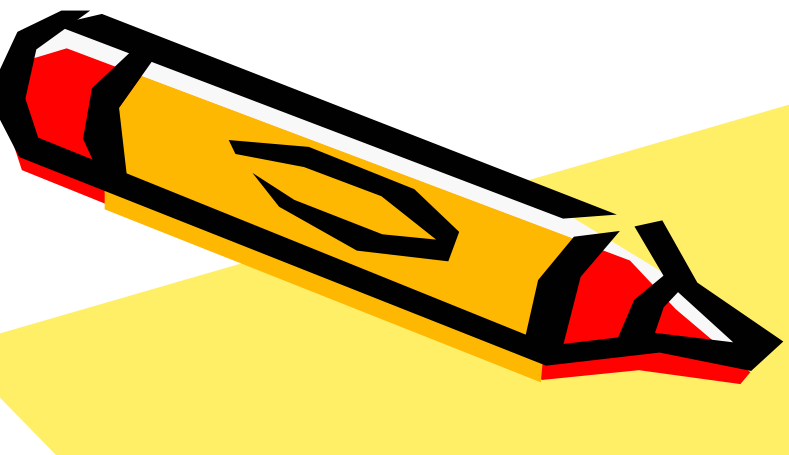
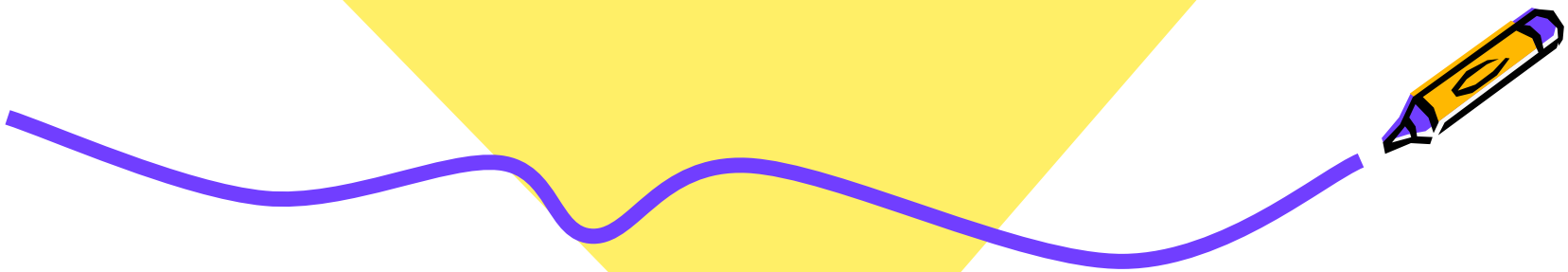


Word Sense Disambiguation

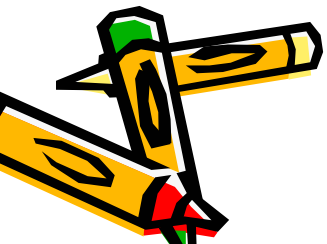
Presented by Jen-Wei Kuo



Reference

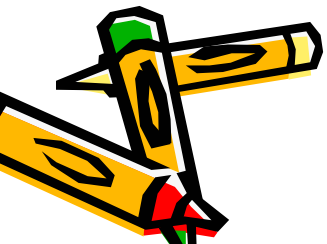
Foundations of Statistical Natural Language Processing,
Chapter 7,
Word Sense Disambiguation

Speech and Language Processing,
Chapter 17.1~17.2,
Word Sense Disambiguation and Information Retrieval



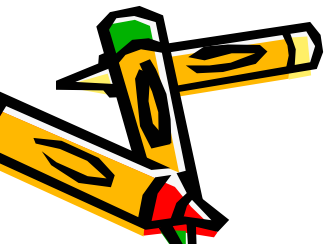
Outline

- Problem
- Task
- Methodological Preliminaries
  - Supervised versus Unsupervised Learning
  - Pseudowords
  - Upper and Lower Bounds on Performance



Outline (cont.)

