Chapter 2: Related Work

Recently, NE extraction, alignment, and translation have been the subject of active research in MT, since NEs convey vital meaning in documents. In this chapter, we will review related work of our study, including:

- NE identification: There have been many studies on NE identification. This task aims at detecting NEs in texts and classifying them into corresponding types. Some NE identification systems, especially for English NE identification, are well established and have been used in commercial products already.
- Machine transliteration: Recently, the study of machine transliteration has been an important research topic. It is challenging to align phonetic equivalents with their source proper names, especially for those language pairs with different alphabets. We will review previous studies on machine transliteration on different language pairs here.
- Bilingual lexicon construction: There have been some studies on bilingual lexicon construction in the past few years. Most previous studies are based on

individual word translation. We will mainly describe the most representative study, IBM models, and explain the differences between models.

- NE alignment/translation: The problem of NE alignment/translation has drawn much attention in recent years. Since the work is most relevant to our study, we will introduce the work in details according to the two alternative approaches, symmetric and asymmetric manners. The state-of-the-art work will be described one by one.

2.1 NE Identification

In the past few years, a lot of research has been conducted in the task of the NE identification within monolingual documents. There are two major approaches to this task: rule-based approaches and statistical approaches. To recognize NEs, rule-based approaches rely heavily on handcrafting rules, while statistical approaches depend largely on learning data. Previous work based on rule-based approaches includes Black et al. (1998), Chen et al. (1998a), Fukumoto et al. (1998), Humphreys et al. (1998), Krupka and Hausman (1998), Mikheev et al. (1998), Kim and Woodland (2000), Chua and Liu (2002). Other studies based on statistical approaches includes hidden Markov models (HMMs) (Miller et al., 1998; Bikel et al., 1999; Zhou and Su, 2002; Sun et al., 2003), decision trees (Sekine et al., 1998), maximum entropy models

(Borthwick, 1999), support vector machines (Isozaki and Kazawa, 2002; McNamee and Mayfield, 2002; Takeuchi and Collier, 2002; Solorio and López, 2004), boosting (Carreras et al., 2002; Collins, 2002; Tsukamoto et al., 2002; Wu et al., 2002), and transformation-based learning (Black and Argyrios, 2002). The rule-based LTG system (Mikheev et al., 1998) and the statistical BBN system (Bikel et al., 1999) are two representative systems for rule-base approaches and statistical approaches respectively. In the LTG system, a 5-stage procedure combining grammar rules and partial matching techniques is applied to identify NEs. In summary, in the first step the system applies “sure-fire rules” based on known corporate designator (Ltd., Inc., etc.), person title (Mr., Dr., etc.), and definite contexts (in/at LOC, etc.). Then the system uses partial match testing, other contextual information, and gazetteers to identify NEs. In the BBN system proposed by Bikel et al. (1999), NE identification is regarded as a classification task modeled by HMM. Each word in a text is assigned to either one of desired NE classes (PER, LOC, ORG, etc.) or the class NOT-A-NAME to indicate “none of the desired classes.” The states of the proposed HMM are organized into regions. Each region corresponds to either one desired class or NOT-A-NAME. Within each of the regions, every word is represented by a state in a bigram language model. The generation of words and name-classes is performed in three steps: (1) select a name-class; (2) generate the first word within the name-class;

(3) generate all subsequent words within the current name-class. The model parameters are automatically trained from large hand labeled corpora. Obviously, the major drawback of the rule-based approaches is that the construction of an NE identifier is labor-intensive. On the other hand, the statistical approaches have the advantages of portability and robustness.

2.2 Machine Transliteration

Although much work on the NE identification within monolingual documents has been reported, little work has been reported on NE translation. NE translation is closely related to machine transliteration, which is the essential part of NE translation systems. Machine transliteration is classified into two types based on transliteration direction. Transliteration is the process that converts an original word in the source language into an approximate phonetic equivalent in the target language, whereas back-transliteration is the reverse process that converts the transliteration back into its original word.

Recently, machine transliteration has been studied by numerous researchers working with various language pairs, including English/Arabic (Stalls and Knight, 1998), English/Chinese (Chen et al., 1998b; Wan and Verspoor, 1998; Lin and Chen, 2002; Lee and Chang, 2003; Lee et al., 2003), English/Japanese (Knight and Graehl,

1998; Tsuji, 2002), and English/Korean (Lee and Choi, 1997; Kang and Choi, 2001; Oh and Choi, 2002). We summarize these studies as follows.

