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”**
- Translations: **“poorly”→“mal”,“pobremente”** / **“ill”→“mal”,“enfermo”** / **“badly”→“mal”,“malamente”**
- 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”**)

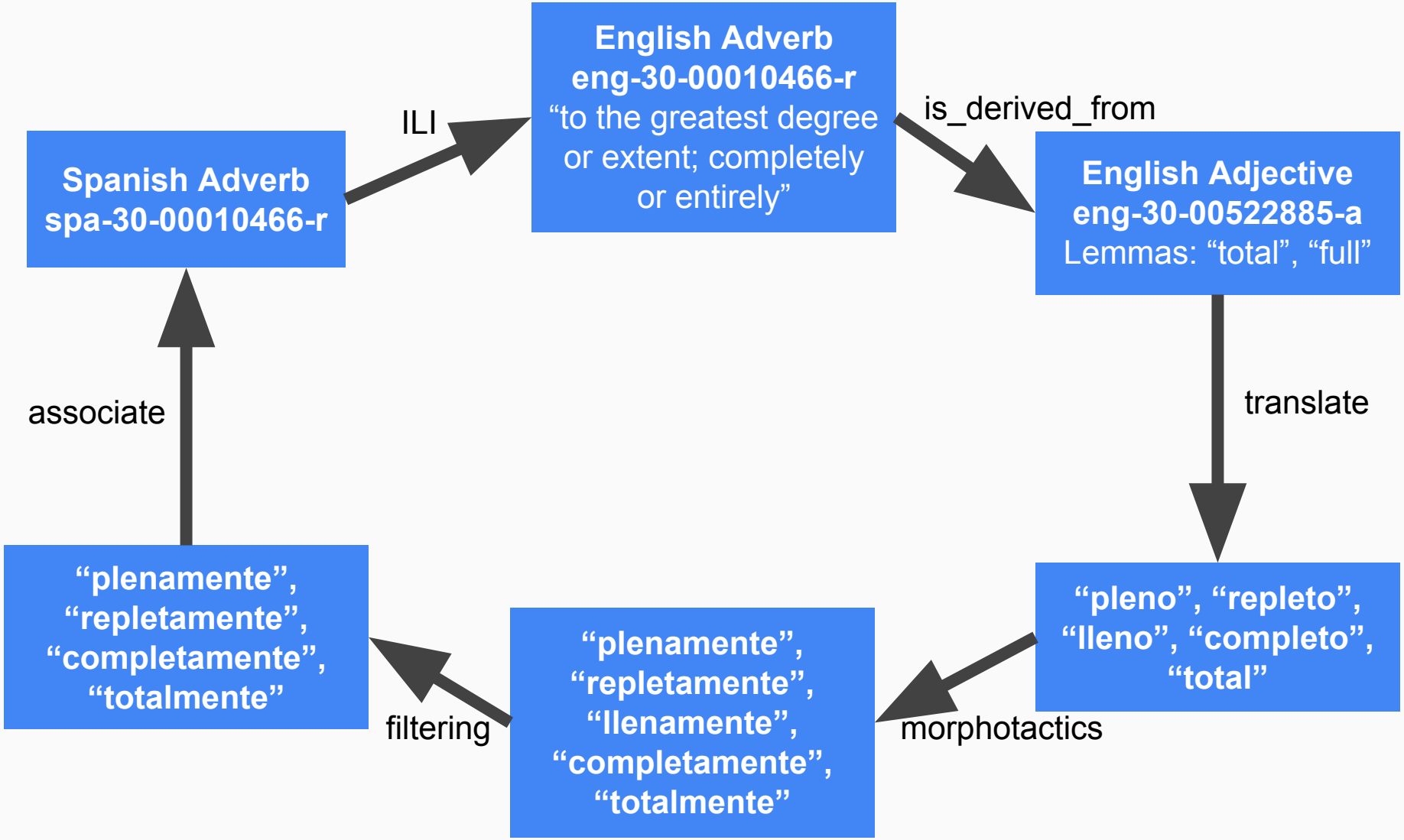
Derived Adverb

But the morphotactic rules not always generate valid adverbs…

azul (*blue*) → ***azulmente** (?*bluely*)

The selector filters adverbs using a corpus of news text.

