Interactive Second Language Learning from News Websites

Tao Chen, Naijia Zheng, Yue Zhao, Muthu Chandrasekaran and Min-Yen Kan

kanmy@comp.nus.edu.sg
Slides available at: dwz.cn/kan-nlptea2
Formal language learning is time-consuming, and learning materials are often limited.
...but we also spend over 35B hours every month keeping up with the news.
WordNews
A browser extension for vocabulary learning when reading online news
Police inspect the debris washed up on the island of Reunion.

Debris on the Indian Ocean island of Reunion is to be transported to France to find out whether it is from the missing flight MH370, Malaysia’s prime minister has said.

Initial reports suggest the 2m-long object is very likely to be from a Boeing 777, Najib Razak said.

The Malaysia Airlines - a Boeing 777 - vanished while travelling from Kuala Lumpur to Beijing in March 2014.

The search has focused on part of the southern Indian Ocean east of Reunion.

Oceanographer David Griffin, of Australia’s national science agency, told the BBC that the location of the find was "consistent with where we think debris might have turned up".

There were 239 passengers and crew on board the plane when it went missing.
News Context

• Identify the news category by URL pattern


7 categories: Entertainment, World, Finance, Sports, Fashion, Technology, Travel

• Classify words based on category document frequency
  E.g., “superstar” belongs to “Entertainment”

• For both English and Chinese news and words
Outline

• Introduction

• Translating
  Word Sense Disambiguation (WSD)

• Testing
  Distractor Generation

• Conclusion
Word Sense Disambiguation

- Expanded College English Test 4 Dictionary
  - English, Chinese (relative frequency), part-of-speech
  - 33,664 English-Chinese pairs and
    ~4k unique English words

- Baseline: always choose the most frequent relative Chinese translation
  - 100% of coverage as it always has a translation
  - Low accuracy as it lacks context modeling
Word Sense Disambiguation

- Approach 1: News Category
  - Pick the Chinese translation with the same category as the news article
  - E.g., “利息” => “interest” in Finance news

- Approach 2: Part-of-Speech (POS)
  - Pick up the Chinese translation with the same POS as the target English word
  - E.g., “book” => “书” (noun) and “预定” (verb)
WSD: Bing Translator Based Methods

- Approach 3: Substring Match

1. Bing Translator
2. Look up dictionary
3. Matched: 俱楽部 (Bing) is a substring of 俱楽部 (dict)

Limited by dictionary coverage!
WSD: Bing Translator Based Methods

1. Bing Translator

… into the world's **top** 40 clubs

2. Look up dictionary

noun: 顶部, 顶端, 顶,
adj: 上面的, 最大的, ...
verb: 盖

3. No output using substring match

…进入世界 **顶级** 40名俱乐部
WSD: Bing Translator Based Methods

• Approach 4: Relaxed Match

1. Bing Translator + Chinese Segmentation

2. Look up dictionary

3. Relaxed Match: 顶级 (Bing) is superset of 顶 (dict)

... into the world's top 40 clubs

... 进入 世界 顶级 40名 俱乐部

noun: 顶部, 顶端, 顶, adj: 上面的, 最大的, ...
verb: 盖
WSD: Bing Translator Based Methods

1. Bing Translator + Chinese Segmentation
   ... state department spokeswomen said ...

2. Look up dictionary
   noun: 态, 国, 州 ...
   verb: 声明, 陈述, ... 发言 ...
   adj: 国家的

3. Two relaxed matches, both wrong
   国家 (Bing) => 国 (dict)
   发言人 (Bing) => 发言 (dict)
WSD: Bing Translator Based Methods

- Approach 5: Bing Alignment

1. Bing Translator + Bing Alignment

2. Look up dictionary

... state department spokeswomen said ...

Better - No output if the alignment is phrase to phrase

... 国家 部门 的 发言人 说 ...

noun: 态, 国, 州,...
verb: 声明, 陈述, ... 发言, ...
adj: 国家的
WSD: Evaluation

- Baseline
- 1. News Category
- 2. POS
- 3. Bing - Substring
- 4. Bing - Relaxed
- 5. Bing - Align

Bar chart showing coverage and accuracy percentages for different methods.
Outline

• Introduction

• Translating
  Word Sense Disambiguation (WSD)

• Testing
  Distractor Generation

• Conclusion
What is a set of suitable distractors?

• Have the same form as the target word
  – Same POS tag
• Fit the sentence context
  – News category
• Have proper difficulty level according to user’s level of mastery
  – Difficult distractors are more semantically similar to the target words
Generating proper distractors

The difficulty level is measured by Lin distance between the target word and candidate distractor in WordNet

$$sim(t, d) = \frac{2 \times logP\left(lso(t, d)\right)}{logP(t) + logP(d)}$$

A distractor is deemed hard when its similarity to target word is above threshold (e.g., 0.1)
Distractor Generation

1. WordNews Hard: Same word form +
   Same news category +
   Semantically Similar

2. Random News: Same word form +
   Same news category

Vary the number of hard distractors based on user’s knowledge level

   – Beginner: two random + one hard
   – Intermediate: three hard
Human Evaluation

• Baseline
  – WordGap System (Knoop and Wilske, 2013)
    • Distractor: target’s synonyms of synonyms in WordNet

• Evaluation 1: WordGap vs. Random News
• Evaluation 2: WordGap vs. WordNews Hard
Human Evaluation

22. Most sex workers that Hail-Jares encounters through street-based outreach are not in it for a _____, or because they lack the drive to succeed, she says. *

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>lark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>frolic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>runaround</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cavort</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>remember</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>film</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>architect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One is the target word, three are from WordGap, and the other three are from WordNews Hard or Random News.
# Human Evaluation

- **WordGap vs. Random News.**

<table>
<thead>
<tr>
<th></th>
<th># of wins</th>
<th>Avg. Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordGap</td>
<td>27</td>
<td>3.84</td>
</tr>
<tr>
<td>Random News</td>
<td>23</td>
<td>4.10</td>
</tr>
</tbody>
</table>

- **WordGap vs. WordNews Hard.**

<table>
<thead>
<tr>
<th></th>
<th># of wins</th>
<th>Avg. Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordGap</td>
<td>21</td>
<td>4.16</td>
</tr>
<tr>
<td>WordNews Hard</td>
<td>29</td>
<td>3.49</td>
</tr>
</tbody>
</table>

Lower scores are better
Conclusion

• **WordNews**: a Chrome extension enabling interactive vocabulary learning when reading online news

  ![Image of WordSense Disambiguation](image1.png)
  ![Image of Distractor Generation](image2.png)

  *Word Sense Disambiguation based on Machine Translation*
  *Distractor Generation based on news context and semantic similarity*

• **Future work**
  – Mobile client and longitudinal user study

Thanks for listening!