Semantic Role Labeling and Lexical Simplification: two samples of NLP applications

Leonardo Zilio
CENTAL - UCL
Semantic Role Labeling for Portuguese

Leonardo Zilio (Instituto de Letras – UFRGS)
Maria José Bocorny Finatto (Instituto de Letras – UFRGS)
Aline Villavicencio (Instituto de Informática – UFRGS)
Objectives

• To understand how the semantic structure of Portuguese works in specialized and non-specialized contexts

• To further describe the Portuguese language in terms of generic and descriptive semantic roles
  • In 2011, there was only one project on semantic role labeling (FrameNet Brasil)
Semantic Roles

Mary broke the vase.

AGENT + PATIENT

The vase broke.

PATIENT
Related Work

- FrameNet
  - Descriptive semantic roles associated to a specific communicative scenario (e.g., PLAYER, REFEREE [in a soccer match context])

- VerbNet
  - Generic descriptive semantic roles (e.g. AGENTE, THEME, PATIENT)

- PropBank
  - Numbered semantic roles (e.g., A1, A2) + roles for adjuncts (e.g., TIME, PLACE)
Our Choice

• VerbNet Semantic Roles + PropBank Adjunct Roles
46 Semantic Roles
Metodology

- Corpus
- Parser PALAVRAS
- SCF Extractor
- Database
- Annotation
# Corpora

<table>
<thead>
<tr>
<th></th>
<th>Diário Gaúcho Newspaper</th>
<th>Cardiology Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td>2008</td>
<td>2005-2007</td>
</tr>
<tr>
<td><strong># of Tokens</strong></td>
<td>1M</td>
<td>1,4M</td>
</tr>
</tbody>
</table>
Parsing

João viu o cachorro. (John saw the dog.)

João [João] <hum> PROP M S @SUBJ> #1->2
viu [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0
o [o] <artd> DET M S @N #3->4
cachorro [cachorro] <AzO> N M S @<ACC #4->2
$. #5->0
</s>
João viu o cachorro. (John saw the dog.)
Extraction of SCFs

- Subcategorization Frames are simpler syntactic representations of sentences.

- SCFs represent sentences in terms of their phrases:
  - NP_V_NP = Martin has a car.
  - NP_V_NP_PP = Martin bought a car from Paul.
  - NP_V_PP = Martin goes to the library.

- For us, SCFs help organizing sentences in the database.
Subcategorization Frames (SCF) Extractor

Adriano Zanette (Instituto de Informática – UFRGS)
Leonardo Zilio (Instituto de Letras – UFRGS)
Reader Module

- Receives the parsed text
- Separates every sentence
Extractor

- For each sentence, it:
  - Recognizes how many conjugated verbs exist
  - Duplicates the sentence for each conjugated verb
  - Extracts dependent phrases for each conjugated verb
  - Recognizes the syntactic category of each argument and attributes a relevance index (for organization purposes)
What is an argument?

• Complicated question
• For the purposes of the subcategorization frames extractor, there is no distinction between argument and adjunct
• It extracts phrases that are directly dependents of the verb, according to a set of rules
Builder Module

- Puts everything together (according to the relevance index)
- Builds the subcategorization frame for each verb and sentence
- Stores information on the database
Filter

• Not mandatory

• It can filter subcategorization frames based on frequency (or frequency-like parameters)
Let’s see how it works
Simple Example (Again)

João viu o cachorro. (John saw the dog.)

João [João] <hum> PROP M S @SUBJ> #1->2
viu [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0
o [o] <artd> DET M S @N #3->4
cachorro [cachorro] <Azo> N M S @<ACC #4->2
$. #5->0
</s>
Reader Module

• Simply recognizes that there is a full sentence; and
• Passes it to the Extractor Module
Extractor

• For each sentence, it:
  • Recognizes how many conjugated verbs exist
  • Duplicates the sentence for each conjugated verb
  • Extracts dependent phrases for each conjugated verb
  • Recognizes the syntactic category of each argument and attributes a relevance index (for organization purposes)
Extractor

• For each sentence, it:
  • Recognizes how many conjugated verbs exist
  • Duplicates the sentence for each conjugated verb
  • Extracts dependent phrases for each conjugated verb
  • Recognizes the syntactic category of each argument and attributes a relevance index (for organization purposes)
Extractor Module

João viu o cachorro. (John saw the dog.)

