Chapter 8: Conclusions and Future Work

In this Chapter, we conclude this thesis by summarizing the main contributions of our work and pointing out some possible future directions for research on named entity alignment.

8.1 Contributions

The main contributions of this thesis can be summarized as follows:

1. We have developed a statistical model-based approach to named entity translation. The parameters of the proposed models can be automatically learned from training data. Based on the proposed models, a framework is proposed to align named entity pairs in parallel corpora. While aligning a NE pair, the proposed phrase translation model can effectively reduce translation candidates to a limited set, instead of generating all permutations of translation candidates for a source NE.

2. We have introduced a transliteration model based on statistical learning approach to deal with common problems in machine transliteration. Compared with previous studies, the proposed approach does not require either a bilingual pronunciation dictionary or handcrafted similarity scores between transliteration units. Thus, the proposed approach can be easily ported to other language pairs as long as we have a bilingual proper name list for training the proposed models. Moreover, the approach directly measures the similarity score on the grapheme level, thus accelerating the matching process without the use of grapheme-to-phoneme phase.

3. We have presented extra language-specific modules to further improve the performance. A Chinese person name model is applied to associate a foreign name (or an abbreviated name) with its equivalent name in the Chinese language. An abbreviation handling method is involved to measure the similarity between a Chinese NE candidate and its abbreviation. An acronym expansion module is developed to expand acronyms to their original forms while aligning bilingual named entities.

4. Our experimental results have confirmed that the proposed statistical approach combining a phrase translation model, a transliteration model, and

language-specific features is highly effective on aligning named entity pairs in the undertaken corpora.

8.2 Future Work

We believe that the use of various kinds of linguistic information and domain-specific knowledge will lead to further improvement on alignment of bilingual named entities. Therefore our future work will focus on integrating other knowledge sources to the proposed framework for more robust alignment of bilingual NE pairs. The relevant linguistic information and knowledge includes

- Syntactic analysis and contextual information: By analyzing the contextual structures of aligned sentences, natural language parsing techniques can be used to improve the alignment of bilingual named entities
- Other language cues: In Chinese a noun phrase quoted by a pair of quotation often stands for a named entity. This language-specific information for NE identification can be used to further enhance the performance.
- Extra dictionaries: The current approach uses limited amount of dictionaries. We believe that the use of dictionaries for domain-specific NEs and common names can improve the performance.

- Repetitive co-occurrence feature: The repetitive co-occurrence feature of NE pairs in parallel corpus can also be utilized to confirm the alignment of NE pairs.

Future research should extend the study of the named entity alignment to other applications, such as bilingual document alignment. Moreover, we can apply the statistical model-based theory to NLP problems, such as noun phrase translation/alignment, and port the proposed method to other language pairs, such as Chinese and Japanese.