A Chinese-to-Chinese Statistical Machine Translation Model for Mining Synonymous Simplified-Traditional Chinese Terms

Jing-Shin Chang and Chun-Kai Kung

Department of Computer Science & Information Engineering
National Chi-Nan University
1, Univ. Road, Puli, Nantou, Taiwan, ROC.
jshin@csie.ncnu.edu.tw & s94321541@ncnu.edu.tw

Abstract

A monolingual Chinese-to-Chinese SMT model as well as a global optimization strategy are proposed in this paper to extract equivalent Chinese terms (such as "雷射" lae-ser and "激光" gi-guan for "laser") which are used in different areas of the various Chinese-speaking communities. Preliminary evaluation shows that the synonymous traditional Chinese (TC) terms for simplified Chinese (SC) terms can be identified with an accuracy of 84% on a small test set. On the other hand, the traditional-to-simplified Chinese term translation achieves 87% accuracy. Furthermore, the global optimization strategy generally improves the performance by decreasing search errors. The main idea behind the model is to create "parallel left/right contexts" out of non-parallel web pages for term alignment. The monolingual SMT model, by its very nature to find translation equivalents can potentially be useful for finding synonym sets (synsets) for any generic monolingual lexicon. The potential for adapting such a model for mining large synsets from non-parallel corpora is therefore expectable.

Introduction & Motivation

Problems with Regional Variations

Regional variation in language usage is an important barrier for communication between people of different regions, even though they may share a common core of one language. This situation had been observed across the Mainland China, Hong Kong, Taiwan, Singapore and some South Asia regions, where the Chinese language is used as the same core but with different variations in vocabularies. For instance, while '雷射' lae-ser is used in Taiwan for 'laser', the same term is expressed as '激光' gi-guan in the Mainland China, which does not only use different characters but also represent the same meaning in a very different way. Similar variations had also been observed even across regions that use Chinese characters in different languages. This is exactly the situation between the Chinese and Japanese languages.

In language processing applications, such regional variation will result in difficulties that are frequently observed in cross-language applications. For instance, to acquire information from a search engine, one may have to provide the above two forms of "laser" in order not to miss any interesting information for things like "laser printer".

On the other hand, such variant terms can be regarded, in some sense, as synonyms. Therefore, language processing tasks using mixed corpora with regional variations as the training materials may have to normalize the corpora to the same canonical form before being used for training. In summary, many information-processing applications will suffer from such regional variation if correct translation between the variant terms cannot be well resolved.

In particular, the current research is interested in the variation of term usage between the Taiwan region and the Mainland China region, which use traditional Chinese (TC) and simplified Chinese (SC) characters in documents, respectively. The rapidly increasing number of Chinese documents on the Web and the booming interaction across these two regions make it clear.

Characteristics of the SC-TC Variations and Related Works

The variations between SC-documents and TC-documents can be categorized into several different levels (Halpern, 1999). The most serious level is the use of different terms in different regions for representing the same meaning. For instance, the 'taxi' is named differently as '出租汽車' chu-ju-chi-che, '計程車' gi-cheng-che and '的士' di-shi (or even '德士' der-shi) in China, Taiwan and Hong Kong, respectively.

Such variations cannot be easily translated between SC and TC without knowing the context, and thus will introduce problems for information processing. Therefore, contextual terms or other hints have to be consulted. For instance, in (Lu, 2002, 2003), anchor tags pointing to the same URL will provide strong support that the embedded anchor texts of the anchor tags are referring to the same entity; if the embedded texts are multilingual (including regional variations), then they may form translation equivalent of each other. As an example, if 'SONY', '索尼' so-ni and '新力' shin-li are enclosed by anchor tags that have the same URL reference to 'http://www.sony.com/,' then these three terms might be translation of each other.

Using this kind of strong structural hints has the advantage that multilingual translation (including regional variations) can be identified at the same time with reasonable precision. Unfortunately, only a small amount of such terms (for important persons, organizations and companies) will be anchor-tagged. The majority of terms that are produced due to regional variations (like 'taxi') might never be tagged in such a strict way. The vocabulary size of the translation lexicon acquired in this way will then be limited. Quantitative analysis in the following paragraphs will make this undesirable limitation even clearer.