Lee and Choi (1997) proposed an automatic learning procedure for English-to-Korean transliteration with limited evaluation. Chen et al. (1998) proposed a method for Chinese-to-English back-transliteration. In that heuristic approach, letters commonly shared between a romanized Chinese word and an original English word are considered. The model is also enhanced with pronunciation rules. Knight and Graehl (1998) explored a generative model for Japanese-to-English back-transliteration based on the source-channel framework. Stalls and Knight (1998) extended that approach to Arabic-to-English back-transliteration. Wan and Verspoor (1998) proposed a method for English-to-Chinese place name transliteration based on heuristic rules for relationships between English phonemes and the Chinese phonetic system. Kang and Choi (2001) proposed a method based on decision trees to learn transliteration and back-transliteration rules between English and Korean. Lin and Chen (2002) proposed a learning algorithm for Chinese-to-English back-transliteration using both a pronunciation dictionary and a speech synthesis system to generate the pronunciation of an English word. Oh and Choi (2002) presented an English-to-Korean transliteration model using a pronunciation dictionary and contextual rules. Al-Onaizan and Knight (2002) presented a spelling-based model

for Arabic-to-English named entity transliteration. Most of the above approaches require a pronunciation dictionary for converting a source word into a sequence of pronunciations. However, words with unknown pronunciations may cause problems for transliteration. In addition, Chen et al. (1998) and Oh and Choi (2002) used a language-dependent penalty function to measure the similarity between a proper name and its corresponding transliteration. For learning the rules of transliteration and back-transliteration, Kang and Choi (2001) used a language-dependent penalty function to perform phonetic alignment between pairs of English words and Korean transliterations. Wan and Verspoor (1998) also used handcrafted heuristic mapping rules. This may lead to problems when porting to other language pairs.

However, NE transformation involves both transliteration and phrase translation. Generally speaking, person names are almost always transliterated, whereas location names and organization names are occasionally partially or entirely transliterated. Thus, the goal to extract NE pairs cannot be achieved via transliteration alone. Unlike previous studies, we proposed a method (Lee and Chang, 2003) for tackling this issue which requires neither a pronunciation dictionary for converting source words into phonetic symbols nor manually assigned phonetic similarity scores between bilingual name pairs (Chapter 4). The parameters of the model are automatically learned from a bilingual proper name list using the Expectation Maximization (EM) algorithm

(Dempster et al., 1977). Thus, the proposed method is easier to port to other language pairs as long as some transliteration training data are available.

2.3 Bilingual Lexicon Construction

Other research work related to extraction of bilingual NEs is the study of automatic bilingual lexicon construction based on bilingual corpora. In recent years, this study has attracted lots of researchers (Brown et al., 1993; Dagan et al., 1993; Kupiec, 1993; Wu and Xia, 1994; Melamed, 1996; Smadja et al., 1996; Fung and Yee, 1998; Rapp, 1999) to tackle this issue based on word co-occurrence statistics in a corpus. We will briefly introduce the IBM translation models (Mode 1, 2, 3, 4, and 5) since they are the most typical word-based approach in the MT literature. Brown et al. (1993) regarded the MT task as a sequence of operations of a noisy channel. The operations are word duplications (copy, insertion, and deletion), word translations, and word movements. The Model 1 simply started with a word-for-word copy and translation. A source word may be copied more than once or may not be copied to the target. The copy operations are performed with uniform probability. Source words are translated independently based on a word translation table. Model 2 augments Model 1 with distortion probabilities, which can be used to find the word movements with probabilities. Model 3 extends Model 2 by introducing word fertilities, which control

the number of target words that a source word is translated into. Models 4 and 5 introduce relative distortions instead of absolute distortions. A source word moved to its target position depends on the relative position of the previous word.

Although Bilingual word alignment in parallel corpora is one of most efficient ways to compile a list of bilingual lexicons, it is reported that the performance using word-based approaches is often degraded when lots of NE phrases are involved in the texts to be aligned. Thus, such study leads to poor multi-word phrase alignment, such as NE alignment. Moreover, as noted by Tsuji (2002), the above work cannot be effectively applied to low-frequency words, such as transliteration. To overcome the problem, we present an approach to extracting bilingual NE pairs based on the proposed statistical phrase translation and transliteration models.

