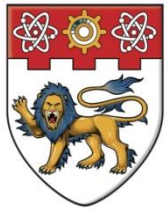


NANYANG TECHNOLOGICAL UNIVERSITY

Computational Lexical Semantics

Cross-lingual sense projection

Giulia Bonansinga
Division of Linguistics and Multilingual Studies

28-08-2014

Outline

- Introduction
  - the manual annotation bottleneck
- MultiSemCor (Bentivogli and Pianta, 2005)
  - Cross-language sense transfer: how to
  - Critical issues
  - Evaluation
  - Feasability on existing parallel corpora
- MultiSemCor+: the Romanian SemCor (Lupu et al., 2005)
  - Browsing the MultiSemCor Web Interface

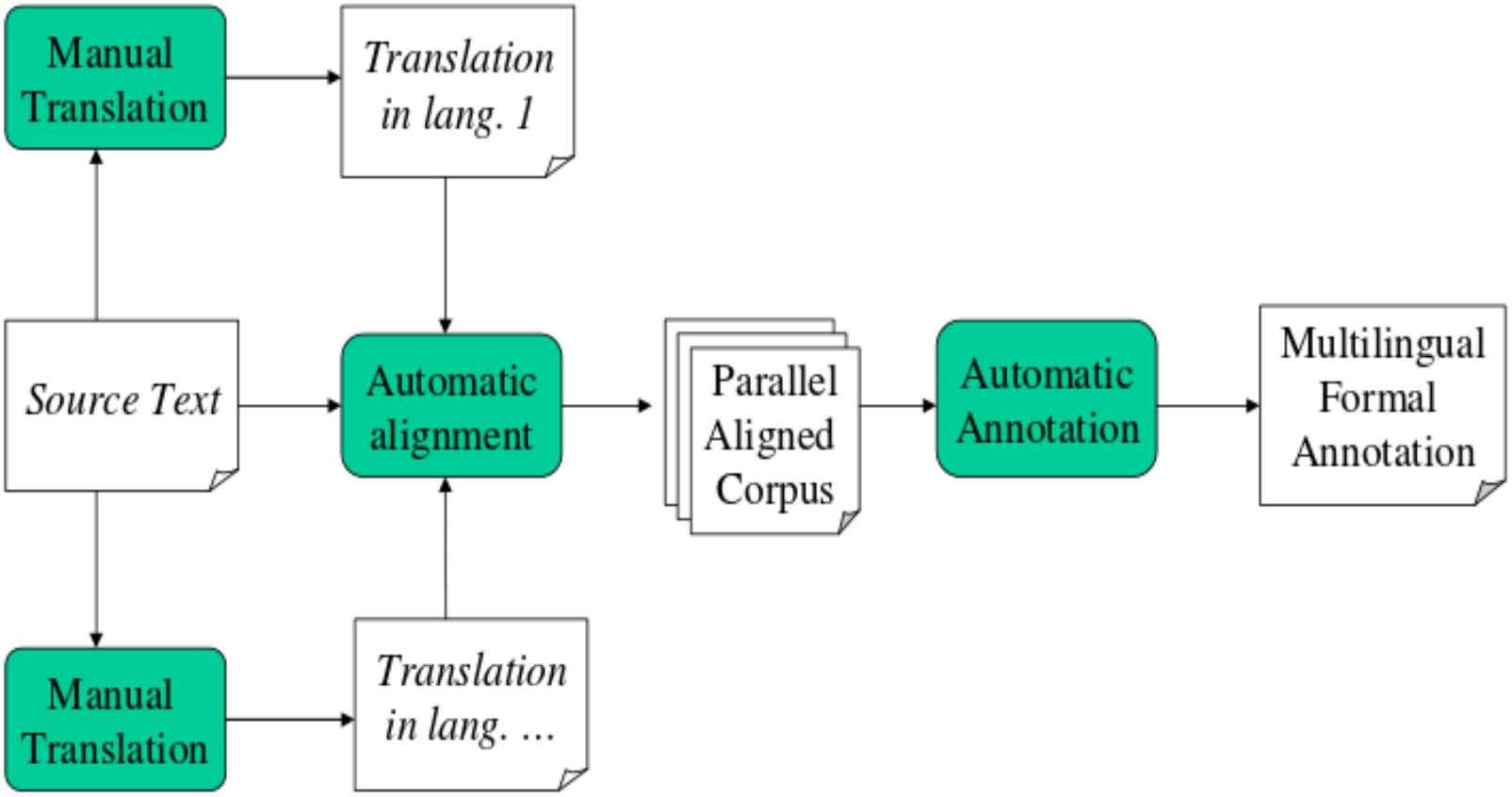
Motivation

- Manual-quality annotated resources are crucial for many NLP tasks
    - but manual annotation is costly and *very* time-consuming
- Alternatives?
    - use less annotated data
    - reduce the cost of manual annotation