- Method
  - Supervised Disambiguation
    - Bayesian Classification.
    - An Information-Theoretic Approach.
  - Dictionary-Based Disambiguation
    - Based on Senses Definition
    - Thesaurus-Based Disambiguation
    - Based on Translations in a second-language corpus.
    - One sense per discourse, one sense per collocation.
  - Unsupervised Disambiguation



Problem

- Many words have several meanings or senses.
- There is thus *ambiguity* about how they are to be interpreted.(不同的解釋方式→ambiguity)

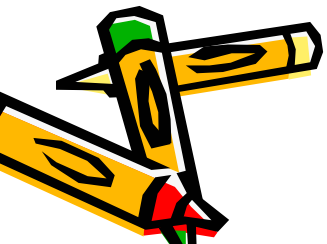
However, the senses are not aloways so well-defined

- For Example：bank
  - The rising ground bordering a lake, river, or sea...(邊坡)
  - As establishment for the custody(保管), loan exchange, or issue of money, for the extension of credit, and for facilitating the transmission of funds.(銀行)

Task

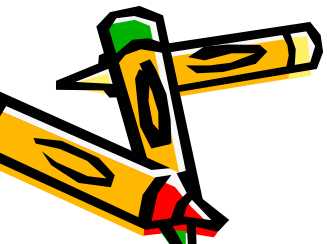
- To determine which of the senses of an ambiguous word is invoked in a particular use of the word.(字義和用法有關)

- How to do：
  - A word is assumed to have a finite number of discrete senses.
  - Look at the context of the word's use.
  - But often the different senses of a word are closely related.



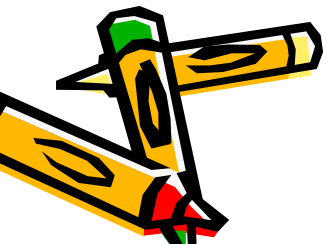
Methodological Preliminaries

- Supervised versus Unsupervised Learing
  - Supervised：
    - Classification task.
    - The sense label of a word is known.
  - Unsupervised：
    - Clustering task.
    - The sense label of a word is unknown.



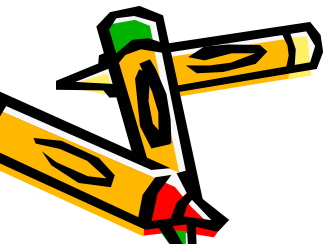
Methodological Preliminaries

- Pseudowords
  - Used to generate artificial evaluation data for comparison and improvements of text-processing alogorithms.
  - Make pseudowords by conflating two or more natural words.
  - For example：Occurrences of banana and door can be replaced by banana-door.
  - The disambiguation algorithm can now be tested on this data to disambiguate the pseudowords.
  - For example：Banana-door into banana and door.



Methodological Preliminaries

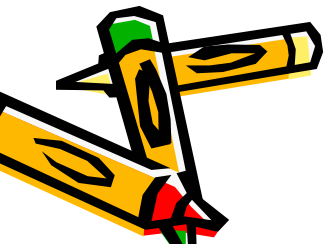
- Upper and Lower Bounds on Performance
  - Used to find out how well an algorithm performs relative to the difficulty of the task.
  - Upper Bounds：
    - Human performance.
  - Lower Bounds：
    - Performance of the simplest (baseline) model.



Method

Supervised Disambiguation

- Training Corpus
  - Each occurrence of the ambiguous word w is annotated with a semantic label (its contextually appropriate sense S_k).
  - Classification problems.
- Approaches
  - Bayesian Classification (Gale et al. 1992)
  - Information Theory (Brown et al. 1991)



Method

Supervised Disambiguation

- Bayesian Classification
  - Bayes Decision Rule：Decide s' if

$$P(s'|c) > P(s_k|c) \quad \text{for } s_k \neq s'$$

  - Look at the words around an ambiguous word in a large context window.
  - Each context 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 to choose the class with highest conditional probability.

