Interactive Second Language Learning from News Websites

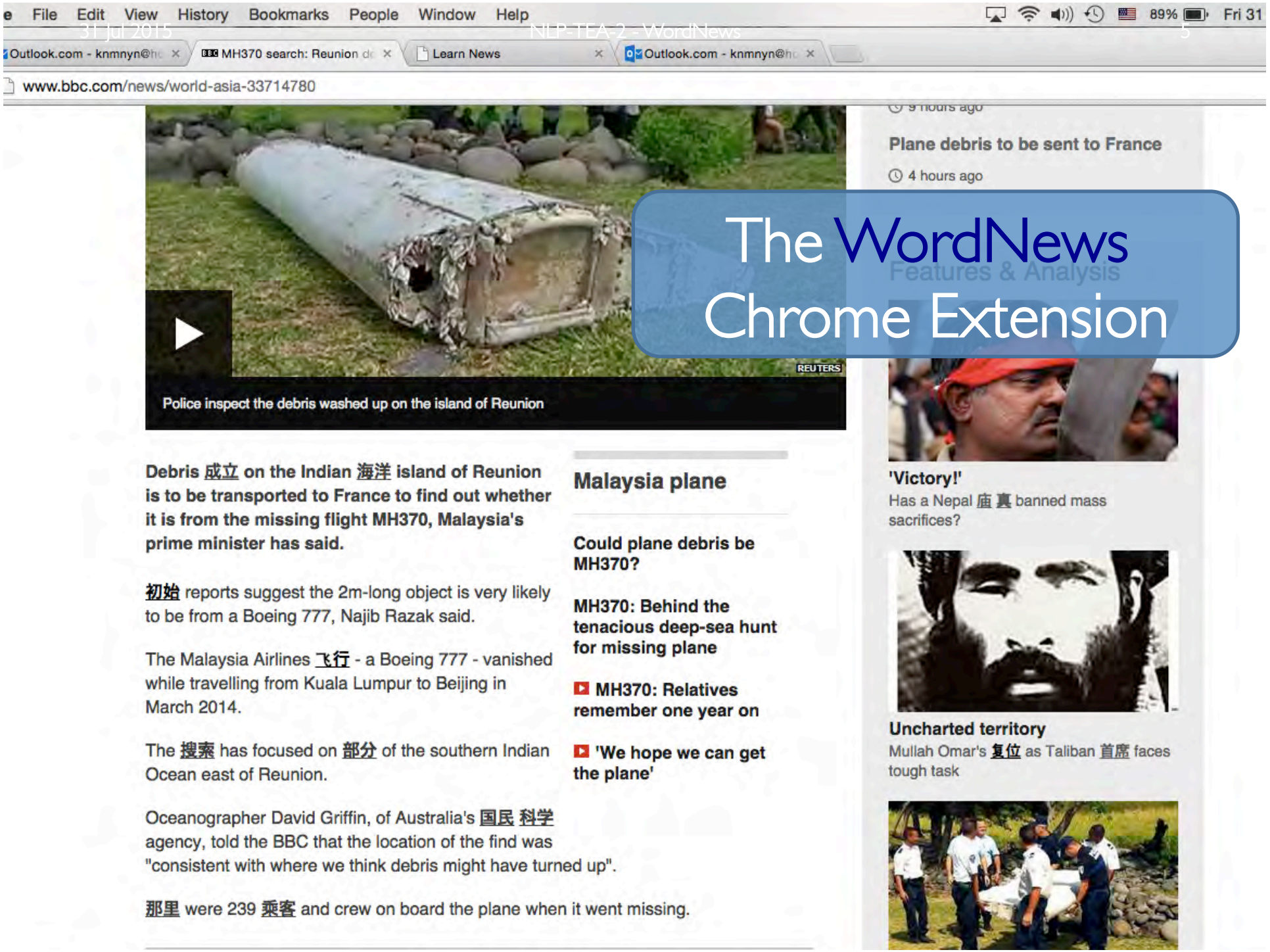
Tao Chen, Naijia Zheng, Yue Zhao,
Muthu Chandrasekaran and Min-Yen Kan
kanmy@comp.nus.edu.sg
Slides available at: dwz.cn/kan-nlptea2