Derived Adverb



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

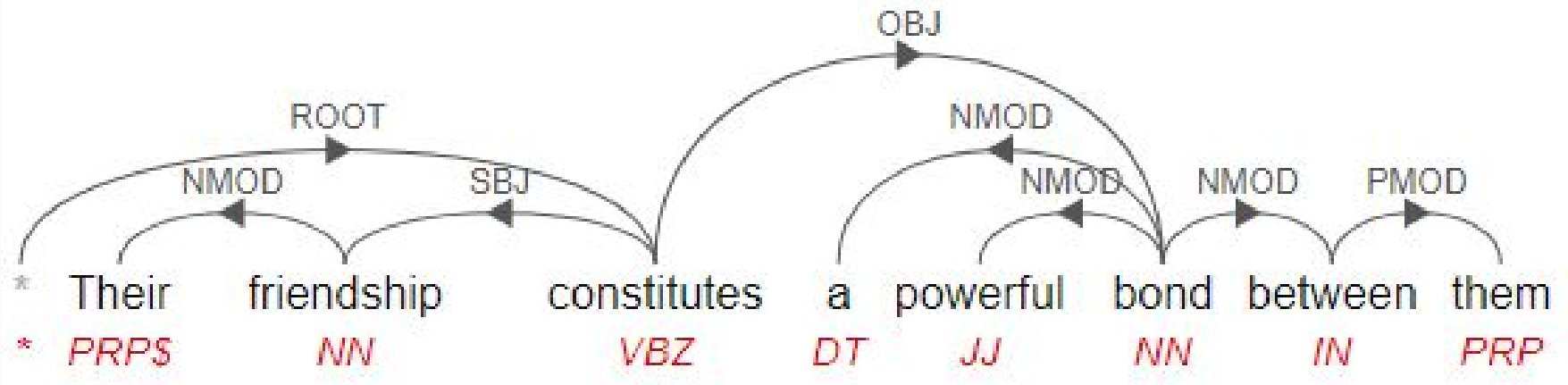
Follow the same path from the root to a node in the Spanish parse tree

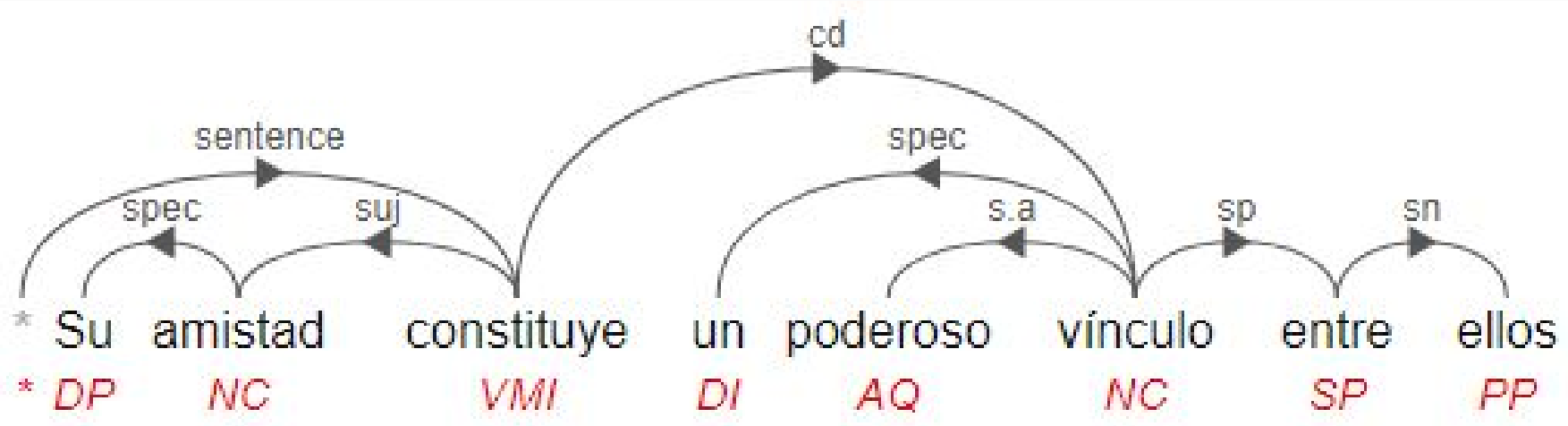
Select the lemma in this node if it has the correct POS and is a valid translation candidate

