Word Sense Disambiguation

Hsu Ting-Wei

Outline

- Introduction
  - 7.1 Methodological Preliminaries
    - 7.1.1 Supervised and Unsupervised learning
    - 7.1.2 Pseudowords
    - 7.1.3 Upper and lower bounds on performance
- Methods for Disambiguation
  - 7.2 Supervised Disambiguation
    - 7.2.1 Bayesian classification
    - 7.2.2 An information-theoretic approach
  - 7.3 Dictionary-based Disambiguation
    - 7.3.1 Disambiguation based on sense definitions
    - 7.3.2 Thesaurus-based disambiguation
    - 7.3.3 Disambiguation based on translations in a second-language corpus
    - 7.3.4 One sense per discourse, one sense per collocation
  - 7.4 Unsupervised Disambiguation

Introduction

- The task of disambiguation is 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.
- Ex: The word "*bank*" ,some senses that we found in a dictionary were:
  - bank [1,noun]: the rising ground bordering a lake, river, or sea…(岸)
  - bank [2, verb]: to heap or pile in a bank (築堤防護)
  - bank [3, noun]: an establishment for the custody, loan, or exchange of money (銀行)
  - bank [4, verb]: to deposit money (存錢)
  - bank [5, noun]: a series of objects arranged in a row (排;組)

    - Reference: Webster's Dictionary online http://www.m-w.com

Introduction (cont.)

- Two ambiguity in a sentence：
  - Tagging
    - Most part of speech tagging models simply use local context (nearby structure)
  - Word sense disambiguation
    - Word sense disambiguation methods often try to use context words in a broader context
- Ex: *You should butter your toast.*
  - Tagging
    - The word “butter” can be a 名詞 or 動詞
  - Word sense disambiguation
    - The word “butter” can mean 塗奶油 or 說甜言蜜語

7.1 Methodological Preliminaries

- 7.1.1 Supervised and Unsupervised learning
  - Supervised learning (classification、function-fitting)
    - The actual status for each piece of data on which we train
    - One extrapolates the shape of a function based on some data points.
  - Unsupervised learning (clustering task)
    - We don't know the classification of the data in the training sample

7.1 Methodological Preliminaries (cont.)

- 7.1.2 Pseudowords
  - Hand-labeling is a time intensive and laborious task
  - Test data are hard to come by
  - It is often convenient to generate artificial evaluation data for the comparison and improvement of text procession algorithms
  - Artificial ambiguous words can be created by conflating two or more natural words
  - Ex: banana-door
  - Easy to create large-scale train/test set

7.1 Methodological Preliminaries (cont.)

- 7.1.3 Upper and lower bounds on performance
  - It's meaningless that only consider numerical evaluation
    - Need to consider how difficult the task is
  - Using upper and lower bounds to estimate
    - Upper bound→ human performance
      - We can't expect an automatic procedure to do better
    - Lower bound (baseline)→ the simplest possible algorithm
      - Assignment of all contexts to the most frequent sense

Methods for Disambiguation

- 7.2 Supervised Disambiguation
  - Disambiguation based on a labeled training set
- 7.3 Dictionary-based
  - Disambiguation based on lexical resources such as dictionaries and thesauri
- 7.4 Unsupervised Disambiguation
  - Disambiguation based on training on an unlabeled text corpora.

Notational conventions used in this chapter

Symbol	Meaning
w	an ambiguous word
$s_1, \ldots, s_k, \ldots, s_K$	senses of the ambiguous word w
$c_1, \ldots, c_i, \ldots c_I$	contexts of w in a corpus
$v_1, \ldots, v_j, \ldots, v_J$	words used as contextual features for disambiguation

7.2 Supervised Disambiguation

- Training corpus: Each occurrence of the ambiguous word w is annotated with a semantic label
- Supervised disambiguation is a classification task.
- We will look at:
  - **Bayesian classification** (Gale et al. 1992).
  - **Information-theoretic approach** (Brown et al. 1991)

7.2 Supervised Disambiguation (cont.)

- 7.2.1 Bayesian classification (Gale et al.1992)
  - The approach treats the context of occurrence as a bag of words without structure, but it integrates information from many words in the context window. (feature)
  - Bayes Decision rule
    - Decide s' if $P(s' \mid c) > P(s_k \mid c)$ for $s_k \neq s'$
  - Bayes decision rule is optimal because it minimizes the probability of error
  - Choose the class (or sense) with the highest conditional probability and hence the smallest error rate.