João [João] <hum> PROP M S @SUBJ> #1->2
viu [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0
o [o] <artd> DET M S @>N #3->4
cachorro [cachorro] <Azo> N M S @<ACC #4->2
$. #5->0
</s>
Extractor Module

João viu o cachorro. (John saw the dog.)

João  [João] <hum> PROP M S @SUBJ> #1->2
viu   [ver] <vH> <fmc> <mv> V PS 3S IND VFIN @FS-STA #2->0
o     [o] <artd> DET M S @>N #3->4
cachorro  [cachorro] <Azo> N M S @<ACC #4->2
$.     #5->0
</s>
Extractor

• For each sentence, it:
  • Recognizes how many conjugated verbs exist
  • **Duplicates the sentence for each conjugated verb**
  • Extracts dependent phrases for each conjugated verb
  • Recognizes the syntactic category of each argument and attributes a relevance index (for organization purposes)
Extractor

• For each sentence, it:
  • Recognizes how many conjugated verbs exist
  • Duplicates the sentence for each conjugated verb
  • Extracts dependent phrases for each conjugated verb
  • Recognizes the syntactic category of each argument and attributes a relevance index (for organization purposes)
Simple Example (Again)

Root (0)

viiu (2) [saw]

João (1) [John]  cachorro (4) [dog]

o (3) [the]
Simple Example (Again)

Root (0)

viu (2) [saw]

Verb

João (1) [John]  cachorro (4) [dog]

o (3) [the]
Simple Example (Again)

Root (0)
viu (2) [saw]
João (1) [John]
cachorro (4) [dog]
o (3) [the]

Arguments
Extractor Module

• For each sentence, it:
  • Recognizes how many conjugated verbs exist
  • Duplicates the sentence for each conjugated verb
  • Extracts dependent phrases for each conjugated verb
  • Recognizes the syntactic category of each argument and attributes a relevance index (for organization purposes)
Rules

• Format:
  • If [tag], then [argument_type]

• If SUBJ, then Subject (Relevance Index: 1)
• If ACC, then Direct Object (Relevance Index: 3)
• If ACC-PASS, then Reflexive Object (Relevance Index: 3)
• Etc.
Builder Module

- Verb: ver
- Sentence: João viu o cachorro.
- SCF: SUBJ_V_NP
- SUBJ: João
- NP: o cachorro
After processing all data, if the number of similar SCFs do not amount to a certain threshold, the SCF is excluded from the database.
Annotation Interface

Exemplos do frame 'SUBJ[NP]_V_NP' do verbo 'encontrar'

Exemplo 1
Encontrei um túmulo destruído, que não tinha dono, com os dois vasos

Mostrar anotação

ARG_1 OCULTO SUJEITO
ARG_2 um túmulo destruído que não tinha dono com os dois vasos OBJETO DIRETO

Exemplo 2

Selector

Theme
Co-Theme
Agent
Co-Agent
Stimulus
Instrument
Patient
Co-Patient
Experimenter
Target
Recipient
Beneficiary
Initial_Time
Moment
Final_Time
Frequency
Duration
Source
Initial_Location
VerbLexPor

<table>
<thead>
<tr>
<th>Diário Gaúcho</th>
<th>Cardiologia</th>
</tr>
</thead>
<tbody>
<tr>
<td>191 verbs</td>
<td>77 verbs</td>
</tr>
<tr>
<td>5,301 instances</td>
<td>1,931 instances</td>
</tr>
<tr>
<td>11,089 arguments</td>
<td>4,192 arguments</td>
</tr>
</tbody>
</table>
Availability

• XML and SQL
• Website Project CAMELEON
  • [http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic%2orole%2olabel%2ocorpus%20-%20Brazilian%20Portuguese](http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic%2orole%2olabel%2ocorpus%20-%20Brazilian%20Portuguese)
Text Simplification
Semantic Relations
Rodrigo Wilkens
Leonardo Zilio
Eduardo Ferreira
Aline Villavicencio
Objective