Method

Supervised Disambiguation

- Bayesian Classification
  - We want to assign the ambiguous word w to the sense s', given context c, where

$$s' = \arg\max_{s_k} P(s_k \mid c)$$

Baye's Rule

$$= \arg\max_{s_k} \frac{P(c \mid s_k)}{P(c)} P(s_k)$$

$$= \arg\max_{s_k} P(c \mid s_k) P(s_k)$$

log

$$= \arg\max_{s_k} [\log P(c \mid s_k) + \log P(s_k)]$$

Method

Supervised Disambiguation

- Bayesian Classification
  - Naive Bayes Assumption：

The attributes (contextual words) used for description are all conditionally independent.

$$P(c \mid s_k) = P(\{v_j \mid v_j \text{ in } c\} \mid s_k) = \prod_{v_j \text{in } c} P(v_j \mid s_k)$$

  - Consequences of this assumption：
    - ***Bag of Words Model***：The structure and linear ordering of words within the context is ignored.
    - The presence of one word in the bag is ***independent*** of another.

Method

Supervised Disambiguation

- Bayesian Classification
  - Decide s' if

$$s' = \arg\max_{s_k} [\log P(s_k) + \sum_{v_j \text{ in } c} \log P(v_j \mid s_k)]$$

  - $P(v_j \mid s_k)$ and $P(s_k)$ are computed from the labeled training corpus, perhaps with appropriate smoothing.

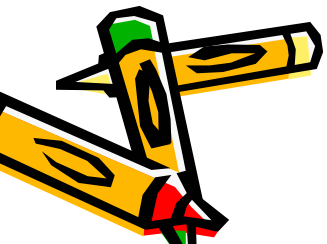
$$P(v_j \mid s_k) = \frac{C(v_j, s_k)}{C(s_k)} \qquad P(s_k) = \frac{C(s_k)}{C(w)}$$

where $C(v_j, s_k)$ is the number of occurrences of v_j in a context of sense s_k in the training corpus, $C(s_k)$ is the number of occurrences of s_k in the training corpus, $C(w)$ is the total number of occurrences of the ambiguous word w.

Method

Supervised Disambiguation

- Information Theoretic Approach
  - **Bayes Classifier** uses information from all words in the context window by using an independence assumption.
  - In the **Information Theoretic Approach** we try to find a single contextual feature that reliably indicates which sense of the ambiguous word is being used.



Method

Supervised Disambiguation

- Information Theoretic Approach
  - **Bayes Classifier** uses information from all words in the context window by using an independence assumption.
  - In the **Information Theoretic Approach** we try to find a single contextual feature that reliably indicates which sense of the ambiguous word is being used.

Ambiguous word	Indicator	Examples: value → sense
prendre	object	*mesure* → *to take* *décision* → *to make*
vouloir	tense	present → *to want* conditional → *to like*
cent	word to the left	*per* → % number → *c.* [money]

Table 7.3 Highly informative indicators for three ambiguous French words.

Method

Supervised Disambiguation

- Information Theoretic Approach
  - Two senses of the word：prendre
    - **Prendre** une measure → **take** a measure
    - **Prendre** une decision → **make** a decision
  - The translations of the ambiguous word $\{t_1,...,t_m\}$ are {take,make} ← meaning
  - The possible indicator words $\{x_1,...,x_m\}$ are {mesure,note,exemple,decision,parole} ← indicate the meaning
  - Find a partition $Q=\{Q_1,Q_2\}$ of $\{x_1,...,x_m\}$ and $P=\{P_1,P_2\}$ of $\{t_1,...,t_m\}$ that maximizes the mutual information：

$$I(P;Q)=\sum_{t\in P}\sum_{x\in Q}p(t,x)\log\frac{p(t,x)}{p(t)p(x)}$$

Method

Supervised Disambiguation

- Information Theoretic Approach

Flip-Flop Algorithm :

find a random partition $P=\{P_1,P_2\}$ for $\{t_1,\ldots, t_m\}$

while (improving) do

find partition $Q=\{Q_1, Q_2\}$ of $\{x_1,\ldots,x_n\}$

that maximizes $I(P;Q)$

find partition $P=\{P_1, P_2\}$ of $\{t_1,\ldots, t_m\}$

that maximizes $I(P;Q)$

end

Method

Supervised Disambiguation

- Information Theoretic Approach
  - Disambiguation：
    - For the occurrence of the ambiguous word, determine the value x_i, of the indicator.
    - If x_i is in Q_1, assign the occurrence to sense 1, if x_i is in Q_2, assign the occurrence to sense 2.

