A Structure-Based Model for Chinese Organization Name Translation

YUFENG CHEN and CHENGQING ZONG
Institute of Automation, Chinese Academy of Sciences

Named entity (NE) translation is a fundamental task in multilingual natural language processing. The performance of a machine translation system depends heavily on precise translation of the inclusive NEs. Furthermore, organization name (ON) is the most complex NE for translation among all the NEs. In this article, the structure formulation of ONs is investigated and a hierarchical structure-based ON translation model for Chinese-to-English translation system is presented.

First, the model performs ON chunking; then both the translation of words within chunks and the process of chunk-reordering are achieved by synchronous context-free grammar (CFG). The CFG rules are extracted from bilingual ON pairs in a training program.

The main contributions of this article are: (1) defining appropriate chunk-units for analyzing the internal structure of Chinese ONs; (2) making the chunk-based ON translation feasible and flexible via a hierarchical CFG derivation; and (3) proposing a training architecture to automatically learn the synchronous CFG for constructing ONs with chunk-units from aligned bilingual ON pairs. The experiments show that the proposed approach translates the Chinese ONs into English with an accuracy of 93.75% and significantly improves the performance of a baseline statistical machine translation (SMT) system.

Categories and Subject Descriptors: F.4.2 [**Mathematical Logical and Formal Languages**]: Grammars and Other Rewriting Systems—*Grammar types; Parallel rewriting systems*; G.4 [**Mathematics of Computing**]: Mathematical Software—*Algorithm design and analysis*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding; Machine Translation*

General Terms: Algorithms, Languages, Experimentation, Performance

Additional Key Words and Phrases: Machine translation, named entity, organization name, structural analysis, chunk, synchronous context-free grammar, hierarchical derivation, alignment, rules extraction

The research work described in this article has been partially supported by the Natural Science Foundation of China under grant no. 60575043, and 60121302, National High-Tech Research and Development Program of China (863 program) under grant no. 2006AA01Z194, National Key Technologies R&D Program of China under grant no. 2006BAH03B02, and Nokia (China) Co. Ltd as well.

Authors' addresses: National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China; email: chenyf@nlpr.ia.ac.cn and cqzong@nlpr.ia.ac.cn.

ACM Reference Format:
Chen, J. and Zong, Z. 2008. A Structure-Based Model for Chinese Organization Name Translation. ACM Trans. Asian Lang. Inform. Process. 7, 1, Article 1 (February 2008), 30 pages. DOI = 10.1145/1330291.1330292. http://doi.acm.org/10.1145.1330291.1330292.

1. INTRODUCTION

Named entity (NE) expressions are words or phrases that name a specific entity [Chinchor and Marsh 1997]. Named entities—especially named persons, locations, and organizations—deliver essential meaning in human languages. Therefore, NE translation plays a very important role in multilingual processing, such as statistical machine translation (SMT) and cross-lingual information retrieval.

Generally, NEs occur very frequently in texts. Although many studies have been conducted on NE recognition and NE alignment, little research has been devoted to NE structure formulation and translation. As an essential component in multilingual processing systems, NE translation deserves greater attention than it receives as it is implemented by the traditional SMT system. One reason is that words or phrases extracted from bilingual sentence pairs are often inapplicable for NE translation, and the words contained in NE are domain specific and often need transliteration. The other reason is that an NE is not as sophisticated as a sentence because its denomination must follow a certain regulation. It is ineffective to translate NEs through exhaustive search during decoding, which would lead to spurious ambiguity.

There are two strategies for NE translation. One is to mine NE translation pairs from the Web and extract NE pairs from the parallel or comparable corpora. This is essentially the same as translating an NE by searching its equivalence. The other is to directly translate an NE by word/phrase translation or transliteration. Most previous studies focused on the first strategy, including the work by Huang et al. [2003], Feng et al. [2004], and Lee et al. [2006], all of which are limited by the coverage of the used corpus and the Web resource. Researchers using the second strategy for translation have commonly adopted approaches similar to SMT without considering the particular structure of NEs. This approach is evident in Zhang et al.'s work [2005a].

Using the second strategy, we focus on directly translating NEs according to their inherent structures. Since different NEs types have different translation structures, different translation models for the different NE types need to be applied in order to achieve the desired outcome. Person name (PN) tends to be transliterated based on phonetic similarity or heterography; location name (LN) tends to be transformed by both semantic translation and phonetic transliteration [Chen et al. 2003]. The keyword of LN is usually translated based on its meaning, and the remaining words require transliteration. Because the structure of PN and LN is relatively simple, the reordering of their inside words is easy to be decided in translation, and the transliteration problem becomes the major problem. However, organization name (ON) is totally different. The structure of ONs is complex and usually nested,

including PN, LN, or sub-ON. Therefore, ON is the most difficult to handle among all the NE types.

In this article, we aim to analyze the ON structure and learn the translation rules from the bilingual ON corpus and subsequently to build a translation model that can perform ON translation between Chinese and English.

In Chinese, ONs are numerous and diverse in their structures. Moreover, new ONs are constantly being created. Based on the statistics of Chinese news text released by the Linguistic Data Consortium (LDC2005T06 corpus), we found that 64% of ONs contained in the text are out of the NE dictionary (LDC2005T34 NE list), whereas these NEs usually contain frequently occurring words and only 8% of the words are out of vocabulary (LDC2002L27 lexicon). Hence, the extraction of NE equivalence from the parallel/comparable corpus is insufficient for ON translation. It often fails to translate ONs that do not appear in the corpus or NE dictionary, even though they may contain very commonly used words. This leads us to believe that it is very important to understand how the Chinese words contained in an ON are organized in the English counterpart. This study is therefore different from traditional SMT by adding ON-specific linguistic information for ON translation.

As for SMT, there have been many effective translation models for sentence translation [Zong and Seligman 2005]. But these models are not suitable for NE translation. The phrase-based model can effectively perform translations but it is restricted to phrases that are common and short (generally, shorter than seven Chinese words), and it fails to satisfactorily deal with long-range reordering. Recently researchers, — for example, Wu [1997], Chiang [2005], and Liu et al. [2006] — have focused on syntax-based methods that attempt to achieve reordering at different levels by modeling the source or target sentence with syntactic information.

Inspired by the syntax-based model for sentence translation, this article addresses the structure-based ON translation by proposing a special ON translation model independent of the translation models used for common sentences. We analyze the parallel ON corpus and find the inherent structure of ONs. “Chunk” is the unit required to construct an ON. Three types of ON chunks are defined according to the alignment of the Chinese-to-English ON pair, and then a chunk-reordering formula is observed. Therefore, we propose a hierarchical structure-based translation model for Chinese organization names. It will be evident that this newly proposed model can be integrated into an SMT system to improve the overall quality of translation.

The structure-based ON translation model is composed of an automatic chunking model and a chunk-based synchronous context-free grammar (CFG) model. In this training process, we aim to estimate the chunking model parameters and extract CFG rules for the translation model.

The remainder of this article is organized as follows: Section 2 reviews related works. In Section 3, we analyze the ON structure by defining three types of chunks and explain the rationale for a structure-based translation model. In Section 4, we propose a structure-based ON translation model. The training process is described in Section 5. Section 6 presents various

experiments as well as the corresponding discussion. Finally, the conclusion is drawn in Section 7.

2. RELATED WORK

In the past few years, researchers have proposed many approaches to NE translation, such as word/subword translation or transliteration [Stalls and Knight 1998]. Unfortunately, applying the word-based source-channel model to NE translation usually leads to unsatisfactory results. As a consequence, recent studies have focused on NE alignment (translingual equivalence extraction). Moore [2003] developed an approach to learning phrase translations from the parallel corpus based on a sequence of cost models. Huang et al. [2003] described a multifeature NE alignment model to extract NE equivalences, based on which NE translation dictionary was constructed. Kumano et al. [2004] proposed a method for extracting English-Chinese NE pairs from a content-aligned corpus. A maximum entropy model for NE alignment was presented by Feng et al. [2004]. Sproat et al. [2006] investigated the Chinese-English NE transliteration equivalence with comparable corpora. Lee et al. [2006] proposed a new approach to aligning bilingual NEs in a bilingual corpus by incorporating a statistical model with multiple sources. The translation approach that depends on NE alignment and dictionary construction has achieved relatively high accuracy on frequently occurring NEs. However, it fails to cover NEs that do not occur in the bilingual corpus. Further, the performance of ON alignment is always unsatisfactory because a distinguishing characteristic of an ON lies in its highly complex structure, which involves great variety.

There still exist many issues surrounding NE translation that call for further investigation. In an effort to assemble rare NEs, Al-Onaizan and Knight [2002] established a Web resource to rescore translation candidates. Huang et al. [2004] developed an approach combining phonetic and semantic similarity with a view to translate rarely occurring NEs. Shao and Ng [2004] presented a hybrid method to mine new translations from Chinese-English comparable corpora, combining both transliteration and context information. Zhang et al. [2005b] mined translations of out-of-vocabulary (OOV) terms from the Web through cross-lingual query expansion. Chen and Chen [2006] developed a three-step approach based on the Google search engine to deal with the backward Chinese-to-English translation.

Obviously, an NE dictionary based on extraction of NE equivalences will never provide a complete rendering of the NE translation from English to Chinese. In order to solve this problem, automatic translation and transliteration methods are adopted in company with the equivalence extraction. For transliteration modeling, Meng et al. [2001] and Gao et al. [2004] studied a phoneme-based transliteration model for the English-Chinese NE equivalence. Transliteration rules were trained from a large bilingual transliteration lexicon [Oh and Choi 2005]. Li et al. [2004, 2007] presented a joint source channel model for transliteration, and automated the semantic transliteration process for personal names. For the automatic translation, Zhang et al.

