Multilingual Sense Intersection in a Parallel Corpus with Diverse Language Families

Giulia Bonansinga[1] and Francis Bond[2]

[1]Dipartimento di Filologia, Letteratura e Linguistica
University of Pisa

[2]Linguistics and Multilingual Studies
Nanyang Technological University

giuliauni@gmail.com, bond@ieee.org

January 29, 2016

Sense mapping approaches to Cross-Lingual Word Sense Disambiguation

- Many approaches to Word Sense Disambiguation need high-quality sense-annotated corpora
- Enrich existing parallel corpora with sense annotation by exploiting the differences and similarities in languages

- We aim to overcome the Knowledge acquisition bottleneck that is still present in many less represented languages

Our contribution for reducing the Knowledge Acquisition Bottleneck

- **Task**: given a multilingual corpus, find the appropriate sense for all content words in each component
- **How?** Retrieve all the senses in WordNet that can be associated with each target word and compare them with all the word senses of the aligned translations
- **What is our target text?** Any parallel corpus, as long its components are word-aligned and there are open WordNets inter-linked together for the languages involved

The data: SemCor and its siblings across the world

- SemCor is a sense-annotated corpus of English (Landes et al., 1998)
- Translated to Italian (Bentivogli and Pianta, 2005), Romanian (Lupu et al., 2005) and Japanese (Bond et al., 2012)
  - Mainly annotated through **sense projection** (SP)
    - Assumption: the translation process tends to preserve the meaning across languages
  - Word alignment for Romanian and any other component was inferred in a sense-based fashion
  - Mapping between different WN versions was necessary for all texts but Romanian

	Texts	Tokens	Target words	After mapping
EN	116	258,499	119,802	118,750
IT	116	268,905	92,420	92,022
RO	82	175,603	48,634	=
JP	116	119,802	150,555	=

Table: Statistics for each component of the SemCor parallel corpus

The shared sense inventory

- WordNets aligned to Princeton WordNet, mainly accessed through the Open Multilingual WordNet (Bond and Paik, 2012)

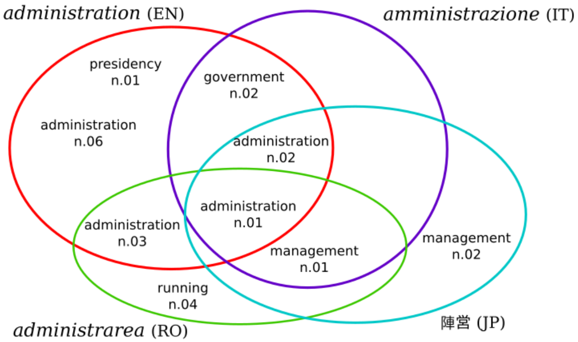
	Synsets	Senses
English	117,659	206,978
Italian	34,728	69,824
Romanian	59,348	85,238
Japanese	57,184	158,069

Table: Coverage of the WordNets for the languages involved

Sense Intersection (SI)

- Assumption: a polysemous word in a certain language is likely to be translated into different words in another language

(EN) *The jury praised the administration and operation of the Atlanta Police Department.*
(IT) *Il jury ha elogiato l'amministrazione e l'operato del Dipartimento di Polizia di Atlanta.*
(RO) *Juriul a lăudat administrarea și conducerea Secției de poliție din Atlanta.*
(JP) 陪審 団 は 、 アトランタ 警察 署 の 陣営 と 働き を 賞賛 し た 。



Sense Intersection (SI) - How the algorithm works

- For each source word and its aligned translation(s), retrieve the set of all its senses in WordNet
- Compute intersection between all sets of candidate senses in order to reduce sense ambiguity
- If the **overlap** (the set resulting from the intersection) contains only one sense, then the source word and its translation(s) are fully **disambiguated**
- Otherwise, use **sense frequency statistics** to disambiguate within the remaining candidate senses

Bringing coarse-grained senses in (I)

- Coverage is very important, but different applications may have different priorities
- A trade-off between the detail of the sense description and its actual usability in real contexts is highly desirable
- Human annotators tend to be as precise as possible, setting a pretty hard threshold to meet
- For our task, we may just be satisfied ignoring minor sense distinctions, as long as the correct sense is conveyed

Bringing coarse-grained senses in (II)

- Navigli et al. (2006) devised an automatic methodology to find a reasonable sense clustering for the senses in WN 2.1 (~30,000)
- We mapped the senses in the clusters found to WN 3.0, losing 101 of them in the process (typically one-element clusters)
- When evaluating, we checked whether the sense chosen by the human annotator belonged to the same cluster as the one selected by the algorithm

Improvement comparing to previous results

Method	English		Italian		Romanian	
	Precision	Coverage	Precision	Coverage	Precision	Coverage
MFS (baseline)	**0.761**	**0.998**	0.599	**0.999**	0.531	**1**
3-way Intersection	0.750	0.778	**0.653**	0.915	**0.590**	1
Coarse-grained MFS	**0.850**	**0.998**	0.687	**0.999**	**0.794**	**1**
Coarse-grained SI	0.849	0.778	**0.761**	0.915	0.661	1

- Resort to sense frequency statistics (SFS) whenever the target word is not yet disambiguated after SI
- SFS are calculated over all texts in the corpus **except** the one being annotated

A more meaningful preliminary evaluation on a small 4-lingual corpus

Method	English	
	Precision	Coverage
Coarse-grained MFS	0.851	**0.998**
Coarse-grained 4-SI	**0.854**	0.788

Table: Coarse-grained evaluation of the results scored with 4-way SI and MFS baseline, computed over the shared subset (49 texts)

Conclusions

- SI beats the MFS baseline for Italian and Romanian in precision, but performs worse for English (whose sense frequencies come from SemCor)
- Coarse-grained evaluation improves all scores, but it really boosts the precision obtained using the MFS baseline with Romanian text
- Error analysis shows that the annotation found by SI is often appropriate, even though it does not match the (very) specific one in the corpus
- Known issues: the corpus is small and the sense frequency statistics are biased

Ongoing and future work

- Produce the alignments for each language pair and compare with sense-based alignment
- Apply SI to different corpora and languages to create new WordNet annotated corpora
- See if using ItalWordnet (Roventini et al., 2002) as well as MultiWordNet (Pianta et al., 2002) helps
- Get more general sense frequency statistics
  - WN Gloss corpus is a good place to start from