

SemEval 2010: Cross-Lingual Word Sense Disambiguation

Els Lefever and Véronique Hoste

LT3 Language and Translation Technology Team
<http://veto.hogent.be/lt3>

December 9, 2009



- 1 Background
 - Word Sense Disambiguation
 - SemEval
- 2 Cross-Lingual WSD
 - Task setup
 - Data creation
 - Evaluation
- 3 Conclusions



Word Sense Disambiguation?

WSD = Select correct sense of ambiguous word in given context



Shortcomings state-of-the-art WSD

- **supervised approaches** to WSD achieve best results, but:
 - heavily rely on **large sense-tagged corpora** (expensive and scarce for languages other than English)
 - fixed **sense inventory**: often arbitrary divisions of word meanings into dictionary senses [Atkins1991].
 - **low inter-annotator agreements** on sense tagging experiments [Fellbaum et al.2001]
- growing feeling in WSD community that WSD should be integrated in **real applications** such as MT or multilingual IR



WSD from parallel corpora

- corpus-based approach: using **translations from a parallel corpus** instead of human-defined sense labels
- **Advantages**
 - easier to integrate in **real applications** (e.g. [Carpuat and Wu2007] show that SMT benefits from dedicated WSD module)
 - implicitly deals with **granularity** problem
 - **language independent** approach



WSD from parallel corpora

- Deduce word senses from parallel corpora \implies hypothesis: **different sense distinctions** of a polysemous word are often **lexicalized** across languages
- many WSD studies have already shown the validity of cross-lingual evidence idea ([Gale et al.1993], [Ide et al.2002], [Ng et al.2003], [Apidianaki2009], etc.)
- the task we propose:
 - differs: independence from an externally defined sense set
 - benchmark data set for cross-lingual WSD



SemEval 2010

- **Senseval**: Evaluation Exercises for the Semantic Analysis of Text, focussing on Word Sense Disambiguation
- **SemEval**: broader scope - include semantic analysis tasks outside of WSD
- Timeline **SemEval 2010**
 - trial data has been released in September
 - deadline for submission of systems in mid Spring 2010 (April or so)
 - probably collocated with ACL (July 11-16, Uppsala)



Organizing the cross-lingual WSD task

- **lexical sample of 25 English nouns** (divided into trial set of 5 nouns and test set of 20 nouns)
- two subtasks:
 - **Bilingual** task: answer consists of translations in one language
 - **multilingual** task: answer consists of translations in five supported languages (viz. French, Dutch, German, Italian and Spanish)
- we developed:
 - **sense inventory** in which the sense distinctions were extracted from multilingual corpus
 - **lexical sample data set** in which the ambiguous words were annotated with the senses from the multilingual sense inventory



Example Bilingual evaluation

- Input: (...) living on the **bank** of the river

Language	Sense label
Dutch (NL)	oever/dijk
French (F)	rives/rivage/bord/bords
German (D)	Ufer
Italian (I)	riva
Spanish (ES)	orilla

- Input: (...) money supply is of direct interest to any **bank**

Language	Sense label
Dutch (NL)	bank/kredietinstelling
French (F)	banque/établissement de crédit
German (D)	Bank/Kreditinstitut
Italian (I)	banca
Spanish (ES)	banco



Example Multilingual evaluation

- (...) living on the **bank** of the river

Language	Sense label
NL,F,D,I,ES	oever/dijk, rives/rivage/bord/bords, Ufer, riva, orilla

- (...) money supply is of direct interest to any **bank**

Language	Sense label
NL,F,D,I,ES	bank/kredietinstelling, banque/ établissement de crédit, Bank/ Kreditinstitut, banca, banco



Parallel corpus

- **Europarl** for 6 languages, viz. English (our target language), Dutch, French, German, Italian and Spanish
- **Sentence-aligned** using a tool based on the Gale and Church algorithm [Gale and Church1991]
- take the **intersection** of all 1 to 1 alignments in order to obtain six-lingual sentence-aligned corpus



Construction of Sense Inventory

- After the selection of all English sentences containing the target nouns and the aligned translations in the five target languages, the following two steps were taken
 - 1 run **word alignment** to find set of possible translations
 - 2 manual **clustering** by meaning (per target word) of the resulting translations



Word Alignment

- We ran GIZA++ to generate initial word alignments for the five language pairs
- Example:

SOURCE: (...) who continue to pursue their normal **occupations** however.

