Using Context to Improve the Spanish WordNet Translation

Alfonso Methol
Guillermo López
Juan Miguel Álvarez
Luis Chiruzzo
Dina Wonsever

Facultad de Ingeniería
Universidad de la República
Montevideo
Uruguay
Agenda

- Introduction
- Translation process
  - Simple selectors
  - Selectors using contextual information
- Results
- Conclusions
Introduction

MCR contains WordNet translations for Spanish, Catalan, Basque, Galician and Portuguese

Same structure as Princeton WordNet, similar synset ids, but the lemmas are translated

It is widely used tool, but it is not complete

We aim to improve the coverage of Spanish MCR by automatically translating synsets
Translation process

Create a list of translation candidates for a Spanish synset based on the lemmas in the English synset.

Use a heuristic process that selects which of the candidates is suitable for the synset.

- These processes are called **selectors**.
- Four simple selectors and two selectors based on contextual information.
Translation candidates

Bilingual dictionaries
- Apertium (42,996 lemmas)
- Wiktionary (47,982 lemmas)
- Eurovoc (2,032 lemmas)

Machine translation systems
- Google Translate
- Microsoft Translate
- Yandex
Monosemy

Select all translation candidates for English lemmas that appear in only one synset

It assumes that these lemmas are unambiguous

Example:

- Lemma: “advisable”
- Only in synset: eng-30-00067038-a “worthy of being recommended or suggested; prudent or wise”
- Translations: “aconsejable”, “recomendable”, “conveniente”
- Associate those lemmas to spa-30-00067038-a
Single Translation

Select the translation candidates of all English lemmas that have only one possible Spanish translation.

It assumes that if there is only one translation, this translation must be valid for all synsets having that lemma.

Example:
- Lemma: “agile” (adjective)
- Synsets: eng-30-00032733-a “moving quickly and lightly” / eng-30-01334833-a “mentally quick”
- Unique translation: “ágil”
- Associate “ágil” to spa-30-00032733 and spa-30-01334833-a
Factorization

Select translation candidates that are translation of all the lemmas in a synset

This selector is executed only for synsets that have more than one lemma

Example:
- Synset: eng-30-00011516-r (adverb) “in a poor or improper or unsatisfactory manner; not well”
- Lemmas: “poorly”, “ill”, “badly”
- One of the translations is valid for all lemmas: “mal”
Derived Adverb

Select Spanish adverbs generated from the translation of English adjectives

It uses the “is_derived_from” relation present in adverbs

It applies morphotactic rules to generate adverbs from Spanish adjectives

- Ends in “r” or “n” → append “amente” (e.g. “alentador” → “alentadoramente”)
- Ends in “o” → substitute by “amente” (e.g. “rápido” → “rápidamente”)
- Else → append “mente” (e.g. “vil” → “vilmente”)
But the morphotactic rules not always generate valid adverbs...

\[
\text{azul (blue)} \rightarrow *\text{azulmente (bluely)}
\]

The selector filters adverbs using a corpus of news text.
Derived Adverb

Spanish Adverb

English Adverb

English Adjective

ILI

associate

is_derived_from

translate

filtering

morphotactics

“plenamente”, “repletamente”, “completamente”, “totalmente”

“pleno”, “repleto”, “lleno”, “completo”, “total”

“plenamente”, “repletamente”, “llenamente”, “completamente”, “totalmente”

“total”, “full”

“totalmente”
Using contextual information

Synset examples contain lemmas used in context

We translated all examples for English lemmas using Google Translate

Only 28% of the synsets have associated examples, upper bound for the performance of these methods
Filtering

Find pairs of <English lemma, Spanish Lemma> where the English lemma appears in an example and the Spanish lemma appears in the translation.

If there is no exact match, use FreeLing to obtain lemmas and POS for both the example and the translation.
Filtering

Example:
- Example: "his last words" associated to synset eng-30-00004296-a
- Only lemma in the synset: "last"
- Translation candidate: "último"
- Translated example: "sus últimas palabras"

There is no exact match, so we use the tagger and lemmatizer:
- Analyzed examples: "[(his,PRP$) (last,JJ) (word,NNS)]" / "[(su,D) (último,A) (palabra,N)]"

There is a match: last/JJ and último/A
Structure

Perform dependency parsing on both the English example and its translation

Obtain the path from the root to the target lemma in the English parse tree

Follow the same path from the root to a node in the Spanish parse tree

Select the lemma in this node if it has the correct POS and is a valid translation candidate
Structure

