

Linguistic Knowledge and Word Sense Disambiguation

Tanja Gaustad
Humanities Computing
University of Groningen, The Netherlands
tanja@let.rug.nl
www.let.rug.nl/~tanja

Overview

- Word Sense Disambiguation (WSD)
- Maximum Entropy WSD System
 - * corpus
 - * classification task
 - * linguistic features
- Lemma-based approach
- Results and Evaluation
- Future work

Overview

- Word Sense Disambiguation (WSD)
- Maximum Entropy WSD System
 - * corpus
 - * classification task
 - * linguistic features
- Lemma-based approach
- Results and Evaluation
- Future work

What problem are we talking about?



“Mijn vader **zagen** we niet meer.”

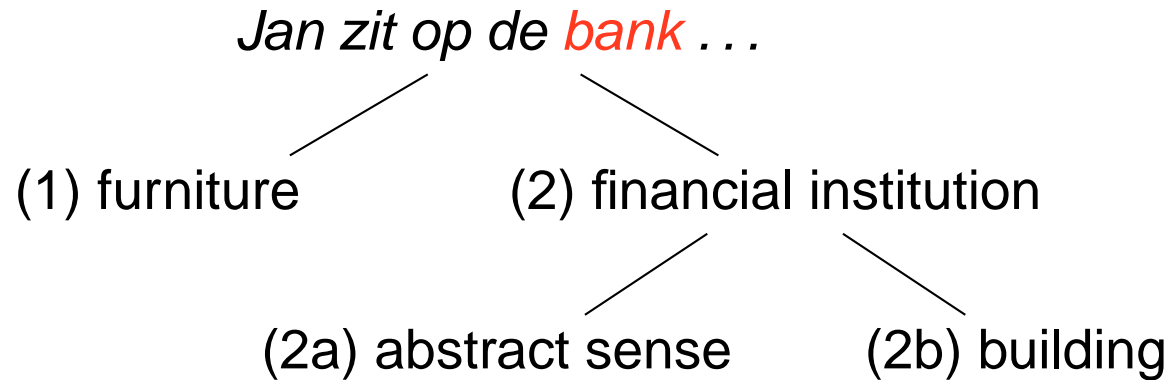
Word Sense Disambiguation

- Lexical semantic ambiguity
 - * is a major problem in Natural Language Processing (NLP)
 - * is largely unsolved
- WSD is the task of attributing the correct sense(s) to words in context
- Ambiguous words in given context need to be resolved for numerous NLP applications, e.g.:
 - * Machine Translation
 - * Information Retrieval
 - * Parsing
 - * Language Understanding

Assigning Senses

- PoS-tagging assigns grammatical categories \Rightarrow level of syntax
- Sense tagging assigns senses to tokens \Rightarrow level of semantics
- Automatic sense tagging: *word sense disambiguation* (WSD)
- Sense-tagged corpus: corpus where one or several words are assigned a semantic tag referring to the meaning of the word in the given context

Example



- ... en kijkt TV. \Rightarrow (1)
- ... en kijkt naar de straat beneden. \Rightarrow (2b)
- ... en geniet de uitzicht. \Rightarrow ambiguous

Sense Inventory

- E.g. dictionary, thesaurus, semantic networks (e.g. WordNet)
- Can differ in number and kind of senses distinguished for each word
- Words are assumed to have a finite number of distinct senses
- Sense inventory language dependent, often also application dependent (all words vs. lexical sample)
- fine- vs. coarse-grained sense distinction: the finer the distinction the more difficult (also for human annotators)

Examples of Sense Inventories

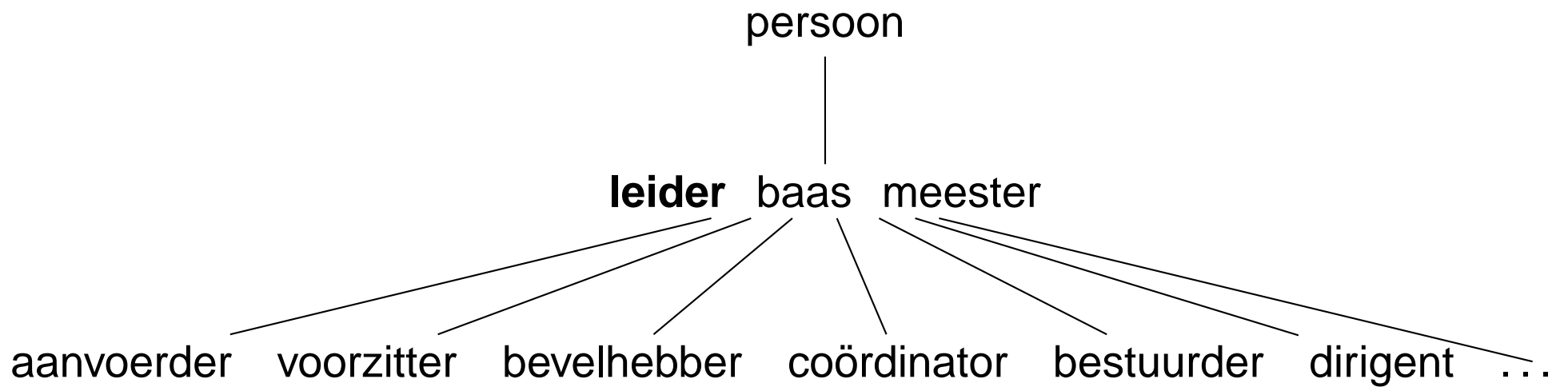
Dictionary (Van Dale Groot Woordenboek Hedendaags Nederlands)

