Empirical Methods in Natural Language Processing
Lecture 11
Word Sense Disambiguation

Philipp Koehn

11 February 2008

School of informatics

Word Senses

- Some words have multiple **meanings**
- This is called **Polysemy**
- Example: *bank*
  - financial institution: *I put my money in the bank.*
  - river shore: *He rested at the bank of the river.*
- How could a computer tell these senses apart?

Homonym

- Sometimes two completely different words are spelled the same
- This is called a **Homonym**
- Example: *can*
  - modal verb: *You can do it!*
  - container: *She bought a can of soda.*
- Distinction between *Polysemy* and *Homonymy* not always clear

How many senses?

- How many senses does the word *interest* have?
  - *She pays 3%* **interest** *on the loan.*
  - *He showed a lot of* **interest** *in the painting.*
  - *Microsoft purchased a controlling* **interest** *in Google.*
  - *It is in the national* **interest** *to invade the Bahamas.*
  - *I only have your best* **interest** *in mind.*
  - *Playing chess is one of my* **interests**.
  - *Business* **interests** *lobbied for the legislation.*

- Are these seven different senses? Four? Three?

Wordnet

- One way to define senses is to look them up in **Wordnet**, a hierarchical database of senses

- According to Wordnet, *interest* has 7 senses:
  - Sense 1: *a sense of concern with and curiosity about someone or something*, Synonym: *involvement*
  - Sense 2: *the power of attracting or holding one's interest (because it is unusual or exciting etc.)*, Synonym: *interestingness*
  - Sense 3: *a reason for wanting something done*, Synonym: *sake*
  - Sense 4: *a fixed charge for borrowing money; usually a percentage of the amount borrowed*
  - Sense 5: *a diversion that occupies one's time and thoughts (usually*

pleasantly), Synonyms: *pastime, pursuit*
  - Sense 6: *a right or legal share of something; a financial involvement with something*, Synonym: *stake*
  - Sense 7: *(usually plural) a social group whose members control some field of activity and who have common aims*, Synonym: *interest group*

- Organization of Wordnet
  - Wordnet groups words into synsets.
  - polysemous words are part of multiple synsets
  - synsets are organized into a hierarchical structure of **is-a** relationships, e.g. a *dog is-a pet*, *pet is-a animal*

- Is Wordnet too fine grained?

Different sense = different translation

- Another way to define senses: if occurrences of the word have different translations, these indicate different sense

- Example *interest* translated into German
  - *Zins*: financial charge paid for load (Wordnet sense 4)
  - *Anteil*: stake in a company (Wordnet sense 6)
  - *Interesse*: all other senses

Languages differ

- Foreign language may make finer distinctions
- Translations of *river* into French
  - *fleuve*: river that flows into the sea
  - *rivière*: smaller river
- English may make finer distinctions than a foreign language
- Translations of German *Sicherheit* into English
  - *security*
  - *safety*
  - *confidence*

One last word on senses

- A lot of research in word sense disambiguation is focused on polysemous words with clearly distinct meanings, e.g. *bank*, *plant*, *bat*, ...

- Often meanings are close and hard to tell apart, e.g. *area*, *field*, *domain*, *part*, *member*, ...

  - *She is a part of the team.*
  - *She is a member of the team.*
  - *The wheel is a part of the car.*
  - * *The wheel is a member of the car.*

Word sense disambiguation (WSD)

- For many applications, we would like to disambiguate senses
  - we may be only interested in one sense
  - searching for *chemical plant* on the web, we do not want to know about chemicals in bananas
- Task: Given a polysemous word, find the sense in a given *context*
- Popular topic, data driven methods perform well

WSD as supervised learning problem

- Words can be labeled with their senses
  - *She pays 3%* **interest/INTEREST-MONEY** *on the loan.*
  - *He showed a lot of* **interest/INTEREST-CURIOSITY** *in the painting.*

- Similar to *tagging*
  - given a corpus tagged with senses
  - define features that indicate one sense over another
  - learn a model that predicts the correct sense given the features

- We can apply similar supervised learning methods
  - **Naive Bayes**, related to *HMM*
  - *Transformation-based learning*
  - *Maximum entropy learning*

Simple features

- Directly neighboring words
  - **plant** *life*
  - *manufacturing* **plant**
  - *assembly* **plant**
  - **plant** *closure*
  - **plant** *species*

- Any content words in a 10 word window (also larger windows)
  - *animal*
  - *equipment*
  - *employee*
  - *automatic*

More features

- Syntactically related words
- Syntactic role in sense
- Topic of the text
- Part-of-speech tag, surrounding part-of-speech tags

Training data for supervised WSD

- **SENSEVAL** competition
  - bi-annual competition on WSD
  - provides annotated corpora in many languages
- Pseudo-words
  - create artificial corpus by artificially conflate words
  - example: replace all occurrences of *banana* and *door* with *banana-door*
- Multi-lingual parallel corpora
  - translated texts aligned at the sentence level
  - translation indicates sense

Naive Bayes

- We want to predict the sense S given a set of features F
- First, apply the Bayes rule

$$\text{argmax}_S p(S|F) = \text{argmax}_S p(F|S)p(F) \quad (1)$$

- Then, decompose $p(F)$ by assuming all features are independent (that's *naive*!)

$$p(F) = \prod_{f_i \in F} p(f_i|S) \quad (2)$$

- The *prior* $p(S)$ and the conditional *posterior* probabilities $p(f_i|S)$ can be learned by maximum likelihood estimation

Decision list

- Yarowsky [1994] uses a **decision list** for WSD
  - two senses per word
  - rules of the form: collocation $\rightarrow$ sense
  - example: *manufacturing* **plant** $\rightarrow$ *PLANT-FACTORY*
  - rules are ordered, most reliable rules first
  - when classifying a test example, step through the list, make decision on first rule that applies

- Learning: rules are ordered by

$$\log\left(\frac{p(sense_A|collocation_i)}{p(sense_B|collocation_i)}\right) \tag{3}$$

  Smoothing is important

Bootstrapping

- Yarowsky [1995] presents **bootstrapping** method

  1. label a few examples
  2. learn a decision list
  3. apply decision list to unlabeled examples, thus labeling them
  4. add newly labeled examples to training set
  5. go to step 2, until no more examples can be labeled

- Initial starting point could also be
  - a short decision list
  - words from dictionary definition

One sense per discourse

- Rules encode the principle:
  One sense per collocation

- Bootstrapping method also uses important principle:
  One sense per discourse
  - in one discourse only one sense of a polysemous word appears
  - text talks either about *PLANT-FACTORY* or *PLANT-LIVING*

- Improved bootstrapping method
  - after labeling examples, one sense per discourse principle is enforced
  - all examples in one document are labeled with the same sense
  - or, examples that are not in the majority sense are un-labeled