[2005a] proposed a phrase-based context-dependent joint probability model for automatic translation, which was similar to phrase-level translation models in SMT.

However, little research has been focused on NE inherent structure that can be utilized to translate NEs. Chen et al. [2003, 2006] studied formulation and transformation rules for English-Chinese NEs. They adopted a frequency-based approach to extract keywords of NEs with or without dictionary assistance, and constructed transformation rules from the bilingual NE corpus. Their study focused on transformation rules with a view to distinguishing translated parts from transliterated parts. But the performance of the rule-application in NE translation was not described where they pointed out that the keyword pairs of organization names were too numerous to allow the formulation of suitable rules. In an attempt to resolve this difficulty, we further investigate the structure of organization names. This, in turn, leads to the next step of extracting the translation rules and developing a structure-based model for ON translation.

3. ANALYSIS OF THE STRUCTURE OF ORGANIZATION NAMES

3.1 Statistical Characteristics of Organization Names

For statistical analysis of ONs, we first follow the division in Chen and Chu [2004] where one ON is divided into two parts: the name part and the keyword.

This article adopts Chinese-to-English name entity lists (LDC2005T34) for analysis that were released by the Linguistic Data Consortium (LDC). The lists contain two ON corpora. One has 54,000 industry proper-name pairs (Indus_corpus is short for "ldc_propernames_industry_ce_v1.beta.txt"), including vast numbers of corporation names, hotel names, and various business entities. ONs of this type are simple in structure and the name part is best served by transliteration, such as "吉百利公司 (Cadbury Company)", in which the name part "吉百利" is transliterated. The second corpus has 30,000 organization proper-name pairs (Org_corpus is short for "ldc_propernames_org_ce_v1.beta.txt"), containing mostly state institution names and names of local authorities. ONs of this type are more complex in structure, because the name part often contains many more modifiers.

Based on the two corpora, we get some raw statistical data whose differences are summarized in Table I. "Chn-ON-Length" denotes the length range and the average length for a Chinese ON in words and in characters. In the case of a single ON pair, if the Chinese ON is translated into English word by word in order, we consider the Chinese-to-English word alignment of the ON pair to be monotonic. Otherwise we consider the alignment to be a reordering list. "Reordering Count" denotes the count of the Chinese ONs that need the reordering operation when they are translated into English. "Transliteration Proportion" denotes the proportion of the words that need to be transliterated among the words contained in ONs.

As shown in Table I, only 6% of ONs in the Indus_corpus need reordering. Moreover, the reordering is relatively simple because the keyword of an

Table I. Statistics of Chinese ONs

Corpus	Total No. of Pairs	Chn-ON-Length	Reordering Count	Transliteration Proportion
Org_corpus	30,800	1∼14 words (2∼25 characters) 4 words on average (about 9 characters)	13,552 (about 44% of ONs are translated with reordering)	5%
Indus_corpus	54,747	1∼12 words (4∼23 characters) 3 words on average (about 8 characters)	3,284 (about 6% of ONs are translated with reordering)	92%

industrial ON usually tends to locate in the rightmost or leftmost positions, that is, at the end or the beginning of the name when translated into English. For example, in "中国银行 (Bank of China)," the keyword "银行 (Bank)" is translated first, followed by "中国 (China)" in translation. Nevertheless, 44% of ONs in the Org_corpus need reordering when they are translated into English. The name parts of ONs in the Org_corpus contain many modifiers and usually need reordering in translation. We hypothesize that the words needing reordering in translation is a consequence of the inherent structure of an ON.

On the other hand, the transliteration proportion in the Indus_corpus is much higher than that in the Org_corpus. The reason is that ONs in the Org_corpus are usually composed of frequently occurring words or phrases that do not need transliteration. But many industrial names in the Indus_corpus usually need transliteration and the transliteration is often heterographic, as in "吉百利 (*JiBaiLi*, Cadbury)", where the italicized word is the Chinese pinyin transcription. Most transliterations of the inclusive words can be referred to an NE dictionary or Web resource.

Drawing upon the statistics, it is possible to find that the ONs in both the Org_corpus and Indus_corpus have inherent structures that can be utilized for translation, even though the structure of the latter is relatively simple. As for an ON, its transliteration problem is a part of the whole structure, which can be separately considered. Since the transliteration model has been given much attention in previous research, it is worthwhile to investigate the ON structure and propose a translation model.

3.2 Structural Analysis of Organization Names

Organization name belongs to a compound noun with the "attribute + keyword" type. Compared with PN and LN, ON is much more complex in structure. At present, there is no uniform definition and translation criterion for organization name. From our analysis, a typical structure of Chinese ON is shown in Backus-Naur Form (BNF) as follows:

ON ::= {[location name] [suborganization name] [ordinal | cardinal number] [person name] [other modifiers]}* <organization keyword>

Here, square brackets "[•]" denotes the optional item[1]; "<•>" indicates the required item; "{•}" indicates that none or several items will be selected. Each item included in {•}* is a modifier for the keyword, and "other modifiers" represents other characteristics of an organization: for example, "工业 (industrial)", "教育 (educational)", etc. Furthermore, "organization keyword" is the last organization appellation. For example:

(a) {[重 庆 医 科 大 学 (Chongqing Medical University)] [第 一 (the first)] [附属(affiliated)]}* <医院(hospital)>
(The First Affiliated Hospital of Chongqing Medical University)

As Example (a) shows,"重庆医科大学 (Chongqing Medical University)" is the inclusive suborganization name;"第一 (the first)" is the ordinal number; "附属 (affiliated)" is the modifier, and"医院 (hospital)" is the keyword.

According to the BNF, Chinese ONs can be classified into two categories. The first category is multi-keyword-ON that contains a suborganization name. Therefore, it usually contains more than one keyword. For instance, Example (a) includes two keywords, "大学 (university)" and "医院 (hospital)". The second category is single-keyword-ON that does not contain a suborganization name. One multi-keyword-ON can be divided into several single-keyword-ONs (the division is described in Section 4.1). The structural analysis below is based on single-keyword-ONs[2].

Based on the word alignment of Chinese-English ON pairs, we analyze the translation rules between Chinese ONs and their English equivalences. Because words contained in Chinese ONs are mostly content words, each of them corresponds to some words in the English part. First, a definition is given for convenience.

Definition 1. Given a source-target ON pair $\{S, T\}$, there is a source word sequence s_i^j that ranges from i to j. For any sequence included in s_i^j ranging from positions k_1 to k_2 ($i \le k_1 \le k_2 \le j$), its corresponding equivalence ranges from k_1' to k_2' in the target part. If for any $k_1 \le k_2$, k_1' $\le k_2$', then s_i^j is considered as a monotone phrase[3]. Meanwhile, both the monotone source phrase s_i^j and its corresponding equivalence t_i^j form a phrase pair $\{s_i^j, t_i^j\}$. In other words, the word alignment within this kind of phrase is monotone.

According to Definition 1, we can extract phrase pairs from one ON pair. Figure 1 gives the word alignment via our pretreatment[4] and the inclusive phrases of Example (b).

---

[1]Few ONs lack keywords, which are usually due to abstract or abbreviation.
[2]For simplicity, ONs referred to in Section 3 are original single-keyword-ONs or divided single-keyword-ONs in multi-keyword-ONs, unless otherwise noted.
[3]Phrases referred to in Section 3 are the defined monotone phrase, other than the commonly used phrase notion in SMT.
[4]The word alignment almost follows the result obtained by the GIZA++ toolkit. For convenience, some English functional words that align to null words are treated as follows: "for" or "of" is connected with the preceding English word, aligning to the same Chinese equivalence; "with" or "the" joins with the back English word, aligning to the same Chinese equivalence.

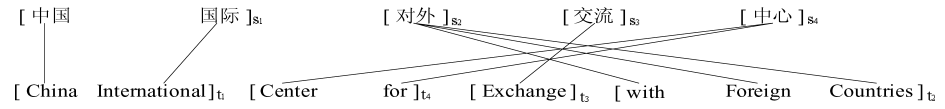

Fig. 1. Phrase alignment of Example (b).

Table II. ON Translation Patterns

Translation Pattern	Example	Frequency
$\{s, t\}$	[北京海事卫星测控站] [Beijing Maritime Satellite Monitoring Station]	56.5%
$\{s_1s_2, t_2t_1\}$	[中国][棒垒球协会] [Baseball and Softball Association of] [China]	15.5%
$\{s_1s_2s_3, t_1t_3t_2\}$	[鞍山][钢铁][学院] [Anshan] [Institute of] [Iron and Steel Engineering]	14.6%
$\{s_1s_2s_3, t_3t_2t_1\}$	[中国][保卫][大同盟][Great League for] [Defense of] [China]	1.7%
$\{s_1s_2s_3, t_2t_1t_3\}$	[电子技术][开放][实验室] [Public-use][Electronic Technology] [Lab]	0.5%
$\{s_1s_2s_3s_4, t_1t_4t_3t_2\}$	[全国][安全][生产][委员会][Safety] [National] [Committee of] [Industrial]	1.0%
$\{s_1s_2s_3s_4, t_1t_3t_2t_4\}$	[中巴][经济贸易][联合][委员会] [Economic and Trade] [Committee] [Sino-Pakistani] [Joint]	0.5%

(b) 中国国际对外交流中心 (China International Center for Exchange with Foreign Countries)

As shown in Figure 1, "[•]" denotes one phrase. Based on the alignment, four phrase pairs are found: {中国国际 , China International}, {对外 , with Foreign Countries}, {交流 , Exchange}, and {中心 , Center for}. Therefore, the alignment based on the phrase pairs can be described as $\{s_1s_2s_3s_4, t_1t_4t_3t_2\}$, which may be treated as a translation pattern. We randomly select 10,000 Chinese-English ON pairs from the Org-corpus, and sum up the translation patterns by calculating their corresponding frequency. The top seven patterns with high frequency are displayed in Table II.