Slides available at:
dwz.cn/kan-nlptea2
Formal language learning is time-consuming, and learning materials are often limited.
Bonjour
привет
Hola
Hello
您好
Ni hao
Olá
Image courtesy: skyscanner.net


…but we also spend over 35B hours every month keeping up with the news
Image courtesy: anid.com.br

WordNews

A browser extension for vocabulary learning when reading online news



The WordNews Chrome Extension

News Context

- Identify the news category by URL pattern

  `http://edition.cnn.com/2015/07/08/entertainment/feat-tom-selleck-droughtshaming-water/index.html`

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

- Classify words based on category document frequency

  E.g., "superstar" belongs to "Entertainment"

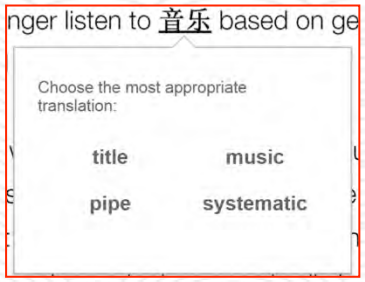
- For both English and Chinese news and words

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

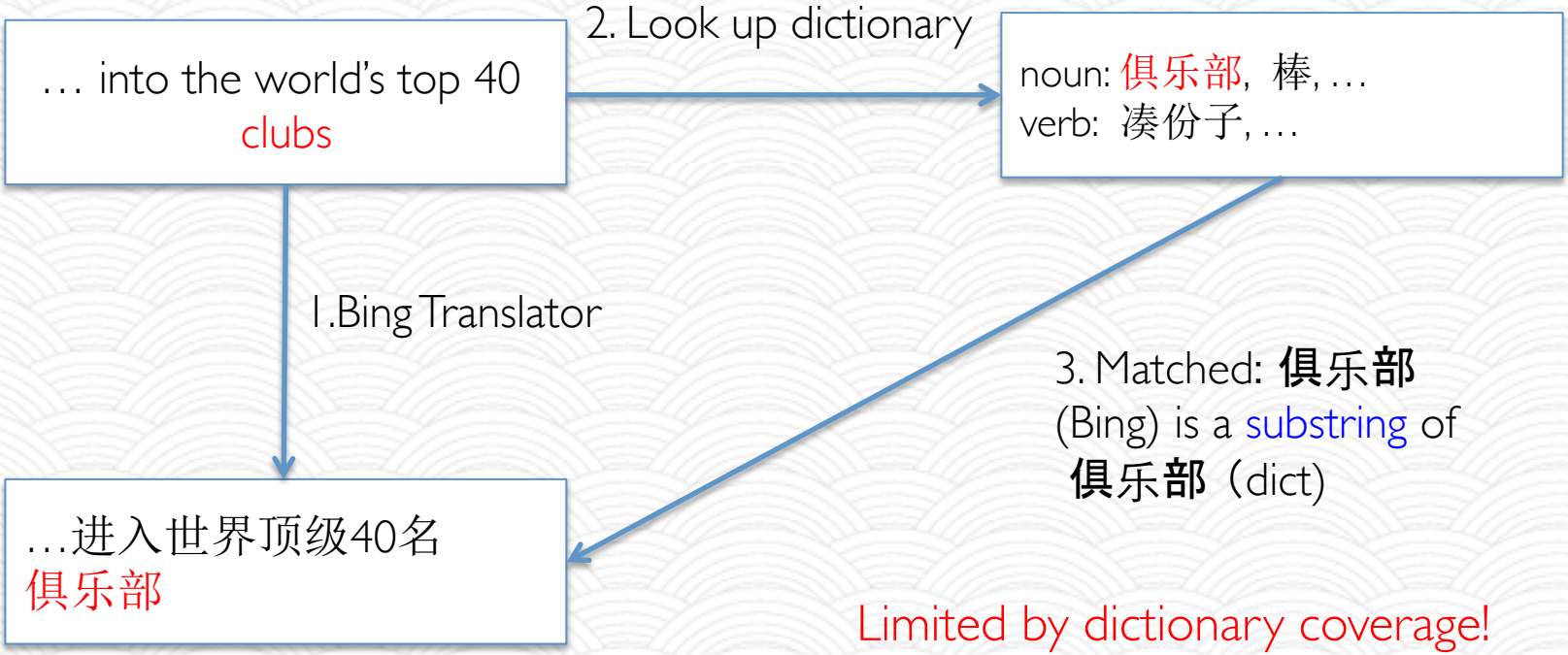
- Baseline: always choose the most frequent relative Chinese translation
  - 100% of **coverage** as it always has a translation
  - Low **accuracy** as it lacks context modeling