SPANISH: (...) que realizan su **actividad profesional** normal.

DUTCH: (...) die ook hun normale **beroepsbezigheden** blijven voortzetten.

GERMAN: (...) die aber ihrer normalen **Berufstätigkeit** nachgehen.

ITALIAN: (...) che tuttavia non cessano di svolgere la loro **attività lavorativa** precedente.

FRENCH: (...) qui poursuivent cependant leur **activité professionnelle** normale.



Word Alignment: Multiwords

- one single target word can lead to **multiword** translations (e.g. *actividad profesional*) and to **compounds** (e.g. *beroepsbezigheden*).
- in both cases we kept the multi-part translation as a valid translation suggestion
- all word alignments were **manually verified** in the five languages:
 - **correction** of wrong word alignments
 - **“NULL”** link for words for which no valid translation could be identified (not translated or very fuzzy link)
 - dedicated **remarks** section where annotators provide comments for *compound* translations, *fuzzy* translations, *wrong input* (different PoS) and *tokenisation* problems.



Manual Clustering

- the resulting word alignments are **clustered by meaning** across languages (translations are linked across languages on the basis of the unique sentence ID)
- clusters are organized in **two levels**:
 - **top level** reflects the main sense categories (e.g for the word *bank* we have (1) financial meaning, (2) supply or stock, (3) sloping land beside water, (4) West Bank and (5) group of similar objects)
 - the **subclusters** represent the finer sense distinctions
- translations that correspond to English multiword units are identified and the different compound parts are separated by §§ in the clustering file



Example manual clustering 1

	English	Dutch	Italian	French	German	Spanish
1. Job, line of work						
1.1 General		beroep	occupazione	travail	Beruf	ocupación
		professionele activiteit	impiego	emploi	Berufsleben	trabajo
		dienstbetrekking	lavoro	profession	Beschäftigung	actividad
		vak	attività professionale	métier	Arbeit	actividad profesional
	(maritime) occupations	zeevaart	professione	profession	Seemanns	profesión
1.2 groups of professions						
	occupation (of road ha	vervoers	settore	profession	Transport	ramo
		sector	settore in cui operare	secteur	Berufs	profesión
		beroeps	categoria professionale	métier	Berufs	campo profesional
1.3 Function at work						
		functie	funzione	fonction	Tätigkeit	cargo
		betrekking	carica	poste	Beschäftigung	ocupación
2. activity, bezigheid						
		bezigheid	attività	affaire	Betätigungs	actividad
		activiteit	sbocco	passee-temps	Tätigkeit	ocupación
		bezighouden	occupazione	activité	beschäftigen	
3. (Military) control						
3.1 occupation of country/people						
		belegering	occupare	occuper	Okkupation	ocupación
		bezet	occupato	occupation	Besetzung	ocupado
		bezet houden	il fatto dell'occupazione	occupé	besetzen	ocupar
3.2 Person(s) that occupy						
	(troops of) occupation/	bezettingen	occupazione	occupation	Besatzungs	ocupación
		betzetter	stato di occupazione	occupant	Fremdherrschaft	dominio
		bezettingen	occupante		Herrschaft	ocupante
		bezettingen			Okkupations	
3.3 Active - non military - occupation						
	land	land	occupazione	occupation	Raum	ocupación
	office	inbeslagname	occupazione	occupation	Besetzung	ocupación
	churches	kerk	occupazione	occupation	Kirchen	ocupación
4. Usage (passive - non aggressive - occupation)						
	space/air	beslag	occupazione	occupation	Sperrung	ocupación
		ruimte	utilizzo		Raum	ordenación
	seats	bezetting	assegnazione	attribution	Besetzung	atribución