leider

1. iem. die leidt, bestuurt → baas, aanvoerder
2. persoon of ploeg die op de eerste plaats staat in een competitie of wedstrijd
3. paal, stijl, lat enz. die iets in de goede richting houdt

Examples of Sense Inventories

EuroWordNet Dutch (1.5)



Difficulties with Assigning Senses

- The more senses, the more difficult
- Local information is not good enough (\neq PoS tagging)
- Useful cues (can) differ for every ambiguous word
- Inter-annotator agreement is often very low

Word Sense Disambiguation II

- Means to recover the correct sense in a given context:
 - * **distributional information** (frequency)
 - * **collocational information** (context words)
 - * **further related information** (morphology, PoS, syntax)
 - * world knowledge
- WSD system presented here is
 - * for Dutch
 - * combination of statistical classification with linguistic information
 - * corpus-based

Main Research Question

- Which linguistic knowledge sources are most useful for word sense disambiguation?

Overview

- Word Sense Disambiguation (WSD)
- Maximum Entropy WSD System
 - * corpus
 - * classification task
 - * linguistic features
- Lemma-based approach
- Results and Evaluation
- Future work

Corpus

- SENSEVAL-2 corpus for Dutch
 - * training: 120,000 tokens, 98,000 words, 55,000 ambiguous words
 - * testing: 40,000 tokens, 32,000 words, 19,000 ambiguous words
 - * publicly available (www.senseval.org)
- Non-hierarchical sense inventory
- Senses not split according to PoS, e.g.

boog BUIGEN_KROMMEN
 BUIGEN_BUIGINGMAKEN
 BOOG_PIJL

Example from Corpus

een/een_lidwoord
oorverdovende/oorverdovend
donderslag/=
deed/doen_maken_dat
de/=
aarde/aarde_planeet
beven/=
./=
<utt>

Overview

- Word Sense Disambiguation (WSD)
- Maximum Entropy WSD System
 - * corpus
 - * classification task
 - * linguistic features
- Lemma-based approach
- Results and Evaluation
- Future work

WSD as Classification Task

- Problem restated: Use statistical information about senses and context words to build model which correctly predicts word senses
⇒ Classify input (ambiguous words) into correct classes (senses).
- Maximum entropy: technique to estimate probability distributions
- Use features extracted from labeled training data to derive constraints for model
- Constraints characterize class-specific expectations for distribution
- Distribution should maximize entropy **and** model should satisfy constraints imposed by training data

Advantages of Maximum Entropy Classification

- Property functions take into account any information which might be useful for disambiguation
- Dissimilar types of information can be combined into single model for WSD
- No independence assumptions (as in e.g. a Naive Bayes algorithm) necessary

Overview

- Word Sense Disambiguation (WSD)
- Maximum Entropy WSD System
 - * corpus
 - * classification task
 - * linguistic features
- Lemma-based approach
- Results and Evaluation
- Future work

Linguistic Features

- Context
- Morphology (\Rightarrow lemma-based approach)
- PoS
- Dependency relations