Theoretically, there exist numerous translation patterns. However, these patterns could be simplified, differing from sentence translation, because an ON is "attribute + keyword" type in a definite structure. The semantic relationship between the contained words determines their corresponding positions in the target ON. By considering the locations of the phrase pairs in each translation pattern, it is possible to find that the span of each phrase pair is restricted within a certain range. Therefore, the source ON can be divided into several chunks, where the word reordering is limited within each chunk during translation, and then the chunks are reordered to obtain the translation output.

Based on the analysis, we propose a "chunk" unit to analyze the inherent ON structure. According to the syntactic function of different phrases in one ON, three types of chunk pairs are designated to sum up all the ON translation patterns. Namely, an ON comprises three potential chunks according to the following definition.

For an ON pair, three types of chunks in the Chinese part are identified in turn according to Definition 2 based on its phrase pairs. Note that one or two types of chunks may be absent in the ON.

Definition 2.

(1) *Regionally Restrictive Chunk (RC).* It is the supreme modifier whose alignment on word level is monotone. Its identification satisfies all of the following conditions: (a) it should stay on the leftmost position of the Chinese ON; (b) it should be contained in the first phrase of the Chinese ON; (c) it should be composed of location names or ordinal numbers in series. For example, "中国国际 (China international)."

(2) *Keyword Chunk (KC).* It states an essential appellation for the ON and its alignment on word level is monotone. It satisfies the following conditions: (a) it is the last part of the Chinese ON; (b) it is contained in the last phrase; (c) if there is only one phrase left after RC, only the organization keyword is identified as the Keyword Chunk; if there is more than one phrase left, the last phrase is identified as the Keyword Chunk. For example, "协会 (association)" and "大同盟 (great league)" in Table II can be identified.

(3) *Middle Specification Chunk (MC).* This is the second important modifier and is what the remaining part in an ON other than RC and KC is considered. The words included usually need to be reordered in translation such that its alignment on word level is monotone or reordered. For example, "海事卫星测控 (maritime satellite monitoring)" is translated word by word in order, whereas "对外交流 (exchange with foreign countries)" is translated by word reordering.

According to Definition 2, the potential chunk sequences of Chinese ONs are identified as follows: (1) $C_{RC}C_{MC}C_{KC}$; (2) $C_{MC}C_{KC}$; (3) $C_{RC}C_{KC}$; (4) $C_{RC}C_{MC}$; (5) C_{RC}; (6) C_{MC}; and (7) C_{KC}, where C_X denotes the chunk in the Chinese part. At the same time, the corresponding chunks E_X in the English part are also identified. Therefore, each ON pair is divided into several chunk pairs (at most three). Generally, an ON contains more than one chunk, yet the instances of (5), (6), or (7) usually appear under the condition that these ONs are abstract, abbreviated, or partially recognized.

Actually, each type of chunk clusters words semantically for translation, according to the words reordering range in the alignment. The decomposition of ON in terms of the "chunk" units covers all different ON translation patterns. For instance, if the order of chunks and the inclusive words are all preserved in translation, this translation pattern is monotone, $\{s, t\}$, that is, there is only one phrase pair which is the dominating translation pattern with the highest frequency. Table III shows the structural analysis of the examples, according to Definition 2.

After identifying the three types of chunks, we consider the positions of the three chunks in different translation patterns. As Table III illustrates, it is possible to find the chunk-reordering formula based on a large number of bilingual ON pairs: the position of RC is the first to be allocated, because it is put at the beginning or the end of the translation result. The position of the MC and the KC is always adjoining. Moreover, if there is reordering between two chunks, the translation usually needs a preposition to link them, such as "for" or "of." This above finding can be adopted as the guideline for the ON translation model.

Table III. Structural Analysis Based on Chunk

Chunk Pairs	Example
$\{C_{RC}C_{MC}C_{KC}, E_{RC}E_{MC}E_{KC}\}$	北京 海事卫星测控 站 Beijing Maritime Satellite Monitoring Station
$\{C_{RC}C_{MC}C_{KC}, E_{MC}E_{KC}E_{RC}\}$	中国 棒垒球 协会 Baseball and Softball Association of China
$\{C_{RC}C_{MC}C_{KC}, E_{RC}E_{KC}E_{MC}\}$	鞍山 钢铁 学院 Anshan Institute of Iron and Steel Engineering
$\{C_{RC}C_{MC}C_{KC}, E_{KC}E_{MC}E_{RC}\}$	中国 保卫 大同盟 Great League for Defense of China
$\{C_{MC}C_{KC}, E_{MC}E_{KC}\}$	电子技术开放 实验室 Public-use Electronic Technology Lab
$\{C_{RC}C_{MC}C_{KC}, E_{RC}E_{KC}E_{MC}\}$	全国 安全生产 委员会 National Committee of Industrial Safety
$\{C_{RC}C_{MC}C_{KC}, E_{RC}E_{MC}E_{KC}\}$	中巴 经济贸易联合 委员会 Sino-Pakistani Joint Economic and Trade Committee

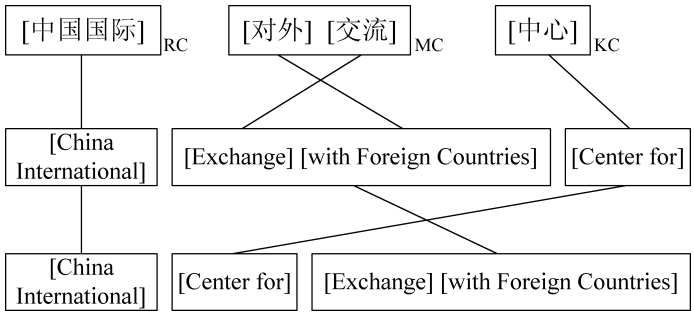

Fig. 2. The chunk-based translation process for Example (b).

3.3 Motivation for the Structure-Based Translation

The definition of the three chunks well describes the structure information of ON and represents the words' division in translation. Different chunks have different translation characteristics, and the chunk-reordering formula for their combination achieves the whole translation. In other words, the reordering of chunks and their words follows a certain formula in the English part.

As shown in Figure 2, the translation process of Example (b) is transformed into three stages: (1) the ON is divided into RC ("中国国际"), MC ("对外交流"), and KC ("中心"); (2) the words are translated and reordered inside each chunk (for example, "对外" is transposed with "交流" in MC); and (3) the chunks are reordered.

On the basis of this observation, we propose a structure-based translation for Chinese ONs by the following steps: and (1) the Chinese ON is automatically chunked; and (2) each chunk is translated and reordered by synchronous CFG rules.

The chunking parameters and CFG rules are learned from the training process. We apply Definition 2 to training data (ON pairs) for structural analysis—that is, identifying chunk pairs—and then estimate chunking parameters as well as extract CFG rules.

The next section introduces the structure-based ON translation model.

4. STRUCTURE-BASED TRANSLATION MODEL FOR ORGANIZATION NAME

On the basis of our discussion thus far, our approach to Chinese-to-English ON translation may be decomposed using two submodels: the chunking model and the chunk-based CFG model. Therefore, our proposed ON translation method first divides the source ON into chunks by the chunking model (Section 4.1); and second, translates and reorders chunks by the chunk-based CFG model (Section 4.2), where the CFG rules are described (Section 4.2.1) and the CFG derivation is presented (Section 4.2.2). The framework of the structure-based translation model will be presented in Section 4.3.

4.1 Chunking Model

As we have mentioned, there are two categories of ONs. A multi-keyword-ON should be divided into single-keyword-ONs according to the keywords it contains. To do this, the decomposition module works on a multi-keyword-ON from left to right. If a keyword is identified, the keyword with previous words is extracted as one single-keyword-ON. In Example (a), "重庆医科大学第一附属医院 (The First Affiliated Hospital of Chongqing Medical University)" is divided into "重庆医科大学 (Chongqing Medical University)" and "第一附属医院 (The First Affiliated Hospital)," according to the two organization keywords "大学 (university)" and "医院 (hospital)".

Let O denote a Chinese single-keyword-ON, which is composed of $n\,(n \geq 1)$ Chinese words (Performed segmentation) or characters (unsegmented), $o_1, o_2, \ldots\ o_n$. The task of chunking is to find the most likely sequence of chunks: $C^* = C_1 \ldots C_m$ ($m \leq 3$, $C_i \in \{C_{RC}, C_{MC}, C_{KC}\}$) that maximizes the probability $p(C|O)$. Here, the Bayesian rule is applied to rewrite $p(C|O)$ as:

$$p(C|O) = \frac{p(O|C)p(C)}{p(O)}$$

Now it is possible to make two assumptions for chunking phrases: (1) the chunk-tag sequence could be modeled with the first-order Markov chain, and (2) words inside each chunk are independent of other chunks and only depend on their associated chunk-tag. So the chunking model can be specified as the following equation:

$$\begin{aligned} C^* &= \arg\max_C [p(O|C)p(C)] \\ &= \arg\max_C [p(o_1 \ldots o_n|C_1 \ldots C_m)p(C_1 \ldots C_m)] \\ &= \arg\max_C [\prod_{i=1}^{m} p(o_{i1} \ldots o_{ij}|C_i)p(C_i|C_{i-1})] \end{aligned} \tag{1}$$

where $p(C)$ is the chunk contextual model. All parameters of Equation (1) are trained from bilingual ON pairs, described later in Section 5.

