Word Sense Disambiguation

(M&S Ch 7)

Overview of the Problem

- ***Problem:*** many words have different meanings or senses ==> there is ambiguity about how they are to be interpreted.
- ***Task:*** to determine which of the senses of an ambiguous word is invoked in a particular use of the word. This is done by looking at the context of the word's use.
- ***Note:*** more often than not the different senses of a word are closely related.

Overview of our Discussion

- ***Methodology***
- ***Supervised Disambiguation***: based on a labeled training set.
- ***Dictionary-Based Disambiguation***: based on lexical resources such as dictionaries and thesauri.
- ***Unsupervised Disambiguation***: based on unlabeled corpora.

Methodological Preliminaries

- ***Supervised versus Unsupervised Learning***: in supervised learning the sense label of a word occurrence is known. In unsupervised learning, it is not known.
- ***Pseudowords***: used to generate artificial evaluation data for comparison and improvements of text-processing algorithms.
- ***Upper and Lower Bounds on Performance***: used to find out how well an algorithm performs relative to the difficulty of the task.

Supervised Disambiguation

- ***Training set***: exemplars where each occurrence of the ambiguous word w is annotated with a semantic label ==> Classification problem.
- ***Approaches***:
  - Bayesian Classification: the context of occurrence is treated as a bag of words without structure, but it integrates information from many words.
  - Information Theory: only looks at informative features in the context. These features may be sensitive to text structure.
  - There are many more approaches (see Chapter 16 and the Senseval competition).

Supervised Disambiguation: Bayesian Classification (I)

- ***(Gale et al, 1992)'s idea:*** to look at the words around an ambiguous word in a large context window. Each content word contributes potentially useful information about which sense of the ambiguous word is likely to be used with it. The classifier does no feature selection. Instead, it combines the evidence from all features.
- ***Bayes decision rule***: Decide s' if $P(s'|C) > P(s_k|C)$ for $s_k \neq s'$.
- $P(s_k|C)$ is computed by Bayes' Rule.

Supervised Disambiguation: Bayesian Classification (II)

- ***Naïve Bayes assumption***:

$P(C|s_k) = P(\{v_j| v_j \text{ in } C\}| s_k) = \Pi_{vj \text{ in } C} P(v_j | s_k)$

- The Naïve Bayes assumption is incorrect in the context of text processing, but it is useful.
- ***Decision Rule for Naïve Bayes***: Decide s’ if $s' = \text{argmax}_{sk} [\log P(s_k) + \Sigma_{vj \text{ in } C} \log P(v_j |s_k)]$
- $P(v_j |s_k)$ and $P(s_k)$ are computed via Maximum-Likelihood Estimation, perhaps with appropriate smoothing, from the labeled training corpus.

Supervised Disambiguation: An Information-Theoretic Approach

- ***(Brown et al., 1991)'s idea:*** to find a single contextual feature that reliably indicates which sense of the ambiguous word is being used.
- The ***Flip-Flop*** algorithm is used to disambiguate between the different senses of a word using the mutual information as a measure.
- $I(X;Y)=\Sigma_{x\in X}\Sigma_{y\in Y}p(x,y) \log p(x,y)/(p(x)p(y))$
- The algorithm works by searching for a partition of senses that maximizes the mutual information. The algorithm stops when the increase becomes insignificant.

Dictionary-Based Disambiguation: Overview

- We will be looking at three different methods:
  - Disambiguation based on sense definitions
  - Thesaurus-Based Disambiguation
  - Disambiguation based on translations in a second-language corpus
- Also, we will show how a careful examination of the distributional properties of senses can lead to significant improvements in disambiguation.

Disambiguation based on sense definitions

- ***(Lesk, 1986)'s idea***: a word's dictionary definitions are likely to be good indicators for the sense they define.
- Express the dictionary sub-definitions of the ambiguous word as sets of bag-of-words and the words occurring in the context of the ambiguous word as single bags-of-words emanating from its dictionary definitions (all pooled together).
- Disambiguate the ambiguous word by choosing the sub-definition of the ambiguous word that has the greatest overlap with the words occurring in its context.

Thesaurus-Based Disambiguation

- ***Idea:*** the semantic categories of the words in a context determine the semantic category of the context as a whole. This category, in turn, determines which word senses are used.
- ***(Walker, 87):*** each word is assigned one or more subject codes which corresponds to its different meanings. For each subject code, we count the number of words (from the context) having the same subject code. We select the subject code corresponding to the highest count.
- ***(Yarowski, 92):*** adapted the algorithm for words that do not occur in the thesaurus but that are very informative. E.g., Navratilova --> Sports

Disambiguation based on translations in a second-language corpus

- ***(Dagan & Itai, 91)'s idea***: words can be disambiguated by looking at how they are translated in other languages.
- Example: the word "interest" has two translations in German: 1) "Beteiligung" (legal share--50% a interest in the company) 2) "Interesse" (attention, concern--her interest in Mathematics).
- To disambiguate the word "interest", we identify the sentence it occurs in, search a German corpus for instances of the phrase, and assign the meaning associated with the German use of the word in that phrase.

One sense per discourse, one sense per collocation

- ***(Yarowsky, 1995)'s Idea***: there are constraints between different occurrences of an ambiguous word within a corpus that can be exploited for disambiguation:
  - ***One sense per discourse***: The sense of a target word is highly consistent within any given document.
  - ***One sense per collocation***: nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship.

Unsupervised Disambiguation

- ***Idea:*** disambiguate word senses without having recourse to supporting tools such as dictionaries and thesauri and in the absence of labeled text. Simply cluster the contexts of an ambiguous word into a number of groups and discriminate between these groups without labeling them.
- ***(Schütze, 1998):*** The probabilistic model is the same Bayesian model as the one used for supervised classification, but the $P(v_j \mid s_k)$ are estimated using the EM algorithm.