Example manual clustering 2

Dutch	Italian	French	German	Spanish
Bank	Cisgiordania	Cisjordanie	West§§jordanland	Cisjordania
Cisjordanie	sponda	rive	Jordan§§ufer	río Jordán
Jordaan-oever	cisgiordano	bande	West-Bank	Franja
Jordaan§§oever	riva occidentale del Giordano	cisjordanien	West§§bank	costa
Transjordanie	sponda occidentale del Giordano	Bank	Bank	orilla
West§§bank	Bank	Banque	West§§jordangebiet	Ribera
West§§oever	striscia di Gaza		West§§jordanien	junto
deel	riva		West§§jordan§§ufer	
oever			West§§küste	
			West§§ufer	
			Ufer	

Table: Translation cluster for the English noun **bank** in the *West Bank* meaning



Annotation of trial and test instances

- the resulting sense inventory is used to **annotate** the trial (100) and test (1000) instances that are selected from JRC-ACQUIS
- For the annotation, we asked our linguists to:
 - pick the contextually appropriate **sense cluster**
 - choose their **three preferred translations** from this cluster, that are used to assign **frequency weights** to the gold standard translations



Annotation Example

SENTENCE 3. Considering the importance of the existing links between the Community and the Palestinian people of the West **Bank** and the Gaza Strip, and the common values that they share

Language	Sense label
French Cluster:	4
French 1:	Cisjordanie
French 2:	rive
French 3:	bande
Italian Cluster:	4
Italian 1:	Cisgiordania
Italian 2:	riva
Italian 3:	sponda

Table: French and Italian annotation example of instance containing **bank** in the *Cisjordan* meaning



Cluster Agreement

The table below illustrates the agreement on the appropriate sense cluster for the five trial words

Target word	Avg Nr diff. clusters per sent	Avg Nr diff. top level clusters	total Nr of sent with cluster consensus	total Nr of sent with top level cluster consensus
Bank	1.15	1.05	18	19
Passage	2	1.25	9	16
Plant	2	1.05	5	19
Occupation	2.15	1.7	5	9
Movement	2.9	1.7	1	9



Gold Standard

- **gold standard** contains for every sentence the chosen cluster number(s) and a set of translations (enriched with frequency information)
- the format is very similar to the format that will be used for the Lexical Substitution task[Sinha et al.2009]

bank.n.fr 3 :: bande 2; cisjordanie 5; cisjordanien 1; rive 3;

bank.n.nl 3 :: cisjordanië 1; jordaanoever 3; oever 2; westbank 3; westoever 3;

bank.n.de 3 :: west-bank 1; westbank 2; westjordanien 2; westjordanland 2;

westjordanufer 3; westufer 2;

bank.n.it 3 :: cisgiordania 3; riva 1; riva occidentale del giordano 2; sponda 1;

sponda occidentale del giordano 1; striscia di gaza 1;

bank.n.es 3 :: cisjordanía 4; franja 3; ribera 1; río jordán 2;



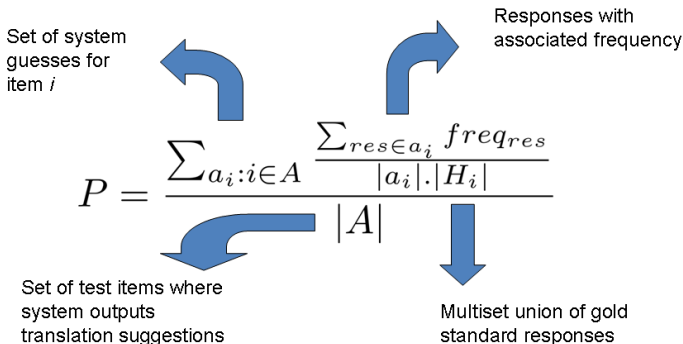
Evaluation strategies

- evaluation scheme that is inspired by Lexical substitution task [McCarthy and Navigli 2007]
- precision and recall
- **best result** evaluation and **more relaxed** evaluation (top five results)