4.2 Chunk-Based CFG Model

4.2.1 *Chunk-based CFG Rules.* The model is formally based on a synchronous CFG grammar [Lewis and Stearns 1968; Aho and Ullman 1969]. Following the synchronous CFG presented in Chiang [2005], the elementary structure of synchronous CFG is rewriting rules with aligned pairs of right-hand sides:

$$\mathrm{X} \rightarrow (\alpha, \beta, \sim) \tag{1}$$

where X is a nonterminal, and α and β are both strings of terminals and nonterminals. In our model, α and β contain terminals and one nonterminal in the chunk unit, and "$\sim$" is a one-to-one correspondence between chunk nonterminal occurrences in α and the one in β. Rewriting begins with a pair of linked start symbols. At each step, two corresponding nonterminals are rewritten using the two components of a single rule.

Thus according to the bilingual alignment, it is possible to formalize the chunks of Example (b) in a synchronous CFG as:

$$\mathrm{X} \rightarrow \ <\text{中国国际})_{RC}\ \mathrm{X},\ (\text{China International})_{RC}\ \mathrm{X}> \tag{2}$$

$$\mathrm{X} \rightarrow \ <\mathrm{X}\ (\text{中心})_{KC},\ (\text{Center for})_{KC}\ \mathrm{X}> \tag{3}$$

$$\mathrm{X} \rightarrow \ <(\text{中国国际})_{RC}\ (\mathrm{X})_{MC}\ (\text{中心})_{KC},\ (\text{China International})_{RC}\ (\text{Center for})_{KC}\ (\mathrm{X})_{MC}> \tag{4}$$

For differentiation, CFG rules (2) and (3) take one chunk as the terminal respectively, which are regarded as common rules, whereas the last rule, (4), takes two chunks as the terminal, which can be considered as a template with a variable Hu et al. [2006]. To highlight the contrast with common rules, we call a rule with two chunk terminals as a *template*.

A log-linear model used here is based on the model of Och and Ney [2002] to estimate the weight of the above rules by minimum error rate training.

$$w(X \rightarrow \ <\alpha, \beta>) = \prod_i \theta_i(X \rightarrow \ <\alpha, \beta>)^{\lambda_i}$$

Let θ_i denote features defined on rules. In this model, four features are used:

—The probability distribution of the rule, $p(\alpha|\beta)$ and $p(\beta|\alpha)$.
—The lexical weights, $p_w(\alpha|\beta)$ and $p_w(\beta|\alpha)$ [Koehn et al. 2003].

Templates and common rules with weights are generated from a bilingual ON corpus without any syntactic information. However, the following five types of rules need to be added:

$$O \rightarrow <O\ X\ ,\ X\ O> \tag{5}$$

$$O \rightarrow <X\ ,\ X> \tag{6}$$

$$X \rightarrow <C_{RC}\ X,\ E^*_{RC}\ X> \tag{7}$$

$$X \rightarrow <C_{MC},\ E^*_{MC}> \tag{8}$$

$$X \rightarrow <X\ C_{KC},\ X\ E^*_{KC}> \tag{9}$$

Rule (5) and Rule (6) are "glue" rules with weight one. Rule (5) is applied to multi-keyword-ONs, of which their contained single-keyword-ONs should be rewritten separately. Native Chinese speakers always regard the preceding ON as more general, so they usually place it at the end of the translation output. For example, Example (a) is translated as "The First Affiliated Hospital of / Chongqing Medical University." Thus Rule (5) describes the inversion. Rule (6) combines a sequence of chunks to form an ON, and "O" is the start-symbol for the adopted CFG.

Rules (7), (8), and (9) are three special decomposed chunk-based rules, applying to the case when no such rule for RC, MC, or KC is found in the training data. In these formulas, the reordering of chunks is manually defined and the translation of each chunk is on the word level. E^*_{RC}, E^*_{MC}, and E^*_{KC} denote the best translation output of C_{RC}, C_{MC}, and C_{KC} on the word level. These three rules suggest that the positions of nonoccurring chunks are monotone in translation by our definition.

Whether the translation of words in the RC is monotone or reordered depends on the translator's preference. For convenience, in our approach it is set as monotone, as shown in Rule (7). The equation is expressed as follows:

$$\begin{aligned} E^*_{RC} &= \arg\max_{E_{RC}} p(E_{RC}|C_{RC}) = \arg\max_{E_{RC}} p(C_{RC}|E_{RC}) p_{LM}(E_{RC}) \\ &= \arg\max_{E_{RC}} \prod_{i=1}^{I} p(c_{RC_i}|\, e_{RC_i})\ p_{LM}(E_{RC}) \end{aligned} \tag{2}$$

Where, $C_{RC} = c_{RC_1}, \ldots c_{RC_1}, E_{RC} = e_{RC_1}, \ldots, e_{RC_1}, e_{RC_1}$ denotes the counterpart of c_{RC_1}, and P_{LM} (•) is the language model, which is trained from English ONs.

Word translation in MC is not always monotone. Hence, it is necessary to reorder the words in the MC according to the reorder model in Rule (8). The model follows the distortion model [Koehn et al. 2003] with an appropriate

Table IV. Four Types of CFG Rules

Rule Types	Characteristics	Weight in Derivation	Hierarchy
"glue" rule	combine the sequences	weight one	rank one
template	two chunks as the terminal	with the highest weight	rank two
common rule	one chunk as the terminal	with the highest weight	rank three
special rule	no chunk is covered	with the highest probability based on word	rank four

value for the parameter φ = exp(-1) in the given experiment. The reorder model is described as the following formula:

$$\begin{aligned} E^*_{MC} &= \arg\max_{E_{MC}} p(E_{MC}|C_{MC}) = \arg\max_{E_{MC}} p(C_{MC}|E_{MC})p_{LM}(E_{MC}) \\ &= \arg\max_{E_{MC}} \prod_{i=1}^{I} p(c_{MC_i}|e_{MC_i})\varphi^{|a_i - b_{i-1}-1|} p_{LM}(E_{MC}) \end{aligned} \tag{3}$$

where a_i is the start position of the Chinese words that are translated into the i_{th} English word e_i in the MC, and b_{i-1} denotes the end position of the Chinese words translated into the (i-1)$_{th}$ English word e_{i-1} in the MC.

In Rule (9), the translation of the KC is also fixed to be monotone as the RC. So, we have:

$$\begin{aligned} E^*_{KC} &= \arg\max_{E_{KC}} p(E_{KC}|C_{KC}) = \arg\max_{E_{KC}} p(C_{KC}|E_{KC})p_{LM}(E_{KC}) \\ &= \arg\max_{E_{KC}} \prod_{i=1}^{I} p(c_{KC_i}|\, e_{KC_I})\; p_{LM}(E_{KC}) \end{aligned} \tag{4}$$

The common rules and templates are generated from the corpus. Moreover, by adding special rules, all possible conditions are satisfactorily covered.

4.2.2 *CFG Derivation.* Most words or chunks that constitute an ON in the bilingual corpus are infrequent, so there is a problem of sparse data severely hampering the language model. Hence, the language model is not adopted in our ON translation model. The CFG derivation is based on each rule with highest probability, according to the chunk-reordering formula observed in the structural analysis. The aim is to choose each rule with the highest-probability weight to yield the highest-probability derivation following certain steps.

As a whole, the entire CFG rule set contains four types of rules: common rule, template, "glue" rule, and special rule. All of them are based on "chunk" unit but they are hierarchical in derivation. For instance, a template contains more items than a common rule and only the translation of the variable needs to be confirmed, the template matching ensures a correct translation more quickly and efficiently, which will be shown in the experiment. In this model, a template is therefore considered more preferential than a common rule in derivation.

As shown in Table IV, the hierarchy gives the model the option during the CFG derivation: beginning with the "glue" rules, it first matches the templates, and then common rules. If they all fail, the model falls into word-level translation by special rules.

ON Derivation Algorithm

Step 1: Combine a sequence of *Xs* to form an ON, or combine single-keyword-ONs to form a multi-keyword-ON by "glue" rules.

Step 2: Match the template set. If there are available templates, use the template with the highest weight and rewrite the residue following steps 3-5; otherwise, turn to the following steps directly.

Step 3: Match the common rule set to fix the position and translation of the RC with the highest weight. If it fails, Rule (7) is used. (In other words, the nonterminal of RC is rewritten by a common rule. If there is no available common rule, it turns to the special rule based on word level).

Step 4: Match the common rule set to fix the position and translation of the KC with the highest weight. If it fails, Rule (9) is used.

Step 5: Match the common rule set to fix the translation of the MC. If it fails, Rule (8) is used.

Fig. 3. ON derivation algorithm.

< O,O >

*step*1 ⇒ < *X*, *X* >

*step*2 ⇒ *template matching fails*

*step*3 ⇒< (中国国际)$_{RC}$ X , (China International)$_{RC}$ X >

*step*4 ⇒< (中国国际)$_{RC}$ X (中心)$_{KC}$, (China International)$_{RC}$ (Center for)$_{KC}$ X >

step5 ⇒< (中国国际)$_{RC}$ (对外交流)$_{MC}$ (中心)$_{KC}$, (China International)$_{RC}$ (Center for)$_{KC}$ (Exchange with Foreign Countries)$_{MC}$ >

Fig. 4. Example (b) derivation of chunk-based synchronous CFG.

Given an ON, we list five steps for its translation by CFG derivation in Figure 3, and these steps are defined as the ON derivation algorithm. The derivation is hierarchical according to different types of rules, and steps 3-5 are based on the chunk-reordering formula described in Section 3.2.

