Domain Adaptation and Training Data Acquisition in Wide-Coverage Word Sense Disambiguation and its Application to Information Retrieval

Zhi Zhong
Supervisor: Prof. Hwee Tou Ng

Department of Computer Science
National University of Singapore
Introduction

• *Word Sense Disambiguation (WSD)*: the task of identifying the correct sense/meaning of a word in a given context.
  – A basic semantic understanding task at the lexical level
  – An intermediate task for many other natural language processing (NLP) tasks

• Approaches
  – Knowledge based approaches (Lesk, 1986)
  – Unsupervised learning approaches (Pham et al., 2005)
  – Supervised learning approaches (Lee and Ng, 2002)
Introduction

Difficulties in Supervised WSD

• Low accuracy
  – Fine-grained vs. Coarse-grained sense inventory
• Lack of sense-annotated data
  – Reusing existing training data (Kohomban and Lee, 2005; Ando, 2006)
  – Generating training data from raw corpora (Mihalcea, 2002; Niu et al., 2005)
  – Active learning (Ng, 1997; Chan and Ng, 2007)
  – Multilingual resources (Resnik and Yarowsky, 1997; Ng et al., 2003; Chan and Ng, 2005a)
• Domain adaptation problem
  – Predicting sense distribution (McCarthey et al., 2004; Chan and Ng, 2005b; Chan and Ng, 2006)
  – Using active learning to pick good target domain instances (Chan and Ng, 2007)
Introduction

Applications

• Machine Translation
  – Selecting the correct translations for ambiguous words (Chan et al, 2007; Carpuat and Wu, 2007; Chiang et al., 2009)

• Information Retrieval
  – Disambiguating terms in queries and documents (Voorhees, 1993; Schutze and Pedersen, 1995; Stokoe et al., 2003; Kim et al., 2004)
  – Expansion of queries and documents (Voorhees, 1994; Liu et al., 2004; Cao, 2005; Fang, 2008; Agirre et al., 2010)

• Text Classification
  – Representing documents as a bag of senses (Kehagias et al., 2003; Bloehdorn and Hotho, 2004)

• Subjectivity Analysis
  – Many words have both subjective and objective senses (Wiebe and Mihalcea, 2006; Akkaya et al., 2009; Balamurali et al., 2011)
Contributions

• An open source supervised WSD system, IMS (It Makes Sense)
• Domain Adaptation, combining the feature augmentation technique with active learning
• Extracting sense annotated examples from parallel corpora without extra human efforts
• Improving information retrieval by incorporating word senses
Outline

• Introduction
• An open source word sense disambiguation system
• Domain adaptation for word sense disambiguation
• Automatic extraction of training data from parallel corpora
• Word sense disambiguation for information retrieval
• Conclusion
An open source word sense disambiguation system

• Motivations:
  – Few publicly available open source WSD systems
  – WSD is required as a component by other applications

• IMS (It Makes Sense):
  – A supervised learning system for WSD
  – Open source
  – Provide an extensible and flexible platform for researchers interested in WSD
  – An English all-words WSD component for other NLP tasks
System Description

Input Document

Preprocessing
• Sentence splitter
• Tokenizer
• POS Tagger
• Lemmatizer

Instance Extraction
• POS Feature Extractor
• Surrounding Word Extractor
• Local Collocation Extractor

Classification
• Machine Learning Toolkit

Knowledge Sources: (Lee and Ng, 2002)
• Part of Speech: P_{-3}, P_{-2}, P_{-1}, P_{0}, P_{1}, P_{2}, P_{3}
• Surrounding Words
• Local Collocations: C_{-1,-1}, C_{1,1}, C_{-2,-1}, C_{2,1}, C_{-1,1}, C_{1,2}, C_{-2,-2}, C_{2,2}, C_{-1,1}, C_{1,2}, C_{-2,1}, C_{2,2}, C_{1,3}

One model for each word type
Support Liblinear, LibSVM, MaxEnt, Weka toolkit
Sense-annotated Data Set

