Building a Bilingual Lexicon Using Phrase-based Statistical Machine Translation via a Pivot Language

Takashi Tsunakawa* Naoaki Okazaki* Jun'ichi Tsujii*/**

*Department of Computer Science, Graduate School of Information Science and Technology, University of Tokyo, Japan

**School of Computer Science, University of Manchester / National Centre for Text Mining, UK

Motivation

- Bilingual lexicons from/to English are usually richer than those within non-English languages

→ **Connecting non-English terms via English** (***pivotal approach***) could be a reasonable approach

- We apply an **SMT framework** for the pivotal approach

→ Characteristics between source and target languages can be directly modeled as the feature functions

Problems of existing methods

- Merging two bilingual lexicon via identical English terms can associate a few bilingual lexicons (**low *utilization ratio***)
  - In the right example, only *Down's syndrome* can be associated
- Characteristics between languages are unused
  - Chinese and Japanese terms share identical/similar characters.

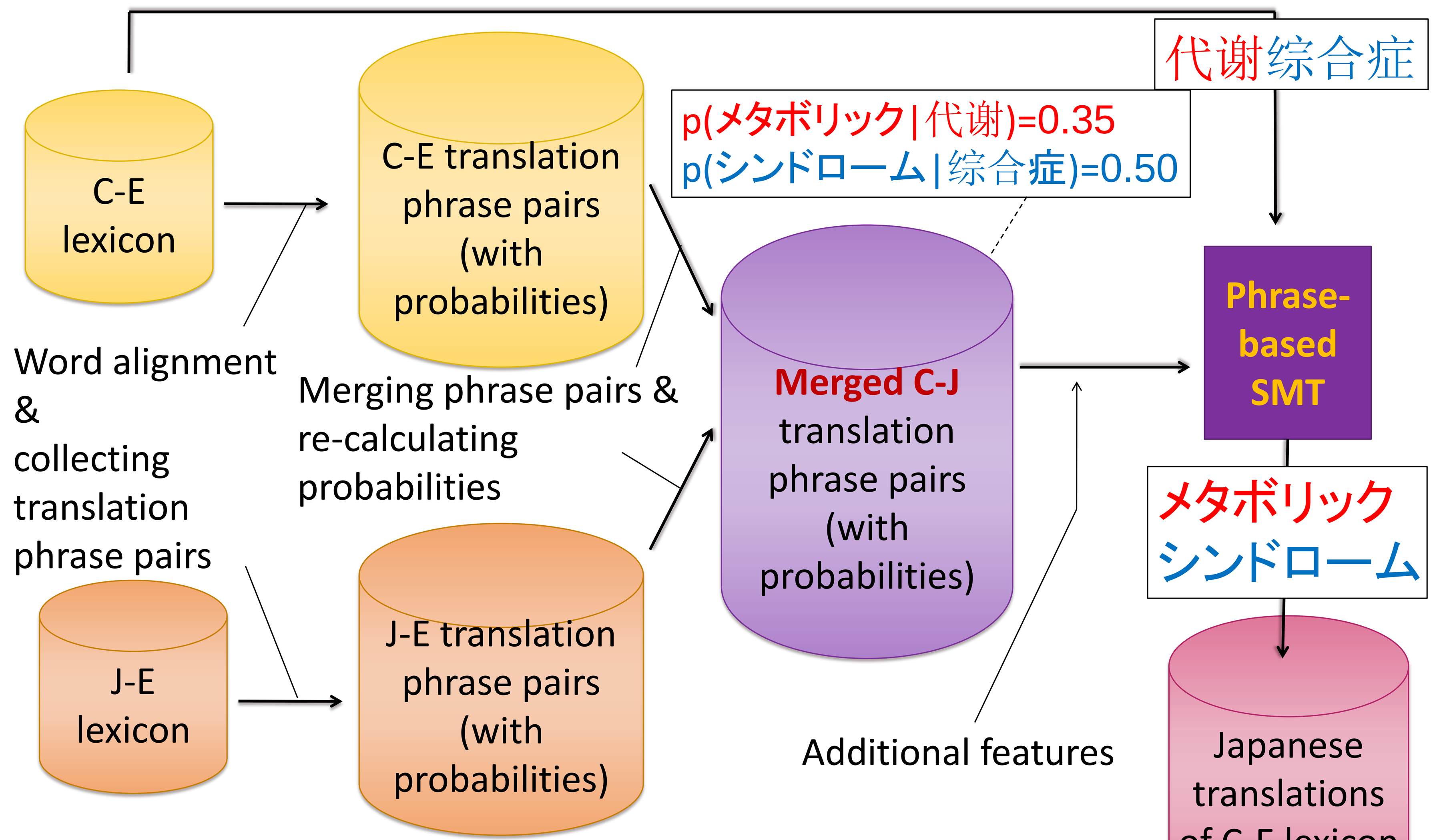
Merging two bilingual lexicons

Chinese	English
代谢综合症	metabolic syndrome
道恩综合症	Down's syndrome
抗体缺乏综合症	antibody-deficiency syndrome

