

Automatic Acquisition of Sense Examples using ExRetriever

Montse Cuadros, Jordi Atserias, Mauro Castillo, German Rigau

February 9, 2005



<http://www.lsi.upc.es/~nlp/meaning>



Montse Cuadros

Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work

Outline

- **Automatic Acquisition of Examples for WSD**
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work

Automatic Acquisition of Examples for WSD

- Current research on WSD uses semantically annotated corpora to train Machine Learning algorithms to WSD
- Recent work is focusing on reducing the acquisition cost and the need for supervision in corpus-based methods for WSD.
- [Leacock *et al.* 98], [Mihalcea & Moldovan 99] and [Agirre & Martinez 00] automatically generate arbitrarily large corpora for unsupervised WSD training, using the knowledge contained in WordNet to formulate search engine queries over large text collections or the Web.

Outline

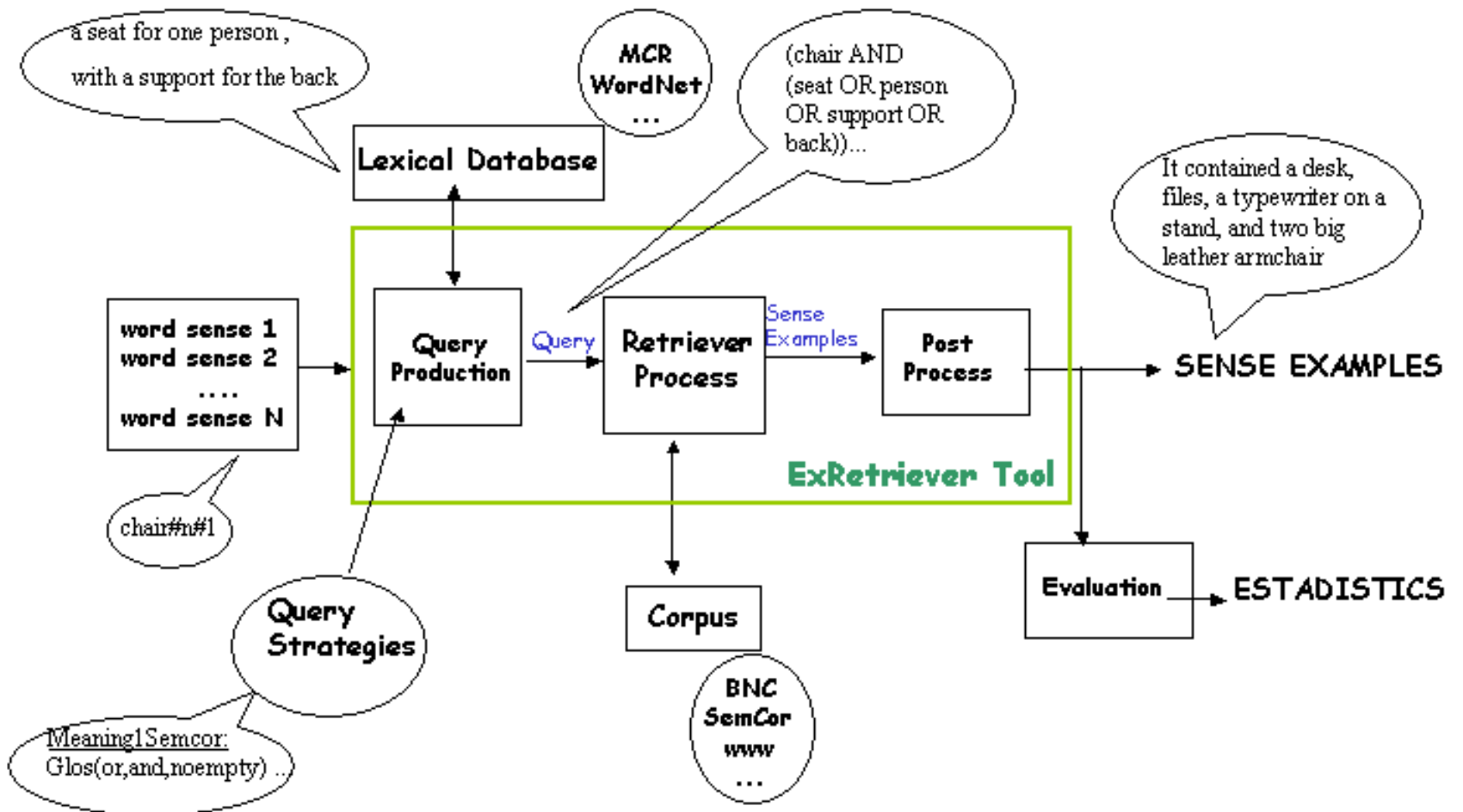
- Automatic Acquisition of Examples for WSD
- **ExRetriever: A Sense Example Retriever Tool**
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work