• Sense-annotated corpus
  – SEMCOR (Miller et al., 1994)
  – The DSO corpus (Ng and Lee, 1996)

• Sense-annotated data extracted from parallel corpora (Chan and Ng, 2005a)
  – 6 English-Chinese parallel corpora
  – The top 60 most frequently occurring polysemous content word in Brown Corpus

<table>
<thead>
<tr>
<th>POS</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
</tr>
</thead>
<tbody>
<tr>
<td># of types</td>
<td>11,445</td>
<td>4,705</td>
<td>5,129</td>
<td>28</td>
</tr>
</tbody>
</table>
## Experiments

<table>
<thead>
<tr>
<th>Task</th>
<th>SE2</th>
<th>SE3</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS (Liblinear)</td>
<td>65.3%</td>
<td>72.6%</td>
</tr>
<tr>
<td>IMS (WEKA)</td>
<td>65.0%</td>
<td>72.0%</td>
</tr>
<tr>
<td>IMS (MaxEnt)</td>
<td>62.2%</td>
<td>69.4%</td>
</tr>
<tr>
<td>Rank 1</td>
<td>64.2%</td>
<td>72.9%</td>
</tr>
<tr>
<td>Rank 2</td>
<td>63.8%</td>
<td>72.6%</td>
</tr>
<tr>
<td>Rank 3</td>
<td>62.9%</td>
<td>72.4%</td>
</tr>
<tr>
<td>MFS</td>
<td>47.6%</td>
<td>55.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Italian</th>
<th>Spanish</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS (Liblinear)</td>
<td>56.9%</td>
<td>87.3%</td>
<td>62.3%</td>
</tr>
<tr>
<td>IMS (WEKA)</td>
<td>57.1%</td>
<td>87.2%</td>
<td>63.3%</td>
</tr>
<tr>
<td>IMS (MaxEnt)</td>
<td>56.6%</td>
<td>84.1%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Rank 1</td>
<td>53.1%</td>
<td>84.2%</td>
<td>60.4%</td>
</tr>
<tr>
<td>Rank 2</td>
<td>51.5%</td>
<td>84.0%</td>
<td>-</td>
</tr>
<tr>
<td>Rank 3</td>
<td>49.8%</td>
<td>82.5%</td>
<td>-</td>
</tr>
<tr>
<td>MFS</td>
<td>18.3%</td>
<td>67.7%</td>
<td>28.5%</td>
</tr>
</tbody>
</table>

English Lexical-sample tasks

SensEval-3 Lexical-sample tasks
## Experiments

<table>
<thead>
<tr>
<th>Task</th>
<th>SensEval-2 Fine-grained</th>
<th>SensEval-3 Fine-grained</th>
<th>SemEval-07 Fine-grained</th>
<th>SemEval-07 coarse-grained</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS (Liblinear)</td>
<td>68.2%</td>
<td>67.6%</td>
<td>58.3%</td>
<td>82.6%</td>
</tr>
<tr>
<td>IMS (WEKA)</td>
<td>67.8%</td>
<td>67.5%</td>
<td>59.1%</td>
<td>82.2%</td>
</tr>
<tr>
<td>IMS (MaxEnt)</td>
<td>67.5%</td>
<td>67.4%</td>
<td>58.9%</td>
<td>82.0%</td>
</tr>
<tr>
<td>Rank 1</td>
<td>69.0%</td>
<td>65.2%</td>
<td>59.1%</td>
<td>82.5%</td>
</tr>
<tr>
<td>Rank 2</td>
<td>63.6%</td>
<td>64.6%</td>
<td>58.7%</td>
<td>81.6%</td>
</tr>
<tr>
<td>Rank 3</td>
<td>61.8%</td>
<td>64.1%</td>
<td>58.3%</td>
<td>81.5%</td>
</tr>
<tr>
<td>WNs1</td>
<td>61.9%</td>
<td>62.4%</td>
<td>51.4%</td>
<td>78.9%</td>
</tr>
</tbody>
</table>