Some facts

- Huge imbalance between the resources available for English and those available for other languages
- Plenty of existing parallel corpora
- What if translation was used as annotation?
  - Bentivogli and Pianta exploit this situation and propose an annotation transfer methodology

Translation as Annotation



Advantages

- 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

Assumption

- 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

Inspiration

- The idea of obtaining linguistic information about a text in one language by exploiting parallel or comparable texts in another language has been explored in the field of **Word Sense Disambiguation** (WSD) since the early 1990s.
  - Brown et al. (1991)
  - Gale at al. (1992)
- Further works:
  - Target word selection (Dagan et al. 1991, Dagan and Itai 1994)
  - Word sense clustering (Ide et al. 2002)
  - Cross-language word sense annotation (Diab 2002)
  - Tag projection + processor induction (Yarowsky et al. 2001)
  - Projection of syntactic relations (Hwa et al. 2002, Cabezas et al. 2001)

Creating the MultiSemCor: the procedure

- Goal: align with the English **SemCor** corpus (Landes, Leacock and Tengi 1998).
- Procedure in 3 steps:
  - manually **translate** the SemCor texts into Italian
  - automatically **align** Italian and English texts at the sentence and word level
  - automatically **transfer** the word sense annotations from English to the aligned Italian words

Creating the MultiSemCor: the result

- An Italian corpus annotated with PoS, lemma and word sense
- An English/Italian parallel corpus lexically annotated with a shared inventory of word senses, the synsets of MultiWordNet (Pianta, Bentivogli and Girardi 2002)

SemCor

- Developed at Princeton University
- Subset of the English Brown Corpus (~700,000 running words, all POS tagged)
    - more than 230,000 content words are also lemmatized and semantically annotated with reference to **WordNet** (Fellbaum, 1998).
- 352 texts:
    - “all-words” component consists of 186 texts, in which all open-class words are POS tagged, lemmatized and semantically annotated
        - 350,732 tokens, 192,639 semantically annotated
    - “only-verbs” component consists of the remaining 166 texts in which only verbs have been annotated with lemma and word sense.

MultiSemCor

- English/Italian parallel corpus created on the basis of the English SemCor corpus
- Uses the original release of SemCor (annotated with reference to WordNet 1.6 version), working on the all-words component

	English	**Italian**
Tokens	258,499	268,905
Semantically annotated tokens	119,802	92,420
Distinct synsets	20,142	14,790
Distinct word senses	25,060	22,025

1. Obtaining Italian translations of SemCor texts

- Professional translators were asked to translate the texts
  - Translating and transferring annotations may be a better option than hand-labeling a new corpus from scratch
- Advantages
  - A parallel corpus aligned at the word level with a shared inventory of senses is produced
  - In the case of a corpus translated on purpose, the translation can be controlled
    - criteria to follow in order to maximize alignment and annotation transfer

Controlled translation criteria

- To facilitate the work of the word aligner:
  - maintain the sentence segmentation of the original English texts
  - mark Italian multiword named entities with an underscore, following SemCor conventions (e.g. Unione_Europea as a translation of European_Union)
  - prefer the same dictionary used by the automatic word aligner
- To maximize the quality of the annotation transfer:
  - choose the most synonymous translation equivalents and, more specifically, prefer those belonging to the same PoS.
- These criteria should *never* be followed to the expense of good Italian prose

2. Aligning the texts at the word level with KNOWA

- English/Italian word aligner developed at ITC-irst (Pianta and Bentivogli, 2004), mostly based on information contained in the Collins bilingual dictionary
- Features:
    - morphological analyzer
    - multiword recognizer for both Italian and English
- The application to the MultiSemCor makes the alignment task easier for KNOWA:
    - all multiwords included in WordNet are explicitly marked in SemCor
    - only content words have word sense annotations in SemCor, so it is more important that KNOWA behaves correctly on those
        - content words are easier to align than function words!