Linguistic Features: Context

- *Een oorverdovende donderslag deed de **aarde** beven.*

- Left context:
donderslag
deed
de

- Right context:
beven
.
=

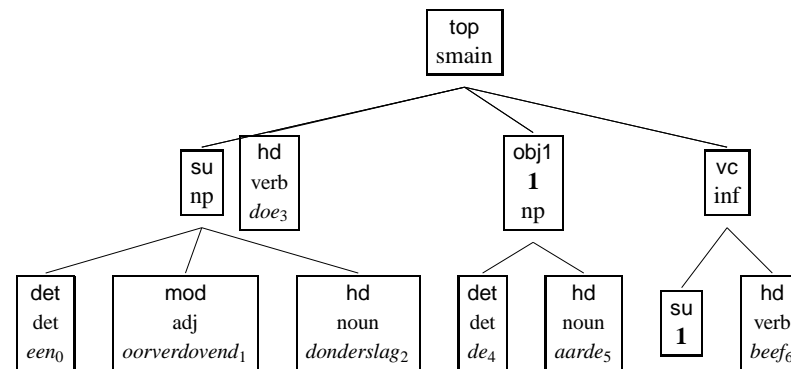
Linguistic Features: Context II

- Context can either be words or lemmas \Rightarrow lemmas work better (more abstraction)
- Tried various context sizes (± 3 , ± 5 , ± 10) \Rightarrow smallest context works best (confirms findings on human subjects)
- Settings: context size of ± 3 words left and right within same sentence
- Clearly outperforms baseline (choosing most frequent sense)

Linguistic Features: PoS

- *Een ART oorverdovende ADJ donderslag N deed V de ART **aarde** N beven V . Punc*
- PoS of ambiguous word
- PoS of context
- Combination of both sources of knowledge significantly increases results

Linguistic Features: Dependency Relations



- *Een oorverdovende donderslag deed de aarde* **DET SU/OBJ1** *beven.*
- Only dependency relations of ambiguous word considered
- Two types of relations distinguished: where **ambiguous word = head** and where **ambiguous word = dependent**

Linguistic Features: Dependency Relations II

- Including dependency relations as only feature already performs better than using context
- Combination of PoS of ambiguous word, context lemmas and dependency relations works well
- Adding PoS of the context degrades performance

Building Classifiers

- Procedure to build classifiers
 - * lemmatize, PoS tag and parse corpus with Alpino
 - * extract all instances for each ambiguous word form
 - * transform instances into feature vectors, e.g.
aarde N donderslag deed de beven . = aarde_planet
 - * build classifier for each ambiguous word form
- Features: ± 3 context lemmas (only within same sentence), PoS, morphological information, syntactic information

Overview

- Word Sense Disambiguation (WSD)
- Maximum Entropy WSD System
 - * corpus
 - * classification task
 - * linguistic features
- Lemma-based approach
- Results and Evaluation
- Future work

