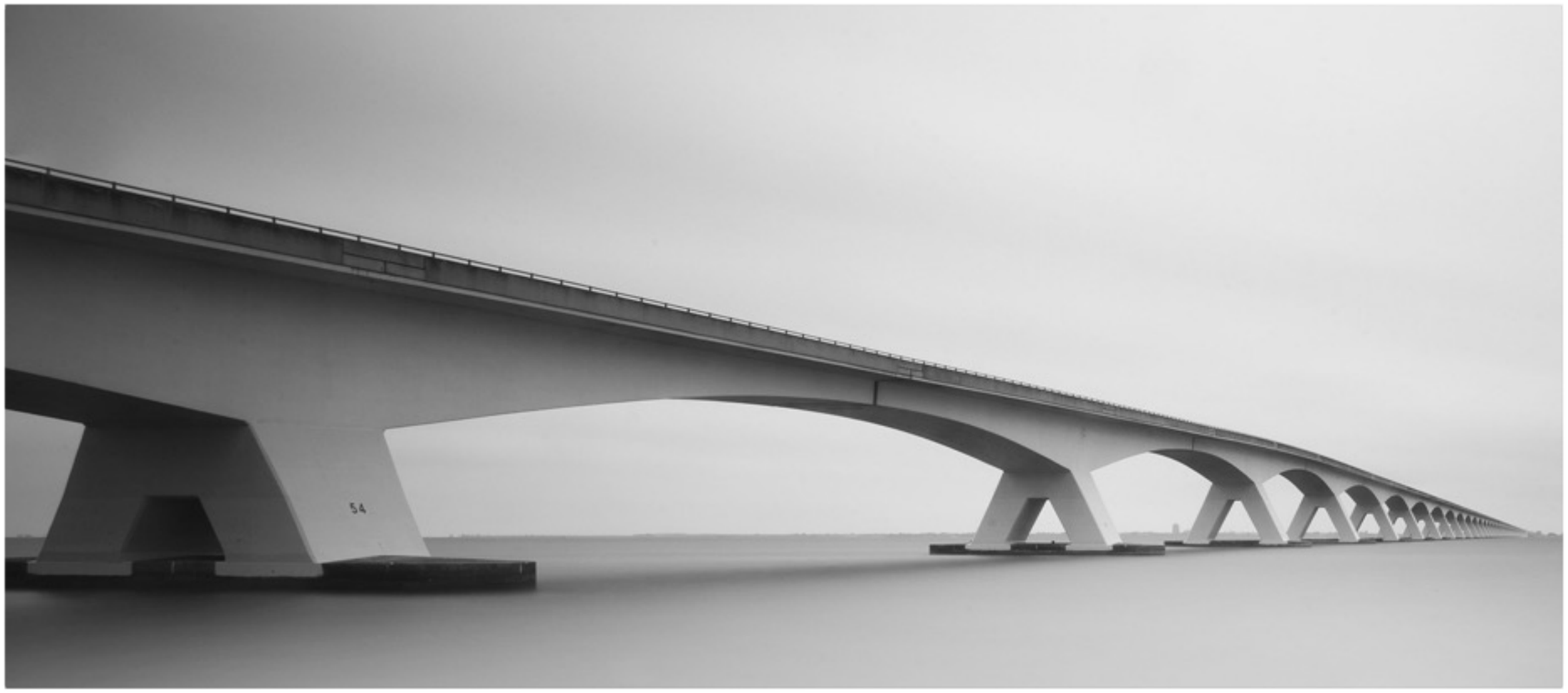


Multilingual Task a start-off

Speaker: Wenjing Fang (Bean)
Contact: fangwenjing_scu@126.com

Content

- Compound nouns equivalents extraction
  - Task description
  - Solution
    - Evaluate
- Brief introduction: bilingual dictionary
- Conclusion
- Our proposal: mining cultural difference

Content

- Compound nouns equivalents extraction
  - Task description
  - Solution
    - Evaluate
- Brief introduction: bilingual dictionary
- Conclusion
- Our proposal: mining cultural difference

Task: Learn from the Title

Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora

- Compound noun
- Def: two or more words join together to make a noun.
- Forms: three forms
  - Closed form: softball, redhead, makeup, and keyboard.
  - Hyphenated form: six-pack, five-year-old, and son-in-law.
  - Open form: post office, upper class, and attorney general.