SensEval/SemEval fine-grained and coarse-grained all-words tasks
Outline

• Introduction
• An open source word sense disambiguation system
• Domain adaptation for word sense disambiguation
• Automatic extraction of training data from parallel corpora
• Word sense disambiguation for information retrieval
• Conclusion
Domain adaptation for word sense disambiguation

- **Target-domain Test Data**
  - OntoNotes WSJ section 23 (Hovy et al., 2006)

- **Target-domain Training Data**
  - OntoNotes WSJ sections 02-21
  - Accuracy of **89.1%**

- **Source-domain Training Data**
  - SEMCOR (mapped from WordNet sense-inventory to OntoNotes sense-inventory)
  - Accuracy of **76.2%**

- **Suffer a substantial decrease in accuracy.**
  - General problem for many NLP tasks (Daume III and Marcu, 2006)
  - Domain adaptation is needed

<table>
<thead>
<tr>
<th>Section</th>
<th># of word-types</th>
<th># of instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual</td>
<td>Cumulative</td>
</tr>
<tr>
<td>02</td>
<td>248</td>
<td>425</td>
</tr>
<tr>
<td>03</td>
<td>79</td>
<td>107</td>
</tr>
<tr>
<td>04</td>
<td>186</td>
<td>389</td>
</tr>
<tr>
<td>05</td>
<td>287</td>
<td>625</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>288</td>
<td>536</td>
</tr>
<tr>
<td>21</td>
<td>262</td>
<td>470</td>
</tr>
<tr>
<td>23</td>
<td>685</td>
<td>3,755</td>
</tr>
</tbody>
</table>
Combine SEMCOR and OntoNotes

WSD Accuracy (%)

Section Numbers

ON  SC+ON

National University of Singapore
The Feature Augmentation Technique

• Daume III 2007
• For an instance \( x \), suppose \( \Phi(x) \) is the original feature vector of \( x \). The augmented feature vector of \( x \) is:

\[
\Phi'(x) = \begin{cases} 
\langle \Phi(x), \Phi(x), 0 \rangle & \text{if } x \in D_s \\
\langle \Phi(x), 0, \Phi(x) \rangle & \text{if } x \in D_t
\end{cases}
\]

Where 0 is a zero vector of length \( |\Phi(x)| \), \( D_s \) and \( D_t \) are the sets of instances from the source and target domain respectively.
The Feature Augmentation Technique

WSD Accuracy (%) vs. Section Numbers

ON | SC+ON | SC+ON Augment

National University of Singapore

76.2% train on SEMCOR
Active Learning with the Feature Augmentation Technique

• Annotating a small number of target domain instances is worth the effort.
• Combining source-domain and target-domain training data via the feature augmentation technique is a good strategy.
• Use active learning to minimize the human effort in annotating target-domain instances. (Chan and Ng, 2007)
Active Learning with the Feature Augmentation Technique

- **Initial:** \( D_S \leftarrow \text{SEMCOR}, D_A \leftarrow \text{WSJ sections 02-21}, D_T \leftarrow \text{empty} \)
- **Loop:**
  - Until \( D_A \) is empty

  ![Diagram](image)
Active Learning with the Feature Augmentation Technique

![Graph showing WSD Accuracy (%) vs Iteration Number for different SemCor and iteration numbers: 50, 100, 150, 200, 300, 400, 500, and all. Each line represents a different dataset size, with iteration numbers ranging from 2 to 34.]
Active Learning with the Feature Augmentation Technique

- **SEMCOR Only**
- **Active Learning with 1,500 OntoNotes Instances**
- **Best Result with 31,114 OntoNotes Instances**

The graph shows the WSD accuracy (%) for different datasets and learning scenarios:

- **SEMCOR Only**
- **Active Learning with 1,500 OntoNotes Instances**
- **Best Result with 31,114 OntoNotes Instances**

Accuracy values:
- 76.2 for SEMCOR Only
- 82.6 for Active Learning with 1,500 OntoNotes Instances
- 89.1 for the Best Result with 31,114 OntoNotes Instances