The ON derivation algorithm covers translation of all the ONs, of which the chunk sequences have seven potential types, as mentioned in Section 3.2. Practically speaking, this hierarchical derivation shows significant advantage for the ONs that contain more than one chunk. However, translation for the ON that contains only one chunk uses the special rules, the same as a normal word-based translation approach. For example, "我们的家园—俄罗斯 (Russia is our home)" is chunked as the "MC", which is translated by Rule (8).

Figure 4 shows the derivation of Example (b) following the ON derivation algorithm, under the condition that template matching fails and there are three available common rules for derivation. If one ON does not contain RC, MC, or KC, its derivation elides their corresponding steps.

Restricted by these steps, the model requires few decoders, but achieves a highly accurate translation, because the process of detecting the three chunks and following the above steps is the same as pruning the search space, with the same aim: finding the best derivation. The chunking and derivation process is somewhat similar to monolingual parsing constraints in other grammar-based models [Wu and Wong 1998]. However, it is based on the chunk-reordering formula of the ON inherent structure.

(1) **Chunking**

(1.1) The ON is divided into several single-keyword-ONs, in case that it is a multi-keyword-ON.

(1.2) Each single-keyword-ON is divided into several chunks (at most three) automatically by the chunking model.

(2) **Chunk-based CFG decoding**

The chunks are translated and reordered by four chunk-based CFG rules: "glue" rules, templates, common rules, and special rules. The proposed ON derivation algorithm gives the model the hierarchy to achieve the whole translation using the CFG rules.

Fig. 5. The process of Chinese ON translation.

As a whole, the chunk-based CFG derivation with five steps is an attempt to effectively determine the optimal route of ON translation according to its typical structure.

4.3 Framework for Structure-Based ON Translation Model

A Chinese ON is translated as shown in Figure 5. Figure 5 summarizes the framework of the overall process for ON translation, which is performed via a two-stage model. The proposed model integrates hierarchical CFG rules, "glue" rules, templates, common rules, and special rules, by a defined ON derivation algorithm. The CFG rules are ranked according to the covered unit, from chunk to word level, which can most completely cover the Chinese ON translation. The CFG model is similar to the hierarchical phrased-based model [Chiang 2005]; however, the model we propose here is based on the ON structure, the proposed chunk-unit, and the hierarchical derivation according to a five-step process.

5. TRAINING

The aim of the training process is to obtain the parameters for both the chunking model and CFG rules. Figure 6 illustrates the overall training architecture. Based on ON pairs, we first perform word alignment, and then identify the three types of chunk pairs according to structural analysis, which we described in Section 3.2. Based on the chunk pairs, we can determine the chunking parameters, and extract templates and common rules. The generation of word-based probability follows the former method [Koehn et al. 2003]. On the whole, there are two main tasks to be handled: (1) word alignment of ON pairs (see Section 5.1); and (2) templates and common rules extraction (see Section 5.2).

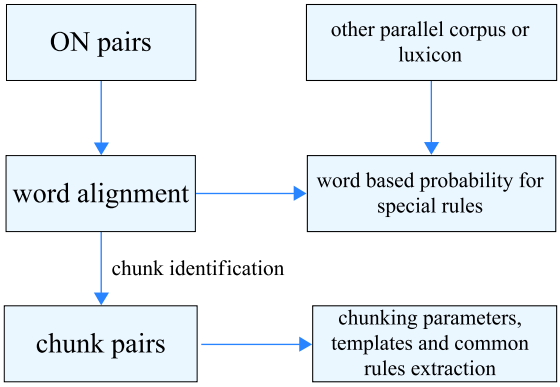

Fig. 6. The overall training architecture.

5.1 Word Alignment of ON Pairs

Word alignment of ON pairs is somewhat different from word alignment of bilingual sentences, mainly due to the difficulties caused by abbreviations and transliteration problems. In this study, we tried two approaches with the aim of achieving high-quality word alignments. Approach I is similar to traditional phrase extraction for SMT, but it adds an alignment hypothesis. Approach II is a frequency-based method that combines the result of Approach I in order to improve the alignment quality.

5.1.1 *Alignment Approach I.* Starting with a bilingual ON pair list, we ran GIZA++ toolkit [Och and Ney 2002] on the corpus in both directions: Chinese-to-English and English-to-Chinese. This produces two word alignments in both directions. As a test, GIZA++ was applied to the 4,000 ON pairs from the Org_corpus as described in Section 3. The result, after a manual double check of the alignments, was that more than 50% alignments made by GIZA++ were wrong. This was especially evident in the English-to-Chinese alignment.

With further analysis it became clear that words included in Chinese ONs are always content words, as was seen in Examples (a) and (b). Each content word has a concrete meaning that should be translated into at least one English word, but in the GIZA++ operation on English-to-Chinese, at most one English word is allowed to be aligned with a Chinese word. So alignment errors are inevitable. However, the alignment in the English-to-Chinese translation is more satisfactory when it deals with words that need transliteration or are wrongly segmented. For example, "小(Xiao) / 汤(Tang) / 山(Shan)," denotes a segmentation symbol. The three Chinese characters should be one word but are segmented with error. In the English-to-Chinese translation, the three characters were correctly aligned as one English word, "Xiaotangshan", whereas in the Chinese-to-English translation they failed.

Because some words used in an ON occur infrequently in the whole corpus, some word alignment errors may result in numerous mistakes. Obviously, selecting correct alignments and discarding incorrect alignments to achieve the utmost precision is a problem that requires attention.

Considering the fact that ONs always contain content words, we propose a pretreatment that will filter the alignments: In the Chinese-to-English translation, a specific alignment should be considered as a hypothetically correct alignment, in which each Chinese word is aligned to one or several consecutive English words. The same holds for English-to-Chinese translation. In Example (b), each Chinese word aligns to one or several consecutive English words. Hypothetically, this is a correct alignment. According to this hypothesis, we need only select the hypothetically correct alignment from the respective results of the Chinese-to-English translation and the English-to-Chinese translation of GIZA++. Thereafter, both alignments can be reconciled.

For an ON, if its alignments in the two translations are both hypothetically correct, it is now possible to apply a heuristic approach [Koehn et al. 2003] to unite the two alignments. This allows greater assurance for the alignment of the pairs, but the quantity of such alignments is insufficient. Consequently, Alignment Approach II based on Alignment Approach I must be applied to improve the overall performance.

5.1.2 *Alignment Approach II.* Segmentation errors and alignment errors are inevitable through GIZA++. Furthermore, the abbreviations that commonly occur in ONs cannot be captured by Approach I, which means we must invoke Alignment Approach II to obtain higher quality. In Approach II, we calculate the scores of all candidate segment alignments within ON pairs, and then obtain the optimal word alignment for a given ON pair with a maximum score.

First, to measure the alignment score of each Chinese segment and its English equivalent, we borrow the $tf \times id$ notion from information retrieval. Following Chen et al.'s [2003] formula, let $\{O, E\}$ denote one ON pair from the bilingual ON corpus. In this case, some Chinese segment $o \in O$ should be aligned to some English segment $e \in E$. If some functional words, such as "the," "for," "of," etc., occur in the English translation, they will be set with their adjoining words as a single segment, in order to ensure that each English word aligns with a Chinese segment. Because Chinese segmentation has problems, it is more efficient to start the frequency computation from the English ON. Let us assume there are N English segments, and the term frequency (tf) of a Chinese ON segment o in e is described by the number of occurrences of o in e. (df) of o is the number of English segments into which o is translated. Preference is given to longer Chinese segments by adding $\log_2(|o| + 1)$, and $|o|$ denotes the length of the Chinese ON segment. In addition, $p(o|e)$ and $p(e|o)$ are the probability results from the hypothetically correct alignment described in Section 5.1.1, and we can add lexicons. Those three factors of frequency featuring within the square bracket of λ_1 are directly adopted from the research of Chen et al. [2003]. Here, in order to combine the frequency

feature and translation probability feature, we use a general log-linear model:

$$score(\{o, e\}) = [f(\{o, e\}) \times idf(o) \times \log_2(|o| + 1)]^{\lambda_1}[p(o|e)]^{\lambda_2}[p(e|o)]^{\lambda_3}$$

$$f(\{o, e\}) = \frac{tf(\{o, e\})}{\max_i(tf\{o_i, e\})}$$

$$idf(o) = \log_2(\frac{N}{df(o)})$$

where λ_1, λ_2, λ_3 are empirically chosen to discriminate correct and incorrect alignment to achieve better accuracy and are set as 0.5, 0.25, and 0.25 respectively in our experiments. o_i is one of all possible Chinese segments in e.

Based on the score of each possible segment pair, the optimal alignment A_{opt} between Chinese and English in one ON pair is obtained by the following greedy approximation algorithm [Huang and Vogel 2002]:

(1) Initialize ON-Aligned as an empty set, and set all possible (o, e) segment-pairs as the list of combinations of Chinese segments and English segments in the given ON pair. To reduce the search space, the length of the Chinese segment is limited to less than 7 characters, and the English segment is limited to fewer than 4 words.
(2) Sort segment pairs (o, e) in descending order according to the *score* ($\{o, e\}$).
(3) Move the topmost pair (o, e), that is the pair with the maximum score, from the segment-pairs to ON-Aligned.
(4) Remove all (o, •) and (•, e) from the segment-pairs.
(5) Repeat from step 3 until the set of segment-pairs is empty. Then the optimal alignment A_{opt} is produced.

Empirically, this algorithm usually finds the word alignment of an ON pair with a maximal score.