Japanese	English
代謝異常	metabolic disorder
ダウン症候群	Down's syndrome
メタボリック	metabolic
シンドローム	syndrome
抗体欠乏	antibody deficiency

Chinese	Japanese
道恩综合症	ダウン症候群
代谢综合症	メタボリックシンドローム
代谢综合症	代謝症候群
抗体缺乏综合症	抗体欠乏症候群

Framework



SMT model for term translation

For each input term $\overline{w}_C$, we obtain the (10-)best translation by the log-linear model:

$$\hat{\overline{w}}_J = \arg\max_{\overline{w}_J} \sum_{m=1}^{M} \lambda_m h_m(\overline{w}_J, \overline{w}_C)$$

where h_m and λ_m denote the feature functions and their weights.

Features for the log-linear SMT modeling

- **Merged translation probabilities** (Utiyama & Isahara, 2007)

— Apply morphological analyzers, and obtain word alignments by GIZA++ (Och and Ney, 2003) for J-E and C-E lexicons

— Collect phrase pairs by *grow-diag-final* method (using Moses, Koehn et al., 2007), and calculate the translation probabilities by relative frequencies

$$h_p(\overline{w}_J, \overline{w}_C) = p(\overline{w}_J \mid \overline{w}_C) = \frac{1}{Z_C} \sum_{\overline{w}_E} p(\overline{w}_J \mid \overline{w}_E) p(\overline{w}_E \mid \overline{w}_C),$$

$$Z_C = \sum_{\overline{w}'_J} \sum_{\overline{w}_E} p(\overline{w}'_J \mid \overline{w}_E) p(\overline{w}_E \mid \overline{w}_C)$$

- **Normalized character-level edit distance** of two phrases

$$h_{ed}(w_J, w_C) = 1 - \frac{\text{Edit distance of Unicode characters}}{\text{Max. of the number of characters}}$$

— h_{ed}(後天性免疫不全症候群, 後天免疫缺乏症候群) = 1 – 3/10 = 0.7

- **Additional lexicon**

— The number of translation word pairs included in the additional bilingual lexicon

— h_{lex}(後天性免疫不全症候群, 获得性免疫缺陷综合症) = 3

if (免疫, 免疫), (不全, 缺陷), and (症候群, 综合症) are in the additional lexicon

Experiment

Utilization ratio: The ratio of Chinese terms translated into the other language

C-to-J	Utilization ratio
Exact matching	26.2%
Our method	72.8%

Translation performance on the test set*

*Test set consists of 500 Chinese and Japanese term pairs

Features	BLEU	NIST	Acc.
Base features	0.4519	7.4060	0.676
with edit distance	0.4670	7.4963	0.682
with additional lexicon	0.4800	7.5907	0.674
All	0.4952	7.7046	0.685

Used lexicons: All lexicons consist of technical terms

C-E: Wanfang Data Dictionary (C: 375,990 terms / E: 429,807 terms)

J-E: JST MT Dictionary (J: 465,563 terms / 418,044 terms)

Additional lexicon: EDR J-E-C lexicon (C: 90,605 terms / J: 94,928 terms)

Translation examples

C	耻骨联合
E	pubis symphysis
J	恥骨結合

C-to-J	恥骨 結合	○
E-to-J	結合 恥 (symphysis shame)	×

C	理想流体动力学
E	ideal fluid dynamics
J	理想流体力学

C-to-J	理想 流体 力学	○
E-to-J	理想 液 力学 (ideal fluid dynamics)	×

C	中间腰淋巴结
E	intermediate lumbar lymph nodes
J	中間腰リンパ節

C-to-J	中間 節 腰淋 (int. nodes [mistokenized])	×
E-to-J	中間 腰 リンパ 節	○

Conclusion

- The experiment demonstrated that our method improved the utilization ratio drastically and performed reasonable translations of given lexicons
- We also showed that features between the source and target languages are effective for improving the performance

- Future work
  - Improve Chinese word segmentation
  - Introduce a sophisticated Chinese character similarity model
  - Extract a source-target phrase table from corpora