- 10 active learning iterations on 150 most frequency word-types
- All target domain training instances

National University of Singapore
Outline

• Introduction
• An open source word sense disambiguation system
• Domain adaptation for word sense disambiguation
• Automatic extraction of training data from parallel corpora
• Word sense disambiguation for information retrieval
• Conclusion
Previous Work

*Chan and Ng (2005)* proposed a method to extract training instances from parallel texts by identifying the sense of the English words with their Chinese translations:

<table>
<thead>
<tr>
<th>#</th>
<th>article.n</th>
<th>Sense descriptions in WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>文章</td>
<td>Nonfictional prose forming an independent part of a publication</td>
</tr>
<tr>
<td>2</td>
<td>物品 物件 货品</td>
<td>One of a class of artifacts</td>
</tr>
<tr>
<td>3</td>
<td>条文 条款 条</td>
<td>A separate section of a legal document</td>
</tr>
<tr>
<td>4</td>
<td>冠词</td>
<td>A determiner that may indicate the specificity of reference of a noun phrase</td>
</tr>
</tbody>
</table>

- The reporter wrote an article about environment protection.
- 记者写了关于环境保护的文章。
- No hawker shall deposit in a private market any goods or other articles whatsoever.
- 任何小贩不得在私营街市内存放任何种类的货品或物品。
Previous Work

• Limitations:
  – Need manual selection of appropriate translations in a second language for every word sense.
  – A huge number of words in a language like English.

• Motivations and Aims:
  – Automatically select Chinese translations for English word senses.
  – The whole process of extracting training instances from parallel texts will be unsupervised.
Our Methods

• WordNet is an English sense inventory.
  – Basic element -- Synset: a set of synonyms
    • *e.g.* SynA -> cognition, knowledge
  – Each word sense has a corresponding synset
    • *e.g.* cognition%1:03:00 \[\rightarrow\] SynA
      knowledge%1:03:00

• **Four** methods to select Chinese translations for WordNet senses by making use of bilingual dictionaries and bilingual corpora.
1. Sinica Bilingual Ontological WordNet

Sinica Bilingual Ontological WordNet (BOW)

• Map WordNet to English-Chinese Translation Equivalents Database (ECTED)
• Functions as an English-Chinese bilingual WordNet
• Each WordNet synset has a set of Chinese translations
• 94,874 Chinese translations for 66,025 WordNet noun synsets. Each synset has 1.4 Chinese translations on average.
2. Common Bilingual Dictionaries

• Common English-Chinese bilingual dictionaries contain Chinese translations, and English/Chinese glosses for each English word sense.
• Sense definitions of these dictionaries are quite different from WordNet.
  – e.g. “interest”:
    3 senses in PowerWord vs. 7 senses in WordNet
2. Common Bilingual Dictionaries

Heuristic I:

If a Chinese translation $c$ shared by two or more synonyms in a WordNet synset, assign $c$ as a Chinese translation for this synset.

*e.g.* SynB -> pause, intermission, break, interruption and suspension

- **pause:** 中止 暫停 休止符...
- **break:** 休息 暫停 破裂 突変...
- **suspension:** 悬浮 暫停 中止 延迟...

*Chinese translations from PowerWord*
2. Common Bilingual Dictionaries

Heuristic I:

If a Chinese translation c shared by two or more synonyms in a WordNet synset, assign c as a Chinese translation for this synset.

*e.g.* SynB -> pause, intermission, break, interruption and suspension

- **pause:** 中止 暂停 休止符...
- **break:** 休息 暂停 破裂 突变...
- **suspension:** 悬浮 暂停 中止 延迟...