Method

Dictionary-Based Disambiguation

- Concept：
  - Sense definitions are extracted from existing sources such as dictionaries and thesaurus.
- Approaches：
  - Based on Sense Definitions. (Lesk,1986)
  - Thesaurus-Based Disambiguation. (Walker,1987) (Yarowsky, 1992)
  - Based on Translations (Dagan et al. 1991&1994)
  - One Sense per Discourse, One Sense per Collocation (Yarowsky, 1995)

Method

Dictionary-Based Disambiguation

- Disambiguation Based on Sense Definition：
  - A word's dictionary definitions are likely to be good indicators of the senses 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(散發) 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.

Method

Dictionary-Based Disambiguation

- Disambiguation Based on Sense Definition：
  - The algorithm:

Given a context c for a word w

For all senses $s_1,\ldots,s_k$ of w do

score (S_k) =

overlap (word set of dictionary definition of sense S_k,

word set of dictionary definition of V_j in context c)

Choose the sense with highest score.

Method

Dictionary-Based Disambiguation

- Disambiguation Based on Sense Definition：
  - Example (Two Senses of *ash*):

Senses		Definition
S_1	tree	a tree of the olive family
S_2	burned stuff	the solid residue left when combustible material is burned

Score		Context
S_1	S_2	
0	1	This cigar burns slowly and creates a stiff ash
1	0	The ash is one of the last trees to com into leaf.

Method

Dictionary-Based Disambiguation

- Thesaurus-Based Disambiguation：
  - This exploits the semantic categorization provided by a thesaurus like Roget's.
  - The semantic categories of the words in a context determine the semantic category of the context as a whole. And this category in turn determines which word senses are used.
  - (Walker,1987)：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.

Method

Dictionary-Based Disambiguation

- Thesaurus-Based Disambiguation：
  - The algorithm:
    - Given a context c for a word w with senses $s_1,\ldots,s_k$.
    - Find the bags of words corresponding to each sense s_k in the dictionary (s_k bags of words).
    - Compare with the bag of words formed by combining the context word definitions. Pick the sense which gives maximum overlap with this bag.

Method

Dictionary-Based Disambiguation

- Thesaurus-Based Disambiguation：
  - (Yarowsky,1992)：Add new words to a category if they occur more often than chance. For example Navratilova can be added to the sports category.
  - The Bayes classifier is used for both adaptation and disambiguation.
  - Adapted the algorithm for words that do not occur in the thesaurus but that are very informative.
    E.g., Navratilova --> Sports

Method

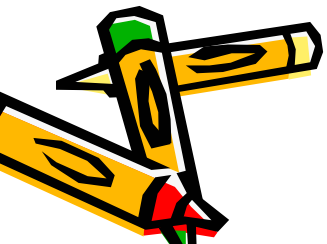
Dictionary-Based Disambiguation

- Disambiguation based on translations in a second-language corpus： (Dagan et al. 1991&1994)
  - 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.
  - Disambiguate words based on translations.
  - Count the number of times a sense translation occurs in a second language corpus along with translations of the context words. Pick the sense with the highest score.

Method

Dictionary-Based Disambiguation

- One Sense per Discourse, One Sense per Collocation：(Yarowsky, 1995)
  - 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.

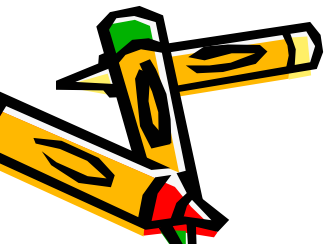


Method

Unsupervised Disambiguation

- (Schutze,1998)
  - Disambiguate word senses without having resourse 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.
  - 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.

?



