

Semantic Tagging and Sense-Tagged Corpora

Tanja Gaustad

15 May 2002

Sense Tagging

- POS-tagging assigns grammatical categories ⇒ level of syntax
- Sense tagging assigns lexical senses, “meaning” to tokens in a corpus ⇒ 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

Sense Tagging

Jan zit op de *bank* ...

(1) furniture

(2) financial institution

(2a) abstract sense

(2b) building

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

First Example of a Sense-tagged Corpus

- Senseval 1

800001

Late on Thursday night it was travelling at about three metres a second in wind blowing at 20 to 25 knots when an empty car fell off just as it reached the top.

The **<tag "532675">accident</>** appeared to have little effect on the Christmas party, except to lengthen it considerably.

Second Example of a Sense-tagged Corpus

- SEMCOR

```
<wf pos=RB wnsn=1 lexs=4:02:02::>>Only</wf>
<wf pos=DT>a</wf>
<wf pos=JJ wnsn=1 lexs=3:00:00::>>relative</wf>
<wf pos=NN wnsn=1 lexs=1:23:01::>>handful</wf>
<wf pos=IN>of</wf>
<wf pos=JJ wnsn=1 lexs=5:00:01:specified:00>>such</wf>
<wf pos=NN wnsn=2 lexs=1:10:00::>>reports</wf>
<wf pos=VB ot=notag>was</wf>
<wf pos=VB wnsn=2 lexs=2:30:01::>>received</wf>
<punc>.</punc>
```

Differences

Senseval 1

SEMCOR

- Lexical sample: only chosen ambiguous words are annotated with a sense tag
- All words: all content words are annotated with a sense tag
- Sense tags correspond to senses in dictionary entries
- Sense tags correspond to WordNet nodes
- Context is fixed to two sentences
- Context is rest of corpus

Available Corpora: Lexical Sample

- “interest” Corpus (Bruce/Wiebe, 1994): English, 2,369 occurrences
- DSO Corpus (Ng/Lee, 1996): English, 121 nouns & 70 verbs, 192,000 occurrences
- Senseval 1 (1998): English (HECTOR database), 15 nouns, 13 verbs, 8 adjectives, 5 indeterminates, 8,448 occurrences, dictionary as sense inventory
- Senseval 2 (2001): Various languages (Basque, English, Italian, Japanese, Korean, Spanish, Swedish), varying between 3,900 occurrences (83 words) and 7,567 occurrences (44 words), WordNet as sense inventory

Available Corpora: All Words

- SEMCOR (Miller et al., 1993; Landes et al., 1998): English, part of Brown Corpus and WordNet project, 200,000 words, all content words annotated according to WordNet
- Senseval 2 (2001): Various languages (Czech, Dutch, English, Estonian), 5,000-6,000 occurrences, WordNet or EuroWordNet (where available) as sense inventory

Use(s) of Sense-tagged Corpora

- Sense tags necessary for semantic interpretation
- Training and testing of Machine learning techniques (e.g. WSD)
- Various applications profit from sense tagging:
 - Information retrieval: disambiguation of keywords
 - Machine translation: proper translation of ambiguous words
 - Speech processing: correct phonetisation of homophones
 - Text processing: spelling correction (e.g. diacritics), lexical access of Semitic languages, etc.

Ingredients for Producing a Sense-tagged Corpus

- Corpus
- Sense inventory
- Annotators and annotation tool

Corpus

- Any collection of text
- Can be raw or annotated (POS, syntactic structure, etc.)