Synset: eng-30-13792183-n
Lemma: “bond”
Structure

Synset: eng-30-00593852-v
Lemma: “read”
<table>
<thead>
<tr>
<th>Selector</th>
<th>Generated</th>
<th>MCR</th>
<th>Intersection</th>
<th>Overlap</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosemy</td>
<td>183,386</td>
<td>146,501</td>
<td>47,632</td>
<td>32.51%</td>
<td>74.03%</td>
</tr>
<tr>
<td>Single Transl.</td>
<td>81,058</td>
<td>146,501</td>
<td>38,505</td>
<td>26.28%</td>
<td>52.50%</td>
</tr>
<tr>
<td>Factorization</td>
<td>111,919</td>
<td>146,501</td>
<td>34,400</td>
<td>23.48%</td>
<td>69.26%</td>
</tr>
<tr>
<td>Derived Adv.</td>
<td>5,161</td>
<td>3,583</td>
<td>1,907</td>
<td>53.22%</td>
<td>63.05%</td>
</tr>
<tr>
<td>All Simple</td>
<td>256,852</td>
<td>146,501</td>
<td>72,674</td>
<td>50.39%</td>
<td>71.71%</td>
</tr>
<tr>
<td>Filtering</td>
<td>22,401</td>
<td>146,501</td>
<td>12,680</td>
<td>8.66%</td>
<td>43.40%</td>
</tr>
<tr>
<td>Structure</td>
<td>12,168</td>
<td>146,501</td>
<td>6,857</td>
<td>4.68%</td>
<td>43.65%</td>
</tr>
<tr>
<td>All Context</td>
<td>25,223</td>
<td>146,501</td>
<td>13,291</td>
<td>9.07%</td>
<td>47.31%</td>
</tr>
<tr>
<td>All</td>
<td>264,105</td>
<td>146,501</td>
<td>75,416</td>
<td>51.48%</td>
<td>71.44%</td>
</tr>
<tr>
<td>Selector</td>
<td>Precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monosemy</td>
<td>65.70%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single Transl.</td>
<td>73.65%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factorization</td>
<td>64.42%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Derived Adv.</td>
<td>73.20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Simple</td>
<td>69.05%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filtering</td>
<td>83.96%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>81.30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Context</td>
<td>82.67%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POS</th>
<th>Simple Sel.</th>
<th>Contextual Sel.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>74.89%</td>
<td>87.34%</td>
</tr>
<tr>
<td>Adverb</td>
<td>73.65%</td>
<td>88.42%</td>
</tr>
<tr>
<td>Noun</td>
<td>57.51%</td>
<td>80.24%</td>
</tr>
<tr>
<td>Verb</td>
<td>52.47%</td>
<td>74.12%</td>
</tr>
</tbody>
</table>
Results: Coverage

Coverage over a corpus of 850M Spanish words from news text

<table>
<thead>
<tr>
<th>POS</th>
<th>Lemmas in corpus</th>
<th>Lemmas in MCR</th>
<th>MCR + new lemmas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>42,604</td>
<td>5,592 (13.12%)</td>
<td>18,063 (42.40%)</td>
</tr>
<tr>
<td>Adverb</td>
<td>10,676</td>
<td>523 (4.90%)</td>
<td>7,105 (66.55%)</td>
</tr>
<tr>
<td>Noun</td>
<td>104,811</td>
<td>11,523 (10.99%)</td>
<td>35,525 (33.90%)</td>
</tr>
<tr>
<td>Verb</td>
<td>37,522</td>
<td>8,821 (23.51%)</td>
<td>22,427 (59.77%)</td>
</tr>
<tr>
<td>All</td>
<td>195,613</td>
<td>26,459 (13.52%)</td>
<td>83,130 (42.50%)</td>
</tr>
</tbody>
</table>
Conclusions

The four simple selectors generated 182,051 nouns, 19,682 verbs, 17,384 adjectives and 8,436 adverbs

But the precision was lower: 69.05%

The two selectors based on contextual information generated 5,339 nouns, 4,441 verbs, 6,444 adjectives and 1,747 adverbs with higher precision: 82.67%
Future work

Expand the translation sources

Improve the context based selectors considering other parsers and other label combinations

Apply the context based selectors to bigger corpora (e.g. SemCor)

Create selectors based on distributed semantics (e.g. word2vec) and WordNet lexical relations

Execute the process for other languages
Thank you!