7.2 Supervised Disambiguation (cont.)

- 7.2.1 Bayesian classification (Gale et al.1992)
  - Computing Posterior Probability for Bayes Classification
    - We want to assign the ambiguous word w to the sense s', given context c, where:

$$s' = \arg\max P(s_k \mid c)$$

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

Bay's Rule

$$= \arg\max P(c \mid s_k) P(s_k)$$

$$= \arg\max[\log P(c \mid s_k) + \log P(s_k)]$$

log

7.2 Supervised Disambiguation (cont.)

- 7.2.1 Bayesian classification (Gale et al.1992)
  - Naive Bayes assumption (Gale et al. 1992)
    - An instance of a particular kind of Bayes classifier

$$P(c \mid s_k) = P(\{v_j \mid v_j \ in \ \mathrm{c}\} \mid \mathrm{s}_k) = \prod_{v_j \ in \ c} P(v_j \mid \mathrm{s}_k)$$

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

7.2 Supervised Disambiguation (cont.)

- 7.2.1 Bayesian classification (Gale et al.1992)
  - Decision Rule for Naive Bayes
    - Decide s′ if

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

    - $P(v_j|s_k)$ and $P(s_k)$ are computed via Maximum-Likelihood Estimation, perhaps with appropriate smoothing, from the labeled training corpus

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

7.2 Supervised Disambiguation (cont.)

- 7.2.1 Bayesian classification (Gale et al.1992)
  - Bayesian disambiguation algorithm

```
1  comment: Training
2  for all senses s_k of w do
3      for all words v_j in the vocabulary do
4          P(v_j|s_k) = C(v_j,s_k) / C(v_i)
5      end
6  end
7  for all senses s_k of w do
8      P(s_k) = C(s_k) / C(w)
9  end
10 comment: Disambiguation
11 for all senses s_k of w do
12     score(s_k) = log P(s_k)
13     for all words v_j in the context window c do
14         score(s_k) = score(s_k) + log P(v_j|s_k)
15     end
16 end
17 choose s' = arg max_k score(s_k)
```

7.2 Supervised Disambiguation (cont.)

- **7.2.1 Bayesian classification (Gale et al.1992)**
  - Example of Bayesian disambiguation algorithm
    - w = drug

Sense (s1..sk)	Clues for sense (v1…vj)
Medication	prices, prescription, patent, increase, consumer, pharmaceutical
Illegal substance	abuse, paraphernalia, illicit, alcohol, cocaine, traffickers

P(prices|'medication') > P(price|'illict substance')

$$P(v_j \mid s_k) = \frac{C(v_j, s_k)}{\sum_t C(v_t, s_k)}$$

  - Bayes Classifier uses information from all words in the context window by using an independence assumption
    - Unrealistic independence assumption

7.2 Supervised Disambiguation (cont.)

- **7.2.2 An information-theoretic approach (Brown et al. 1991)**
  - The approach looks at only one informative feature in the context, which may be sensitive to text structure. But this feature is carefully selected from a large number of potential "informants."

indicator x1…xn | t1…tm

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

Highly informative indicators for three ambiguous French words

French	English
– Prendre une mesure	→ take a measure
– Prendre une decision	→ make a decision

7.2 Supervised Disambiguation (cont.)

- **7.2.2 An information-theoretic approach (Brown et al. 1991)**
  - Flip-Flop Algorithm (Brown et al., 1991)
    - The algorithm is used to disambiguate between the different senses of a word using the mutual information as a measure.
    - The algorithm is an efficient linear-time algorithm for computing the best partition of values for a particular indicator.
    - Categorize the informant (contextual word) as to which sense it indicates.

```
1 find random partition P = {P1, P2} of {t1, ..., tm}
2 while (improving) do
3         find partition Q = {Q1, Q2} of {x1, ..., xn}
4             that maximizes I(P;Q)
5         find partition P = {P1, P2} of {t1, ..., tm}
7             that maximizes I(P;Q)
8 end
```