Word Sense Disambiguation

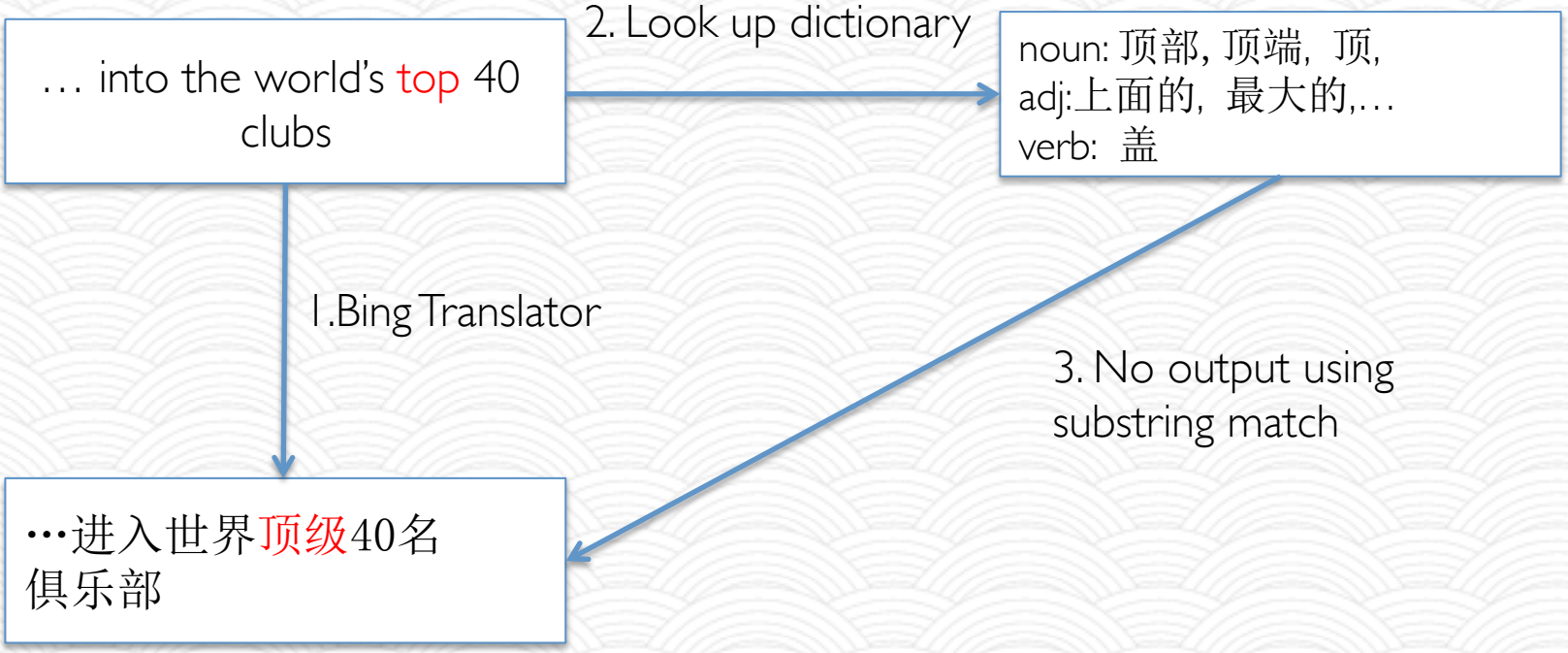
- Approach 1: News Category
  - Pick the Chinese translation with the same category as the news article
  - E.g., "利息" =>"interest" in Finance news
- Approach 2: Part-of-Speech (POS)
  - Pick up the Chinese translation with the same POS as the target English word
  - E.g., "book" => "书" (noun) and "预定" (verb)