*Chinese translations from PowerWord*
Heuristic II:

For a monosemous word \( e \) in WordNet, assign its Chinese translations from common bilingual dictionaries to the WordNet synset corresponding to \( e \)'s only sense.

\[ e.g. \text{SynC} \rightarrow \text{blessing, boon (monosemous)} \]

- blessing: 祝福
- boon: 恩惠 实惠 福利

Chinese translations from PowerWord
3. Shorten Chinese Translations

The Chinese translation $c$, which is selected as sense $s$ of $e$ but has no occurrence in parallel texts aligned to $e$ -- *Wasted.*

- Shorten the Chinese translation $c$:
  - The longest prefix or suffix that has occurrences aligned to $e$ in parallel texts.
  - Not a substring of any Chinese translations from dictionaries for a different sense $s'$ of $e$. 
3. Shorten Chinese Translations

e.g.

revenue sense 2: 尤指国家的税收
\[(especially \, referring \, to \, federal \, tax)\]

value sense 6: 价值观念
\[(value \, concept)\]
3. Shorten Chinese Translations

e.g.

revenue sense 2: 尤指国家的税收
(especially referring to federal tax)

value sense 6: 价值观念
(value concept)
4. Word Similarity Measure

The Chinese words aligned to \( e \) in parallel texts, but not selected -- \textit{Wasted}.

- Calculate their similarities to the Chinese translations those are already selected for some senses of \( e \).
  - Distribution similarity measure based on syntactic relations. \textit{(Lin, 1998)}

- Assign them to the senses corresponding to their most similar Chinese translations.
# 4. Word Similarity Measure

<table>
<thead>
<tr>
<th>#</th>
<th>Selected Chinese translations for <strong>revenue</strong></th>
<th>Sense description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>收益 收入 总收入</td>
<td>the entire amount of income before any deductions are made</td>
</tr>
<tr>
<td>2</td>
<td>尤指国家之税收 税收</td>
<td>government income due to taxation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Candidate for revenue</th>
<th>Most similar</th>
<th>Score</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>税款 (taxation)</td>
<td>税收</td>
<td>0.1995</td>
<td></td>
</tr>
<tr>
<td>征税 (levy)</td>
<td>税收</td>
<td>0.1254</td>
<td></td>
</tr>
<tr>
<td>收支 (income and expense)</td>
<td>收益</td>
<td>0.1006</td>
<td></td>
</tr>
</tbody>
</table>
4. Word Similarity Measure

<table>
<thead>
<tr>
<th>#</th>
<th>Selected Chinese translations for <strong>revenue</strong></th>
<th>Sense description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>收益 收入 总收入</td>
<td>the entire amount of income before any deductions are made</td>
</tr>
<tr>
<td>2</td>
<td>尤指国家之税收 税收</td>
<td>government income due to taxation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Candidate for revenue</th>
<th>Most similar</th>
<th>Score</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>税款 (taxation)</td>
<td>税收</td>
<td>0.1995</td>
<td>2</td>
</tr>
<tr>
<td>征税 (levy)</td>
<td>税收</td>
<td>0.1254</td>
<td>2</td>
</tr>
<tr>
<td>收支 (income and expense)</td>
<td>收益</td>
<td>0.1006</td>
<td>1</td>
</tr>
</tbody>
</table>
Experiments

• Training Data Set:
  – All the examples from SEMCOR and up to 1000 examples extracted from parallel texts according to SEMCOR distribution.

• Test Set: Nouns in OntoNotes 2.0
  – 605 noun types with 29,510 examples.
  – 257 nouns among the top 60% words in Brown Corpus, which have manually selected Chinese translations from (Chan and Ng, 2005a)

<table>
<thead>
<tr>
<th></th>
<th>T60Set</th>
<th>All nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td># of word type</td>
<td>257</td>
<td>605</td>
</tr>
<tr>
<td># of examples</td>
<td>22,353</td>
<td>29,510</td>
</tr>
<tr>
<td>Avg. # of senses per word</td>
<td>4.3</td>
<td>3.5</td>
</tr>
</tbody>
</table>
Experiments

WordNet Sense 1 Baseline: a strong baseline in WSD by always picking the first sense in WordNet