ExRetriever: A Sense Example Retriever Tool

- ExRetriever characterises each sense of a word as a specific query.
- Using a query construction strategy, defined *a priori*.
- Strategies can take into account the information related to words from lexical database in order to automatically generate the set of queries.
- The resulting specific queries are used to retrieve particular sense examples from a large text collection.

ExRetriever: A Sense Example Retriever Tool (2)

- Able to use different lexical databases e.g.the Multilingual Central Repository of MEANING [Atserias *et al.* 04]
- Different corpora (SemCor, BNC, the Web, etc.)
- ExRetriever has been powered with a declarative language to define different query construction strategies.
- Postprocess module (e.g. tagging, lemmatizing, recognizing WordNet multiwords). [Arranz *et al.* 04]



Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- **The Query Language**
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work

The Query Language

- **Operators** "and", "or" and "not".
- **Constants:**
 - **noempty** a parameter of **Glos** function to remove empty words.
 - **senses** *lemma#POS#sense number* (e.g. church#n#2)
 - **relations** names of the lexical relationships used as parameters to "rel" and "nrel" (e.g. *hyponym*).
- **Functions** Currently,
 - **Glos** to build expressions from the words in the gloss.
 - **rel** to look up the different relations in the Lexical database
 - **nrel** similar to *rel*, establishes the maximum polysemy of the returned senses.

Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- **Examples**
- WSC measure
- Experiments and Results
- Conclusions and Future Work

Example

Meaning1Semcor: `Glos(or,and,noempty)` OR `or(nrel(1,syns))` OR `or(nrel(1,hypo))`

The first function `Glos(or,and,noempty)` returns a logical formula which is the target word (e.g. *chair*) and the union set with *or* of the *noempty* words of the *gloss* of the sense considered (e.g. *char#n#3*).

The second function, `or(nrel(1,syns))` returns the union set with *or* of the monosemous synonyms.

Finally, `or(nrel(1,hypo))` returns the union set of the monosemous hyponyms.

Example: the noun *Chair*

chair#n#1:
<p>(<i>chair</i> AND (<i>seat</i> OR <i>person</i> OR <i>support</i> OR <i>back</i>)) OR (<i>barber chair</i> OR <i>chaise longue</i> OR <i>folding chair</i> OR <i>highchair</i> OR <i>feeding chair</i> OR <i>ladder-back chair</i> OR <i>lawn chair</i> OR <i>garden chair</i> OR <i>rocking chair</i> OR <i>straight chair</i> OR <i>side chair</i> OR <i>swivel chair</i> OR <i>tablet-armed chair</i> OR <i>wheelchair</i>)</p>

Table 1: Queries for *chair* noun using **Meaning1SemCor**

sense	gloss	hypo	syn
n#1	<i>a seat for one person , with a support for the back</i>	<i>armchair (2) barber_chair ...</i>	

Table 2: Sense of *chair* noun in wordNet 1.6

Sense II

char#n#2:
(<i>chair</i> AND (position OR professor)) OR (professorship)