Task: Learn from the Title

Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora

- Compound noun
- Def: two or more words join together to make a noun.
- Forms: three forms
- Patterns:

  - noun/adjective - snow white
  - noun/noun - toothpaste, football, fish tank
  - noun/preposition (adverb) - love-in, hanger on, passer-by
  - noun/verb - haircut, browbeat, rainfall
  - preposition/adjective - over-ripe
  - preposition (adverb)/noun - underground, underworld, bystander, onlooker

Task: Learn from the Title

Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora

- Compound noun
  - Def: two or more words join together to make a noun.
  - Forms: three forms
  - Patterns: e.g
- Non-parallel: parallel case (subtitle)
- Similarity: method used

Task: why

- Why compound noun
  - Common case in multiple words expressions
  - NL feature: adapt all the time :)
- Why non-parallel
  - Cross-lingual limits: position and grammar
  - Short of parallel corpus
- Description:
  - Locate translation equivalents by similarity

Content

- Compound nouns equivalents extraction
  - Task description
  - Solution
    - Evaluate
- Brief introduction: bilingual dictionary
- Conclusion
- Our proposal: mining cultural difference

Solution

- Intuition & steps
- Co-occurrence
- Word embedding
- Similarity

Intuition

- Think about a case:
  - How we guess an unknown word
  - Why we can understand free order languages
- Power of context
  - Representation: co-occurrence
  - Example: 営業利益 *eigyo rieki*, → *operating profits*, *business interest*

- ... its fourth-quarter *operating profit* will fall short of expectations ...
- ... the powerful coalition of *business interests* is pumping money into advertisements ...

Steps

- Raw corpus

POS Patterns

- Extract compound nouns

Component word

- Select candidates

Context Vector

- Similarity ranking

Solution

- Intuition & steps
- Co-occurrence
- Word embedding
- Similarity

Co-occurence

- Two types
- Sentence-level
  - word(syntacticly independent): noisy
  - syntactic dependence
- Semantic attributes
  - Abstract & representation
  - Lexicon corpus

Solution

- Intuition & steps
- Co-occurrence
- Word embedding
- Similarity

Word embedding

- Mapping words to context vectors of real numbers
- Two elements
  - What dimension
  - How to measure a real number

Word embedding

- Mapping words to context vectors of real numbers
- Two elements
- Sentence-level
- vocabulary space

$$\boldsymbol{c}_{w1}(t) = (\mu_w(t, r_1), ..., \mu_w(t, r_n)) \qquad (4)$$

Word embedding

- Mapping words to context vectors of real numbers
- Two elements
- Sentence-level
  - vocabulary space
  - measure

log likelihood ratio

$$\mu_w(t,r) = \begin{cases} L(t,r) & : f(t,r) \neq 0 \\ 0 & : f(t,r) = 0 \end{cases} \quad (1)$$

$$\begin{aligned} L(t,r) &= \sum_{i,j \in 1,2} k_{ij} \log \frac{k_{ij} N}{C_i R_j} \\ &= k_{11} \log \frac{k_{11} N}{C_1 R_1} + k_{12} \log \frac{k_{12} N}{C_1 R_2} \\ &+ k_{21} \log \frac{k_{21} N}{C_2 R_1} + k_{22} \log \frac{k_{22} N}{C_2 R_2} \end{aligned} \quad (2)$$

$$\begin{aligned} k_{11} &= f(t,r) \\ k_{12} &= f(t) - k_{11} \\ k_{21} &= f(r) - k_{11} \\ k_{22} &= N - k_{11} - k_{12} - k_{21} \\ C_1 &= k_{11} + k_{12} \\ C_2 &= k_{21} + k_{22} \\ R_1 &= k_{11} + k_{21} \\ R_2 &= k_{12} + k_{22} \end{aligned} \quad (3)$$

Word embedding

- Mapping words to context vectors of real numbers
- Two elements
- Sentence-level
- vocabulary space
- measure: log likelihood ratio
- Syntactic dependence

$$f'(t,r) = wf(t,r)$$

$$w = 1 + \frac{f_d(t,r)}{f(t,r)} * const$$

Solution

- Intuition & steps
- Co-occurrence
- Word embedding
- Similarity

Similarity

- Similarity : vector
  - Cosine similarity
- By the way: distance
  - Euclidean distance

Content

- Compound nouns equivalents extraction
  - Task description
  - Solution
    - Evaluate
- Brief introduction: bilingual dictionary
- Conclusion
- Our proposal: mining cultural difference

Evaluate

- Manually choose 393 pairs
- Precision table

sets	corpora	word1 (c_{w1})	word2 (c_{w2})	attr (c_a)
[1H]	NIK-WSJ	73.4	74.2	66.4
[1L]	NIK-WSJ	53.3	53.3	43.0
[1]	NIK-WSJ	63.0	63.4	54.2
[2]	NIK-WSJ	71.1	72.6	65.9
[2]	NIK-REU	71.9	71.9	66.7
[2]	MAI-WSJ	58.5	58.5	63.7
[2]	MAI-REU	57.0	56.3	65.2

Content

- Compound nouns equivalents extraction
  - Task description
  - Solution
    - Evaluate
- Brief introduction: bilingual dictionary
- Conclusion
- Our proposal: mining cultural difference

Bilingual dictionary

- Task: word-word translation
- Method: word embedding
  - Syntacticly independent v.s. Syntacticly dependent
  - Real number: probability of word co-occurrence

Bilingual dictionary

- Task: word-word translation
- Method: word embedding
- Evaluation

- MMR: [0,1] the ↑ the better, metric for ranking

$$MMR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}, \quad rank_i = \begin{cases} r_i, \text{ if } r_i < n \\ 0, \text{ otherwise} \end{cases}$$

n means top n evaluation

r_i means the rank of correct translation in top n ranking

N means the total number of words for evaluation

Content

- Compound nouns equivalents extraction
  - Task description
  - Solution
    - Evaluate
- Brief introduction: bilingual dictionary
- Conclusion
- Our proposal: mining cultural difference

Conclusion

- Multi-lingual task
  - corpus: mass media & lexicon
  - representation: word embedding
- Word vector: syntacticly dependent v.s word-base
- Utilising wiki: process tips
  - title&links
  - articles: gloss
  - process: stemmer, stop words

Content

- Compound nouns equivalents extraction
  - Task description
  - Solution
    - Evaluate
- Brief introduction: bilingual dictionary
- Conclusion
- Our proposal: mining cultural difference

Think more ?