SemCor + Manu 80.3%

SemCor + Manu 77.0%

SemCor + Similarity + Shorten + PowerWord + BOW
Outline

• Introduction
• An open source word sense disambiguation system
• Domain adaptation for word sense disambiguation
• Automatic extraction of training data from parallel corpora
• Word sense disambiguation for information retrieval
• Conclusion
Information Retrieval

• Information Retrieval (IR): rank documents based on relevance to keyword queries

• WSD for IR:
  – Identify the meanings of ambiguous query words
  – Make use of semantic relations between senses
The Language Modeling Approach to IR

Language models are constructed for each query and each document in a text collection $C$

- **Query** $q$: $\theta_q$
  
  $$p(t \mid \theta_q) = \frac{tf(t,q)}{\sum_{t' \in q} tf(t',q)}$$
  
  Maximum Likelihood Estimation

- **Document** $d$: $\theta_d$
  
  $$p(t \mid \theta_d) = \frac{tf(t,d) + \mu \cdot p(t \mid \theta_C)}{\sum_{t' \in d} tf(t',d) + \mu}$$
  
  Dirichlet-prior smoothing

Documents are ranked by their distances to $q$ according to their language models

- Negative Kullback-Leibler (KL) divergence
Pseudo Relevance Feedback

• Pseudo Relevance Feedback (PRF)
  – Retrieve documents from a text collection $C$ with the original query $q$
  – Assume that the top $k$ retrieved documents $D_q$ are relevant
  – Interpolate the original query model with $m$ expansion query terms selected from $D_q$
  – Retrieve documents with the expanded query

• Collection Enrichment (CE)
  – Improve the quality of the feedback documents by using an additional external document collection $X$ in the first retrieval step in PRF
Word Sense Disambiguation System

- Word Sense Disambiguation System Induced from Parallel Corpora
  - Use translations as senses
  - Extract training data from parallel corpora
- Word Sense Disambiguation for Short Queries
  - Propose a method to disambiguate terms in short queries
Word Sense Disambiguation System

• The meaning of a word can be disambiguated by its translation in a second language
• Induce a supervised WSD system from parallel corpora, with translations as senses (*no* manual labeling of senses):
  – Align parallel corpora with GIZA++
  – For an English morphological root $e$, extract its occurrences and corresponding translations from parallel corpora
  – Training data: English part
  – Sense: English morphological root $e$ + Chinese translation
    • To differentiate English words sharing the same Chinese translations
    • E.g., Sense of *girls*: girl_女子, sense of *women*: woman_女子
  – Employ IMS to train the WSD models
Word Sense Disambiguation System

Both terms in query and documents can be ambiguous

• Documents are full articles with sufficient context
  – Apply WSD system on documents directly
  – $p(w, s, d)$: probability of assigning sense $s$ to word occurrence $w$ in document $d$
  – Each word occurrence in a document is assigned its list of senses with probabilities

• Queries are usually short, often with only two or three terms in a query.
  – Challenge: Insufficient context for WSD systems
  – Solution: Find some relevant text fragments that contain the query terms
Word Sense Disambiguation for Short Queries

• Assumption of pseudo relevance feedback (PRF):
  – The top $k$ retrieved documents $D_q$ are relevant
• $D_q$ can be considered as relevant text fragments of $q$
• For each query term $t$ in $q$, utilize the sense distribution of the words with the same stem form as $t$ in $D_q$ as a proxy to estimate the senses for $t$
  – $S(t, q)$: the union of the senses assigned to word occurrences in $D_q$ with the same stem form as $t$
  – $p(t, s, q)$: probability of tagging $t$ as sense $s$ is proportional to the sum of the probabilities of tagging the above word occurrences as sense $s$
Incorporating Senses into LM
Approach

With the above method, we have assigned senses to terms in both queries and documents

- \( p(w, s, d) \): probability of tagging word occurrence \( w \) in \( d \) as sense \( s \)
- \( p(t, s, q) \): probability of tagging query term \( t \) in \( q \) as sense \( s \)

Adjust term frequencies with sense information
Incorporating Senses into LM
Approach