Table 3: Queries for *chair* noun using **Meaning1SemCor**

sense	gloss	hypo	syn
n#2	<i>the position of professor</i>		professorship

Table 4: Sense of *chair* noun in wordNet 1.6

Sense III

chair#n#3:
(<i>chair</i> AND (<i>officer</i> OR <i>presides</i> OR <i>meetings</i> OR <i>organization</i>)) OR (<i>chairman</i> OR <i>chairwoman</i> OR <i>chairperson</i>) OR (<i>vice chairman</i>)

Table 5: Queries for *chair* noun using **Meaning1SemCor**

sense	gloss	hypo	syn
n#3	<i>the officer who presides at the meetings of an organization</i>	<i>vice_chairman</i>	<i>president</i> (6) <i>chairman</i> <i>chairwoman</i> <i>chairperson</i>

Table 6: Sense of *chair* noun in wordNet 1.6

Sense IV

chair#n#4:
(<i>chair</i> AND (instrument OR death OR electrocution OR resembles)) OR (electric chair OR death chair OR hot seat)

Table 7: Queries for *chair* noun using **Meaning1SemCor**

sense	gloss	hypo	syn
n#4	<i>an instrument of death by electrocution that resembles a chair</i>		electric_chair death_chair hot_seat

Table 8: Sense of *chair* noun in wordNet 1.6

Example of the Post Processing

<**Example** Sentences="1" src="brownv/tagfiles/br-a10#43160" > Seeking this two-year <MEANING synsetPOS="v" baseSense="1" baseLema="call" origPOS="n" rel="hypo" synsetSense="1" synsetLema="term" basePOS="v" > term </MEANING> are James_Culbertson, Dwight_M._Steeves, James_C._Piersee, W._M. Sexton and Theodore_W._Heitschmidt. </**Example**>

lemaTAG: term lemaORIG: term
posTAG: n posORIG: v

Example of the Post Processing

<**Example** Sentences="1" src="brownv/tagfiles/br-e16#52577" > Since the sides are also covered up to the spray <MEANING synsetPOS="n" baseSense="3" baseLema="bar" origPOS="n" rel="hypo" synsetSense="1" synsetLema="rails" basePOS="n" > rails </MEANING> , they are also rough sanded in_that area. </**Example**>

lemaTAG: rail lemaORIG: rails
posTAG: n posORIG: n

Example of an extracted sentence

<**Example** Sentences="1" src="brown2/tagfiles/br-l15#104577" > It contained a desk, files, a typewriter on a stand, and two big leather <MEANING origPOS="n" rel="hypo" synsetSense="1" synsetLema="armchair" synsetPOS="n" baseSense="1" baseLema="chair" basePOS="n" origSense="1" > armchairs </MEANING> . </**Example**>

Results for *chair* against SemCor

Sense	Ok	Ko	NoTag	#Sense	P	R	F1
n#1	16	2	1	34	89	44	59
n#2	1	1	0	3	50	25	33
n#3	7	0	32	11	100	64	78
n#4	0	1	0	0	0	0	0
Totals	24	4	33	48	86	24	38

Table 9: Results of *chair#n* applying **Meaning1SemCor**

Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- **WSC measure**
- Experiments and Results
- Conclusions and Future Work

WSC measure

- Precision, Recall and F1 don't show if the examples retrieved cover all the sense of a word.
- This is a crucial issue if we want to use the acquired examples to train supervised WSD systems.
- We have defined a new measure, WSC (word sense coverage).

WSC measure :

$$WSC = 100 \sum_{w=1}^n \times \frac{SensesWithinRetrievedExamples(w)}{SensesWithinCorpus(w)}$$

Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- **Experiments and Results**
- Conclusions and Future Work

Experiments and Results

- MCR as lexical Database.
- Semcor as corpus.
- 6 different query construction strategies.
- Precision, Recall, F1 and WSC measure.