Structure

Synset: **eng-30-13792183-n**

Lemma: **“bond”**

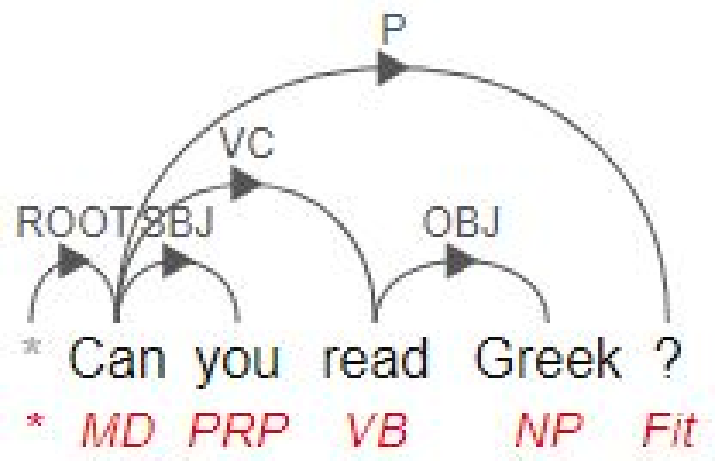


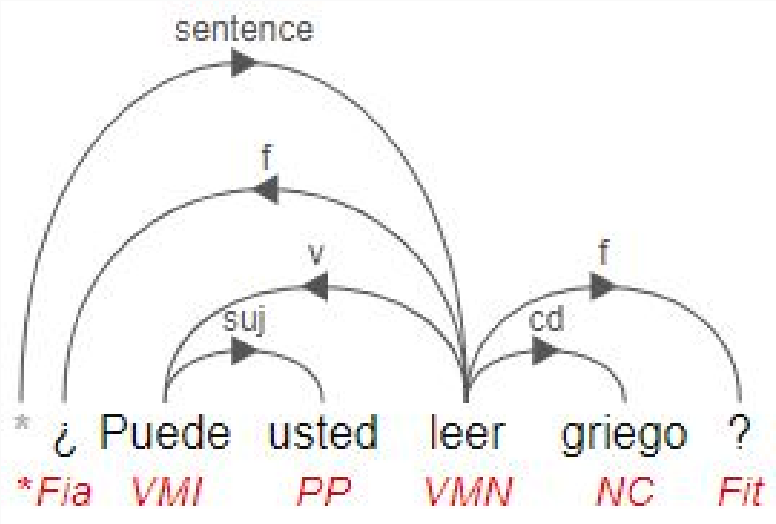


Structure

Synset: **eng-30-00593852-v**

Lemma: **“read”**





Results: Overlap

Selector	Generated	MCR	Intersection	Overlap	New
Monosemy	183,386	146,501	47,632	32.51%	74.03%
Single Transl.	81,058	146,501	38,505	26.28%	52.50%
Factorization	111,919	146,501	34,400	23.48%	69.26%
Derived Adv.	5,161	3,583	1,907	53.22%	63.05%
All Simple	256,852	146,501	72.674	50.39%	71.71%
Filtering	22,401	146,501	12,680	8.66%	43.40%
Structure	12,168	146,501	6,857	4.68%	43.65%
All Context	25,223	146,501	13,291	9.07%	47.31%
All	264,105	146,501	75,416	51.48%	71.44%

Results: Precision

Selector	Precision
Monosemy	65.70%
Single Transl.	73.65%
Factorization	64.42%
Derived Adv.	73.20%
All Simple	69.05%
Filtering	83.96%
Structure	81.30%
All Context	82.67%

POS	Simple Sel.	Contextual Sel.
Adjective	74.89%	87.34%
Adverb	73.65%	88.42%
Noun	57.51%	80.24%
Verb	52.47%	74.12%

Results: Coverage

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

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

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!