Table 1. shows the training set vocabulary sizes of the two dictionaries used in the First and Second SIGHAN word segmentation bakeoffs (Sproat, 2003; Emerson, 2005) from the traditional Chinese and simplified Chinese communities. The sizes of TC-specific, SC-specific and the common vocabularies are also shown (excluding non-lexical items like numerical expressions.).

It is easy to see that the non-overlapping parts of the two dictionaries are large (in comparison with the common vocabulary). About 23K SC-only terms (46% of SC terms) and 104K TC-only terms (79% of TC terms) cannot be seen in another community. Using the above anchor tag approach is unlikely to align all such unused terms, since not all these ordinary terms deserve a special hyperlink in structured documents.

Vocabulary Set	#Words
TC	131,615
SC	51,133
TC ∩ SC	27,437
TC - SC	104,178
SC - TC	23,696

Table 1. Vocabulary Sizes for the Academia Sinica (TC) and PKU (SC) Lexicons.

Tables 2~3 further analyze the distribution of those non-overlapping parts. It can be seen that most non-overlapping words belongs to the noun and verb classes. Among the non-overlapping nouns which are not used in another community, ordinary nouns, personal names, and location names have the largest percentages among all. The organization (company) names (such as 'SONY') and transliterated names (like 'Bush'), which might be enclosed by anchor tags, have only a small percentage. This implies that SC-TC term alignment may not be fully resolved by relying only on annotated tags on web pages. We have to create other kinds of "anchors" from text corpora or find other lexical hints in order to fully align the 23K/104K SC/TC-specific terms. Actually, the ability for large-scale term alignment over those non-overlapping terms will be the main focus of the current research, which had not yet been explored in the literature.

Class	SC + TC	SC-only	TC-only
Function words	4%	0%	0%
Nouns (*Table 3)	45%	81%	67%
Verbs	28%	7%	11%
Adjectives	10%	3%	11%
Verb & Noun	2%	7%	0%
Adj. & Adv.	1%	1%	0%
Idioms	5%	0%	5%
Quantifiers	2%	1%	2%
Ancient terms	0%	0%	1%
others	3%	0%	3%

Table 2. Distribution of Simplified and Traditional Chinese Terms (Sample Size=100)

Nouns	SC + TC	SC-only	TC-only
Location	1%	9%	4%
Person Name	0%	27%	13%
Companies	0%	0%	2%
Transliteration	1%	7%	3%
Ordinary Nouns	43%	38%	45%

Table 3. Distribution of Nouns in Simplified and Traditional Chinese Lexicons

Statistical Machine Translation Model for SC-TC Term Alignment

To remove the above-mentioned language barrier, a large-scale translation lexicon for two different regions might be required. In the current research, the regional variation problem is modeled as a special language translation problem. In particular, a Chinese-to-Chinese Statistical Machine Translation (C2C SMT) model is proposed to "align" terms that are specific to the Mainland China region (where simplified Chinese, SC, characters are used) and those specific to the Taiwan region (where traditional Chinese, TC, characters are adopted.)

By modeling the problem as a C2C machine translation problem, one might quickly jump to the though that the state-of-the-art SMT models (Brown, 1990, 1993) and tools, such as GIZA++ (Och, 2000a, 2000b, 2004), could be used if parallel TC-SC corpora are available. Indeed, there are special websites (like http://www.sogi.com.tw/, '手機王' so-gi-wang 'handset king') which provide SC-TC parallel web pages of limited sizes and domains, whose translation is beyond the orthographic conversion level. Our preliminary tests using the SMT training tool GIZA++ and such a parallel SC-to-TC corpora did achieve term alignment with high precision.

Unfortunately, parallel TC-SC corpora have rarely been constructed, and it might seem weird to construct parallel corpora for a single language. Under such constraints, comparable or even free texts might have to be used for training; and, the model has to pay some attention to create "pseudo" anchors for aligning simplified Chinese terms with their traditional counterparts. For this purpose,

we had actually use a contextual sub-window around the SC-term (or TC-term) in question for matching their counterpart in the TC-corpus (or SC-corpus.) In the following section, the detailed formulation of our SMT model for SC-TC translation/alignment will be discussed.