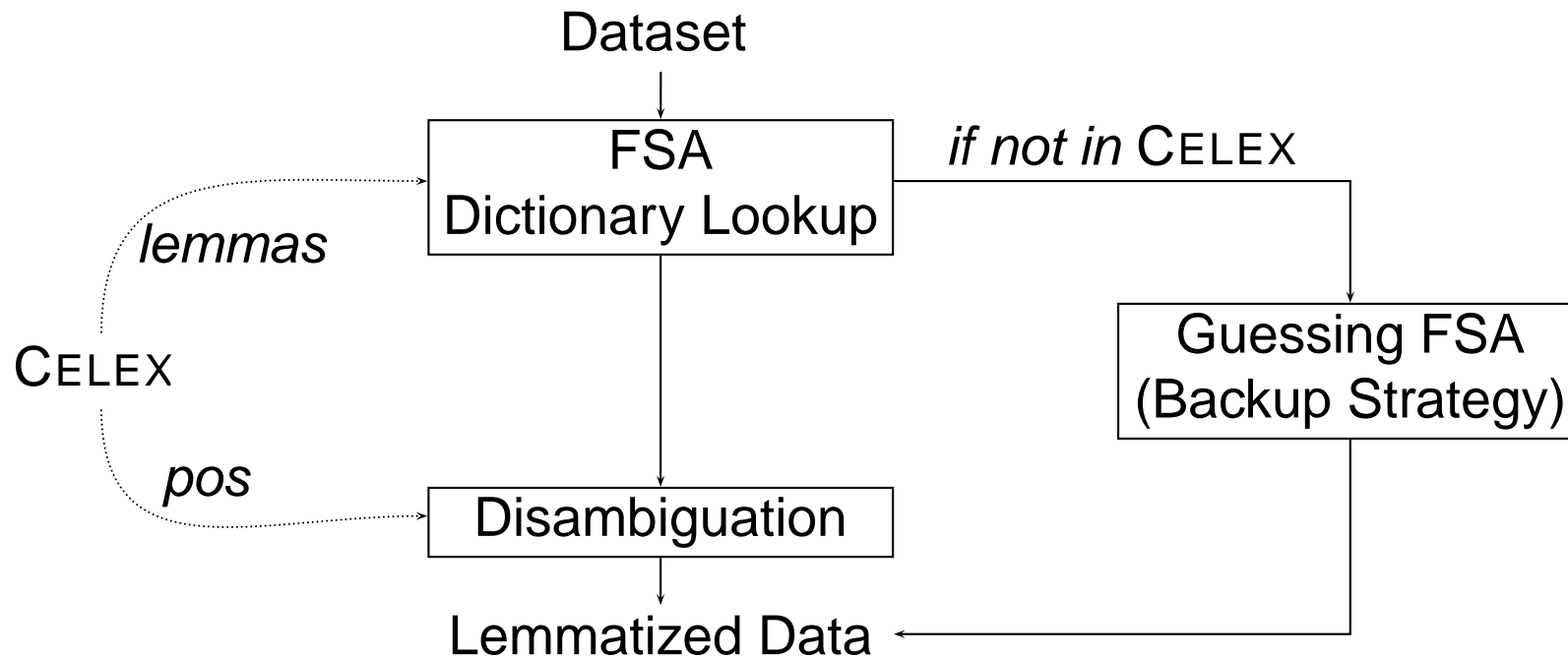
Lemma-Based Approach

- Previous research built a separate classifier for each ambiguous word form, e.g. *voet* ('foot') and *voeten* ('feet')
- Lemma-based approach builds a separate classifier for each ambiguous **lemma**, e.g. *voet* subsumes *voet* and *voeten*
- Advantage: All inflected forms are clustered together
⇒ the more inflection in a language, the more lemmatization will compress and generalize the data
- Higher accuracy expected with lemma-based approach

Dictionary-Based Lemmatizer for Dutch

- Corpora contain many different, often infrequent words
- Lemmatizer reduces all inflected forms of a word to their lemma
- Consequently, # of different lemmas $<$ # of different word forms
⇒ more reliable estimation of probabilities
- Accurate and fast lemmatizer is a prerequisite for lemma-based approach to work
- Combination of lexical database (CELEX) and finite-state automata

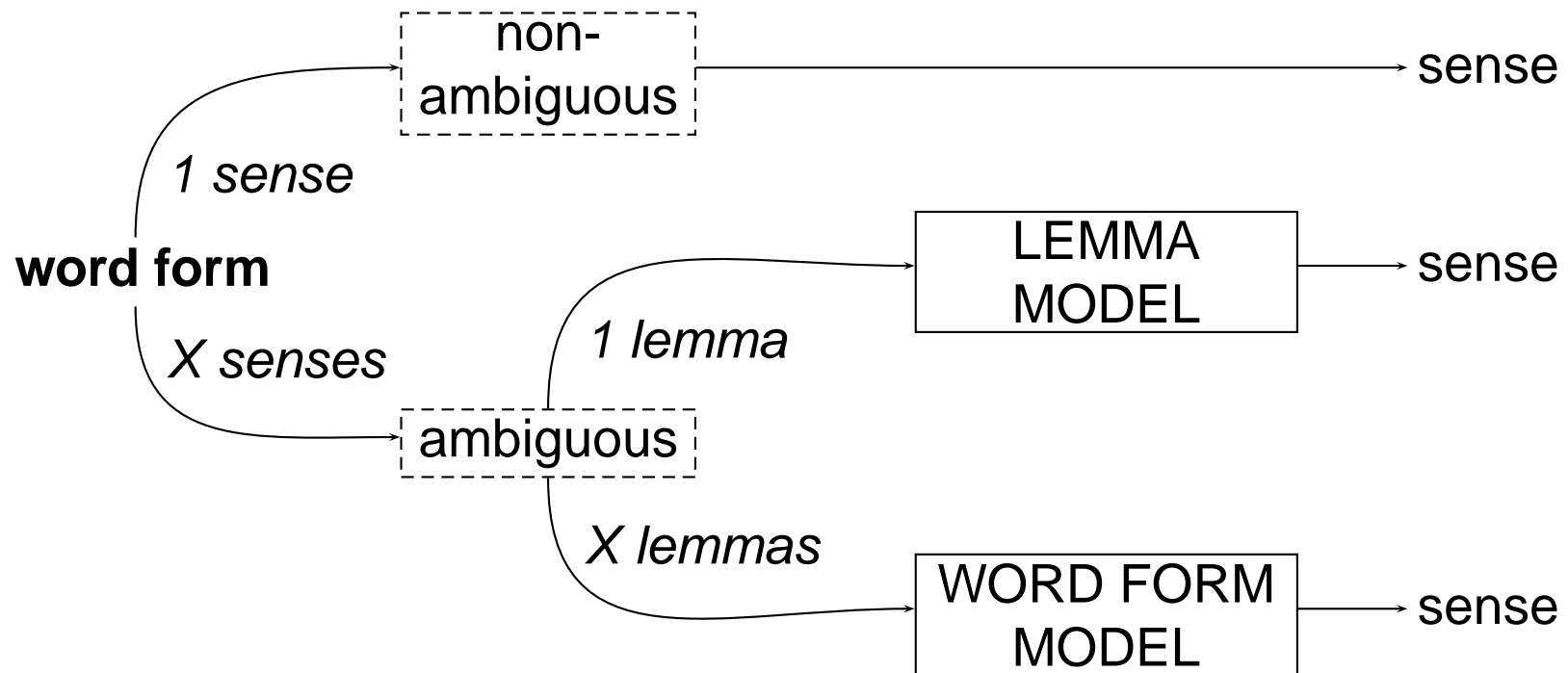
Dictionary-Based Lemmatizer for Dutch II