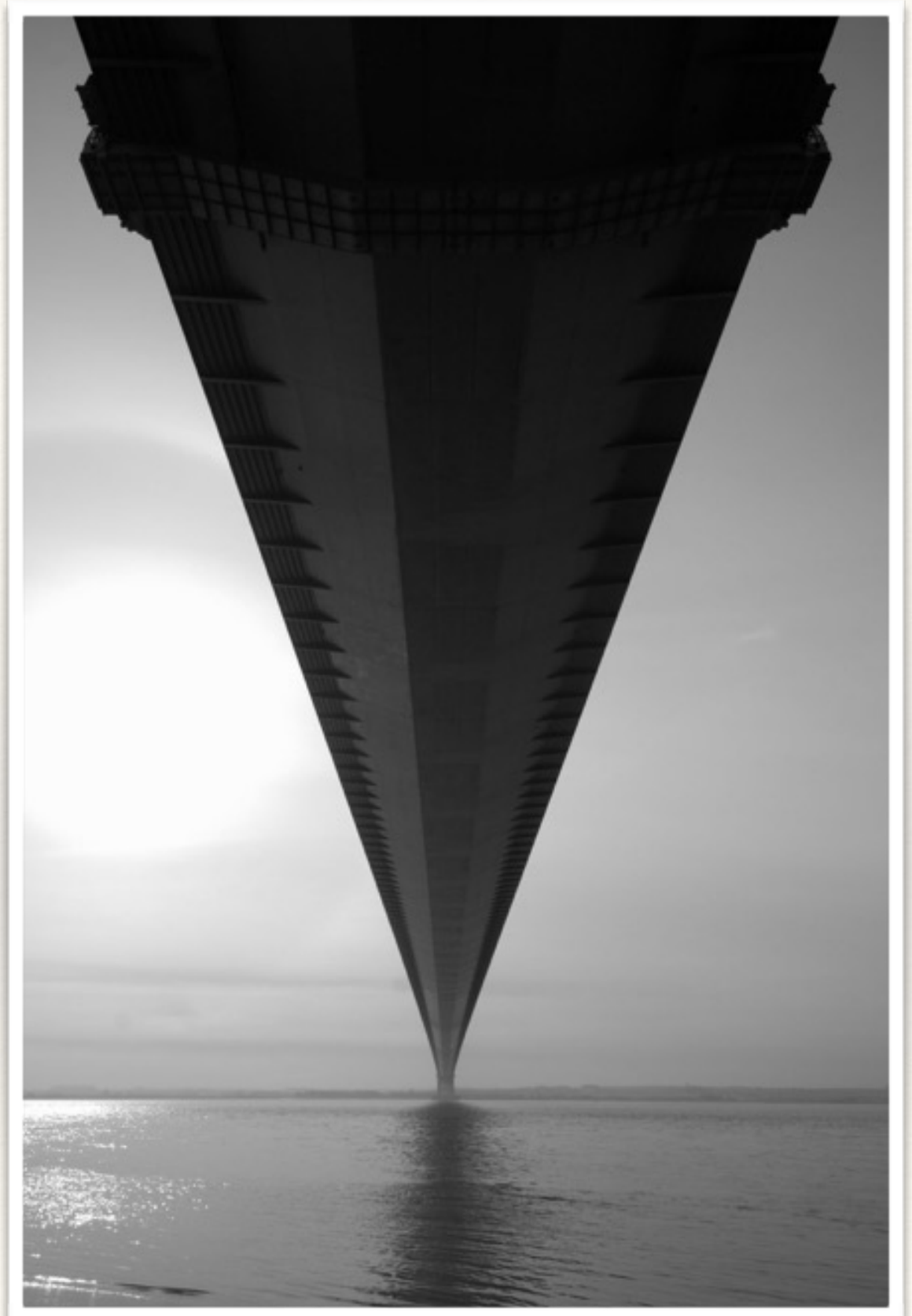
- Mining similarity
- Mining changes ?
  - Find new terms
  - Use trend of a term
- Mining difference: our proposal
  - case 1: translation 啤酒炸鸡 & 白酒花生
  - case 2: connotation 龙 & dragon

Q & A

Quiz 1:
What is word embedding?

Quiz 2:
MMR: which is better? 0.1 or 0.3.

Quiz 3:
How to measure similarity of vectors ?



Thanks.

Think up your case ,

start off with your multi-lingual journey~