Translation-based Word Sense Disambiguation

Gunn Inger Lyse
University of Bergen

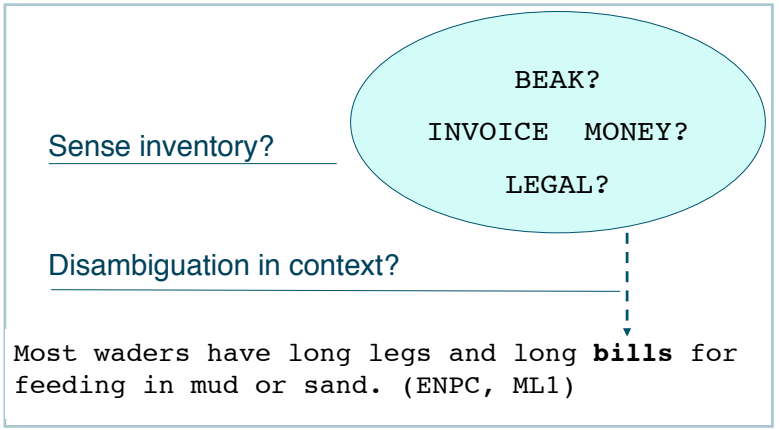
CLARA lexicon course
Bergen, June 2011

- PhD project in affiliation with the **LOGON** project (Machine Translation)
- LOGON project description: ”The biggest single challenge in computational linguistics is ambiguity”.

Background—Word Sense Disambiguation

Stemmen lød plutselig interessert
?? His **vote** all of a sudden sounded interested.
??His **voice** all of a sudden sounded interested.

Background—Word Sense Disambiguation



Background—Word Sense Disambiguation

- Most promising WSD-approach:
  Corpus-based, supervised machine learning techniques

Background—Word Sense Disambiguation

- Most promising WSD-approach:
  Corpus-based, supervised machine learning techniques

waved for the	**bill**	
called for his	**bill**	
wo n't pay the	**bill**	any longer
with its duck-like	**bill**	beaver-like tail and webbed feet
long legs and long	**bills**	for feeding in mud
uses its strong	**bill**	to drill holes into the bark

Background—Word Sense Disambiguation

- Most promising WSD-approach:
  Corpus-based, supervised machine learning techniques

waved for the	**bill**	,INVOICE
called for his	**bill**	,INVOICE
wo n't pay the	**bill**	any longer,INVOICE
with its duck-like	**bill**	beaver-like tail and webbed feet,BEAK
long legs and long	**bills**	for feeding in mud,BEAK
uses its strong	**bill**	to drill holes into the bark,BEAK

Background—Word Sense Disambiguation

- ”The sparse data problem”: the need for training data that are

  (i) sense-labelled prior to learning

  (ii) sufficiently informative for statistical methods.

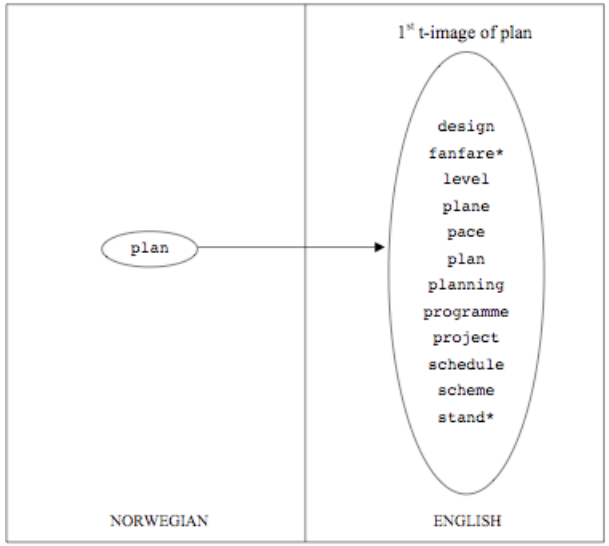
Goal

- Develop and test a method for automatic sense-tagging
- Attempt to alleviate the sparse data problem by generalizing from the seen instances.
- Evaluation: WSD as a practical task to evaluate the key knowledge source: The Mirrors Method

