MultiSemCor in a nutshell

- Created on the basis of the English SemCor corpus (Landes et al., 1998)
  - sense-annotated with reference to Princeton WordNet 1.6
  - 352 texts (~700,000 running words)
- Two perspectives, multiple uses:
  1. An Italian corpus annotated with PoS, lemma and word sense
  2. An English/Italian parallel corpus aligned at the word level
- Annotated with a shared inventory of word senses, MultiWordNet (Pianta et al., 2002)

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

Creation and automatic annotation via sense projection

- Procedure in 3 steps:
  1. **manually** translate the SemCor texts into Italian
  2. **automatically** align Italian and English texts at the sentence and word level
  3. **automatically** transfer the word sense annotations from English to the aligned Italian words
- Example sense projection by exploiting the alignment between words:

The	discontinuity	can	either	be	that	of	war
	n_10344737			v_01775973			n_10071856
↓	↓	↓	↘	↙	↓	↓	↓
La	discontinuità	può	essere	sia	quella	della	guerra
to	destruction	or	that	of	diplomatic	policy	.
	n_0141128				a_02557914	n_04536028	
↓	↓	↓	↓	↓	↘	↙	↓
di	distruzione	sia	quella	della	politica	diplomatica	.

Did it work?

- Evaluation of the annotation on the Gold Standard

Precision (%)	Recall (%)	Coverage (%)
87.9	67.2	76.4

- 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

A few applications

- Given the parallel corpus, automatic sense annotation was performed
- As a nice side effect, MultiSemCor was used to **detect missing lemmas and senses** in the Italian component of MultiWordNet

And much more:

- Apply the transfer procedure to other existing parallel corpora to **obtain new multilingual sense-annotated corpora** and solve the **knowledge acquisition bottleneck** for resource-poor languages
- Word Sense Disambiguation, Multilingual lexical acquisition, Machine Translation, . . .

References

- Bentivogli, Luisa and Emanuele Pianta (2005). «Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus». en. In: *Natural Language Engineering* 11.03, p. 247. ISSN: 1351-3249, 1469-8110. DOI: 10.1017/S1351324905003839. URL: http://www.journals.cambridge.org/abstract_S1351324905003839 (visited on 08/25/2014).
- Bond, Francis et al. (2012). «Japanese SemCor: A sense-tagged corpus of Japanese». In: *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pp. 56–63. URL: http://people.eng.unimelb.edu.au/tbaldwin/pubs/gwn2012.pdf (visited on 02/04/2015).
- Landes, Shari, Claudia Leacock, and Randee I Tengi (1998). «Building Semantic Concordances». In: *WordNet: An Electronic Lexical Database*. Ed. by Christiane Fellbaum. Cambridge, MA: MIT Press, pp. 199–216.
- Lupu, Monica, Diana Trandabat, and Maria Husarciuc (2005). «A Romanian SemCor aligned to the English and Italian MultiSemCor». In: *1st ROMANCE FrameNet Workshop at EUROLAN*. Citeseer, pp. 20–27.
- Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi (2002). «MultiWordNet: Developing an Aligned Multilingual Database». In: *In Proceedings of the First International Conference on Global WordNet*. Mysore, India, pp. 293–302.

Describing MultiSemCor with ISLRN Metadata

Title	MultiSemCor
Full official name	MultiSemCor
Resource Type	Other - Annotated text
Source/Url	http://multisemcor.fbk.eu
Format/MIME Type	text/xml
Size/Duration	527,404 tokens (116 texts in English and Italian)
Access medium	Online. Available upon request (CC-BY-3.0 license)
Description	MultiSemCor is an English/Italian parallel corpus, aligned at the word level and annotated with PoS, lemma and word sense. The parallel corpus is created by exploiting the SemCor corpus, which is a subset of the English Brown corpus containing about 700,000 running words. In SemCor all the words are tagged by PoS, and more than 200,000 content words are also lemmatized and sense-tagged with reference to the Princeton WordNet lexical database.
Version	1.1
Media Type	Text
Language(s)	English, Italian
Resource Creator	Emanuele Pianta and Luisa Bentivogli
Distributor	ITC-irst (now FBK), Italy
Rights Holder	bentivo@fbk.eu
Relation	SemCor (Landes et al., 1998) MultiSemCor+ (Lupu et al., 2005) JSemCor (Bond et al., 2012)