WSD: Bing Translator Based Methods

- Approach 3: Substring Match

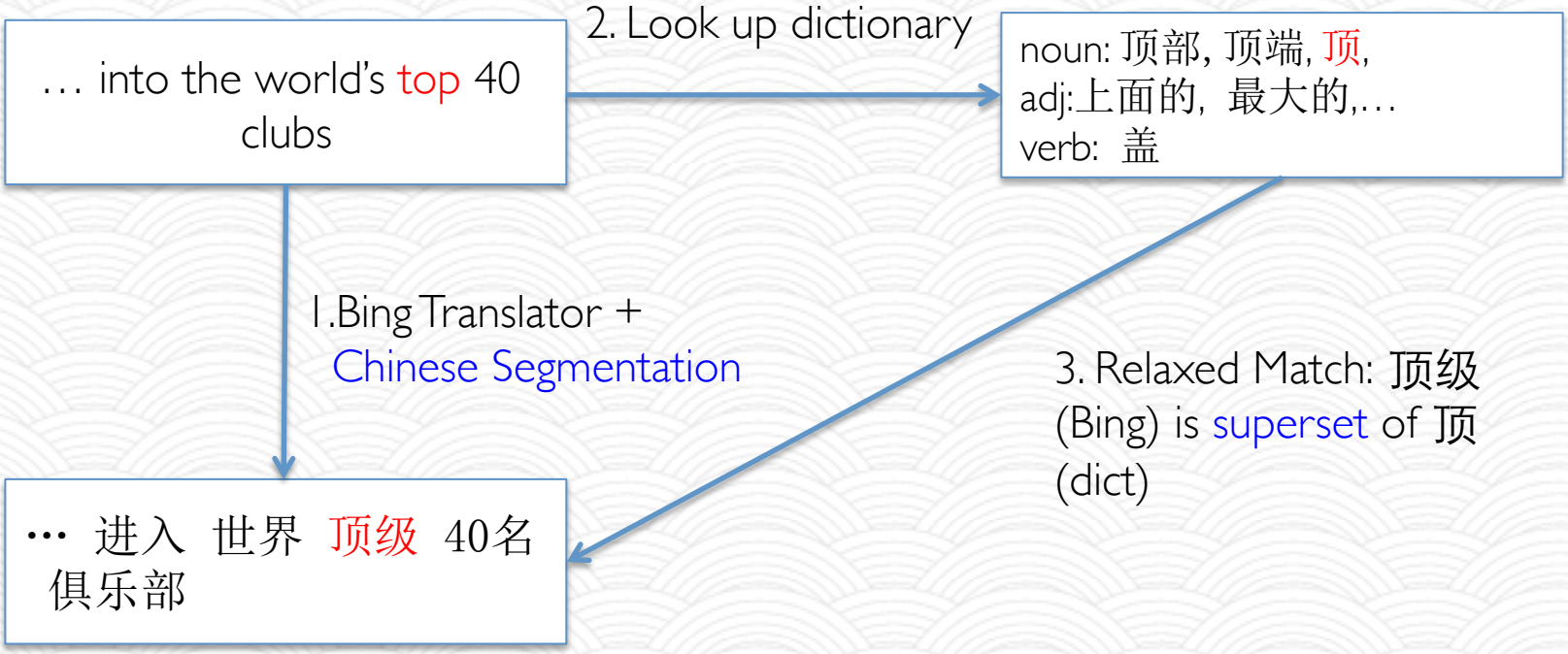


WSD: Bing Translator Based Methods

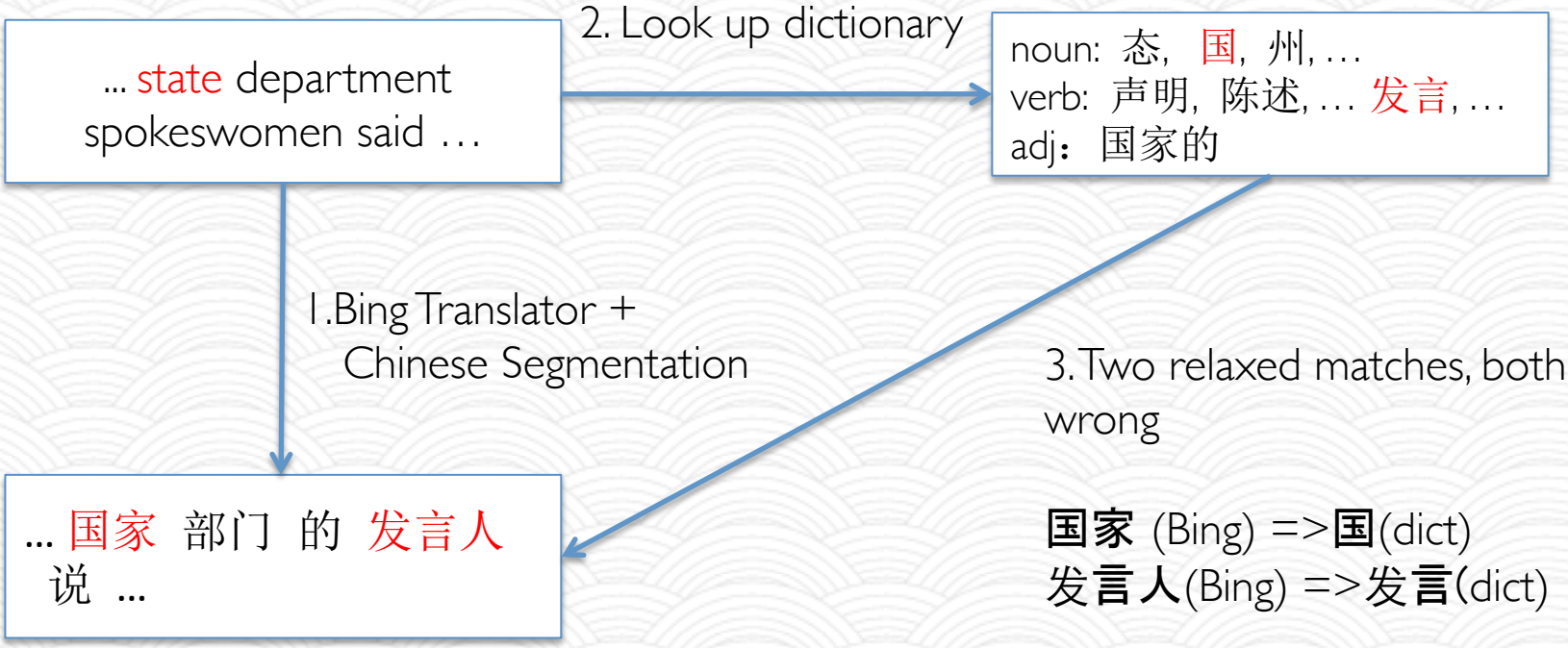


WSD: Bing Translator Based Methods

- Approach 4: Relaxed Match

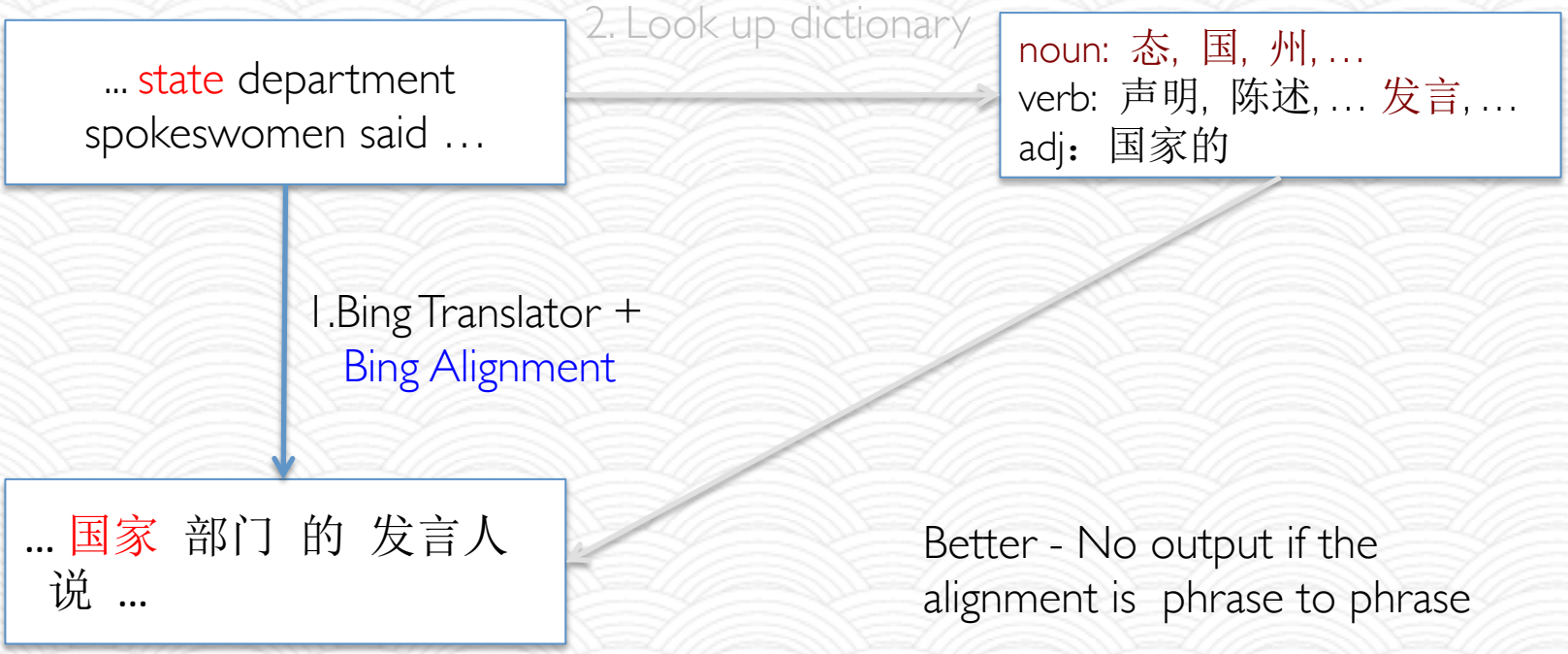


WSD: Bing Translator Based Methods

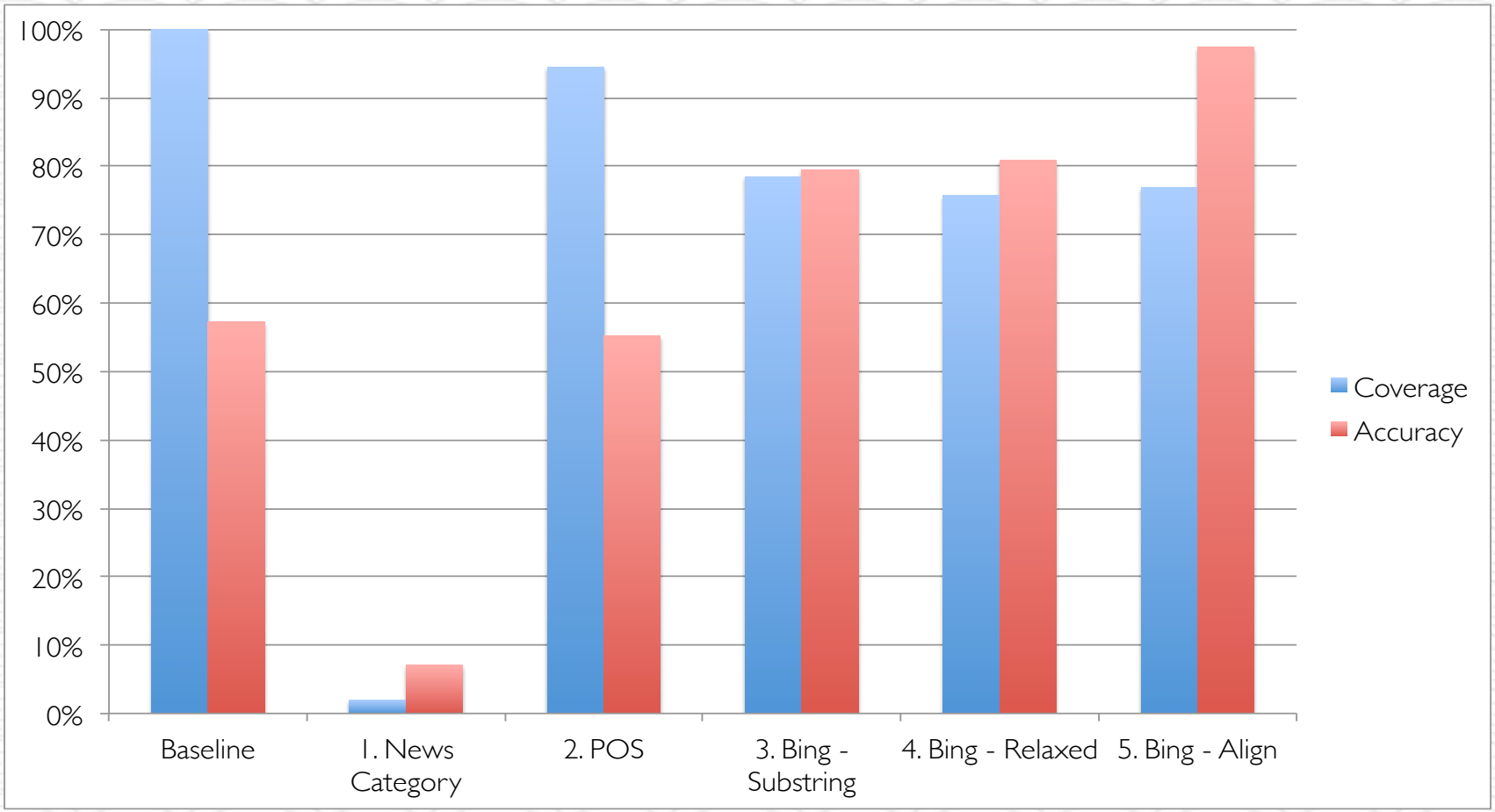


