Learning Source-Target Surface Patterns for Web-based Terminology Translation

Jian-Cheng Wu
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road,
Hsinchu, 300, Taiwan
D928322@oz.nthu.edu.tw

Tracy Lin
Dep. of Communication Eng.
National Chiao Tung University
1001, Ta Hsueh Road,
Hsinchu, 300, Taiwan
tracylin@cm.nctu.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road,
Hsinchu, 300, Taiwan
jschang@cs.nthu.edu.tw

Abstract

This paper introduces a method for learning to find translation of a given source term on the Web. In the approach, the source term is used as a query and part of patterns to retrieve and extract translations in Web pages. The method involves using a bilingual term list to learn source-target surface patterns. At runtime, the given term is submitted to a search engine then the candidate translations are extracted from the returned summaries and subsequently ranked based on the surface patterns, occurrence counts, and transliteration knowledge. We present a prototype called *TermMine* that applies the method to translate terms. Evaluation on a set of encyclopedia terms shows that the method significantly outperforms the state-of-the-art online machine translation systems.

1 Introduction

Translation of terms has long been recognized as the bottleneck of translation by translators. By reusing prior translations a significant time spent in translating terms can be saved. For many years now, Computer-Aided Translation (CAT) tools have been touted as very useful for productivity and quality gains for translators. CAT tools such as Trados typically require up-front investment to populate multilingual terminology and translation memory. However, such investment has proven prohibitive for many in-house translation departments and freelancer translators and the actual productivity gains realized have been insignificant except for a few, very repetitive types of content. Much more productivity gain could be achieved by providing translation service of terminology.

Consider the job of translating a textbook such as "Artificial Intelligence – A Modern Approach." The best practice is probably to start by translating the indexes (Figure 1). It is not uncommon for these repetitive terms to be translated once and applied consistently throughout the book. For example, A good translation *F* = "聲學模型" for the given term *E* = "acoustic model," might be available on the Web due to the common practice of including the source terms (often in brackets, see Figure 2) when using a translated term (e.g. "**…**訓練出語音聲學模型（Acoustic Model） 及語言模型 **…**")**.** The surface patterns of co-occurring source and target terms (e.g., "*F*（*E*") can be learned by using the Web as corpus. Intuitively, we can submit *E* and *F* to a search engine

Figure 1. Some index entries in "Artificial intelligence – A Modern Approach" page 1045.

academy award, 458
accessible, 41
accusative case, 806
Acero, A., 580, 1010
Acharya, A., 131, 994
achieves, 389
Ackley, D. H., 133, 987
acoustic model, 568

Figure 2. Examples of web page summaries with relevant translations returned by Google for some source terms in Figure 1.

1. **...** 奧斯卡獎 Academy Awards. 柏林影展 Berlin International Film Festival. **...**
2. **...** 有兩個「固有格位」(inherent Case)，比如一個賓格 (accusative Case)、一個與 **...**
3. **...** 有一天，當艾克禮牧師(Alfred H. Ackley) 領完佈道會之後，有一猶太青年來問艾牧師說**..**
4. **..**語音辨識首先 先藉由大量的語料，求取其特徵參數，訓練出語音聲學模型（Acoustic Model）及語言模型**...**

Proceedings of the ACL Interactive Poster and Demonstration Sessions,
pages 37–40, Ann Arbor, June 2005.

and then extract the strings beginning with *F* and ending with *E* (or vice versa) to obtain recurring source-target patterns. At runtime, we can submit *E* as query, request specifically for target-language web-pages. With these surface patterns, we can then extract translation candidates *F*s from the summaries returned by the search engine. Additional information of occurrence counts and transliteration patterns can be taken into consideration to rank *F*s.

Table 1. Translations by the machine translation system *Google Translate* and *TermMine*.

Terms	***Google Translate***	***TermMine***
academy award	*學院褒獎	奧斯卡獎
accusative case	*對格案件	賓格
Ackley	-	艾克禮
acoustic model	*音響模型	聲學模型

For instance, among many candidate translations, we will pick the translations "聲學模型" for "acoustic model" and "艾克禮" for "Ackley, " because they fit certain surface-target surface patterns and appears most often in the relevant webpage summaries. Furthermore, the first morpheme "艾" in "艾克禮" is consistent with prior transliterations of "A-" in "Ackley" (See Table 1).

We present a prototype system called *TermMine*, that automatically extracts translation on the Web (Section 3.3) based on surface patterns of target translation and source term in Web pages automatically learned on bilingual terms (Section 3.1). Furthermore, we also draw on our previous work on machine transliteration (Section 3.2) to provide additional evidence. We evaluate *TermMine* on a set of encyclopedia terms and compare the quality of translation of *TermMine* (Section 4) with a online translation system. The results seem to indicate the method produce significantly better results than previous work.

2 Related Work

There is a resurgent of interested in data-intensive approach to machine translation, a research area started from 1950s. Most work in the large body of research on machine translation (Hutchins and Somers, 1992), involves production of sentence-by-sentence translation for a given source text. In our work, we consider a more restricted case where the given text is a short phrase of terminology or proper names (e.g., "acoustic model" or "George Bush").