Method

Unsupervised Disambiguation

- EM algorithm
  - Initialize $p(v_j \mid s_k) \rightarrow$ random
  - Compute likelihood $l(C \mid \mu)$

$$l(C \mid \mu) = \log \prod_{i=1}^{I} \sum_{k=1}^{K} p(c_i \mid s_k) p(s_k) = \sum_{i=1}^{I} \log \sum_{k=1}^{K} p(c_i \mid s_k) p(s_k)$$

  - While $l(C \mid \mu)$ is improving repeat:

$$p(c_i \mid s_k) = \prod_{v_j \in c_i} p(v_j \mid s_k)$$

  - E step :

$$h_{i,k} = \frac{p(c_i \mid s_k)}{\sum_{k=1}^{K} p(c_i \mid s_k)}$$

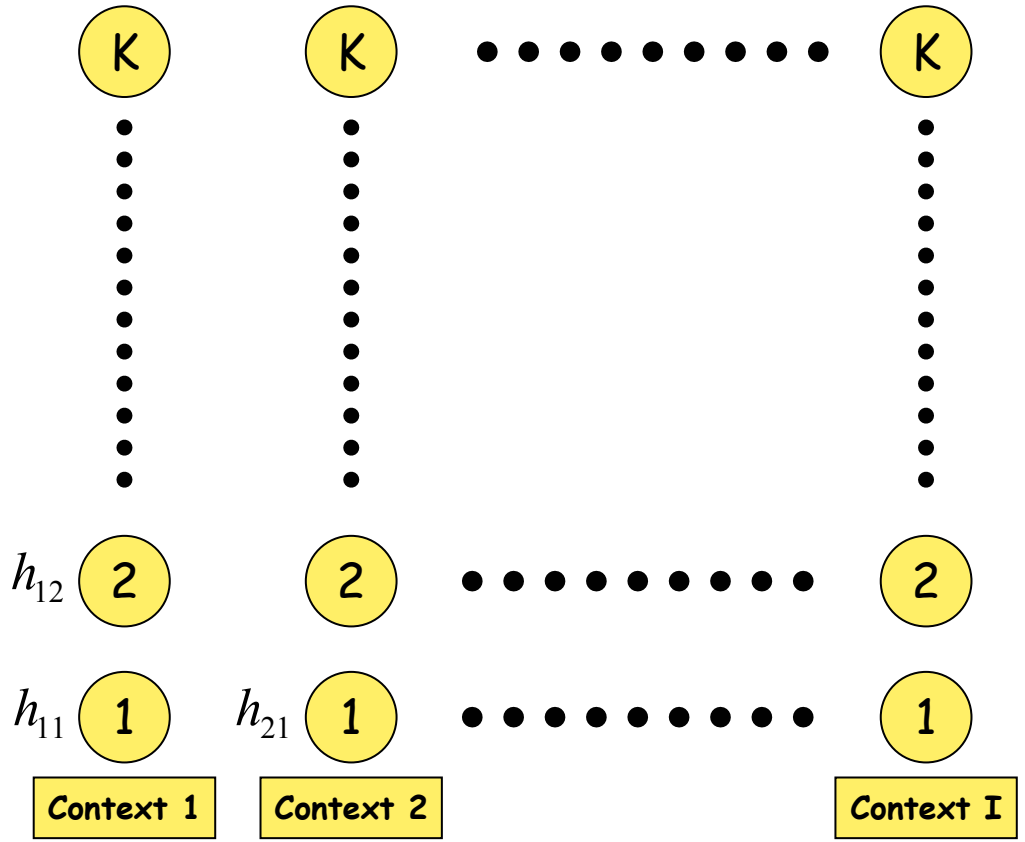
  - M step : Re-estimate

$$p(v_j \mid s_k) = \frac{\sum_{\{c_i : v_j \in c_i\}} h_{i,k}}{\sum_{k=1}^{K} \sum_{\{c_i : v_j \in c_i\}} h_{i,k}} \quad p(s_k) = \frac{\sum_{i=1}^{I} h_{i,k}}{\sum_{k=1}^{K} \sum_{i=1}^{I} h_{i,k}}$$

Method

Unsupervised Disambiguation

- Diagram :



Application

- Tagging
- Information Retrieval

An Application of Word Sense Disambiguation to Information Retrieval (1999) Jason M. Whaley

Word Sense Disambiguation and Information Retrieval Mark Sanderson Department
of Computing Science, University of Glasgow, Glasgow G12 8QQ United Kingdom –SIGIR94