3. Transferring annotations from English to Italian

- For each English-Italian word pair

1. project word sense annotation (if any) from SemCor to the Italian text

    - In MultiSemCor English and Italian correspondent synsets have the same identifier

2. add lemma and PoS as selected during the alignment process

- The transfer of annotations from English to Italian is based on the assumption that translation keeps word meaning across languages

Quality issues

- To what extent are the lexica of different languages comparable?
- Bentivogli and Pianta (2000) investigated the comparability of English and Italian lexica
    - the vast majority of English words have an Italian cross-language synonym
    - only 7.8% of the English words correspond to lexical gaps in Italian
- There will be a relatively small number of cases in which the transfer will not be possible

More practical issues

- What's the goal?
  - High-quality Italian annotation
- At each step of the annotation transfer process we run the risk of degradation of the quality of the Italian annotation

SemCor quality: annotation errors can be found in the original English texts

Word Alignment quality: the word aligner may align words incorrectly

Transfer quality: some annotations may not be transferable

Annotation transfer methodology: evaluation

- A gold standard is created, consisting of 4 unseen English texts (br-f43, br-g11, br-l10, br-j53) from the SemCor corpus
- For each English text, both a free and a controlled translation were made
- The resulting gold standard includes 8,877 English tokens, and 9,224 Italian tokens in controlled translations

Create a gold standard (I)

- To evaluate the performance of the word alignment system, the eight pairs of texts in the gold standard were **manually aligned**
- Annotators were asked:
  - to **align** different kinds of units (simple words, segments of more than one word, parts of words)
  - to **mark** different kinds of semantic correspondence between the aligned units
    - full correspondence (synonymic), non-synonymic correspondence, changes in lexical category and phrasal correspondence

Kind of different alignments

	English	Italian
Simple words	health	salute
Segments (multiwords)	rain dance	danza della pioggia
Segments (generic phrases)	open-mouthed	con la bocca spalancata
Parts of words	clasping *him*	afferrando*lo*

Kinds of semantic correspondences between aligned units

	English	Italian
Full (synonymic)	science	scienza
Non-synonymic	meaning	motivo (*reason, grounds*)
Trans-PoS non synonymic	dream previsions	sogni premonitori (*premonitory dreams*)
Fuzzy	the dreamer sees	una persona sogna (*a person dreams*)
Involving extra-grammatical elements	my hands	(le) mie mani
	(he) wants	vuole
omissions	the (ocean of) mankind	il genere umano

Create a gold standard (II)

- The four controlled Italian translations were manually semantically annotated, taking into account the annotations of the English words
    - if the English synset is appropriate for the Italian word, then transfer the annotation
    - otherwise, look for the right synset in MultiWordNet
- Explicit distinction for
    - errors in SemCor annotation
    - non-transferable annotations

Create a gold standard (III)

- 4,313 Italian lexical annotations produced, compared to the original 4,101 English annotations
- Inter-annotator agreement on **word alignment** was 87% for free translations and 92% for controlled translations
- Inter-annotator agreement on **sense annotation** was 81.9% (higher than the score calculated for the original SemCor annotation task)

$$\text{Agreement} : \frac{2 \times \text{N of common words}}{\text{N of aligned/annotated words}}$$

SemCor quality

- Even though the SemCor corpus was manually annotated, a non-negligible percentage of the annotations turns out to be wrong

he put	his hands	on his pockets
lui mise	(le) sue mani	nelle sue tasche
si mise	le mani	in tasca

`{pouch, sac, sack, pocket -- an enclosed space}`

instead of `{pocket -- a small pouch in a garment for carrying small articles}`

- The 117 English annotations considered wrong by the annotators were **explicitly marked** in the gold standard (2.8% of the total English annotations)

  - Note that wrongly annotated English words only cause annotation errors in the Italian text if they are aligned

Word alignment quality (I)

- Good performance of the English/Italian aligner is crucial
- The performance of KNOWA on MultiSemCor was compared to the gold standard alignments, and measured in terms of alignment precision, recall and coverage

$$\text{Precision} \quad \frac{\text{Knowa correct alignments}}{\text{Knowa alignments}}$$

$$\text{Recall} \quad \frac{\text{Knowa correct alignments}}{\text{Gold Standard alignments}}$$

$$\text{Coverage} \quad \frac{\text{Knowa alignments}}{\text{Gold Standard alignments}}$$

Word alignment quality
Evaluation results (I)

Table 1. *Evaluation of KNOWA on full text*

Translation	Precision (%)	Recall (%)	Coverage (%)
Free	85.9	61.8	70.0
Controlled	89.2	69.4	76.1

Word alignment quality Evaluation results (II)