Identifying Synonyms by Context

The basic property of two synonyms is that they can be used interchangeably in all contexts. Synonyms can therefore be identified by matching their left and right contexts. (It is true that shared contexts do not always imply synonym, but they do provide useful hints to find synonyms.) The more matches one can find the more likely a synonym pair is identified. To identify the correspondence between a simplified Chinese term and a traditional Chinese term, we may check simplified and traditional Chinese text corpora to see if they appear in almost all the same contexts. That is, we can check the context of the SC-term in a simplified Chinese corpus, and then scan over a traditional Chinese corpus for matching patterns that are highly similar to the SC-term and its left/right neighbors. If the same left/right neighbors are also found around the traditional Chinese term, then the embedded traditional Chinese term is likely to be the equivalent of the simplified term. Furthermore, if more than one TC-terms match the contextual patterns of the SC-term, then the one that matches more contextual patterns will be more likely to be the translation equivalent of the SC-term.

For instance, in the simplified Chinese corpus, we may find a short phrase like '一部 數碼 相機' yi-bu su-ma shang-gi ('a digital camera'); if we also find that '一部 數位 相機' yi-bu su-wei shang-gi ('a digital camera') appears in the traditional Chinese corpus, then the term '數碼' su-ma 'digital' in the simplified Chinese texts is very likely to be the translation equivalent of '數位' su-wei 'digital' in a traditional Chinese corpus. Sometimes, even though the left/right neighbors are not so similar, partial similarity might still suggest that they are equivalent. For example, '一部 數碼 相機' yi-bu su-ma shang-gi ('a digital camera') in a simplified Chinese corpus and '一台 數位 相機' yi-tai su-wei shang-gi ('a digital camera') in a traditional Chinese corpus may be sufficient to suggest the equivalence relationship. The more we have such kinds of contextual patterns for '數碼' su-ma and '數位' su-wei, the more likely they should be aligned as translation equivalent of each other. The implication is that we can 'disambiguate' different translation equivalents by their contexts, and accumulate small pieces of matching contexts to enforce the degree of equivalence. We thus have the following general Chinese-o-Chinese machine translation model for identifying translation equivalents between simplified and traditional Chinese terms.

C2C SMT Model for Term Alignment

Assume that a TC-specific lexicon and a SC-specific lexicon are available, the SC-TC term alignment problem is defined as finding the best mapping between terms in the two specific lexicons using TC text corpus and SC text corpus for checking contextual constraints. (In the current state, we actually use the World Wide Web as the corpora for the two language variants.) The common vocabulary for both simplified and traditional Chinese in various contextual windows will then serve as the "anchor terms" or "anchor tags" for matching an SC-term to a TC-term.

To find the best traditional Chinese term *t*, corresponding to a simplified Chinese term *s* in the simplified Chinese-only lexicon, D_S, is equivalent to finding the one with the highest translation probability, $P(t \mid s)$, among those terms in the traditional-specific dictionary D_T. That is,

$$\begin{aligned} t^* &= \arg\max_{t \in D_T} P(t \mid s) \qquad \forall s \in D_S \\ &= \arg\max_{t \in D_T} P(t, s) \\ t &: \text{TC-specific term} \\ s &: \text{SC-specific term} \\ D_T &: \text{TC-specific dictionary} \\ D_S &: \text{SC-specific dictionary} \end{aligned} \tag{1}$$

Since there are so many terms that are specific to the traditional Chinese texts and simplified Chinese texts respectively, it is not easy to identify the correct correspondence without consulting the contexts of the simplified terms and the traditional terms. Intuitively, *s* and *t* are likely to be translation equivalent of each other if a context-window $\langle l_s, s, r_s \rangle$ for *s* is "similar" to a context-window $\langle l_t, t, r_t \rangle$ for *t*, where l_s and r_s represent the respective left and right contexts of the simplified Chinese term *s*, and $\langle l_t, r_t \rangle$ are the left/right neighbors of the traditional Chinese term *t*. In other words, Equation (1) can be modeled as