Adjust the term frequency of \( t \) in \( d \) \( tf(t, d) \) with the sense information

\[
\begin{align*}
\text{tf}_{\text{sen}}(t, d) &= tf(t, d) + \text{sen}(t, q, d) \\
\text{sen}(t, q, d) &= \alpha^{\Delta \cos(t, q, d)} \text{stf}(S(t, q), d)
\end{align*}
\]

- Sense similarity weight
- \( \alpha \geq 1 \)
- More similar, larger weight
- Sum of frequencies of \( t \)'s senses in \( d \)
- Documents with more senses of \( t \) will have higher term frequency

Update the probability of \( t \) in \( d \) accordingly

\[
p(t \mid \theta_{d}^{\text{sen}}) = \frac{\text{tf}_{\text{sen}}(t, d) + \mu \cdot p(t \mid \theta_{C}^{\text{sen}})}{\sum_{t' \in d} \text{tf}_{\text{sen}}(t', d) + \mu}
\]
Expanding with Sense Synonym Relations

Senses in our WSD system consist of two parts: English morphological root and Chinese translation, e.g., lady_女子

• Senses with the same Chinese translation are assumed to be synonyms

• $R(s)$: senses with the same Chinese translation as $s$

• For example:

  girl_女子  female_女子  woman_女子

  $R(girl_女子) = \{ \text{female}_女子, \text{woman}_女子 \}$
Expanding with Sense Synonym Relations

Synonym relation is commonly used to improve IR performance

- $R(s)$: senses that are synonymous with sense $s$
- $S(q)$: union of senses assigned to terms in $q$

$$\begin{align*}
tf_{syn}(t,d) &= tf_{sen}(t,d) + \text{syn}(t,q,d) \\
\text{syn}(t,q,d) &= \sum_{s \in S(t,q)} \beta(s,q) \cdot p(t,s,q) \cdot stf(R(s) - S(q),d)
\end{align*}$$

- Scaling function
- Control the impact of synonyms
- Sum of frequencies of synonyms of $s$ in $d$
- Senses already in $q$ are excluded
- More synonyms, higher term frequency

Update the probability of $t$ in $d$ accordingly
Experimental Setting

Conduct experiments on the TREC collection:

• **Text collection C**
  – TREC disk 4 and 5, minus the CR corpus,
  – 528,155 documents

• **External collection X**
  – The other documents in TREC disk 1 to 5

• **Total:** about 1.6 million documents

• **Tuning data**
  – 50 queries from TREC6 Ad Hoc task

• **Test data**
  – 199 queries from TREC7 & 8 Ad Hoc tasks and Robust03 & 04 tasks
Experiments

Baseline IR approach: Stem$^{prf}$ (using Lemur toolkit)
• Stem-based unigram LM approach
• Dirichlet-prior smoothing for document model
• Negative KL-divergence
• Pseudo relevance feedback with collection enrichment

WSD methods:
• Supervised WSD system
  – About 700 hours to disambiguate 1.6 million documents in $C$ and $X$
• Most frequent sense (MFS) baseline
• Even probability baseline
Experimental Results

- Retrieve the top-ranked 1,000 documents for each query
- Mean average precision (MAP) as evaluation metric