Table 2. *Evaluation of KNOWA on sense tagged words only*

Translation	PoS	Precision (%)	Recall (%)	Coverage (%)
Free	Nouns	95.5	82.7	86.6
	Verbs	87.6	71.3	81.5
	Adjectives	95.9	66.5	69.3
	Adverbs	89.4	42.8	47.9
	Total	**92.8**	**70.3**	**75.8**
Controlled	Nouns	96.9	84.5	87.2
	Verbs	91.4	77.4	84.7
	Adjectives	96.0	72.3	75.3
	Adverbs	91.0	54.4	59.8
	Total	**94.7**	**76.2**	**80.5**

Transfer quality

- Sometimes annotation transfer is not applicable, even if the original English annotations and the word alignment are correct

- An annotation is not transferable from the source to the target language when the translation equivalent does not preserve the lexical meaning of the source word:

  – translation equivalents that are not cross-language synonyms of the source language words
  – translation equivalents that *are* cross-language synonyms, but not lexical units

Incorrect transfer (I)

Translation equivalents that are not cross-language synonyms of the source language words

1) meaning

motivo (reason, grounds)

Suitable in the context, but not a synonymic translation of the English word

2) the possibility for man to coexist with animals

la possibilità per l'uomo di coesistere con gli animali

le possibilità di coesistenza tra gli uomini e gli animali

(the possibility of coexistence between men and animals)

The translation equivalent does not belong to the same lexical category as the source word

3) a dreamer sees

un sognatore vede

una persona sogna *(a person dreams)*

The target phrase has globally the same meaning as the corresponding source phrase, but the single words of the phrase are not cross-language synonyms of their corresponding source words

Incorrect transfer (II)

The translation equivalent is indeed a cross-language synonym of the source expression, but not a lexical unit

- If the target expression is not a lexical unit, it cannot be annotated with one sense as a whole.

1) successfully

con successo (with success)

2) empirically

empiricamente

in modo empirico (in an empirical manner)

This usually happens with **lexical gaps**

Due to translator choice

What impact do non-transferable annotations have?

- Annotators were asked to mark translations pairs in which the English annotation could not be transferred to the Italian translation equivalent
- Non-transferable annotations amount to 692 (16.9% of the English Annotations):
    - 591 (85.4%) due to translation equivalents which are lexical units but are not cross-language synonyms
    - 101 (14.6%) due to translation equivalents that are not lexical units

Final results of the annotation transfer procedure

- Out of the 4,101 SemCor English annotations, the automatic procedure was able to transfer 3,297. Among these, 2,897 are correct and 400 are incorrect for the Italian words

Table 3. *Evaluation of the Italian annotation*

	Precision (%)	Recall (%)	Coverage (%)
Italian *controlled* texts	87.9	67.2	76.4

Table 4. *Source of incorrect transfer*

	#	Error %	Annotation %
English annotation errors	109	27.2	3.3
Word alignment errors	95	23.8	2.9
Non-transferable annotations	196	49.0	5.9
Total Incorrect transfer	**400**	**100**	**12.1**

Incorrect transfer: English annotation errors

- During the creation of the gold standard, 117 annotation errors have been found in the English source (2.8%)
- Almost all of the source errors have been transferred, contributing in a consistent way to the overall Italian annotation error rate.

Table 4. *Source of incorrect transfer*

	#	Error %	Annotation %
English annotation errors	109	27.2	3.3
Word alignment errors	95	23.8	2.9
Non-transferable annotations	196	49.0	5.9
Total Incorrect transfer	**400**	**100**	**12.1**

Incorrect transfer: word alignment errors

- The number of errors in the Italian annotation due to wrong alignments made by KNOWA (2.9%) does not affect the overall Italian annotation in an important way
- Numbers refer to word alignment errors on transferable annotations only

Table 4. *Source of incorrect transfer*

	#	Error %	Annotation %
English annotation errors	109	27.2	3.3
Word alignment errors	95	23.8	2.9
Non-transferable annotations	196	49.0	5.9
Total Incorrect transfer	**400**	**100**	**12.1**

Incorrect transfer: annotation transfer errors

- Words which have been aligned but whose word sense annotation cannot be transferred
- In practice, given the difficulty in deciding what is and what is not a lexical unit, only the lack of synonymy at lexical level has been considered an annotation error
  - Only 196 of the 591 non-synonymous translations marked in the gold standard have been aligned by the word alignment system (33.2%)