WSD: Bing Translator Based Methods

- Approach 5: Bing Alignment



WSD: Evaluation

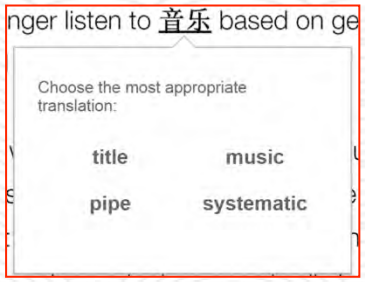


Outline

- 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

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

Lowest common subsumer synset

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. *

	1	2	3	4	5	6	7
lark							
frolic							
runaround							
cavort							
remember							
film							
architect							

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.

	# of wins	Avg. Score
WordGap	27	3.84
Random News	23	4.10

Lower scores are better

- WordGap vs. WordNews Hard.

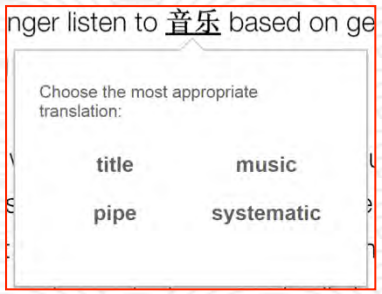
	# of wins	Avg. Score
WordGap	21	4.16
WordNews Hard	29	3.49

Slides available at:
dwz.cn/kan-nlptea2

Conclusion

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

Word Sense Disambiguation based on Machine Translation



Distractor Generation based on news context and semantic similarity

- Future work
  - Mobile client and longitudinal user study

Image Courtesy: www.heley-int.com