A number of systems aim to translate words and phrases out of the sentence context. For example, Knight and Graehl (1998) describe and evaluate a multi-stage method for performing backwards transliteration of Japanese names and technical terms into English by the machine using a generative model. In addition, Koehn and Knight (2003) show that it is reasonable to define noun phrase translation without context as an independent MT subtask and build a noun phrase translation subsystem that improves statistical machine translation methods.

Nagata, Saito, and Suzuki (2001) present a system for finding English translations for a given Japanese technical term by searching for mixed Japanese-English texts on the Web. The method involves locating English phrases near the given Japanese term and scoring them based on occurrence counts and geometric probabilistic function of byte distance between the source and target terms. Kwok also implemented a term translation system for CLIR along the same line.

Cao and Li (2002) propose a new method to translate base noun phrases. The method involves first using Web-based method by Nagata et al., and if no translations are found on the Web, backing off to a hybrid method based on dictionary and Web-based statistics on words and context vectors. They experimented with noun-noun NP report that 910 out of 1,000 NPs can be translated with an average precision rate of 63%.

In contrast to the previous research, we present a system that automatically learns surface patterns for finding translations of a given term on the Web without using a dictionary. We exploit the convention of including the source term with the translation in the form of recurring patterns to extract translations. Additional evident of data redundancy and transliteration patterns is utilized to validate translations found on the Web.

3 The *TermMine* System

In this section we describe a strategy for searching the Web pages containing translations of a given term (e.g., "Bill Clinton" or "aircraft carrier") and extracting translations therein. The proposed method involves learning the surface pattern

knowledge (Section 3.1) necessary for locating translations. A transliteration model automatically trained on a list of proper name and transliterations (Section 3.2) is also utilized to evaluate and select transliterations for proper-name terms. These knowledge sources are used in concert to search, rank, and extract translations (Section 3.3).

3.1 Source and Target Surface patterns

With a set of terms and translations, we can learn the co-occurring patterns of a source term *E* and its translation *F* following the procedure below:

(1) Submit a conjunctive query (i.e. *E* AND *F*) for each pair (*E*, *F*) in a bilingual term list to a search engine.
(2) Tokenize the retrieved summaries into three types of tokens: I. A punctuation II. A source word, designated with the letter "w" III. A maximal block of target words (or characters in the case of language without word delimiters such as Mandarin or Japanese).
(3) Replace the tokens for *E*'s instances with the symbol "*E*" and the type-III token containing the translation *F* with the symbol "*F*". Note the token denoted as "*F*" is a maximal string covering the given translation but containing no punctuations or words in the source language.
(4) Calculate the distance between *E* and *F* by counting the number of tokens in between.
(5) Extract the strings of tokens from *E* to *F* (or the other way around) within a maximum distance of *d* (*d* is set to 3) to produce ranked surface patterns *P*.

For instance, with the source-target pair ("California," "加州") and a retrieved summary of "...亞州簡介. 北加州 Northern California. ...," the surface pattern "FwE" of distance 1 will be derived.

3.2 Transliteration Model

TermMine also relies on a machine transliteration model (Lin, Wu and Chang 2004) to confirm the transliteration of proper names. We use a list of names and transliterations to estimate the transliteration probability function $P(\tau | \omega)$, for any given transliteration unit (TU) ω and transliteration character (TC) τ. Based on the Expectation Maximization (EM) algorithm. A TU for an English name can be a syllable or consonants which corresponds to a character in the target transliteration. Table 2 shows some examples of sub-lexical alignment between proper names and transliterations.

Table 2. Examples of aligned transliteration units.

Name	*transliteration*	Viterbi alignment
Spagna	斯帕尼亞	s-斯 pag-帕 n-尼 a-亞
Kohn	孔恩	Koh-孔 n-恩
Nayyar	納雅	Nay-納 yar-雅
Rivard	里瓦德	ri-里 var-瓦 d-德
Hall	霍爾	ha-霍 ll-爾
Kalam	卡藍	ka-卡 lam-藍

Figure 3. Transliteration probability trained on 1,800 bilingual names (λ denotes an empty string).

τ	ω	P(τ|ω)	τ	ω	P(τ|ω)	τ	ω	P(τ|ω)
a	亞	.458	b	布	.700	ye	耶	.667
	阿	.271		λ	.133		葉	.333
	艾	.059		伯	.033	z	茲	.476
	λ	.051		柏	.033		λ	.286
an	安	.923	an	安	.923		士	.095
	恩	.077		恩	.077		芝	.048

3.3 Finding and locating translations

At runtime, *TermMine* follows the following steps to translate a given term *E*:

(1) **Webpage retrieval**. The term *E* is submitted to a Web search engine with the language option set to the target language to obtain a set of summaries.
(2) **Matching patterns against summaries**. The surface patterns *P* learned in the training phase are applied to match *E* in the tokenized summaries, to extract a token that matches the *F* symbol in the pattern.
(3) **Generating candidates**. We take the distinct substrings *C* of all matched *F*s as the candidates.
(4) **Ranking candidates**. We evaluate and select translation candidates by using both data redundancy and the transliteration model. Candidates with a count or transliteration probability lower than empirically determined thresholds are discarded.
   I. **Data redundancy.** We rank translation candidates by numbers of instances it appeared in the retrieved summaries.
   II. **Transliteration Model.** For upper-case *E*, we assume *E* is a proper name and evaluate each candidate translation *C* by the likelihood of *C* as the transliteration of *E* using the transliteration model described in (Lin, Wu and Chang 2004).