Table 4. *Source of incorrect transfer*

	#	Error %	Annotation %
English annotation errors	109	27.2	3.3
Word alignment errors	95	23.8	2.9
Non-transferable annotations	196	49.0	5.9
Total Incorrect transfer	**400**	**100**	**12.1**

Further improvements on the automatic transfer methodology

- Given these results, there is actually little room to improve on precision
  - the only way to reduce wrong annotation transfer would be to manually correct annotation errors in the original SemCor
  - the issue of non-synonymous translation equivalents cannot be solved
- In principle, only the errors caused by KNOWA can be addressed - but they amount to only 2.9% of all annotations!
- On the other hand, coverage is particularly low for adjectives and adverbs
  - Solution: improve the multiword recognition component of KNOWA

MultiSemCor in a nutshell

- 116 English texts aligned at the word level with their corresponding Italian translations
- The final precision of the cross-language annotation transfer methodology is 87.9%
- Being coverage of 76.4%, after the application of the methodology 23.6% of Italian words still need to be annotated
  - The manual annotation of the remaining text would be cost-effective, compared to annotating the corpus from scratch
- Freely distributed for research purposes in XML-based standard compliant format

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

Cross-language annotation transfer methodology in a nutshell

- An approach to the creation of high quality semantically annotated corpora based on the exploitation of parallel texts

– exploits existing (mostly English) annotated resources

– creates corpora in new (resource-poor) languages

– reduces human effort

Free translation vs controlled translation

- How well does the annotation transfer methodology perform with existing parallel corpora?
  - To simulate this scenario, the MultiSemCor gold standard has been extended by semantically annotating also the free translation of text br-g11 (2,016 words, text-category: *belles-lettres*)

Table 5. *Italian annotation quality in free and controlled translations*

	Precision (%)	Recall (%)	Coverage (%)
Free translation	84.8	63.1	74.4
Controlled translation	87.7	70.8	80.7

- Not suprisingly, annotation of the controlled translation is better
  - the gap between the two ranges from 2.9% for precision to 7.7% for recall

Further possibilities

- MultiSemCor can be used both as a monolingual semantically annotated corpus and as a parallel aligned corpus
- It has been used to automatically enrich the Italian component of MultiWordNet
  - 9.6% of the Italian words automatically sense-tagged were not present in MultiWordNet
- The Italian component can used as a gold standard for the evaluation of WSD systems (Gliozzo, Ranieri and Strapparava 2005).
- Besides NLP applications, MultiSemCor is also suitable for consultation by humans through a Web interface (Ranieri, Pianta and Bentivogli 2004)

Future work

- Apply the methodology to the remaining 70 SemCor texts
- Enlarge the evaluation gold standard
- Extend the methodology to other languages for which a WordNet exists and can be aligned with MultiWordNet
  - The Romanian MultiSemCor is currently aligned with English, but not with Italian
- Explore the possibility of transferring syntactic annotation
  - Brown Corpus (of which SemCor is part) has been annotated within the Penn Treebank, so the syntactic annotations of the SemCor texts are also available
- Explore the full exploitation of parallel corpora by projecting other types of linguistic annotation
  - anaphoric reference
  - discourse-level information such as rhetorical relations

Enriching MultiSemCor with frame information

- MultiSemCor has been automatically enriched with frame labels pointing to the synsets, both for Italian and for English
- 27,793 annotated frame instances for English, 23,872 for Italian
- Accuracy was 0.75 for English and 0.70 for Italian
- For the annotation process and its evaluation Tonelli and Pighin (2009)

The Romanian SemCor (Lupu et al. 2005)

- "MultiSemCor+" contains the Romanian SemCor
- Similar approach
    - translation of 34 English SemCor texts (65,9256 tokens, 3,871 sentences)
    - preprocessing and alignment
    - sense information transfer
- Mapping issues: the SemCor used refers to WordNet 2.0, while MultiSemCor refers to WordNet 1.6
- Currently only 12 texts have been aligned

Towards more multi-lingual sense-tagged corpora

- Japanese SemCor is another translation of the English SemCor, whose senses are projected across from English
    - same texts as in MultiSemCor
- Of the 150,555 content words, 58,265 are sense tagged either as monosemous words or by projecting from the English annotation (Bond et al., 2012)

The MultiSemCor Web Interface

- Intended for:
    - Lexicography
    - Translation studies
    - Linguistic teaching
    - Multilingual browsing
- Showing:
    - Linguistic annotation
    - Bilingual sentence alignment
    - Bilingual semantic concordancing
    - Integration between corpora and lexical resources (WordNet)