Q	Ok	Ko	NoTag	#Sense	P	R	F1	WSC
Lea1	851	10	371	23254	98,84	3,66	7,06	23
Mol1	153	1	83	3241	99,35	4,72	9,01	10
Mol3	1987	22474	1303	7611	8,12	26,11	12,39	47
Mea1	2314	22617	1415	9490	9,28	24,38	13,44	54
Mea2	4513	37688	2986	17171	10,69	26,28	15,20	58

Table 10: Overall figures

Experiments and Results (2)

- **Moldo1** and **Lea1** strategies obtain the best precision (around 99%), but poor coverage and WSC.
- **Meaning1**, **Meaning2**, **Moldo3** methods obtain much better recall (about 25% vs 5%) and WSC but less precision.
- **Meaning2**, the best WSC obtaining examples for 58% of the senses.
- **Moldo2** strategy do not provide results in SemCor, as it looks for the complete synset gloss.

Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- **Conclusions and Future Work**

Conclusions

- ExRetriever, a query-based system to extract sense examples from corpus has been described.
- Some preliminar experiments have been presented. They have been used to evaluate the performance of different types of query construction strategies.
- Using ExRetriever, new strategies can be easily defined, executed and evaluated.

Future Work

- Experiment other strategies. (e.g. performing full parsing on the glosses could help discarding irrelevant words from glosses).
- Using the knowledge already contained into the MCR (e.g., selectional preferences, domain information, etc.) to better model sense words as queries.
- Use alternative schemata for building queries, such as the incremental process performed by [Leacock *et al.* 98].
- Follow [Widdows 03]. It seems that most of the errors produced because of the substitution of the target word for their relatives can be avoided.

Future Work (2)

- Use other sense tagged corpora for direct comparisons of ExRetriever (e.g. DSO).
- Perform indirect evaluations using supervised WSD systems on the acquired sense examples.
- Once acquired a sense tagged corpus using ExRetriever, we will use several Machine Learning algorithms to perform several cross-comparisons with respect to other sense tagged resources (SemCor, DSO and those resources provided by Senseval).

Available here:

You can download it here:

<http://www.lsi.upc.es/~nlp/meaning/downloads.html>

Thanks for your attention



<http://www.lsi.upc.es/~nlp/meaning>

This research has been partially funded by the Spanish Research Department (HERMES TIC2000-0335-C03-02) and by the European Commission (MEANING IST-2001-34460).

Bibliography

References

- [Agirre & Martinez 00] (Agirre & Martinez 00) E. Agirre and D. Martinez. Exploring Automatic Word Sense Disambiguation With Decision Lists and the Web. In *Proceedings of the COLING workshop on Semantic Annotation and Intelligent Annotation*, Luxembourg, 2000.
- [Arranz *et al.* 04] (Arranz *et al.* 04) Victoria Arranz, Jordi Atserias, and Mauro Castillo. Multiword Expressions for Word Sense Disambiguation. Technical report, of the LSI Department. LSI-04-47-R. Universitat Politècnica de Catalunya, 2004.
- [Atserias *et al.* 04] (Atserias *et al.* 04) Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. The Meaning Multilingual Central Repository. In *Second International WordNet Conference-GWC 2004*,

pages 23–30, Brno, Czech Republic, January 2004. ISBN 80-210-3302-9.

[Leacock *et al.* 98]

(Leacock *et al.* 98) C. Leacock, M. Chodorow, and G. Miller. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–166, 1998.

[Mihalcea & Moldovan 99]

(Mihalcea & Moldovan 99) R. Mihalcea and I. Moldovan. An Automatic Method for Generating Sense Tagged Corpora. In *Proceedings of the 16th National Conference on Artificial Intelligence*. AAAI Press, 1999.

[Widdows 03]

(Widdows 03) D. Widdows. Orthogonal Negation in Vector Spaces for Modelling Word-Meanings and Document Retrieval. In *Proceedings of 41th annual meeting of the Association for Computational Linguistics (ACL'2003)*, Sapporo, Japan, 2003.