5.2 Generating Chunk-Based CFG Rules from ON Pairs

5.2.1 *Chunk Identification.* To extract chunk-based CFG rules (templates and common rules), chunk pairs should be generated in advance. This process is the same as ON structural analysis, detailed in Section 3.2. Based on the word alignment of an ON pair, we obtain its monotone phrase alignment according to Definition 1, and then identify the three types of chunk pairs according to Definition 2. To determine the Chinese location name and keyword, we match a location name list and a keyword list, which are extracted from the LDC corpus and lexicons using a frequency-based approach [Chen et al. 2006], partially via manual checking.

Actually, within all the training data of ON pairs, Chinese chunks are identified and their English equivalent, are respectively fixed in the ON pair according to Definition 2. If words in the English chunk are not inconsecutive,

we utilize these ON pairs to extract templates only. ON pairs of this type are less than 0.1% of all the training data.

Based on the chunk pairs, the parameters, $p(o_i|C_i)$ and $p(C_i|C_{i-1})$, of the chunking model are generated.

5.2.2 *Extraction of CFG Rules.* Based on the chunk-identified ON pairs, we can generate the common rules and templates by applying Procedure 1 and Procedure 2 respectively.

Procedure 1. Given an ON pair $\{O, E\}$after chunk identification, $O = C_{RC}C_{MC}C_{KC}$, E is the combination of E_{RC}, E_{MC}, and E_{KC}. Here, C_{RC}, C_{MC}, and C_{KC} denote the RC, MC, and KC, which are possibly contained in the source ON, and E_X denotes the corresponding English translation of C_X.

(1) If C_{RC} exists, $\{C_{RC}, E_{RC}\}$ is established as the initial RC pair, then all the following chunks are considered as nonterminal X, thus $X \rightarrow <C_{RC}X, E_{RC}X>$ or $X \rightarrow <C_{RC}X, XE_{RC}>$ becomes a common rule.
(2) If C_{MC} exists, $\{C_{MC}, E_{MC}\}$ is the initial MC pair, then $X \rightarrow <C_{MC}, E_{MC}>$ is a common rule.
(3) If $\{C_{KC}, E_{KC}\}$ is an initial pair of the KC, then the preceding chunk is to be set as nonterminal X, thus $X \rightarrow <XC_{KC}, E_{KC}X>$ or $X \rightarrow <XC_{KC}, XE_{KC}>$ is a common rule.

Procedure 2. Given an ON pair $\{O, E\}$ after chunk identification, $O = C_{RC}C_{MC}C_{KC}$, E is the combination of E_{RC}, E_{MC}, and E_{KC}. If $\{C_X, E_X\}$ ($X \in (RC, MC, KC)$ is an initial chunk pair, such that $\alpha = \alpha_1 C_X \alpha_2$ and $\beta = \beta_1 E_X \beta_2$, then $X \rightarrow <\alpha_1 X \alpha_2, \beta_1 X \beta_2>$ is a template. Here, α_1, α_2, β_1, and β_2 indicate chunk sequences or null.

For example, there is an ON pair $\{C_{RC}C_{MC}C_{KC}, E_{MC}E_{KC}E_{RC}\}$, where $\{C_{RC}, E_{RC}\}$, $\{C_{MC}, E_{MC}\}$, and $\{C_{KC}, E_{KC}\}$ are the three contained initial chunk pairs, then three templates, $X \rightarrow <XC_{MC}C_{KC}, E_{MC}E_{KC}X>$, $X \rightarrow <C_{RC}XC_{KC}, XE_{KC}E_{RC}>$, and $X \rightarrow <C_{RC}C_{MC}X, E_{MC}XE_{RC}>$, are extracted respectively.

Templates are not as flexible as common rules, but in some cases, generating templates will avoid errors introduced by segmentation or alignment. For example, {北京/$_{RC}$中/$_{MC}$医学院/$_{KC}$, Beijing/$_{RC}$College of/$_{KC}$ Traditional Chinese Medicine/$_{MC}$} is an ON pair after chunk identification, where "中医 (Traditional Chinese Medicine)" should be one word, but were aligned in error. So the chunk identification and extracted CFG common rules are both wrong:

$X \rightarrow <X$ 医学院, College of X>
$X \rightarrow <$中, Traditional Chinese Medicine>

By following Procedure 2, the template is generated as $X \rightarrow <$X 中医学院, X College of Traditional Chinese Medicine>. It turns out to be right, avoiding the alignment error.

According to the above scheme, the common rules and templates are generated from ON alignments. Such rules based on chunking are refined without any ambiguity. Therefore, there is no filter for the grammar. The probability

Table V. Rules Extraction Statistics

	Alignment Approach I	Alignment Approach II
NE-Pair Records	68,960	68,960
Common Rules	52,998	189,682
Templates	12,464	142,028

and lexical weight of the extracted rules are estimated by the maximum likelihood estimation algorithm following Koehn et al.'s [2003] method.

6. EXPERIMENTS AND DISCUSSION

Two set of experiments were conducted for Chinese-to-English ON translation. In the first, we evaluated the proposed ON structure-based model by evaluating the translation accuracy in a blind test set, using two training methods respectively. In the second, we assessed the improvement of the SMT system's translation quality by adding the ON translation model.

6.1 Translation Model for Organization Names

The first experiment used Chinese-English bidirectional Name Entity Lists v1.0 (LDC2005T34) released by the LDC. (The bilingual NE corpus is compiled from Xinhua's database, including the person names, location names, and organization names, etc.) Since person and place names are mostly transliterated, we only extracted the categories of organization, industry, press, and international organization to form a training set. First, we performed a quick proofreading to correct some errors and remove the following types of entries:

—the duplicated entry

—the entries whose English translation contains one or more non-English words

We thus obtained a total of 68,960 ON Chinese-English pairs as the final training set. The number of Chinese words contained in one ON ranges from one to fourteen, most of which fall in the range of two to seven. We ran the training process described in Section 5 on the training set and then compared the two alignment approaches. In Approach I, source ONs were presegmented using an in-house segment system with an 81,000-word list. Table V shows the results using the two approaches. GIZA++ outputs are proved so unsatisfactory that only hypothetically correct alignments were considered, hence the extracted rules based on Alignment Approach I are considerably fewer. Consequently, Approach II predominates. The alignment result would directly affect the translation performance.

Moreover, we obtained a keyword list from the training data based on Alignment Approach II, which contained 2,840 unique entries. A location name list included 150,200 entries, which were extracted mostly from both the training data and the place name entries in the LDC2005T34 corpus. The two lists, covering almost all the keywords and the location names in the training data partially via manual checking, were prepared for chunk identification in training. As entities are countless and new words occur all the time, the lists could

Table VI. Meaning Adequacy Criteria

Degree	Metric
1	At least one word is not translated; incomplete meaning or meaningless
2	Each word is translated; partial meaning
3	Almost the same meaning
4	Exactly the same output as the reference translation

not cover all keywords and location names; consequently, some chunking errors occur. However, these errors are minimal and affect the translation performance only indirectly because the ON translation model is hierarchically based on word, chunk, or template.

For evaluation purposes, we used the meaning adequacy metric for subjective evaluation. Here, "adequacy" refers to the degree to which the translation preserves the original information or meaning in the source ON. Four degrees of the metric are shown in Table VI with short explanations.

For the Chinese ONs in the test data, only one reference translation is available, and it is very difficult to achieve a completely identical result in English even using manual translation. It means the 4th degree alone is not capable of assessing the accuracy. Results achieved with the 3rd or 4th degree can both be regarded as satisfactory. Results of the 1st and 2nd degrees are considered wrong. Two human evaluators to judge the result according to the meaning adequacy metric in Table VI. The kappa coefficient (K) for human agreement rate is 0.92, which is adequate.

There were 432 ONs randomly selected from the LDC2005T34 corpus as the test data, not overlapping with the training data. Each of the ONs contained between 2 and 9 words. The ONs consisted of a total of 1,506 Chinese words. Although the ON structure is finite in chunk units, the vocabulary it contains is large, making it necessary to add the LDC Chinese-English Translation Lexicon (Version 3.0) to cover more word translations via special rules. Next, we tested the ON translation model, and measured each English output according to the four degrees. Finally, we calculated the proportion of each degree, resulting in the average of the measurements.

First, we chunked the test data, with an accuracy of 92% (segmented) and 96% (unsegmented). Second, we conducted experiments on the data with CFG rules produced by Alignment Approach I and Approach II. The phrase-based MT system developed by a local laboratory was our baseline system so we could show the superiority of the proposed ON translation model. The baseline system was based on a log-linear model with the following features, analogous to Pharaoh's default feature set: Phrase translation probabilities, lexical weights, language model trained by SRILM toolkit[5], distortion model, and length penalty [Koehn et al. 2003]. The parameters of this system were trained by a minimum error rate training method using the same training data. We made a comparison by gauging the difference between the translated

[5]http://www.speech.sri.com/projects/srilm/

Table VII. Translation Model Performance Using Alignment Approach I

Model	**Accuracy** (%)		**Error Rate** (%)	
Baseline+Lex (Seg)	4^{th} degree	3^{rd} degree	2^{nd} degree	1^{st} degree
	21.99	57.87	16.44	3.70
	79.86		20.14	
ON (Seg)	4^{th} degree	3^{rd} degree	2^{nd} degree	1^{st} degree
	21.76	54.16	5.56	18.52
	75.92		24.08	
ON+Lex (Seg)	4^{th} degree	3^{rd} degree	2^{nd} degree	1^{st} degree
	25.00	63.66	7.87	3.47
	88.66		11.34	

Table VIII. Translation Model Performance Using Alignment Approach II

Model	**Accuracy** (%)		**Error Rate** (%)	
Baseline+Lex (Seg)	4^{th} degree	3^{rd} degree	2^{nd} degree	1^{st} degree
	21.99	57.87	16.44	3.70
	79.86		20.14	
ON (Seg)	4^{th} degree	3^{rd} degree	2^{nd} degree	1^{st} degree
	30.55	59.72	4.64	5.09
	90.27		9.73	
ON+Lex (Seg)	4^{th} degree	3^{rd} degree	2^{nd} degree	1^{st} degree
	31.94	60.42	5.32	2.32
	92.36		7.64	
ON+Lex (Unseg)	4^{th} degree	3^{rd} degree	2^{nd} degree	1^{st} degree
	32.87	60.88	5.09	1.16
	93.75		6.25	

Table IX. Performance With and Without Template Preference

Model	**Accuracy** (%)	**Error Rate** (%)
Without Template	90.74	9.26
With Template	93.75	6.25

result with ON word segmentation and the one without word segmentation. Furthermore, we considered the preference of templates to use in the experiment, to validate the efficiency of hierarchical derivation.