$t_1,\ldots,t_m$ be the translation of the ambiguous word
$x_1,\ldots,x_n$ the possible values of the indicator

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

7.2 Supervised Disambiguation (cont.)

- **7.2.2 An information-theoretic approach (Brown et al. 1991)**
  - Flip-Flop Algorithm
  - Example
    - $P=\{t_1,..,t_m\}$ = {take,make,rise,speak}
      $Q=\{x_1,\ldots,x_n\}$ = {mesure,note,exemple,decision,parole}
    - Step1
      - Initial: find random partition P
        - » P1={take,rise} , P2={make,speak}
    - Step2
      - Find partition Q of the indicator values would give us maximum I(P;Q)
        - » Q1={measure,note,exemple} , Q2={decision,parole}
      - Repartition P and also maximum I(P;Q)
        - » P1={take} , P2={make,rise,speak}
    - If improving repeat step2

7.3 Dictionary-based Disambiguation

- If we have no information about the sense categorization of specific instance of a word, we can fall back on a general charaterization of the senses.
- Sense definitions are extracted from existing sources such as dictionaries and thesaurus (類語詞典，表示一串有著相似、相關或相反意思的單字)
- The different types of informational method have been used:
  - 7.3.1 Disambiguation based on sense definitions
  - 7.3.2 Thesaurus-based disambiguation
  - 7.3.3 Disambiguation based on translations in a second-language corpus
  - 7.3.4 One sense per discourse, one sense per collocation

7.3 Dictionary-based Disambiguation (cont.)

- 7.3.1 Disambiguation based on sense definitions (Lesk, 1986)
  - A word's dictionary definitions are likely to be good indicators of the senses they define.
  - Lesk's dictionary-based disambiguation algorithm
    - Ambiguous word w
      - Senses of w : $S_1 \ldots S_k$ (bags of words)
      - Dictionary definition of senses : $D_1 \ldots D_k$
      - E_{vj} : the set of words occurring in the dictionary definition ($D_1 \ldots D_k$) of word v_j (bags of words)

```
1 comment: Given: context c
2 for all senses s_k of w do
3       score(s_k) = overlap(D_k, U_{vj in c} E_vj)
4 end
5 choose s' s.t. s' = argmax_Sk score(s_k)
```

7.3 Dictionary-based Disambiguation (cont.)

- **7.3.1 Disambiguation based on sense definitions (Lesk, 1986)**
  - Lesk's dictionary-based disambiguation algorithm
    - Ex: Two senses of *ash*

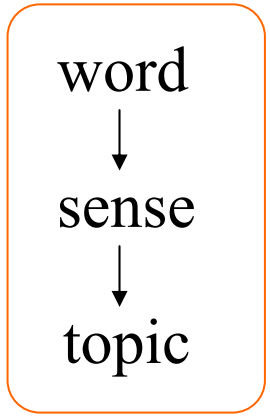
Sense	Definition
S_1 tree	a tree of the olive family
S_2 burned stuff	The solid residue left when combustible material is burned

Scores		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 come into leaf

7.3 Dictionary-based Disambiguation (cont.)

- 7.3.2 Thesaurus-based disambiguation
  - Simple thesaurus-based algorithm **(Walker,1987)**
    - Each word is assigned one or more subject codes in the dictionary
    - If the word is assigned several subject codes, then we assume that they corresponds to the different senses of the word.
    - $t(s_k)$ is the subject code of sense s_k
    - $\delta(t(s_k), v_j)=1$ iff $t(s_k)$ is one of the subject codes of v_j and 0 otherwise

```
1 comment: Given: context c
2 for all senses s_k of w do
3       score(s_k) = Σ_{vj in c} δ(t(s_k),v_j)
4 end
5 choose s' s.t. s' = argmax_{Sk} score(s_k)
```



7.3 Dictionary-based Disambiguation (cont.)

- 7.3.2 Thesaurus-based disambiguation
  - Simple thesaurus-based algorithm **(Walker,1987)**
    - Problem
      - A general categorization of words into topics is often inappropriate for a particular domain
        - » Mouse → mammal, electronic device
        - » When in a computer manual ……
      - A general topic categorization may also have a problem of coverage
        - » Navratilova (網球運動員) → sports
        - » When "Navratilova" is not found in the thesaurus…..