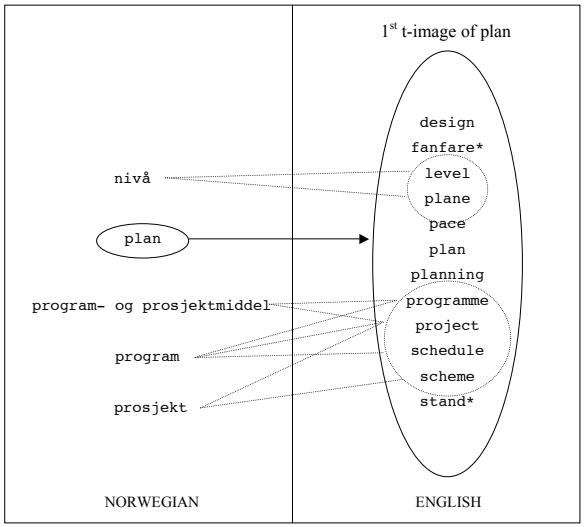
The Mirrors method

- Developed by Helge Dyvik
- Mirrors hypothesis:
- The translational relation as a theoretical primitive for deriving:
  - Sense distinctions
  - Semantic relations between word senses

The Mirrors method



The Mirrors method



The Mirrors method

The Mirrors method

- Problem: how to evaluate the Mirrors method?
- Three main solutions:
  - Comparison against a 'gold standard'
  - Manual verification
  - Validation within a practical NLP task
    - a well-defined end-user application may provide a stable framework to demonstrate the benefits and drawbacks of a resource/system.

The Mirrors method and WSD

– WSD as a practical task to evaluate the Mirrors:

Vary the knowledge source to learn from but maintain the same experimental framework (classification algorithm, data sets, lexical sample and sense inventory).

(Ng & Lee, 1996; Stevenson & Wilks, 2001; Yarowsky & Florian, 2002; Specia et al., 2009)

The Mirrors and WSD

”Using translations from a corpus instead of human defined (e.g. WordNet) sense labels, makes it easier to integrate WSD in multilingual applications, solves the granularity problem that might be task-dependent as well, is language-independent and can be a valid alternative for languages that lack sufficient sense-inventories and sense-tagged corpora”.

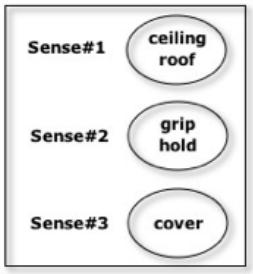
(From the description of the SEMEVAL 2010 task #3: Cross-Lingual Word Sense Disambiguation1)

Method

- Sense-tag a corpus automatically with Mirrors senses
- Select a lexical sample
- Train WSD classifiers
  - the traditional way (context words)
  - using Mirrors-derived information about context words

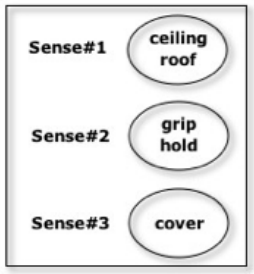
Automatic sense-tagging

Mirrors sense partitions for *takN*



Automatic sense-tagging

Mirrors sense partitions for *takN*



Word-aligned parallel corpus

English	*Norwegian*
... ceiling ...	... tak|**#1** ...
... roof ...	... tak|**#1** ...
... roof ...	... tak|**#1** ...
... ceiling ...	... tak|**#1** ...
... grip ...	... tak|**#2** ...
... grip ...	... tak|**#2** ...
... hold ...	... tak|**#2** ...
... cover ...	... tak|**#3** ...
... roof ...	... tak|**#1** ...
... ceiling ...	... tak|**#1** ...