Table VII shows the accuracy and error rate for each degree on experiments with Alignment Approach I. Similarity, Table VIII shows the accuracy and error rate for each degree on experiments with Alignment Approach II. Table IX compares the translation performance with and without template preference. In these tables, "Baseline + Lex" denotes the baseline performance adding the lexicon; "ON" means using only the ON translation model; "ON + Lex" means using the ON translation model and the lexicon to assist in word translation; "Seg" means performing segmentation on the ON before translation, whereas "Unseg" means an unsegmented translation.

From the results presented in Tables VII, VIII, and IX, the following facts are clear:

(1) The performance of the proposed model is measurably better than the baseline system, that is, the phrase-based SMT system. This is because the common SMT system is less suited for ON translation. According to our

analysis on translation results, 34 reordering errors and 12 word translation errors from the baseline system are fixed by the structure-based model with Alignment Approach II, which demonstrably shows the superiority of reordering in ON translation.

(2) The performance of the model with Alignment Approach II is remarkably better than Approach I, because Approach II produces CFG rules of high quantity and quality, whereas CFG rules extracted based on Alignment Approach I are too few to cover the entire target vocabulary.

(3) Comparing the results between the ON model and ON + Lex model, it can be seen that adding the LDC lexicon can improve the accuracy slightly and reduce the error rate, while simultaneously increase the rate of the 2nd degree. We investigated the reason that most entries—apart from person names or location names—in the LDC lexicon are not directly available for the ON translation. It appears that the part-of-speech (POS) for a word in an ON is usually a noun or an adjective. However, in the lexicon there are usually many kinds of POS tags and different explanations of a given entry. Thus it is difficult to choose the appropriate meaning for ON translation, and this difficulty reduces the effectiveness of using the lexicon. Hence, CFG rules and word entries of high quality should be primarily extracted from a bilingual NE corpus.

(4) Performance without word segmentation is better than that with word segmentation. This is due to the reduction of segmentation error propagation, especially for ONs that often contain abbreviations.

(5) The model with preferential template matching achieves better performance, because it can efficiently translate ONs that have a similar structure with available templates, for example,“北京中医学院(Beijing College of Traditional Chinese Medicine)” and “山东中医学院(Shandong College of Traditional Chinese Medicine).” Moreover, templates can avoid some segmentation or word alignment errors.

As shown in Table VIII, the ON translation model reaches an accuracy of 93.75%, assisted by lexicon. We analyzed 27 erroneous translation results, and Table X shows the top four error types with their corresponding percentages.

According to our analysis, primarily the following error types attribute the wrong results:

In Example 1 of Table X, the proposed model tends to translate literally, but ONs cannot always be translated literally. There may be some insertion or deletion operations in translation. In the reference translation, “开发” is removed and “in Charge of” is inserted. This type of error is the majority of all translation errors.

In Example 2, “国家实验室” should be the Keyword Chunk, so “National”, the translation of “国家,” should be connected with “Laboratory”, whose translation is “实验室.” However, the chunking model incorrectly set “国家” as the Middle Chunk. An ambiguity sometimes occurs between chunks, and then a chunking error will affect the subsequent CFG derivation and, consequently, the final output.

Table X. Examples of Translation Errors

Example Type	Input Chinese ON	Output by our model	Reference Translation	Error type	Percentage
1	扶贫开发领导小组	Leading Group of Aid-the-Poor and Development	Leading Group in Charge of Aid-the-Poor Projects	ONs are not always translated literally	26%
2	分子反应动力学国家实验室	Laboratory of Molecular Reaction Dynamics National	National Laboratory of Molecular Reaction Dynamics	Chunking error	19%
3	闽台协会	Association for Fujian Station	Fujian-Taiwan Association	Word ambiguity	22%
4	上海化学新材料中试基地	Shanghai Chemistry New Materials Pilot-Plant Base	Shanghai New Chemical Materials Pilot-Plant Base	Word reordering error	19%

In Example 3, "台" is the abbreviation of "台湾", but it was translated to "Station" by error. Many ONs contain OOVs or common words that cannot be accurately translated with their precise meaning. To illustrate with another example, the common word "部(ministry/ department)" in "国家司法部(State Ministry of Justice)" is translated as "ministry", but the same word in "销售部(Sales Department)" should be translated as "department". This phenomenon is very common in NE translation and deserves further investigation.

In Example 4, "化学新材料" was correctly set as the Middle Chunk, but the included words have been wrongly reordered due to the distortion model. This is similar to word reordering errors found in the phrase-based SMT system.

The vocabulary range of the training data and lexicon limits the ON translation model. Furthermore, alignment errors that occur in the training process bring errors to the final translation. Most of errors are caused either by implicit or unknown expressions at the word level or by an ambiguity between chunks, rather than being caused by the whole chunk-based structure. The structure-based synchronous CFG derivation, limited to five steps, has proved to be effective.

To eliminate some of the errors, we will follow the latest SMT research to improve the distortion model with insertion or deletion on the word level. Moreover, we need pay more attention to distinguishing different expressions of word in an NE and try to confirm its real meaning by considering the word relationships. In other words, the structure-based translation model could be combined with further minor structure information at the word level to achieve a better performance.

In addition, nonobjective ONs, journal names, or some transliterated parts are more appropriate for equivalence extraction than direct translation because they usually contain special meanings, for example,

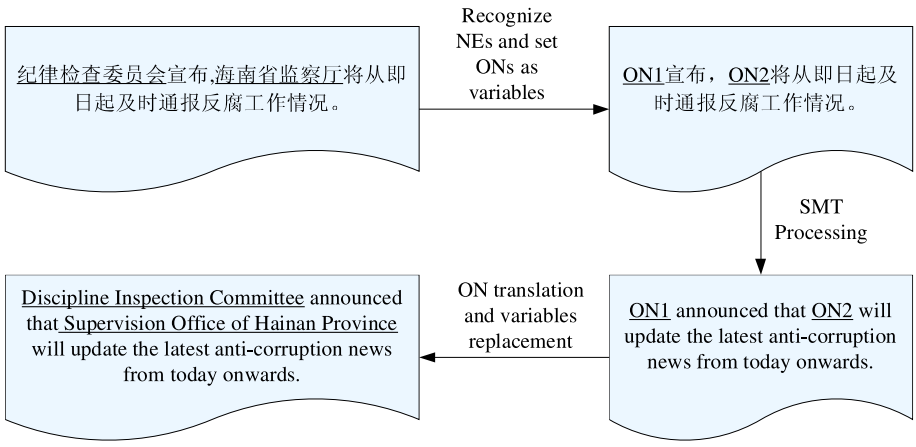

Fig. 7. Flowchart of integrating ON translation model into SMT.

"海湾之花(zahia al-khaleej)." In future work, we will specifically address the approach to translating such types of ONs.

6.2 Improving Translation Quality by Adding a Translation Model for the Organization Name

The ON translation model is integrated into an SMT system in the following process: First, it requires the word segmentation and named entity recognition toolkit developed by our lab to segment the sentences and recognize NEs, in which ONs are labeled as variables in the sentences. Then the SMT system is run to translate the sentences with ON variables. Finally, ONs are translated by the structure-based translation model, and the variables in the sentences are replaced by the outputs of ON translation.

We take a sentence for example: "纪律检查委员会宣布,海南省监察厅将从即日起及时通报反腐工作情况。(Discipline Inspection Committee announced that Supervision Office of Hainan Province will update the latest anti-corruption news from today onward)." Here, "纪律检查委员会 (Discipline Inspection Committee)" and "海南省监察厅 (Supervision Office of Hainan Province)" are recognized as ONs and replaced with variables "ON1" and "ON2" respectively. "Discipline Inspection Committee" and "Supervision Office of Hainan Province" are the translation outputs of the ON translation model. Figure 7 shows the flowchart of integrated translation. The underlined words denote the location of ONs.

To clearly reflect how much the translation quality could be improved by using our proposed approach when an ON is embedded — besides applying the test on a widely adopted NIST data set to be specified next — an additional 120 sentences have been collected specifically for this purpose and tested. This small data set has been selected from the evaluation corpus of the National High-Tech Program of China (named the 863 Program) in 2004. The selection criterion is that each sentence includes at least one ON, and the ON contains more than two words. The set includes more than 3,000 Chinese words in all,

Table XI. Translation Quality Improvement by Adding ON Translation Model

System	BLEU Score	NIST Score
Phrase-based SMT	0.1132	5.4887
+ON model	0.1571	5.9492

Table XII. Performance on the NIST 2005 Data Set With and Without ON Translation Model

Language Model	ON Translation Model	BLEU	NIST
3-gram	Not used	0.1832	7.0944
3-gram	Used	0.2186	7.7462
4-gram	Not used	0.1889	7.2524
4-gram	Used	0.2207	7.7669

and the average length of the sentences is 53 characters, around 24 words. There are 127 ONs included in total, of which the average length is 7.5 Chinese characters. The baseline system that we use for comparison is a phrase-based SMT system. Our evaluation metrics are fully automatic, including the BLEU (Bilingual Evaluation Understudy) score [Papineni et al. 2002] with default settings. Adding the ON model yields a statistically significant BLEU score beyond the baseline system.