- To build a lexical resource with synonyms, antonyms and hypernyms
  - Distributional thesaurus + BabelNet

- Evaluate the resource against a gold standard
Distributional Thesaurus

- Distributional hypothesis:
  - You can know a word by the company it keeps

- Words can be represented as vectors in a multidimensional space
Distributional Thesaurus

- Presents pairs of words, indicating how related they are to each other

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Word 2</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joy</td>
<td>Happiness</td>
<td>48,5%</td>
</tr>
<tr>
<td>Joy</td>
<td>Smile</td>
<td>32,8%</td>
</tr>
<tr>
<td>Joy</td>
<td>Scream</td>
<td>15,0%</td>
</tr>
<tr>
<td>Joy</td>
<td>Brick</td>
<td>3,9%</td>
</tr>
</tbody>
</table>
Metodology

Corpora → Word2Vec → 2 Thesauri

Automatic Evaluation ← Gold Standard
Gold Standard

- AC/DC corpus = Word frequency list
- BabelNet = Resource similar to WordNet
  - Word polysemy
  - Semantic relations
• Which of these words is most related to "author"?
  • Poet
  • Parts
  • Patron
  • Board
Methodology

• Groups of words containing: 1 target word, 1 related word, and 3 non-related words

• TOEFL e WordNet-Based Synonymy Test (WBST)

| Target Word | Alternative 1: Related Word | Alternative 2 | Alternative 3 | Alternative 4 |
Methodology

• Each word from AC/DC corpus was annotated:
  • with frequency (from AC/DC); and
  • with polysemy (from BabelNet)

• Words that were not in BabelNet were excluded
Methodology

• Target word: medium frequency in AC/DC
• Related word: closest to the target word in terms of frequency and polysemy
• Non-related words: farthest average distance from the target word
Initial Resource

<table>
<thead>
<tr>
<th></th>
<th>Synonym</th>
<th>Antonym</th>
<th>Hipernym</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbs</td>
<td>500</td>
<td>200</td>
<td>500</td>
<td>1200</td>
</tr>
<tr>
<td>Nouns</td>
<td>1667</td>
<td>200</td>
<td>1667</td>
<td>3534</td>
</tr>
<tr>
<td>Total</td>
<td>2167</td>
<td>400</td>
<td>2167</td>
<td>4734</td>
</tr>
</tbody>
</table>
Automatic Validation

• All instances of target and related word were validated against Onto.PT
Manual Validation

- All instances that were NOT automatically validated were manually verified
BabelNet-Based Semantic Gold standard (B²SG)

<table>
<thead>
<tr>
<th></th>
<th>Antonym</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>V</td>
<td>N</td>
<td>V</td>
<td>N</td>
<td>V</td>
</tr>
<tr>
<td>Initial</td>
<td>200</td>
<td>200</td>
<td>1667</td>
<td>500</td>
<td>1667</td>
<td>500</td>
</tr>
<tr>
<td>Onto.PT</td>
<td>40</td>
<td>51</td>
<td>676</td>
<td>244</td>
<td>191</td>
<td>0</td>
</tr>
<tr>
<td>Human Judges</td>
<td>105</td>
<td>116</td>
<td>495</td>
<td>191</td>
<td>568</td>
<td>198</td>
</tr>
<tr>
<td>Total Validated</td>
<td>145</td>
<td>167</td>
<td>1171</td>
<td>435</td>
<td>759</td>
<td>198</td>
</tr>
<tr>
<td>% Correct</td>
<td>72.5</td>
<td>83.5</td>
<td>70.2</td>
<td>87.0</td>
<td>45.5</td>
<td>39.6</td>
</tr>
</tbody>
</table>
BabelNet-Based Semantic Gold standard (B²SG)

<table>
<thead>
<tr>
<th></th>
<th>Antonym</th>
<th>Synonym</th>
<th>Hypernym</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>V</td>
<td>N</td>
</tr>
<tr>
<td>Initial</td>
<td>200</td>
<td>200</td>
<td>1667</td>
</tr>
<tr>
<td>Onto.PT</td>
<td>40</td>
<td>51</td>
<td>676</td>
</tr>
<tr>
<td>Human Judges</td>
<td>105</td>
<td>116</td>
<td>495</td>
</tr>
<tr>
<td>Total Validated</td>
<td>145</td>
<td>167</td>
<td>1171</td>
</tr>
<tr>
<td>% Correct</td>
<td>72.5</td>
<td>83.5</td>
<td>70.2</td>
</tr>
</tbody>
</table>
Corpora