Lemma-Based Approach II

- Constructing classifiers based on lemmas, not word forms
⇒ reduces number of classifiers
- Lemmas produce more concise and generic evidence than inflected forms (already noted by Yarowsky (1994))
⇒ more training data available per classifier
- E.g. all instances of one verb are clustered in a single classifier instead of several (one for each inflected form found in the data)
- Remember: Dutch SENSEVAL-2 Data is ambiguous with regard to meaning **and** part-of-speech (PoS)

Schematic Overview of Lemma-Based Approach



Overview

- Word Sense Disambiguation (WSD)
- Maximum Entropy WSD System
 - * corpus
 - * classification task
 - * linguistic features
- Lemma-based approach
- Results and Evaluation
- Future work

Results with Word Form and Lemma-Based Approach

Model	Accuracy	# classifiers
baseline all ambiguous words	78.47%	953
word form classifiers	83.66%	953
lemma-based classifiers	84.15%	669

- Settings: PoS of ambiguous word and context, context lemmas
- Baseline: choose most frequent sense for each ambiguous word
- Comparison of word form-based and lemma-based approach
- Lemma-based approach works significantly better

Number of Classifiers Used During Testing

	lemma-based	word forms
unique ambiguous word forms classifiers used	512	512
based on word forms	230	410
based on lemmas	70	0
word forms subsumed	208	0
word forms seen 1st time	74	102

- Less classifiers need to be built with lemma-based approach
⇒ more material per classifier
- More ambiguous words are treated with lemma-based approach

Detailed Comparison of Results

Model	Accuracy
baseline	76.77%
word form classifiers	78.66%
lemma-based classifiers	80.39%

- Comparison of word form-based and lemma-based approach for word forms with different classifiers only
- Clear gain from lemmatization
 - ⇒ error rate reduction 8%
 - ⇒ fewer classifiers, smaller system
 - ⇒ more word forms classified

Comparison of Different WSD Systems

	ambiguous	all
baseline test data	78.5%	89.4%
word form classifiers	83.7%	92.4%
lemma-based classifiers	84.1%	92.5%
Hendrickx et al. 2002	84.0%	92.5%

- MBL system (Hendrickx et al. 2002) uses
 - * extensive parameter optimization per classifier
 - * frequency threshold of min. 10 training instances
(frequency baseline used for words below threshold)
- Lemma-based system scores same without extensive “per classifier” parameter optimization (better results may be possible)

Comparison of Different WSD Systems: The Impact of Deep Syntactic Information

	ambiguous	all
baseline test data	78.5%	89.4%
word form classifiers	83.7%	92.4%
incl. syntactic information	84.8%	92.8%
lemma-based classifiers	84.1%	92.5%
incl. syntactic information	85.7%	93.4%
Hendrickx et al. 2002	84.0%	92.5%

Evaluation and Conclusion

- System using lemma-based approach
 - * is smaller
 - * is more robust
 - * has higher accuracy (best results to date)
- Compared to earlier results for WSD of Dutch, lemma-based approach performs the same involving less work
- Addition of deep linguistic knowledge (in the form of dependency relations) works best

Overall Conclusion

- Not a single best source of linguistic knowledge, but **combination** of linguistic features yields best results
- Especially deep linguistic information (in the form of dependency relations) contains important cues for disambiguation
- WSD system presented here is state-of-the-art for Dutch WSD

Future Work

- Include semantic information
 - * Topic information
 - * Domain information
 - * Co-occurrence information
 - * Selectional restrictions
- Acquire more data semi-automatically using EuroWordNet
- Apply to other languages
- Test current system in concrete applications