<table>
<thead>
<tr>
<th>Method</th>
<th>TREC7</th>
<th>TERC8</th>
<th>Robust03</th>
<th>Robust04</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1</td>
<td>0.2530</td>
<td>0.3063</td>
<td>0.3704</td>
<td>0.4019</td>
</tr>
<tr>
<td>Top 2</td>
<td>0.2488</td>
<td>0.2876</td>
<td>0.3065</td>
<td>0.4008</td>
</tr>
<tr>
<td>Top 3</td>
<td>0.2427</td>
<td>0.2853</td>
<td>0.3037</td>
<td>0.3514</td>
</tr>
<tr>
<td>Stem_{prf} (BL)</td>
<td>0.2634</td>
<td>0.2944</td>
<td>0.3586</td>
<td>0.3781</td>
</tr>
<tr>
<td>BL+MFS</td>
<td>0.2655</td>
<td>0.2971</td>
<td>0.3626*</td>
<td>0.3802</td>
</tr>
<tr>
<td>BL+Even</td>
<td>0.2655</td>
<td>0.2972</td>
<td>0.3623*</td>
<td>0.3814</td>
</tr>
<tr>
<td>BL+WSD</td>
<td>0.2679*</td>
<td>0.2986*</td>
<td>0.3649*</td>
<td>0.3842</td>
</tr>
<tr>
<td>BL+MFS+Syn</td>
<td>0.2756*</td>
<td>0.3034*</td>
<td>0.3649*</td>
<td>0.3859</td>
</tr>
<tr>
<td>BL+Even+Syn</td>
<td>0.2713*</td>
<td>0.3061*</td>
<td>0.3657*</td>
<td>0.3859*</td>
</tr>
<tr>
<td>BL+WSD+Syn</td>
<td>0.2762*</td>
<td>0.3126*</td>
<td>0.3735*</td>
<td>0.3891*</td>
</tr>
</tbody>
</table>

*: two-tailed t-test with p < 0.05 over baseline
**:two-tailed t-test with p < 0.01 over baseline
Analysis

• BL+{MFS, Even, WSD} improve over baseline Stem_{prf}
  – The morphological roots of senses overcome variation due to inflection (better recall)
  ▸ For example, in topic 326 {ferry sinkings}
    Stem form of “sinkings” is sink
    Inflection forms of “sink”: “sunk”, “sank” and “sunken”
Analysis

• BL+{MFS, Even, WSD} improve over baseline Stem_{prf}
  – The morphological roots of senses overcome variation due to inflection (better recall)

• BL+WSD outperforms BL+{MFS, Even}
  – Pinpoint the sense of query terms (better precision)
  ‣ For example, in topic 357 {territorial \textit{waters} dispute}
    water_水域 0.795 (body of water)
    water_水 0.047 (H_{2}O)
    water_供水 0.025 (provide with water)
  ...
  ‣ BL+WSD is better than BL+{MFS, Even} on 121 queries and 119 queries, respectively
Analysis

• BL+{MFS, Even, WSD} improve over baseline Stem_{prf}
  – The morphological roots of senses overcome variation due to inflection (better recall)

• BL+WSD outperforms BL+{MFS, Even}
  – Pinpoint the sense of query terms (better precision)

• Expanding with synonym relations further improves the performance
  – WSD helps to choose the appropriate synonyms for expansion (better recall)

  For example, in topic 648 {family \textit{leave} law}
  
  \begin{itemize}
  \item leave_假期 0.371 (leave of absence)
  \item leave_离开 0.198 (go away)
  \end{itemize}

...
Outline

• Introduction
• An open source word sense disambiguation system
• Domain adaptation for word sense disambiguation
• Automatic extraction of training data from parallel corpora
  • Word sense disambiguation for information retrieval
• Conclusion
Conclusion

• Develop and release an open source supervised WSD system, IMS. (Chapter 3)
• Examine the domain adaptation problem in WSD by combining active learning and the feature augmentation technique. (Chapter 4)
• Propose a method to extract sense-annotated examples from parallel corpora without extra human effort. (Chapter 5)
• Improve the performance of IR by integrating senses. (Chapter 6)
Future Work

• Investigate the use of a semi-supervised extension of the feature augmentation technique (Daume III et al., 2010)
• Gather WSD training data from parallel corpora for other languages
• Apply WSD to question answering
Publications

Zhi Zhong, Hwee Tou Ng, and Yee Seng Chan (EMNLP 2008)
Word sense disambiguation using OntoNotes: An empirical study.

Zhi Zhong and Hwee Tou Ng (IJCAI 2009)
Word sense disambiguation for all words without hard labor.

Zhi Zhong and Hwee Tou Ng (ACL 2010)
It Makes Sense: A wide-coverage word sense disambiguation system for free text.

Zhi Zhong and Hwee Tou Ng (ACL 2012)
Word sense disambiguation improves information retrieval.