<table>
<thead>
<tr>
<th></th>
<th>TOKENS</th>
<th>TYPES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>1.5G</td>
<td>3.7M</td>
</tr>
<tr>
<td>Lemma</td>
<td>409M</td>
<td>1.5M</td>
</tr>
</tbody>
</table>

Corpus Brasileiro was not used in the lemmatized corpus, because it is not annotated with lemmata.
Distributional Thesauri

- Word2Vec
- Strict Evaluation (target word and all alternatives must be in the corpus)
Evaluation

• Which of these words is most related to "author"?

<table>
<thead>
<tr>
<th>Target</th>
<th>Alternative</th>
<th>Relatedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>Poet</td>
<td>24.8%</td>
</tr>
<tr>
<td>Author</td>
<td>Parts</td>
<td>0.3%</td>
</tr>
<tr>
<td>Author</td>
<td>Patron</td>
<td>0.6%</td>
</tr>
<tr>
<td>Author</td>
<td>Board</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

- Poet
- Parts
- Patron
- Board
### Strict Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Surface</th>
<th></th>
<th></th>
<th>Lemma</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instances</td>
<td>Correct</td>
<td>% Correct</td>
<td>Instances</td>
<td>Correct</td>
<td>% Correct</td>
</tr>
<tr>
<td><strong>Antonym</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Antonym</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>105</td>
<td>90</td>
<td>85.7</td>
<td>98</td>
<td>82</td>
<td>83.7</td>
</tr>
<tr>
<td>V</td>
<td>143</td>
<td>100</td>
<td>69.9</td>
<td>141</td>
<td>110</td>
<td>78.0</td>
</tr>
<tr>
<td><strong>Hipernym</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Hipernym</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>545</td>
<td>432</td>
<td>79.3</td>
<td>525</td>
<td>425</td>
<td>81.0</td>
</tr>
<tr>
<td>V</td>
<td>167</td>
<td>115</td>
<td>68.9</td>
<td>166</td>
<td>118</td>
<td>71.1</td>
</tr>
<tr>
<td><strong>Synonym</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Synonym</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>861</td>
<td>726</td>
<td>84.3</td>
<td>832</td>
<td>721</td>
<td>86.7</td>
</tr>
<tr>
<td>V</td>
<td>366</td>
<td>275</td>
<td>75.1</td>
<td>366</td>
<td>267</td>
<td>73.9</td>
</tr>
</tbody>
</table>
Next Step

• Mixing Distributional Thesaurus with BabelNet for creating a larger dictionary of synonyms
Dictionary of Complex Words

Leonardo Zilio
Susana Bautista
Objective

• Dictionary of complex words from Classic Literature
  • Simpler and complexer alternatives
# Corpus

<table>
<thead>
<tr>
<th>Author – Book</th>
<th>Tokens</th>
<th>Types</th>
<th>Type/Token Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluísio Azevedo – O Cortiço</td>
<td>81.8K</td>
<td>11.2K</td>
<td>0.137</td>
</tr>
<tr>
<td>Joaquim Manuel de Macedo – A Moreninha</td>
<td>47.2K</td>
<td>6.9K</td>
<td>0.147</td>
</tr>
<tr>
<td>José de Alencar – Lucíola</td>
<td>46K</td>
<td>7.8K</td>
<td>0.169</td>
</tr>
<tr>
<td>Machado de Assis – Memorial de Aires</td>
<td>51.1K</td>
<td>6.3K</td>
<td>0.123</td>
</tr>
<tr>
<td><em>(Ce que les Hommes Appellent Amour)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methodology

Corpus File → Sentence Splitter → POS Tagging

Filter (Frequency in AC/DC Corpus) → Interface → Onto.PT (Dictionary) → Output File