Figure 4. The distribution of distances between source and target terms in Web pages.

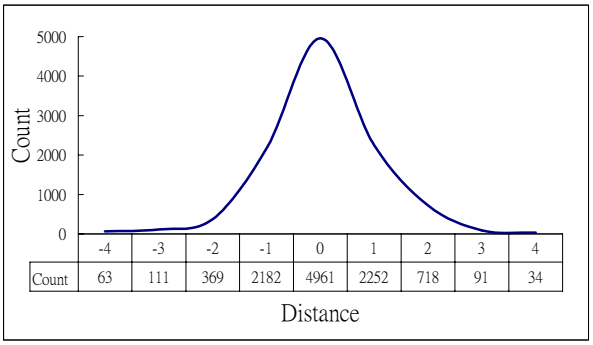

Figure 5. The distribution of distances between source and target terms in Web pages.

Pattern	Count	Acc. Percent	Example	distance
FE	3036	28.1%	亞特拉斯 ATLAS	0
EF	1925	45.9%	Elton John 艾爾頓強	0
E(F	1485	59.7%	Austria(奧地利	-1
F（E	1251	71.2%	亞特拉斯（Atlas	1
F(E	361	74.6%	亞特拉斯(Atlas	1
F.E	203	76.5%	Peter Pan. 小飛俠	1
EwF	197	78.3%	加州 Northern California	-1
E,F	153	79.7%	Mexico, 墨西哥	-1
F》（E	137	81.0%	鐵達尼號》（Titanic	2
F」（E	119	82.1%	亞特拉斯」（Atlas	2

(5) **Expanding the tentative translation**. Based on a heuristics proposed by Smadja (1991) to expand bigrams to full collocations, we extend the top-ranking candidate with count *n* on both sides, while keeping the count greater than *n*/2 (empirically determined). Note that the constant *n* is set to 10 in the experiment described in Section 4.

(6) Final ranking. Rank the expanded versions of candidates by occurrence count and output the ranked list.

4 Experimental results

We took the answers of the first 215 questions on a quiz Website (www.quiz-zone.co.uk) and hand-translations as the training data to obtain a of surface patterns. For all but 17 source terms, we are able to find at least 3 instances of co-occurring of source term and translation. Figure 4 shows distribution of the distances between co-occurring source and target terms. The distances tend to concentrate between - 3 and + 3 (10,680 out of 12,398 instances, or 86%). The 212 surface patterns obtained from these 10,860 instances, have a very skew distribution with the ten most frequent surface patterns accounting for 82% of the cases (see Figure 5). In addition to source-target surface patterns, we also trained a transliteration model (see Figure 3) on 1,800 bilingual proper names appearing in Taiwanese editions of *Scientific American* magazine.

Test results on a set of 300 randomly selected proper names and technical terms from Encyclopedia Britannica indicate that *TermMine* produces 300 top-ranking answers, of which 263 is the exact translations (86%) and 293 contain the answer key (98%). In comparison, the online machine translation service, *Google translate* produces only 156 translations in full, with 103 (34%) matching the answer key exactly, and 145 (48%) containing the answer key.

5 Conclusion

We present a novel Web-based, data-intensive approach to terminology translation from English to Mandarin Chinese. Experimental results and contrastive evaluation indicate significant improvement over previous work and a state-of-sate commercial MT system.

References

Y. Cao and H. Li. (2002). *Base Noun Phrase Translation Using Web Data and the EM Algorithm*, In Proc. of COLING 2002, pp.127-133.

W. Hutchins and H. Somers. (1992). *An Introduction to Machine Translation*. Academic Press.

K. Knight, J. Graehl. (1998). *Machine Transliteration*. In Journal of Computational Linguistics 24(4), pp.599-612.

P. Koehn, K. Knight. (2003). *Feature-Rich Statistical Translation of Noun Phrases*. In Proc. of ACL 2003, pp.311-318.

K. L. Kwok, *The Chinet system*. (2004). (personal communication).

T. Lin, J.C. Wu, J. S. Chang. (2004). *Extraction of Name and Transliteration in Monolingual and Parallel Corpora*. In Proc. of AMTA 2004, pp.177-186.

M. Nagata, T. Saito, and K. Suzuki. (2001). *Using the Web as a bilingual dictionary*. In Proc. of ACL 2001 DD-MT Workshop, pp.95-102.

F. A. Smadja. (1991). *From N-Grams to Collocations: An Evaluation of Xtract*. In Proc. of ACL 1991, pp.279-284.