Best result evaluation

- for the **best result** evaluation, systems can propose as many guesses as system believes are correct, but resulting score is divided by number of guesses


$$P = \frac{\sum_{a_i: i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|a_i| \cdot |H_i|}}{|A|}$$

Set of system guesses for item i

Responses with associated frequency

Set of test items where system outputs translation suggestions

Multiset union of gold standard responses

Best result example

For following example:

Gold standard: happy.a.nl :: gelukkig 3; uitzinnig 2; tevreden 1; blij 1;

System output: happy.a :: gelukkig; tevreden;

The credit in the numerator of precision and recall would be:

$$\frac{3 + 1}{2.7} = .286 \quad (1)$$



Relaxed evaluation

- for the more **relaxed** evaluation, systems can propose up to five guesses, the resulting score is not divided by the number of guesses

$$P = \frac{\sum_{a_i: i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|A|} \quad (2)$$

$$R = \frac{\sum_{a_i: i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|T|} \quad (3)$$



Cluster evaluation

- the same precision and recall formulas are used to perform a **cluster evaluation**, where all translations belonging to the gold standard cluster are considered as valid reference translations
- in the cluster evaluation, we distinguish between **coarse-** (top level clusters) and **fine-grained** (subclusters) performance scores



Baseline Systems

- frequency-based baseline that is based on the GIZA++ word alignments and returns the **most frequent translation** of the target word (used for the best result evaluation)
- frequency-based baseline that returns the **five most frequent translations** according to the GIZA++ word alignments (used for the more relaxed evaluation)
- baseline that returns the **most frequent EuroWordNet sense** (available in five target languages)



Conclusions

- we presented a multilingual Word Sense Disambiguation task for a sample of English nouns
- our sense inventory is built up on the basis of translations retrieved from the Europarl corpus in five target languages
- systems can participate in a bilingual or multilingual evaluation and are asked to provide correct translations for new instances of the selected polysemous target nouns





M. Apidianaki.

2009.

Data-driven semantic analysis for multilingual wsd and lexical selection in translation.

In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Athens, Greece.



B. Atkins.

1991.

Building a lexicon: the contribution of lexicography.
International Journal of Lexicography, 3:167–204.



M. Carpuat and D. Wu.

2007.

Improving statistical machine translation using word sense disambiguation.



In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic.



C. Fellbaum, M. Palmer, H. Trang Dang, L. Delfs, and S. Wolf.

2001.

Manual and automatic semantic annotation with wordnet.
In *Proceedings of the Workshop on WordNet and Other Lexical Resources*.



William A. Gale and Kenneth W. Church.

1991.

A program for aligning sentences in bilingual corpora.
In *Computational Linguistics*, pages 177–184.



W.A. Gale, K.W. Church, and D. Yarowsky.

1993.



A method for disambiguating word senses in a large corpus.
In *Computers and the Humanities*, volume 26, pages 415–439.



N. Ide, T. Erjavec, and D. Tufis.

2002.

Sense discrimination with parallel corpora.

In *Proceedings of ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54–60.



D. McCarthy and R. Navigli.

2007.

Semeval-2007 task 10: English lexical substitution task.

In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.



H.T. Ng, B. Wang, and Y.S. Chan.

2003.



Exploiting parallel texts for word sense disambiguation: An empirical study.

In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, Santa Cruz.



R. D. Sinha, D. McCarthy, and R. Mihalcea.
2009.

Semeval-2010 task 2: Cross-lingual lexical substitution.
In *Proceedings of the NAACL-HLT 2009 Workshop: SEW-2009 - Semantic Evaluations*, Boulder, Colorado.

