTO" TIPO="DISCIPLINA|EFEMERIDE">
Contra-Reforma </EM> ,
na qual os
<EM CATEG="PESSOA" TIPO="GRUPOMEMBRO"> Jesuítas </EM>
tiveram um papel de liderança

Detecção de Elipse

Elipse é a omissão de um ou mais termos numa sentença, identificável tanto por elementos gramaticais presentes na própria oração, quanto pelo contexto.

Entrada: As rosas florescem em maio, as margaridas em agosto.

Saída: As rosas florescem em maio, as margaridas [*florescem*] em agosto.

Resolução de Anáforas

Encontrar o referente de uma expressão (*deixis*) em geral pronome ou sintagma nominal.

- [João]$_i$ chegou completamente bêbado. [Ele]$_i$ foi muito desagradável. [O idiota]$_i$ não se manca.

Problema sintático, semântico ou pragmático?

Semântica

Nota

A divisão entre problemas de "semântica" e de "pragmática" é totalmente arbitrária

No fundo, é tudo pragmática.

Semântica Composicional

$$
\begin{aligned}
\textit{João} &\mapsto \textbf{joão} \\
\textit{ama} &\mapsto \lambda xy(\textbf{ama}(x)(y)) \\
\textit{Maria} &\mapsto \textbf{maria}
\end{aligned}
$$

$$
\begin{aligned}
\textit{João ama Maria} &\mapsto \lambda xy(\textbf{ama}(x)(y))(\textbf{maria})(\textbf{joão}) \quad \twoheadrightarrow_\beta \\
&\mapsto \textbf{ama}(\textbf{maria})(\textbf{joão})
\end{aligned}
$$

Embedding Clássico

Embedding ou **inserção num espaço n-dimensional**

f : **Word Types** $\rightarrow \mathbb{R}^n$

x_1	$\rightarrow$	P(palavra é substantivo)
x_2	$\rightarrow$	P(palavra é verbo)
$\vdots$		$\ldots$
x_{n-1}	$\rightarrow$	P(palavra é núcleo do sujeito)
x_n	$\rightarrow$	P(palavra é núcleo do objeto)

word2vec

Embedding sem semântica conhecida para as posições dos vetores

João	*ama*	*Maria*	$\langle pad \rangle$	$\langle pad \rangle$	$\langle pad \rangle$
27.1	38	26.8	0	0	0
- 6.8	-2	-7.2	0	0	0
27.5	-9	27.8	0	0	0
0.12	78	0.16	0	0	0

doc2vec

Embedding de documentos em um espaço m-dimensional

Em geral, se

$$doc = w_1, w_2, \ldots, w_d$$

então

$$emb(doc) = f(emb(w_1), \ldots, emb(w_d))$$

Pragmática

Desambiguação de palavras

Ex: Eu sento no **banco**, eu entro no **banco**.

Desambiguação de sentenças

Ex: Eu vi o menino com o telescópio

Parseamento Semântico (semantic parsing)

Traduzir sentenças para comandos de máquina

Entrada: Quais os estados cortados pelo Rio São Francisco
Saída:

Select *ESTADO.Nome*
From *ESTADO e, RIO r, GeoIntercept g*
Where *r.Nome= "São Francisco"*
And *g.Id1 = r.Id*
And *g.Id2 = e.Id*

Tradução

Entrada: Time flies like an arrow
Saída: O tempo voa como uma flecha

Tradução

Entrada: Time flies like an arrow
Saída: O tempo voa como uma flecha
Saída: Moscas do tempo gostam de uma seta

Problemas Éticos

- Geração automática de respostas odiosas a tweets de um determinado grupo
- Geração de textos falsos de um determinado autor, "no estilo dele"
- Identificação de autoria de texto, nas mãos da polícia politica
- etc