Word Selection

Complexity Annotation
Como quiserem, continuou Filipe, pondo-se em hábitos menores; mas, por minha vida, que a carraspana de hoje ainda me concede apreciar devidamente aqui o meu amigo Fabrício, que talvez acaba de chegar de alguma visita diplomática, vestido com esmero e alinho, porém, tendo a cabeça encapuzada com a vermelha e velha carapuça do Leopoldo; este, ali escondido dentro do seu robe-de-chambre cor de burro quando foge, e sentado em uma cadeira tão desconjuntada que, para não cair com ela, põe em ação todas as leis de equilíbrio, que estudou em Pouillet; acolá, enfim, o meu romântico Augusto, em ceroulas, com as fraldas à mostra, estirado em um canapé em tão bom uso, que ainda agora mesmo fez com que Leopoldo se lembrasse de Bocage.
-Como quiserem, continuou Filipe, pondo-se em hábitos menores; mas, por minha vida, que a carraspana de hoje ainda me concede apreciar devidamente aqui o meu amigo Fabrício, que talvez acaba de chegar de alguma visita diplomática, vestido com esmero e alinho, porém, tendo a cabeça encapuzada com a vermelha e velha carapuça do Leopoldo; este, ali escondido dentro do seu robe-de-chambre cor de burro quando foge, e sentado em uma cadeira tão desconjuntada que, para não cair com ela, põe em ação todas as leis de equilíbrio, que estudou em Pouillet; acolá, enfim, o meu romântico Augusto, em ceroulas, com as fraldas à mostra, estirado em um canapé em tão bom uso, que ainda agora mesmo fez com que Leopoldo se lembresse de Bocage.
Annotation

• **carraspana:** 2 - carraspana, 1 - bebedeira, 1 - porre
• **esmero:** 3 - aprumo, 3 - asseio, 2 - alinho, 2 - esmero, 1 - elegância, 1 - perfeição, 1 - primor
• **alinho:** 3 - apuro, 3 - asseio, 2 - alinho, 2 - esmero, 1 - decência, 1 - dignidade
• **fraldas:** 2 - fralda, 1 - aba
• **canapé:** 2 - canapé, 1 - sofá
Instance of Annotation Output

• 16, 16, Bocage, quando tomava carraspana, descompunha os médicos. = [carraspana, 779, 789, o, arregaço, 2, carraspana, o, carão, o, chegadela, o, esbregue, o, esfrega, o, pito, o, ralhação, o, ralho, o, repreensão, o, reprimenda, o, tosa, o, tunda, , carraspana, 1, bebedeira, 1, porre, o, , o, ]}
Instance of Annotation Output

Sentence number in the corpus file

• 16,\{Bocage, quando tomava carraspana, descompunha os médicos.=[carraspana, 779, 789, o, arregaço, 2, carraspana, o, carão, o, chegadela, o, esbregue, o, esfrega, o, pito, o, ralhação, o, ralho, o, repreensão, o, reprimenda, o, tosa, o, tunda, , carraspana, 1, bebedeira, 1, porre, o, , o, ]\}
Bocage, quando tomava carraspana, descompunha os médicos.
Instance of Annotation Output

Selected word and complexity annotation

• 16, {Bocage, quando tomava carraspana, descompunha os médicos. = [carraspana, 779, 789, o, arregaço, 2, carraspana, o, carão, o, chegadela, o, esbregue, o, esfrega, o, pito, o, ralhação, o, rolho, o, repreensão, o, reprimenda, o, tosa, o, tunda, , carraspana, 1, bebedeira, 1, porre, o, , o, ]}
Instance of Annotation Output

Word position in the corpus file

• 16, {Bocage, quando tomava carraspana, descompunha os médicos.=[carraspana, 779, 789, o, arregaço, 2, carraspana, o, carão, o, chegadela, o, esbregue, o, esfrega, o, pito, o, ralhação, o, ralho, o, repreensão, o, reprimenda, o, tosa, o, tunda, , carraspana, 1, bebedeira, 1, porre, o, , o, ]}
Results

• Dictionary of Complex Words:
  • 3720 annotations: 790 different word senses

• Simplification gold standard for Literary Texts
Merci!

Leonardo Zilio
leonardo.zilio@uclouvain.be