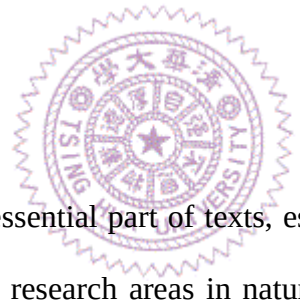
Chapter 1: Introduction

This dissertation introduces a new statistical model-based approach to aligning named entities in parallel corpora. This chapter contains the problem definition of named entity translation, practical difficulties of the task, and a brief description of our approach to solving the problem.



1.1 Motivation

Named Entities (NEs) are an essential part of texts, especially news texts. Extracting and translating NE is vital for research areas in natural language processing (NLP), such as machine translation (MT), cross-language information retrieval (CLIR), bilingual question answering (QA), and bilingual lexicon construction. There are three types of NEs (Chinchor, 1997):

- Entity names: organizations (ORG), persons (PER), and locations (LOC).
- Temporal expressions (dates, times).
- Number expressions (monetary values, percentages).

Temporal expressions and number expressions, being more common than entity names, can be generally described by grammar rules. On the contrary, PER, LOC, and ORG named entities are difficult to be handled with a fixed set of rules, since new entity names are created constantly. Thus, there is an increasing need to investigate

techniques for NE extraction and translation. In this dissertation, we will focus on extracting bilingual pairs of entity names[1].

Transforming NEs between two languages involves both translation and transliteration. In general, it is difficult for human translators to translate/transliterate unfamiliar person names, place names, and organization names. Specifically, person names are almost always transliterated. In this case, words can be decomposed into transliteration units (TUs) and transliterated (Chapter 4). For example, in the NE pair (Ada, 艾妲 "Ai Ta"), the alignments of the TU matching pairs are "a-ai," "d-t," and "a-a." Transformation of location names and organization names between two languages is typically performed via a combination of translation and transliteration. For example, in the NE pairs (Little Smoky River, 小斯莫基河) and (Carnegie Mellon University, 卡內基麥隆大學), "Little," "River," and "University" are translated as "小," "河," and "大學," respectively, and "Smoky," "Carnegie," and "Mellon" are transliterated as "斯莫基 (Ssu Mo Chi)," "卡內基 (Ka Nei Chi)," "麥隆 (Mai Lung)," respectively. There are no rules to tell which word should be translated or transliterated. For example, compared with (Little Smoky River, 小斯莫基河), "Smoky" in (Great Smoky Mountains, 大煙山脈) is translated into "煙," not transliterated into "斯莫基." Moreover, word order is not preserved when transforming organization names, as in the case of (University of California, 加州大學), where "大學" and "加州" are the translation of "University" and "California," respectively. From the above observations, it is obviously advantageous to combine phrase translation and transliteration while extracting or translating NEs.

[1] For simplicity, NEs referred in the rest of this paper are in fact entity names, unless it is otherwise notified.

1.2 Our Approach to the Acquisition of NE Pairs

Extracting bilingual NE pairs is an essential step toward retrieving NE translation knowledge from bilingual documents. Currently, English NE identifiers are well established and are already being used in commercial products, such as BBN's IdentiFinder (Bikel et al., 1999), whereas Chinese NE identifiers are still immature, partly due to the difficulty of Chinese word segmentation (Chen and Liu, 1992; Chien et al., 1999). To achieve the goal of extracting bilingual NEs from parallel corpora, we introduce here a new approach that aims to identify the NEs in an English sentence first and then extract corresponding Chinese equivalents from the aligned sentence by integrating a phrase translation model, a transliteration model, and extra language-specific sources (Lee et al., 2004a; Lee et al., 2004b; Lee et al., 2006b).

At run time, for an English NE identified in the source sentence, we extract NE pairs from aligned sentences as follows:

(1) transform the source NE into a set of possible translation candidates;

(2) find the set of candidates occurring in the target sentence to extract a set of possible NE equivalents;

(3) evaluate the set of possible NE equivalents;

(4) align the source and target NE pair with the highest probability.

For illustration purpose, we use an aligned sentence pair from magazine *Sinorama* (Sinorama, 2002), as shown in the following:

Source sentence:

"Forest policy has gradually turned its focus to conservation, the Taiwan Forestry Bureau this year completely prohibited the cutting of natural forest,

and great numbers of trees have been planted on the mountain slopes of grasslands where there is an extreme lack of vegetation."

Target sentence:

"雖然林業政策逐漸以保育為師，林務局也在今年初下令全面禁止砍伐天然林，且積極在缺乏植被的山坡地、草生地大規模種樹；"

In the above example, there are two English NEs, "Taiwan" and "Forestry Bureau." Although we aim at extracting bilingual NE pairs from parallel corpora, the translation is not always literal. In this example, the NE equivalent "台灣" of "Taiwan" is not in the target sentence at all. In the alignment of "Forestry Bureau" with "林務局," the set of translation candidates for "Forestry Bureau" can be {林業局，林務局，林業社，林務社，林業館，林務館，局林業，局林務,...}. After exploring the above set from the target sentence, the set of possible NE equivalents in Chinese is {林業，林務局，林，局}. Then, we evaluate the members in the equivalent set and find that the most likely translation equivalent of "Forestry Bureau" is "林務局." Therefore, the NE pair (Forestry Bureau, 林務局) can be correctly aligned. Formal description of the proposed approach will be given in Chapter 3.

1.3 A Preview of Subsequent Chapters

The remainder of this dissertation is organized as follows:

- Chapter 2 gives an overview of the recent literature related to our work, including NE identification, machine transliteration, bilingual lexicon construction, and NE alignment/translation.
- Chapter 3 describes the proposed phrase translation model for translating NE phrases in details. While aligning an NE pair, the proposed phrase translation model can effectively construct a limited set of NE translations, without generating all permutations of the combination of translations for each word in a source NE. The translation model can also incorporate other extra language-specific features to further improve the performance.
- Chapter 4 describes the proposed transliteration model for transliterating proper names in details. The proposed transliteration model can match TUs of bilingual proper nouns directly, without converting source words into phonetic symbols using a pronunciation dictionary or grapheme-to-phoneme rules for the source words.
- Chapter 5 explains a framework based on the proposed modules to acquire bilingual NE pairs from parallel corpora. We describe the overall process of the framework in detail with some illustrated examples.
- Chapter 6 presents the experimental setup and a quantitative assessment of the performance of NE alignment.

- Chapter 7 presents the experimental setup and a quantitative assessment of performance of transliteration alignment.
- Chapter 8 summarizes the dissertation and gives concluding remarks. Some future work will also be discussed.