Browsing MultiSemCor

- Two browsing modalities
  - Text oriented → sentence alignment
    - Alignment at sentence and word level
    - Dictionary
  - Word oriented → semantic concordancer
    - search for all the occurrences of a word form, lemma, or word sense (according to MultiWordNet)
    - specify a certain PoS
  - Always possible to switch from one modality to another
- Integration with the reference lexicon, MultiWordNet

Useful links

- MultiSemCor
  - http://multisemcor.fbk.eu/index.php
- MultiWordNet
  - http://multiwordnet.fbk.eu/online/multiwordnet.php
- WordNet
  - http://wordnetweb.princeton.edu/perl/webwn

Applications

- Parallel corpora
- Contribution to multilingual resources by way of annotations available in another language
- And much more:
    - Multilingual lexical acquisition
    - Machine translation
    - Cross-language Information Retrieval

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, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012.** *Japanese semcor: A sense-tagged corpus of Japanese.* In Proceedings of the 6th International Conference of the Global WordNet Association (GWC)
- **Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. and Mercer, R. L. (1991)** *Word-sense disambiguation using statistical methods.* Proceedings of ACL '91, Berkeley, CA.
- **Cabezas, C., Dorr, B. and Resnik, P. (2001)** *Spanish language processing at University of Maryland: Building infrastructure for multilingual applications.* Proceedings of the 2nd International Workshop on Spanish Language Processing and Language Technologies,Jaen, Spain.
- **Dagan, I. and Itai, A. (1994)** *Word sense disambiguation using a second language monolingual corpus.* Computational Linguistics 20 (4): 563-596.

References

- **Dagan, I., Itai, A. and Schwall, U. (1991)** *Two languages are more informative than one.* Proceedings of ACL '91, Berkeley,CA,USA.
- **Diab, M. (2000)** *An unsupervised method for multilingual word sense tagging using parallel corpora: a preliminary investigation.* Proceedings of the ACL 2000 SIGLEX Workshop on Word Senses and Multi-linguality, Hong Kong.
- **Diab, M. and Resnik, P. (2002)** *An unsupervised method for word sense tagging using parallel corpora*. Proceedings of ACL 2002, Philadelphia, USA.
- **Gale, W. A., Church, K. W. and Yarowsky, D. (1992)** *Using bilingual materials to develop word sense disambiguation methods.* Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada.
- **Gliozzo, R., Ranieri, M. and Strapparava, C. (2005)** *Crossing parallel corpora and multilingual lexical databases for WSD.* Proceedings of CICLing-2005, Mexico City, Mexico.
- **Hwa, R., Resnik, P. and Weinberg, A. (2002)** *Breaking the resource bottleneck for multilingual parsing.* Proceedings of the LREC 2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data, Las Palmas, Canary Islands, Spain.

References

- **Ide, N., Erjavec, T. and Tufis, D. (2002)** *Sense discrimination with parallel corpora.* Proceedings of the ACL 2002 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, USA
- **Landes S., Leacock, C. and Tengi, R. I. (1998)** Building semantic concordances. In: Fellbaum, C. (ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press. Resource Acquisition. *Proceedings of LREC-2002 Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data.* Las Palmas, Canary Islands, Spain.
- **Lupu, M.,Trandabat, D. and Husarciuc H.**. *A Romanian SemCor aligned to the English and Italian MultiSemCor*. In 1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School, Proceedings, pages 20{27, Cluj-Napoca, Romania, July 2005.
- **Pianta, E., Bentivogli, L. and Girardi, C. (2002)** *MultiWordNet: developing an aligned multilingual database.* Proceedings of the 1st Global WordNet Conference, Mysore, India
- **Pianta, E. and Bentivogli, L. (2004)** *Knowledge intensive word alignment with KNOWA.* Proceedings of Coling 2004 Geneva, Switzerland

References

Ranieri, M., Pianta, E. and Bentivogli, L. (2004). *Browsing multilingual information with the MultiSemCor web interface.* Proceedings of the LREC-2004 Workshop on "The Amazing Utility of Parallel and Comparable Corpora", Lisbon, Portugal.

Sara Tonelli, Daniele Pighin (2009). *New features for FrameNet – WordNet mapping.* In Proceedings of the Thirteenth Conference on Computational Language Learning (CoNLL), Boulder, Colorado.

Acknowledgements

- Some slides of this presentation have been borrowed from the talk given by Emmanuele Pianta at the workshop “Multi-lingual semantic annotation: Theory and applications” (Saarland University, Saarbruecken, Germany, June 26th and 27th 2006)