NANYANG TECHNOLOGICAL UNIVERSITY

IN SUPREMÆ DIGNITATIS 1343

Workshop for the WordNet Bahasa

# Cross-lingual mapping: projection versus intersection

Giulia Bonansinga
Division of Linguistics and Multilingual Studies

26-10-2014

# Outline

- Introduction
  - Motivation for Cross-lingual Word Sense Disambiguation
- Experiments on MultiSemCor (Bentivogli and Pianta, 2005)
  - Conversion to WordNet 3.0
  - Sense projection
  - Intersection
- Preliminary results for English and Italian

# Cross-lingual Word Sense Disambiguation

- Word Sense Disambiguation (WSD) aims to automatically select the correct sense of a word in its context
- Cross Language WSD makes use of parallel corpora and exploits differences in language to use one language to disambiguate another
  - Still unsolved problems

# Motivation for Cross-lingual WSD

- Many approaches for WSD require large amounts of high-quality sense-annotated data

- But manual annotation is costly and *very* time-consuming...

- Some facts:

  – Many languages still lack lexical resources and annotated corpora

  – Abundance of resources for English

# Intuition behind sense projection

- Existing parallel corpora and existing English annotated resources can be exploited to bootstrap the creation of annotated corpora in new languages
    - Human effort is reduced
    - New multilingual resources become available!
- Solution to the Knowledge Acquisition bottleneck via **projection** of annotations available in other languages

# Sense projection: how-to

- Given a text and its translation into another language, we assume that the translation preserves the meaning
- Hypothesis:
    - If a source text has been semantically annotated and aligned to its translation, then it is possible to **transfer** the annotation from the source text to its translation using **word alignment** as a bridge
- Aligned parallel corpora can be exploited to create annotated resources

# MultiSemCor in a nutshell

- 116 English texts from the SemCor corpus aligned at the word level with their corresponding Italian translations
- Uses the original release of SemCor, annotated with reference to WordNet 1.6 version
  - Precision 87.9%
  - Coverage 76.4%
- Freely distributed for research purposes and available online

| | English | Italian |
|---|---|---|
| Tokens | 258,499 | 268,905 |
| Semantically annotated tokens | 119,802 | 92,420 |

# Experiments on the MultiSemCor texts

- 4 texts from the MultiSemCor corpus

| | | English | Italian |
|---|---|---|---|
| Tokens | | 8,877 | 9,224 (*) |
| Words to be annotated in Bentivogli and Pianta | | 4,101 | 4,313 (*) |
| Annotations after conversion | nouns | 1,719 | 1,465 |
| | verbs | 1,035 | 817 |
| | adjectives | 765 | 558 |
| | adverbs | 575 | 384 |
| | others | - | 20 |
| | Total | 4,055 | 3,244 |

# Sense Inventory

- MultiSemCor is annotated with reference to MultiWordNet, a multilingual database linked to the English Princeton WordNet 1.6
- We convert the annotations to WordNet 3.0 and use Open Multilingual WordNet (OMW)
  - Access to WordNet 3.0 and OMW with NLTK
  - Larger coverage:

| | English | Italian |
|---|---|---|
| Synsets | 117,659 | 34,728 |
| Senses | 206,978 | 61,558 |
| Synsets in EOMW | 117,659 | 38,938 |
| Senses in EOMW | 213,480 | 69,824 |

# Conversion

- Can be easily applied to the whole MultiSemCor
- Easy for English, as senses are encoded with sense keys
- A bit more challenging for Italian, that uses offset encoding
  - Need for mappings WordNet 1.6 → WordNet 3.0
- Problems encountered:
  - dropped lemmas

    that_is, regard_to, out_of_focus, consist_of, as_a_whole, ...
  - dropped synsets

    ```
    nltk.corpus.reader.wordnet.WordNetError: No synset found for key
       'kind%5:00:00:benign:00'
    ```
  - multiple annotations
- Most frequent Sense (MFS) as back-off strategy

# Cross-lingual Sense Projection

Goal: creation of high quality semantically annotated corpora by using parallel text

- exploits existing (mostly English) annotated resources
- creates corpora in new (resource-poor) languages
- reduces human effort

Requirements:

- an alignment at the word level
- a shared sense inventory
- one side of the parallel corpus must be annotated

# Intuition behind Intersection

- A polysemous word in a language is likely to be translated in different words in another languages
- Example:

(EN) *Try talking to some of the **fellows** he works with, friends, anyone.*

(IT) *Cerca di parlare con alcuni dei **compagni** con I quali lavora, con degli amici, con qualcuno.*

**fellow**

```
Synset('chap.n.01')
Synset('colleague.n.02')
Synset('fellow.n.05')
Synset('boyfriend.n.01')
Synset('companion.n.01')
Synset('mate.n.06')
Synset('fellow.n.06')
```

**compagno**

```
Synset('brother.n.04')
Synset('partner.n.03')
Synset('companion.n.01')
Synset('comrade.n.02')
```

- `companion.n.01` (“a friend who is frequently in the company of another”) is the only sense shared
- Most times, we will find more than one sense in common, so we will need a back-off strategy

# Intersection, how-to

- Both sides of a parallel corpus can be disambiguated by only exploiting the alignment between words
- For each word, we retrieve all its possible senses
- If there is an alignment with its translation, then we retrieve the translation's set of candidate senses and we compute the intersection
  - If the overlap consists of one sense only, then the translation pair has been disambiguated
  - Otherwise, MFS is used as back-off strategy if it appears in the overlap or overlap is an empty set
  - If MFS is not in the overlap, the most frequent sense in the overlap is selected

# Preliminary results

| Method | English | | Italian | |
|---|---|---|---|---|
| | Precision | Coverage | Precision | Coverage |
| Projection | - | - | **0.971** | **0.927** |
| B&P 2005 | - | - | 0.879 | 0.764 |
| Intersection | 0.737 | 0.733 | 0.574 | 0.834 |
| MFS | **0.737** | **0.985** | 0.297 | 0.879 |

- More on the results for intersection:

| | English | Italian |
|---|---|---|
| Disambiguated | 24.73% | 30.92% |
| MFS | 39.33% | 26.51% |
| MFS-Ambiguous | 7.57% | 3.02% |
| MFS-Overlap | 2.49% | 23.67% |
| No alignment | 18.77% | 12.08% |
| No match | 7.10% | 0.65% |
| No synset found | - | 3.14% |

# Future work

- Creation of new WordNet annotated corpora
- Convert the whole MultiSemCor to WN 3.0 and experiment with sense projection and intersection
    - The Romanian MultiSemCor is currently aligned with English, but not with Italian
- Try to use more general statistics on sense frequency to overcome the bias on MFS
- Apply context-wise methods after intersection at a reduced cost
- Experiment with other parallel corpora
    - Bentivogli and Pianta found promising results with free translations (precision 85%, coverage 74%)

# References


- **Luisa Bentivogli and Emanuele Pianta. 2005**. *Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus*, In Natural Language Engineering, Special Issue on Parallel Texts, Volume 11, Issue 03, September 2005, pp. 247-261.

- **Francis Bond and Ryan Foster. 2013.** *Linking and extending an Open Multilingual WordNet*. I*n 51st Annual Meeting of the Association for Computational Linguistics: ACL-2013. Sofia. 1352–1362*

- **Francis Bond and Kyonghee Paik. 2012**. *A survey of wordnets and their licenses.* In Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue. 64–71

- **Emmanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002.** *MultiWordNet: developing an aligned multilingual database.* Proceedings of the 1st Global WordNet Conference, Mysore, India