Corpus-based Chinese–Hungarian dictionary building

Judit Ács

judit@aut.me.hu
Mathematical Linguistics Group

Nov 21, 2014

Motivation

Motivation

- Machine readable dictionaries

Motivation

- Machine readable dictionaries
- Manual building takes too much effort

Motivation

- Machine readable dictionaries
- Manual building takes too much effort
- Zhongxiu Liu (Aurora) and Yidi Zhang under the supervision of Attila Zséder

Method

Method

1. Data collection

Method

1. Data collection
2. Preprocessing

Method

1. Data collection
2. Preprocessing
3. Sentence alignment

Method

1. Data collection
2. Preprocessing
3. Sentence alignment
4. Dictionary extraction

Data collection and preprocessing

Data collection and preprocessing

- Freely available books

Data collection and preprocessing

- Freely available books
- Books become public domain after the author's death + 50/70 years

Data collection and preprocessing

- Freely available books
- Books become public domain after the author's death + 50/70 years
- 50+ books collected in Chinese and Hungarian

Data collection and preprocessing

- Freely available books
- Books become public domain after the author's death + 50/70 years
- 50+ books collected in Chinese and Hungarian
- Chinese text was segmented to words

Data collection and preprocessing

- Freely available books
- Books become public domain after the author's death + 50/70 years
- 50+ books collected in Chinese and Hungarian
- Chinese text was segmented to words
- Hungarian text was stemmed

Sentence alignment

Sentence alignment

- One sentence is often translated to several sentences

Sentence alignment

- One sentence is often translated to several sentences
- Hunalign (Varga et al. 2005)

Sentence alignment

- One sentence is often translated to several sentences
- Hunalign (Varga et al. 2005)
- Many-to-many alignment between corresponding Chinese and Hungarian segments

Sentence alignment

- One sentence is often translated to several sentences
- Hunalign (Varga et al. 2005)
- Many-to-many alignment between corresponding Chinese and Hungarian segments
- hu–zh segment pairs with appr. the same meaning

Dictionary extraction

Dictionary extraction

- Similar words often appear in corresponding segments

Dictionary extraction

- Similar words often appear in corresponding segments
- Frequent words appear in many segments

Dictionary extraction

- Similar words often appear in corresponding segments
- Frequent words appear in many segments
- Word co-occurrence statistics: Dice coefficient

Dictionary extraction

- Similar words often appear in corresponding segments
- Frequent words appear in many segments
- Word co-occurrence statistics: Dice coefficient
- Hundict (Ács et al. 2013)

Dice coefficient

Dice coefficient

Between the Hungarian word w_{hu} and the Chinese word w_{zh}:

Dice coefficient

Between the Hungarian word w_{hu} and the Chinese word w_{zh}:

$$Dice(w_{hu}, w_{zh}) = \frac{2|W_{hu} \cap W_{zh}|}{|W_{hu}| + |W_{zh}|},$$

Dice coefficient

Between the Hungarian word w_{hu} and the Chinese word w_{zh}:

$$Dice(w_{hu}, w_{zh}) = \frac{2|W_{hu} \cap W_{zh}|}{|W_{hu}| + |W_{zh}|},$$

- where W_{hu} is the set of segment pairs in which w_{hu} appears

Dice coefficient

Between the Hungarian word w_{hu} and the Chinese word w_{zh}:

$$Dice(w_{hu}, w_{zh}) = \frac{2|W_{hu} \cap W_{zh}|}{|W_{hu}| + |W_{zh}|},$$

- where W_{hu} is the set of segment pairs in which w_{hu} appears
- W_{zh} is the set of segment pair in which w_{hu} appears

Example

于是 三百 亩 小麦 ， 一百 亩 马铃薯 ， 一百五 十 亩 苜蓿 ， 没有 一 亩 地 荒废 了 。

Akkor lesz háromszáz gyeszjatyina búzája , száz burgonyája , százötven lóheréje , s egyetlen kimerült gyeszjatyinája sem .

Results

Results

- 34k words

Score	Hungarian	Chinese	English
1.0	szenátor	枢密官	senator
1.0	csillagász	天文学家	astronomer
0.93	mérnök	工程师	engineer
0.75	huszonnegyedik	二十四日	(on the) 24th

Conclusions

Conclusions

- Medium to high quality translations

Conclusions

- Medium to high quality translations
- Suitable for machine translation, information retrieval etc.

Conclusions

- Medium to high quality translations
- Suitable for machine translation, information retrieval etc.
- Evaluation is needed

Conclusions

- Medium to high quality translations
- Suitable for machine translation, information retrieval etc.
- Evaluation is needed
- Larger corpora: web pages, Wikipedia

Conclusions

- Medium to high quality translations
- Suitable for machine translation, information retrieval etc.
- Evaluation is needed
- Larger corpora: web pages, Wikipedia
- Comparable corpora

Bibliography

- Judit Ács, Katalin Pajkossy, and András Kornai.

  Building basic vocabulary across 40 languages.

  In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

- Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor Tron.

  Parallel corpora for medium density languages.

  In N Nicolov, K Bontcheva, G Angelova, and R Mitkov, editors, *Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05*, pages 247–258. Benjamins, Amsterdam, 2007.