Word Sense Disambiguation

Shih-Hsiang Lin

Outline

- Introduction
- Methodological Preliminaries
  - Supervised and Unsupervised learning
  - Pseudowords
  - Upper and lower bounds on performance
- Methods for Disambiguating
  - Supervised Disambiguation
  - Dictionary-based
  - Unsupervised Disambiguation

Before we starting……

- 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)

5/28 definitions pulled from Webster’s Dictionary online http://www.m-w.com

Introduction

- Because of many words have several meaning or senses
  - there is ambiguity about how they are to be interpreted
- The task of disambiguation is to determine which of the senses of an ambiguous word is invoked in a particular use of the word.
- Types of problem
  - Syntactic ambiguity
    - differences in syntactic categories
  - Semantic ambiguity
    - homonymy(同形/音異義)or polysemy(一詞多義)

Methodological Preliminaries
Supervised and Unsupervised learning

- Supervised learning (classification、function-fitting)
  - know the actual status for each piece of data on which we learn
  - each element in a training set is paired with an acceptable response
- Unsupervised learning (clustering)
  - we don’t know the classification of the data in the training sample
  - adjusts through direct confrontation with new experiences (self organization)

Methodological Preliminaries
Pseudowords

- In order to test the performance of algorithms on a natural ambiguous word
  - a large number of occurrence has to be disambiguated by hand ← time-intensive
    laborious task
- Generate artificial evaluation data
  - pseudowords can be created by conflating two or more natural words
    - create pseudoword banana-door and replaces all occurrence of banana and door in the corpus
- Easy to create large-scale train/test set

Methodological Preliminaries

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
    - K We can't expect an automatic procedure to do better
  - Lower bound→ assign all contexts to the most frequent sense
    - A way to make sense of performance figures
    - A good idea for those which have no standardized evaluation sets for comparing systems

Methods for Disambiguating

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

Notational conventions

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

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)

Bayesian Classification

- Bayes Decision rule
  - Decide s' if $P(s'|c) > P(s_k|c)$ 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.

Computing Posterior Probability for Bayes Classification

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

$$
\begin{aligned}
s' &= \arg\max P(s_k \mid c) \\
&= \arg\max \frac{P(c \mid s_k)}{P(c)} P(s_k) \\
&= \arg\max P(c \mid s_k) P(s_k) \\
&= \arg\max [\log P(c \mid s_k) + \log P(s_k)]
\end{aligned}
$$

Bay's Rule

log

*Each context word Contributes potentially useful information about which sense of the ambiguous word is likely to be used with it

Naive Bayes (Gale et al. 1992)

- An instance of a particular kind of Bayes classifier
- Naive Bayes assumption: The attributes (contextual words) used for description are all conditionally independent

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

- 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

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

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

Bayesian disambiguation algorithm

comment: Training
for all senses s_k of w **do**
    for all words v_j in the vocabulary **do**
        $P(v_j|s_k) = \frac{C(v_j, s_k)}{C(v_j)}$
    end
end
for all senses s_k of w **do**
    $P(s_k) = \frac{C(s_k)}{C(w)}$
end
comment: Disambiguation
for all senses s_k of w **do**
    $\text{score}(s_k) = \log P(s_k)$
    for all words v_j in the context window c **do**
        $\text{score}(s_k) = \text{score}(s_k) + \log P(v_j|s_k)$
    end
end
choose $s' = \arg\max_{s_k} \text{score}(s_k)$

Example of Bayesian disambiguation algorithm

Sense	Clues for sense
Medication	prices, prescription, patent, increase, consumer, pharmaceutical
Illegal substance	abuse, paraphernalia, illicit, alcohol, cocaine, traffickers

Clues for two senses of drug used by a Bayesian classifier

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

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

-unrealistic independence assumption

An Information-Theoretic Approach

- In the information theoretic approach 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	measure → to take decision → to make
vouloir	tense	present → to want conditional → to like
cent	word to the left	per → % number → c.[money]

Hihgly informative indicators for three ambiguous French words

Prendre une decision→make a decision | Prendre une mesure→take a measure

Flip-Flop Algorithm (Brown et al., 1991)

- The Flip-Flop algorithm is used to disambiguate between the different senses of a word using the mutual information as a measure.
- Categorize the informant (contextual word) as to which sense it indicates.

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

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

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

Example of Classification based on Information-Theoretic Approach

- $P=\{t_1,..,t_m\}$ = {take,make,rise,speak}
  $Q=\{x_1,\ldots,x_n\}$ = {mesure,note,exemple,decision,parole}
- Initial: find random partition P
  – P1={take,rise} , P2={make,speak}
- 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

Dictionary-Based Disambiguation

- If we have no information about the sense categorization of a word
  - Relying on the senses in dictionaries and thesauri.
- Sense definitions are extracted from existing sources such as dictionaries and thesauri(同屬詞典)
- Use distributional properties to improve disambiguation
  - Ambiguous words are only used with one sense in any given discourse and with any given collocate

Disambiguation Based on Sense Definitions (Lesk, 1986)

- A word's dictionary definitions are likely to be good indicators of the senses they define.
- 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

Example of Disambiguation Based on Sense Definitions

1 **comment**: Given: context c
2 **for** all senses s_k of w **do**
3 score(s_k) = overlap(D_k, $\bigcup_{v_j \text{ in } c} E_{v_j}$)
4 **end**
5 choose s' s.t. s' = argmax_{S_k} score(s_k)

$D_1, \ldots, D_k$ the dictionary definitions of the senses $s_1, \ldots, s_k$ of the ambiguous word w, represented as the bag of words occurring definition.

v_j is the word occurring in the context c of w

E_{v_j} is the dictionary definition of v_j (union of all the sense definitions of v_j)

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 S_1	S_2	Context
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

Thesaurus-Based Disambiguation (Walker, 1984)

- The semantic categories of the words in a context determine the semantic category of the context as a whole.
  - decide the semantic category of the context
  - then decide which word sense are used
- 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

Thesaurus-Based Disambiguation (cont.)