Automatic sense-tagging: coverage

ENPC automatically sense-tagged tokens (Norwegian side)				
Word class	sense-tagged	untagged	total	coverage
Nouns	155,567	138,291	293,858	.53
Verbs	145,428	94,528	239,956	.61
Adjectives	45,749	66,386	112,135	.41
Adverbs	-	-	66,992	-
Closed-class	-	-	552,356	-
Total	346,744	299,205	1,265,297	

ENPC automatically sense-tagged tokens (English side)				
Word class	sense-tagged	untagged	total	coverage
Nouns	133,742	203,393	337,135	.40
Verbs	145,296	107,509	252,805	.57
Adjectives	43,996	55,108	99,104	.44
Adverbs	-	-	102,569	-
Closed-class	-	-	548,700	-
Total	323,034	366,010	1,340,313	

Automatic sense-tagging

PROS

- sense-tags corpus instances with perfect precision (..as perfect as the automatic word alignment and the Mirrors sense partitions)
- applicable for any language pair for which word-aligned corpus material exists
- May be applied on both language sides.

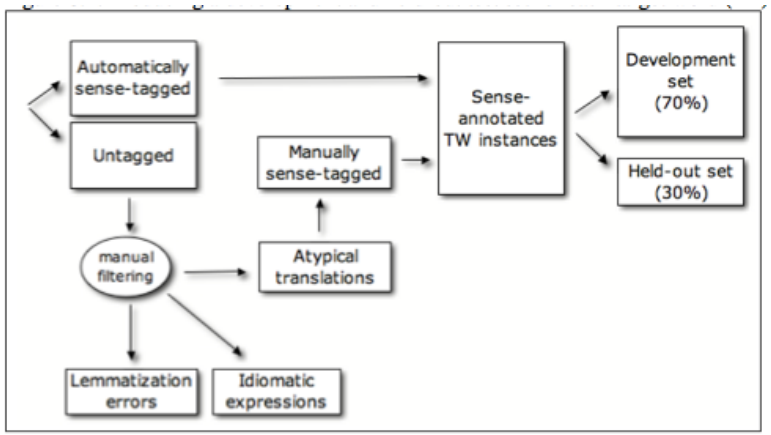
CONS

- intrinsically limited by the need for an existing, identifiable translational correspondent.

Lexical sample

- 15 words with as uncontroversial sense distinctions as possible
  - 4039 instances totally; average training set=188 examples; average test set=80 examples.
- The Swedish lexical sample (SENSEVAL-2) contained 40 lemmas; average training set=218 examples, average test set=38 instances.
- the SEMEVAL-2007 English lexical sample task had 65 verbs and 35 nouns; average training set=222 examples, average test set= 49 examples

Lexical sample: 15 words



Machine Learning algorithm

- Naive Bayes model for learning and classification (well-documented and well-understood in WSD)
- Evaluation:

$$\text{Recall} = \frac{\text{\# correct classifications}}{\text{\# Total classifications to be made}}$$

- Statistical test of significance: McNemar's (when the no. of changed outcomes exceeds 25) and the sign test (when the no. of changed outcomes < 26)

Train on context words vs Mirrors-derived inf. about these context words

- Basic idea:

  Keep experimental framework stable, and test systematically the effect of using different knowledge sources
  - WORDS (W)
  - SEMANTIC-FEATURES (SF)
  - RELATED-WORDS (REL-W)

A WORDS (W) model

- Collect the *n* nearest open-class words

Example with a [±5] context window:

What was it really that they fussed over there in town, in their big flat with all its appliances that regularly broke down (so-called conveniences that demanded both thought and money), meetings, work, appointments, parties, telephones, theatres, **bills3**, fixed times...

Mirrors-derived information about context words

Sense-tagged (bold-face) version of sentence

What was it really that they **fussed1** over there in **town2**, in their **big1 flat3** with all its **appliances1** that regularly broke down (**so-called2 conveniences1** that **demanded1** both **thought2** and money), meetings, **work1**, appointments, **parties3**, **telephones2**, **theatres4**, **bills3**, fixed times... (BV1T)

SEMANTIC-FEATURES (SFs) model

a sense-tagged context word is replaced by the SFs associated with this word sense in the Mirrors word bases.