7.3 Dictionary-based Disambiguation (cont.)

- 7.3.2 Thesaurus-based disambiguation
  - Adaptation thesaurus-based algorithm **(Yarowsky,1987)**
    - Adapted the algorithm for words that do not occur in the thesaurus but that are very Informative
    - Using Bayes classifier for both adaptation and disambiguation

```
comment: Categorize contexts based on categorization of words
for all contexts c_i in the corpus do
    for all thesaurus categories t_l do
        score(c_i, t_l) = log P(c_i|t_l)/P(c_i) P(t_l)
    end
end
t(c_i) = {t_l | score(c_i, t_l) > α}
comment: Categorize words based on categorization of contexts
for all words v_j in the vocabulary do
    Vj = {c | v_j in c}
end
for all topics t_l do
    T_l = {c | t_l ∈ t(c)}
end
for all words v_j, all topics t_l do
    P(v_j|t_l) = |V_j ∩ T_l| / Σ_j |V_j ∩ T_l|
end
for all topics t_l do
    P(t_l) = (Σ_j |V_j ∩ T_l|) / (Σ_l Σ_j |V_j ∩ T_l|)
end
comment: Disambiguation
for all senses s_k of w occurring in c do
    score(s_k) = log P(t(s_k)) + Σ_{v_j in c} log P(v_j|t(s_k))
end
choose s' s.t. s' = argmax_{s_k} score(s_k)
```

7.3 Dictionary-based Disambiguation (cont.)

- 7.3.3 Disambiguation based on translations in a second-language corpus (Dagan et al. 1991; Dagan and Itai 1994)
  - Words can be disambiguated by looking at how they are translated in other languages
  - This method use of word correspondences in a bilingual dictionary
    - First Language
      - The one for which we want to disambiguation
    - Second Language
      - Target language in the bilingual dictionary
  - For example, if we want to disambiguate English based on German corpus, then English is the 1st language, and the German is the 2nd language.

7.3 Dictionary-based Disambiguation (cont.)

- **7.3.3 Disambiguation based on translations in a second-language corpus (Dagan et al. 1991; Dagan and Itai 1994)**
  - Ex : w = *interest*
  - To disambiguate the word “interest”, we identify the phrase 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

	Sense 1	**Sense 2**
Definition	legal share (利息)	attention, concern (興趣)
Translation	*Beteiligung*	*Interesse*
English collocation	acquire an interest	show interest
Translation	*Beteiligung erwerben*	*Interesse zeigen*

7.3 Dictionary-based Disambiguation (cont.)

- **7.3.3 Disambiguation based on translations in a second-language corpus (Dagan et al. 1991; Dagan and Itai 1994)**
  - Disambiguation based on a second-language corpus
    - S is the second-language corpus
    - $T(s_k)$ is the set of possible translations of sense s_k
    - T(v) is the set of possible translations of v

1 **comment:** Given: a context c in which w occurs in relation $R(w, v)$
2 **for** all senses s_k of w **do**
3 $\quad \text{score}(s_k) = |\{c \in S | \exists w' \in T(s_k), v' \in T(V) : R(w', v') \in c\}|$
4 **end**
5 choose $s' = \text{argmax}_{s_k} \text{score}(s_k)$

轉換成第二語言的sense　與轉換過的sense相關的文字

R(Interesse,zeigen) would be higher than count of *R(Beteiligung,zeigen)*

7.3 Dictionary-based Disambiguation (cont.)

- 7.3.4 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

7.3 Dictionary-based Disambiguation (cont.)

- 7.3.4 One sense per discourse, one sense per collocation (Yarowsky,1995)
  - Look one sense per discourse