1 **comment**: Given: context c
2 **for** all senses s_k of w **do**
3 score(s_k) = $\sum_{vj \text{ in } c} \delta(t(s_k), v_j)$
4 **end**
5 choose s' s.t. s' = argmax_{Sk} score(s_k)

$t(s_k)$ 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

The score is the number of words that are compatible with the subject code of sense sk

Problem：A general categorization of words into topics is often inappropriate for a particular domain
Mouse → mammal, electronic device
A general topic categorization may also have a problem of coverage
Navratilova→sports

Thesaurus-Based Disambiguation Creating New Categories(Yarowsky, 1992)

- Add new words to a category if they occur more often than chance
- Adapted the algorithm for words that do not occur in the thesaurus but that are very Informative
  - For example *Navratilova* can be added to the sports category

Thesaurus-Based Disambiguation
Creating New Categories (cont.)

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$(c_i, t_l) = \log \frac{P(c_i|t_l)}{P(c_i)} P(t_l)$
    end
end
$t(c_i) = \{t_l | \text{score}(c_i, t_l) > \alpha\}$
comment: Categorize words based on categorization of contexts
for all words v_j in the vocabulary do
    $Vj = \{c | v_j \text{ in } c\}$
end
for all topics t_l do
    $T_l = \{c | t_l \in t(c)\}$
end
for all words v_j, all topics t_l do
    $P(v_j|t_l) = |V_j \cap T_l| / \sum_j |V_j \cap T_l|$
end
for all topics t_l do
    $P(t_l) = (\sum_j |V_j \cap T_l|) / (\sum_l \sum_j |V_j \cap T_l|)$
end
comment: Disambiguation
for all senses s_k of w occurring in c do
    score$(s_k) = \log P(t(s_k)) + \sum_{v_j \text{ in } c} \log P(v_j | t(s_k))$
end
choose s' s.t. s' = argmax,, score(&)

Disambiguations Based on Translations (Dagan et al. 91 & 94)

- 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.

Disambiguations Based on Translations (cont.)

- Example: the word “interest” has two translations in German:
  - “Beteiligung” (legal share--50% a interest in the company)
  - “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

Disambiguations Based on Translations (cont.)

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

- Step1
  - Count the number of times that translations of the two senses of interest occur with translations of show in the second language corpus
- Step2
  - Compare the counts of the two different senses
- Step 3
  - Choose the sense that has the higher counts as a corresponding sense

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.

One Sense per Discourse, One sense per Collocation (cont.)

comment: Initialization
for all senses s_k of w **do**
    F_k = the set of collocations in s_k's dictionary definition
end
for all senses s_k of w **do**
    $E_k = \emptyset$
end
comment: One sense per collocation
while (at least one E_k changed in the last iteration) **do**
        for all senses s_k of w **do**
            $E_k = \{c_i \mid \exists f_m : f_m \in c_i \wedge f_m \in F_k\}$
        end
        for all senses s_k of w **do**
            $F_k = \{f_m \mid \forall n \neq k \frac{P(s_k \mid f_m)}{P(s_n \mid f_m)} > \alpha\}$
        end
end
comment: One sense per discourse
for all documents d_m **do**
    determine the majority sense s_k of w in d_m
    assign all occurrences of w in d_m to s_k
end

One Sense per Discourse, One sense per Collocation (cont.)

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

Unsupervised Disambiguation

- Sense tagging? Sense discriminate?
- 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 Bayesian model but the $P(v_j \mid s_k)$ are estimated using the EM algorithm

✓Initialize

$P(v_j \mid s_k)$

$P(s_k)$

✓Calculate

$P(c_i \mid s_k)$

✓Calculate

$P(v_j \mid s_k)$

$P(s_k)$

Unsupervised Disambiguation
EM Algorithm

- **Initialize** the parameters μ of model. These 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: $l(C|\mu) = \log \Pi_i \Sigma_k P(c_j | s_k) P(s_k)$
- while $l(C|\mu)$ is improving repeat:
  - **E-step:** $h_{ik} = P(c_j | s_k) P(s_k) / \Sigma_k P(c_j | s_k) P(s_k)$ (use Naive bayes to compute $P(c_j | s_k)$)
  - **M-step:** reestimate the parameters $P(v_j | s_k)$ and $P(s_k)$ by MLE:

    $P(v_j | s_k) = \Sigma_{ci} h_{jk}/Z_j$ where the sum is over all contexts c_i in which v_j occurs, Z_j a normalizing constant.

    $P(s_k) = \Sigma_i h_{jk} / \Sigma_k \Sigma_i h_{jk} = \Sigma_i h_{jk}/I$

END