2.4 NE Alignment/Translation

Two major approaches to automatically harvesting bilingual translation pairs from corpora have received much attention recently. One approach is to mine translation pairs from the Web, whereas the other is to extract them from bilingual corpora. Several studies have been conducted on the web-based approach (Kraaij et al., 2003; Cheng et al., 2004; Lu et al., 2004; Zhang and Vines, 2004). Most of the above studies focused on crawling large numbers of web pages to gather sufficient statistics; hence,

more computer disk space (for storing and indexing the crawled pages) and more time-consuming work (due to bottlenecks in the Internet and the need to remove noisy web pages of crawled data) are required, compared with the approach of using existing parallel corpora. Our work, on the other hand, addresses the task of extracting bilingual pairs from bilingual corpora. Relevant studies on this topic will be described next.

Previous work on the investigations of mapping between bilingual NEs in parallel corpora has been more closely related to our study presented in this thesis. As mentioned by Moore (2003), two different strategies, asymmetry and symmetry, can be applied to deal with the mapping problem. The asymmetric strategy assumes that NEs in the source part are given and that the task is to identify the translation equivalents in the target part. On the other hand, the symmetric strategy tries to find NEs in both languages and then establish the associations between NE pairs. Obviously, the symmetric approach is more difficult to apply since it requires that NEs be identified in both languages. Moreover, the errors and inconsistency induced by NE identification are, subsequently, propagated to NE alignment. Therefore, in this thesis, we adopt the asymmetric approach to extracting bilingual NE equivalents from parallel corpora. Studies based on the symmetric strategy include those of Chen et al. (2003), Huang et al. (2003), and Kumano et al. (2004), while studies based on the

asymmetric strategy include those of Al-Onaizan and Knight (2002), Moore (2003), Lee et al. (2004a; 2004b; 2006b), and Feng et al. (2004).

Research work based on the symmetric strategy will be briefly described in the following. Chen et al. (2003) investigated formulation and transformation rules for English-Chinese NEs. In that study, they use a frequency-based method to construct rules for identifying keywords of NEs from phrase-aligned corpora. Their study focused on constructing transformation rules for NE mapping between languages. However, the performance of NE alignment was not well reported in that study. Huang et al. (2003) proposed a method for acquiring English-Chinese NE pairs from a parallel corpus. Their method is based on a linear combination of the transliteration cost, translation cost, and tagging cost. Kumano et al. (2004) proposed a method for acquiring English-Japanese NE pairs from content-aligned corpora. Their approach tries to find the correspondences between bilingual NE groups based on the similarity of the appearance order in each document. However, the methods proposed by both Huang et al. (2003) and Kumano et al. (2004) require that NEs on both sides be identified beforehand, which is not suitable for our task, since we only identify NEs on the source side.

Another line of research based on the asymmetric strategy will be briefly described in the following. Al-Onaizan and Knight (2002) proposed an algorithm for

translating NEs from Arabic to English using monolingual and bilingual resources. Given an Arabic NE, they use transliteration models (including a phonetic-based model and a spelling-based model), a bilingual dictionary, and an English news corpus to generate a list of English candidates. Then, the list is re-ranked using monolingual cues, such as web counts. However, the accuracy achieved in their experiment left much to be desired. Moore (2003) proposed an approach to choosing English-French NE pairs in parallel corpora based on a sequence of refined models. This approach heavily depends on linguistic information, such as the same NE phrase occurring in the source and target parts, and cues from capitalization. Thus, the approach is not suitable for language pairs of different families, such as English/Chinese. In our previous work (Lee et al., 2004a), we proposed an approach that uses phrase translation and transliteration models to extract English-Chinese NE pairs from parallel corpora. The parameters of the proposed models are automatically estimated using the EM training algorithm in an unsupervised manner. This approach can be further improved by incorporating language-specific knowledge sources (Lee et al., 2004b; Lee et al., 2006b), which will be explained in this thesis. Feng et al. (2004) proposed an approach to English-Chinese NE alignment in parallel corpora. After recognizing English NEs, they use a maximum entropy model that integrates the translation score, transliteration score, co-occurrence score, and distortion score to

extract corresponding Chinese equivalents from the aligned sentences. In order to train the maximum entropy model, a supervised learning method with a bootstrapping strategy is adopted in their method.

In contrast to the previous research in bilingual NE processing, we present here a framework for aligning bilingual NEs in parallel corpora by incorporating proposed statistical models, i.e., a phrase translation model and a transliteration model, along with multiple knowledge sources, including abbreviation handling, person name recognition, and acronym expansion. To reduce the errors and inconsistency that could be propagated by NE identifiers in both the source and target languages, we only require that NEs be identified on the source side.