The BLEU score of the baseline system is a little low because most of the sentences in the test data are too long. From Table XI, we can see that our integrated system achieves a relative improvement of 38.8% over the baseline system. Clearly, the ON translation model gives significant assistance to the SMT system.

To validate whether the ON translation model is sensitive to the domain, we also conducted experiments on the 2005 NIST test set (1,082 sentences) except for 108 sentences, which we used as the development set. ON recognition and variable substitution were performed on both the test set and the development set. There are 473 recognized ONs with a total of 1,811 words. As Table XII shows, the system with the 4-gram and ON translation model achieves the best performance, and the BLEU score increases by 0.03 when ON identification and translation were introduced.

7. CONCLUSION

Traditional NE translation focuses on extracting NE pairs from a parallel/comparable corpus or Web resource. Such an approach cannot yield satisfactory results for organization names that contain high levels of compacted information and are complex in structure. Taking into account the characteristics of Chinese organization names, in this article we have adopted a “chunk” unit to analyze the ON structure based on the alignments of ON pairs, and proposed a structure-based approach for translating ONs directly. The structure-based model includes a chunking model and a chunk-based CFG model with defined derivation. The CFG rules are classified into four types: “glue” rules, templates, common rules, and special rules according to their hierarchical derivation rank. Templates and common rules are learned from bilingual ONs without any syntactically annotated training data; special rules, which are defined in CFG format based on word level, are applied to cover more ON translation.

The contributions of this article are summarized as follows. First, by analyzing the inherent ON structure, we showed that ON components follow a definite formula that allows the designation of three types of chunks. This chunk-unit can sum up different ON translation patterns and provide the rationale for a chunk-based translation model. Second, the model contributes to flexibility and accuracy for ON translation, because a whole ON has been divided into chunks. The defined derivation of five steps follows the hierarchy of CFG rules, and conforms to the reordering formula of three chunks. Our experiments have proved that the proposed structure-based translation model achieves high levels of accuracy. Third, in the training process we developed a novel alignment approach for the word alignments of ON pairs and then generated numerous, high-quality synchronous CFG rules. Finally, the ON translation model demonstrated a significant improvement in translation quality when it was integrated into an SMT system.

We have built a translation foundation for ON according to its structure. It is important to note that the proposed structure-based translation model is flexible by hierarchical derivation. Moreover, lexicons and a transliteration model can be added via special rules to improve the performance. But it remains clear that the model requires more high-precision synchronous CFG rules and explicit word meanings for ON translation. In this article, we proposed a training architecture and compared two alignment approaches. Approach II proved to be more effective. Our future research will focus on how to obtain high-quality rules representing semantic clusters and relations, in order to achieve acceptable English expressions for ONs.

We believe that recognizing and understanding the ON's inherent structure is necessary for its translation; it requires a special translation model other than traditional SMT. The structure-based model presented here translates Chinese ONs into English with an accuracy of 93.75% and achieves significant improvement over the baseline SMT system. The experiments verify the validity of the model as well as the performance improvement when the ON translation model is added to an SMT system.

ACKNOWLEDGMENT

The authors extend sincere thanks to professor Keh-Yih Su for his keen insights and suggestions. Thanks are also given to the authors' associate, Chunguang Chai, for his great help on the comparison work using the phrase-based SMT system, and to Dr. Fei Huang for his careful proofreading and beneficial suggestions.

REFERENCES

AHO, A.V. AND ULLMAN, J. D. 1969. Syntax directed translations and the pushdown assembler. *J. Comput. Syst. Sci.* 3, 37–56.

AL-ONAIZAN, Y. AND KNIGHT, K. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* 400–408.

CHEN, C. AND CHEN, H.-H. 2006. A high-accurate Chinese-English NE backward translation system combining both lexical information and Web statistics. In *Proceedings of the 21st*

International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL) Main Conference Poster Sessions. Sydney, Australia. 81–88.

CHEN, H.-H., LIN, W.-C., YANG, C., AND LIN, W.-H. 2006. Translating/transliterating named entities for multilingual information access. *J. Amer. Soc. Inform. Sci. Tech.* (Special Issue on Multilingual Information Systems) *57*, 5, 645–659.

CHEN, H.-H. AND CHU, Y.-L. 2004. Pattern discovery in named organization corpus. In *Proceedings of 4th International Conference on Language, Resources and Evaluation.* Lisbon, Portugal. 301–303.

CHEN, H.-H., YANG, C., AND LIN, Y. 2003. Learning formulation and transformation rules for multilingual named entities. In *Proceedings of the ACL Workshop on Multilingual and Mixed-Language Named Entity Recognition.* 1–8.

CHIANG, D. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL.* 263–270.

CHINCHOR, N. AND MARSH, E. 1997. MUC-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference.* http://www.itl.nist.gov/iaui/894.02/relatedprojects/muc.

FENG, D., LV, Y., AND ZHOU, M. 2004. A new approach for English-Chinese named entity alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'04).* 372–379.

GAO, W., WONG, K.-F., AND LAM, W. 2004. Phoneme-based transliteration of foreign names for OOV problem. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP'04).* Sanya, Hainan. 374–381.

HU, R., ZONG, C., AND XU, B. 2006. An approach to automatic acquisition of translation templates based on phrase structure extraction and alignment. *IEEE Trans. Audio Speech, and Lang. Proces. 14,* 16560–1663.

HUANG, F. AND VOGEL, S. 2002. Improved named entity translation and bilingual named entity extraction. In *Proceedings of the 4th IEEE International Conference on Multimodal Interface.* Pittsburgh, PA. 253–258.

HUANG, F., VOGEL, S., AND WAIBEL, A. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the Annual Conference of the ACL, Workshop on Multilingual and Mixed-Language Named Entity Recognition.*

HUANG, F., VOGEL, S., AND WAIBEL, A. 2004. Improving named entity translation combining phonetic and semantic similarities. In *Proceedings of the HLT/NAACL.* Boston, MA.

KOEHN, P., OCH, F. J., AND MARCU, D. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL.* 127–133.

KUMANO, T., KASHIOKA, H., TANAKA, H., AND FUKUSIMA, T. 2004. Acquiring bilingual named entity translations from content-aligned corpora. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP'04).* Sanya, Hainan Island, China.

LEE, C.-J., CHANG, J. S., AND JANG, J.-S. R. 2006. Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. *ACM Trans. Asian Lang. Inform. Process. 5*(2), 121-145.

LEWIS, P. M. AND STEARNS, R. E. 1968. Syntax-directed transduction. *J. ACM 15,* 465–488.

LI, H., SIM, K.C., KUO, J.-S., AND DONG, M. 2007. Semantic transliteration of personal names. In *Processings of 45th Annual Meeting of the Association of Computational linguistics (ACL).* Prague, Czech Republic. 120–127.

LI, H., ZHANG, M., AND SU, J. 2004. A joint source channel model for machine transliteration. In *Processings of 42nd ACL.* 159-166.

LIU, Y., LIU, Q., AND LIN, S. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the COLING/ACL'06.* Sydney, Australia. 609–616.

MENG, H., LO, W.K., CHEN, B., AND TANG, K. 2001. Generating Phonetic Cognets to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop.* Trento, Italy.

MOORE, R.C. 2003. Learning translations of named-entity phrases from parallel corpora. In *Proceedings of 10th Conference of the European Chapter of ACL.* Budapest, Hungary.

OCH, F. J. AND NEY, H. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL.* 295–302.

OCH, F. J. AND NEY, H. 2003. A systematic comparison of various statistical alignment models. *Computat. Linguist. 29,* 1, 19–51.

OH, J.-H. AND CHOI, K.-S. 2005. An ensemble of grapheme and phoneme for machine transliteration. In *Proceedings of IJCNLP.* 450–461.

PAPINENI, K., ROUKOS, S., WARD, T., AND ZHU, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).* Philadelphia, PA. 311–318.

SHAO, L. AND NG, H. T. 2004. Mining new word translations from comparable corpora. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING).* Geneva, Switzerland, 618–624.

SPROAT R., TAO, T., AND ZHAI, C. 2006. Named entity transliteration with comparable corpora. In *Proceedings of the COLING/ACL'06.* Sydney, Australia. 73–80.

STALLS, B. AND KNIGHT, K. 1998. Translating names and technical terms in Arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Language.* Montreal, Quebec, Canada.

WU, D. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computat. Linguist. 23*, 3, 377–404.

WU, D. AND WONG, H. 1998. Machine translation with a stochastic grammatical channel. In *Proceedings of the COLING/ACL'98.* 1408-1415.

ZHANG, M., LI, H., SU, J., AND SETIAWAN, H. 2005a. A phrase-based context-dependent joint probability model for named entity translation. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'05).* Jeju Island, Korea. 600–611.

ZHANG, Y., HUANG, F., AND VOGEL, S. 2005b. Mining translations of OOV terms from the Web through cross-lingual query expansion. In *Proceedings of the 28th ACM SIGIR.* Salvador, Brazil. 669–670.

ZONG, C. AND SELIGMAN, M. 2005. Toward practical spoken language translation. *Mach. Trans. 19*, 2, 113–137.

Received December 2007; revised August 2007; accepted April 2007