Example: *telephone2*

[conversation2|telefonsamtale1]

(telephone2 conversation2)

[call1|telefon1]

(telephone2 phone1 call1)

[telephone2|telefonnummer1]

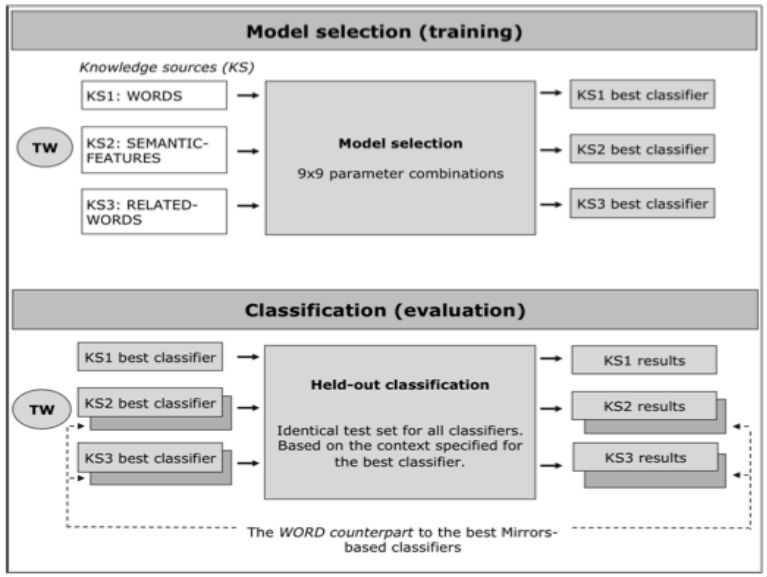
(telephone2 phone1)

A RELATED-WORDS (REL-W) model

- Builds on the defintions of hyperonyms, synonyms and hyponyms of a sense in the Mirrors method.
- Neutralises the original Mirrors distinction between hypero-/hyponymy and synonymy.
- Rrestricts the definition of relatedness to avoid too many RELATED-WORDS.

Example: telephone2

call1 conversation2 phone1 telephone2



- EXP1: how well may a traditional WORD classifier perform?
- EXP2: Replace context words with Mirrors-derived SFs.
- EXP3: Replace context words with Mirrors-derived REL-Ws.
- EXP4: Combine EXP1, EXP2 and EXP3 in a voting scheme where the most confident gets to vote (more confident and more correct classifications?)

Results

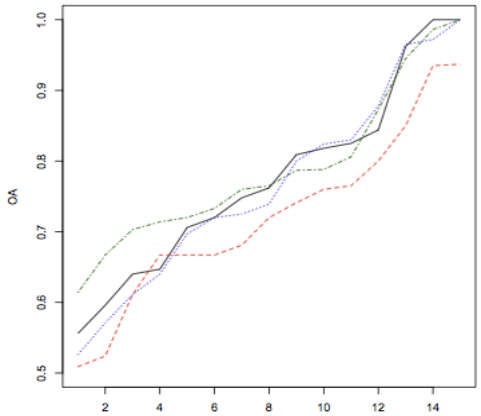


Figure 9.6: Overall Accuracy from Table (9.15) sorted by each model independently. Legend: black=W, red=SF, green=REL-W, blue=combination.

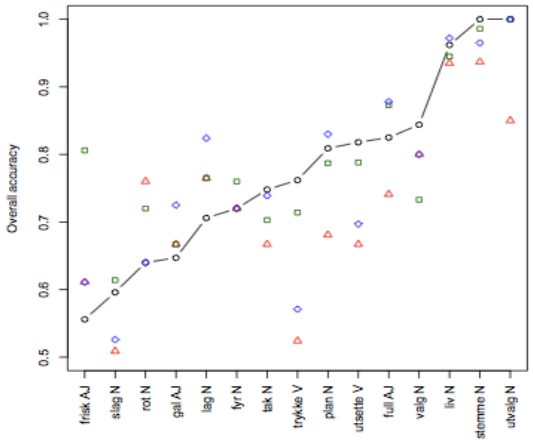


Figure 9.5: Overall Accuracy from Table (9.15) sorted by W. Legend: black=W, red=SF, green=REL-W, blue=combination.