Discourse	Initial label	Context
d_1	living	the existence of plant and animal life
	living	classified as either *plant* or animal
	?	Although bacterial and plant cells are enclosed
d_2	living	contains a varied plant and animal life
	living	the most common *plant* life
	living	slight within Arctic plant species
	factory	are protected by plant parts remaining from

? will be "living"

7.3 Dictionary-based Disambiguation (cont.)

- 7.3.4 One sense per discourse, one sense per collocation (Yarowsky,1995)
  - One sense per collocation : Most sensed are strongly correlated with certain contextual features like other words in the same

```
1  comment: Initialization
2  for all senses s_k of w do
3      F_k = the set of collocations in s_k's dictionary definition
4  end
5  for all senses s_k of w do
6      E_k = ∅
7  end
8  comment: One sense per collocation
9  while (at least one E_k changed in the last iteration) do
10         for all senses s_k of w do
11             E_k = {c_i | ∃f_m : f_m ∈ c_i ∧ f_m ∈ F_k}
12         end
13         for all senses s_k of w do
14             F_k = {f_m | ∀n ≠ k  P(s_k|f_m) / P(s_n|f_m) > α}
15         end
16 end
17 comment: One sense per discourse
18 for all documents d_m do
19     determine the majority sense s_k of w in d_m
20     assign all occurrences of w in d_m to s_k
21 end
```

F_k contains characteristic collocations. E_k is the set of contexts of the ambiguous word w that are currently assigned to s_k.

Collocational features are ranked according to the ratio (similar with information-theoretic method 7.2.2)

This is a surprisingly good performance given that the algorithm does not need a labeled set of thaining examples.

7.4 Unsupervised Disambiguation

- Cluster the contexts of an ambiguous word into a number of groups
- Discriminate between these groups without labeling them
  - Probabilistic model is same the same with section 7.2.1
    - Word : w
    - Senses : s1...sk
    - Estimate $P(v_j|s_k)$,
  - In contrast to Gale et al.'s Bayes classifier, parameter estimation in unsupervised disambiguation is not based on a labeled training set.
  - Instead, we start with a random initialization of the parameters $P(v_j|s_k)$. The $P(v_j|s_k)$ are then reestimated by the EM algorithm.

7.4 Unsupervised Disambiguation (cont.)

- EM Algorithm
  - Learning a word sense clustering.
    - K : number of desired senses
    - $c_1 \ldots, c_i, \ldots c_I$ are the contexts of the ambiguous word in the corpus
    - $v_1 \ldots, v_j, \ldots v_J$ are the words being used as disambiguating features
  - 1. **Initialize**
    - Initialize the parameters of the model μ randomly. The parameters are $P(v_j | s_k)$ and $P(s_k)$, $j = 1,2,\ldots J$, $k = 1,2,\ldots K$.
    - **Compute** the log likelihood of corpus C given the model μ as the product of the probabilities $P(c_i)$ of the individual contexts c_i (where $P(c_i) = \Sigma_k P(c_i | s_k) P(s_k)$) :

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

7.4 Unsupervised Disambiguation (cont.)

- EM Algorithm
  - 2. While l(C|μ) is improving repeat:
    - **E-step**: estimate h_{ik} ,the posterior probability that s_k generated c_i, as follows:

$$h_{ik} = \frac{P(s_k)P(c_i|s_k)}{\sum_{k=1}^{K} P(s_k)P(c_i|s_k)}$$

      To compute $P(c_i|s_k)$, we make the by now familiar Naïve Bayes assumption:

$$P(c_i|s_k) = \prod_j P(v_j|s_k)^{C(v_j \text{ in } c_i)}$$

    - **M-step**: **Re-estimate** the parameters $P(v_j|s_k)$ and $P(s_k)$ by way of maximum likelihood estimation:

$$P(v_j|s_k) = \frac{\sum_i C(v_j \text{ in } c_i) \cdot h_{ik}}{\sum_j \sum_i C(v_j \text{ in } c_i) \cdot h_{ik}}$$

      Recompute the probabilities of the senses as follows:

$$P(s_k) = \frac{\sum_{i=1}^{I} h_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{I} h_{ik}} = \frac{\sum_{i=1}^{I} h_{ik}}{I}$$

Thanks for your attention !