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

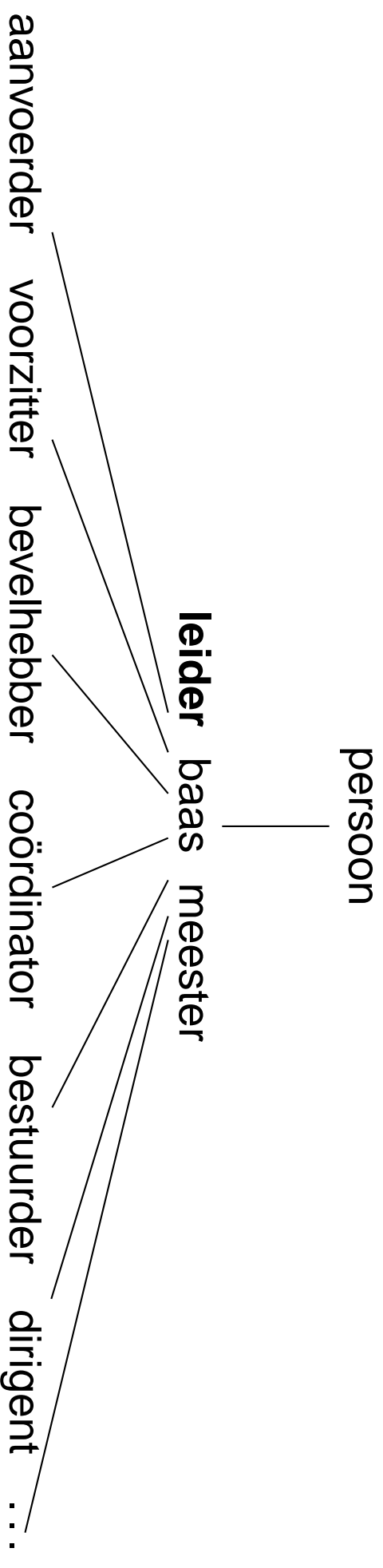
Dictionary (Van Dale Groot Woordenboek Hedendaags Nederlands)

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)



Annotation process

- Hand-annotation \Rightarrow time, cost and expertise-intensive
- Semi-supervised annotation
- Unsupervised annotation \Rightarrow not good enough yet

Inter-Tagger Agreement (ITA)

- Quality measure for (manually) sense-tagged corpora
- Prerequisite: more than one person must have tagged the same text
- ITA = percentage of words assigned same sense by all sense taggers
- Upper bound for WSD systems

Difficulties with Sense-Tagging

- The more senses, the more difficult
- More possible sense tags than POS tags
- 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 (WSD)

- Problem: Lexical semantic ambiguity
- Goal: Recover correct sense in a given context
- Means:
 - Distributional information (frequency)
 - Collocational information (context words)
 - Further related information (morphology, syntax, topic)
 - World knowledge
- Approach: Combine statistics, corpus and linguistics

WSD: Example “accident”

crash (Sense 1): Unfortunate or disastrous incident not caused deliberately; a mishap causing injury or damage; in particular, a crash involving road vehicles.

*Fears that fog could cause a serious **accident** on the M40 have united members of the District Council.*

chance (Sense 2): Something that happens without apparent or deliberate cause; a chance event of set of circumstances.

*We planned the first two children, but our third was an **accident**.*

WSD: Example “accident”

- Frequency: *crash* 0.82 *chance* 0.18
- Context words: *car* *good cue for crash*
 happy *good cue for chance*
 the *no cue*
- Syntax: “. . . our third was an accident.”
 - Construction not likely with *crash*
 - Same words could occur with *crash*
 - Syntax can be a good cue

WSD as Classification Problem

- Problem restated: Use statistical information about senses and contextwords to build model which correctly predicts word senses
 - ⇒ Classify input (ambiguous words) into correct classes (senses).
- Algorithm: e.g.
 - Naive Bayes
 - Maximum Entropy

WSD as Classification Problem

- Binary classification task in the case of “accident”:
class 1 \Rightarrow sense 1
class 2 \Rightarrow sense 2
- Baseline: depends in sense inventory, but usually a lot lower than with POS tagging

More than one word ...

- So far: One word \leftrightarrow one sense
- *Nu ben jij aan de beurt.*
- *Ik wil graag op de hoogte blijven.*
- *Hij komt de klap te boven.*
 - \Rightarrow a unit of more than one word can have one sense
 - word₁ word₂ word₃ \Rightarrow one sense
- Collocations not only difficult for sense tagging, but in general