WORD (W)	SEMANTIC-FEATURE (SF)	RELATED-WORDS (REL-W)
thoughtN	[*consideration1*|*omtanke1*] [*idea1*|*tanke1*] [*thought2*|]	{*consideration1 idea1 thought2*}
workN	[*business2*|*arbeid1*] [*work1*|*forhold1*]	{*business2 work1*}
partyN	[*year2*|*parti1*] [*side2*|*side1*] [*party3*|*selskap1*] [*party3*|*gruppe1*]	{*party3 side2 year2*}
telephoneN	[*conversation2*|*telefonsamtale1*] [*call1*|*telefon1*] [*telephone2*|*telefonnummer1*]	{*call1 conversation2 phone1 telephone2*}
theatreN	[*theatre4*|*teater1*]	{*theatre4*}

A theoretical evaluation of the loss or gain in using Mirrors-derived information

- EXP5: A traditional context words model, but only with those words that are also sense-tagged.
- EXP6: replace the words in EXP5 by SFs
- EXP7: replace the words in EXP6 by REL-Ws.
- EXP8: The quality of the Mirrors senses:

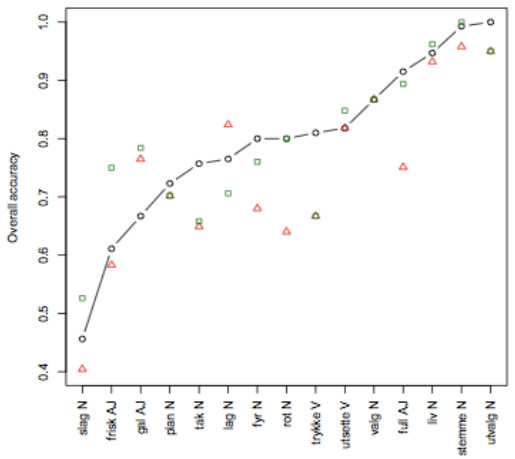


Figure 10.1: Overall Accuracy from Table (10.4) sorted by W. Legend: black=W, red=SF, green=REL-W.

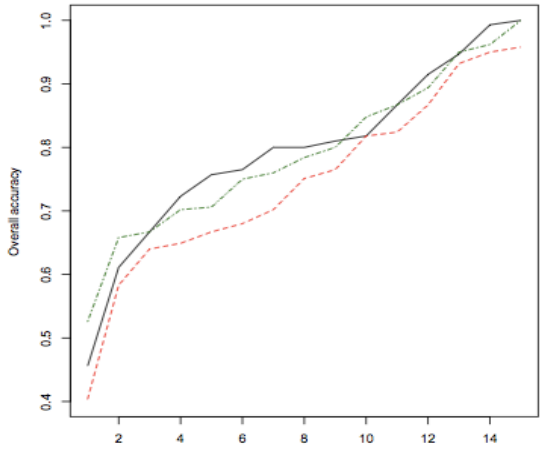


Figure 10.2: Overall Accuracy from Table (10.4) sorted by each model independently. Legend: black=W, red=SF, green=REL-W.

Testing sense distinctions

- The best results are given when using sense-specific information, i.e. when trusting the Mirrors senses that are predicted in the context according to the Mirrors-based automatic sense-tagger.

Conclusion

- Approximately half of the lemmas in the ENPC are sense-tagged automatically.
- The work has shown that poor quality input to the Mirrors is unfortunate, since the method is vulnerable to noise
- Wrt. WSD classification and the hope to improve the results by adding Mirrors-derived knowledge, the missing gain may appear disappointing.
- But wrt. the plausibility of the Mirrors method, the missing difference means that no findings indicate serious drawbacks of the principles underlying the Mirrors method.

Future work

- It is not clear how the Mirrors method would perform with significantly larger data material than the presented use of the ENPC. Testing on an independent, larger sample might shed light on this.

- Experiment with feature selection: (prune away apriori context features that do not co-occur *significantly* with a given word sense)