$$\begin{aligned} P(t,s) &= \sum_{\substack{\langle l_t, r_t \rangle : \langle l_t, t, r_t \rangle \in T_T \\ \langle l_s, r_s \rangle : \langle l_s, s, r_s \rangle \in T_S}} P\left(\langle l_t, t, r_t \rangle, \langle l_s, s, r_s \rangle\right) \\ t^* &= \arg\max_{t \in D_T} \sum_{\substack{\langle l_t, r_t \rangle : \langle l_t, t, r_t \rangle \in T_T \\ \langle l_s, r_s \rangle : \langle l_s, s, r_s \rangle \in T_S}} P\left(\langle l_t, t, r_t \rangle, \langle l_s, s, r_s \rangle\right) \\ T_T &: \text{traditional Chinese text corpus} \\ T_S &: \text{simplified Chinese text corpus} \end{aligned} \tag{2}$$

Note that, the degree of "similarity" between $\langle l_s, s, r_s \rangle$ and $\langle l_t, t, r_t \rangle$ might be easier to estimate than the degree of "equivalence" between *s* and *t*, which cannot be judged directly without any contextual information. For example, if $l_s = l_t$ then the two triples will be similar to some extent; if, in addition, $r_s = r_t$, then the "similarity" will be further enhanced.

The above formulation therefore provides a feasible way to divide the complicated context-free s-to-t translation problem into a large number of context-dependent translation sub-problems, each of which is easier to resolve. Also, the summation operation, over all contextual windows, suggests that the probability $P(t,s)$ is contributed by each $\langle l_s, s, r_s \rangle$ and $\langle l_t, t, r_t \rangle$ pair; those pairs with higher similarity will contribute more than those that are unrelated. With such formulation, one may hopefully estimate $P(t,s)$ easier with the contextual information.

Sometimes, the terms in the triple pairs need not be in the same word order to support the equivalence relationship. The equivalent terms '數碼' su-ma and '數位' su-wei ('digital'), for instance, may appear in different contexts like '一部 數碼 相機' yi-bu su-ma shang-gi and '數位 相機 一部' su-wei shang-gi yi-bu respectively. Given such triple pairs, the equivalent relationship for '數碼' su-ma and '數位' su-wei ('digital') is still valid. Therefore, we may want to take into consideration all different word orders and word alignment patterns, especially for the Chinese language, which is relatively free in word order. For this reason, Equation (2) can further be expressed as

$$\begin{aligned} P(t,s) &= \sum_{\substack{\langle l_t, r_t \rangle, \langle l_t, t, r_t \rangle \in T_T \\ \langle l_s, r_s \rangle, \langle l_s, s, r_s \rangle \in T_S}} \sum_{A} P\left(A, \langle l_t, t, r_t \rangle, \langle l_s, s, r_s \rangle\right) \\ &= \sum_{\substack{\langle t_{-1}, t_1 \rangle, \langle t_{-1}, t_0, t_1 \rangle \in T_T \\ \langle s_{-1}, s_1 \rangle, \langle s_{-1}, s_0, s_1 \rangle \in T_S}} \sum_{a_{-1}, a_0, a_1} P\left(A, \langle t_{-1}, t_0, t_1 \rangle, \langle s_{-1}, s_0, s_1 \rangle\right) \end{aligned} \quad (3)$$

where $A = \langle a_{-1}, a_0, a_1 \rangle$ is an alignment vector associated with the simplified Chinese terms in the triple, such that $a_j = i$ if and only if s_j and t_{a_j} (i.e., t_i) are potential translation pair. To simplify the indexing scheme, the left/right terms are re-indexed in the second equality with the subscripts –1 and 1 respectively, and the simplified or traditional Chinese terms in focus are indexed with 0 in Equation (3) (i.e., $\langle l_s, s, r_s \rangle \equiv \langle s_{-1}, s_0, s_1 \rangle$ and $\langle l_t, t, r_t \rangle \equiv \langle t_{-1}, t_0, t_1 \rangle$).

Web-based Training

The above alignment probability represents an estimation on how two contextual windows are similar to each other when the alignment pattern is known. If the training corpus is small, we may have to simplify the model further and use a standard EM algorithm (Dempster, 1977) for reliably estimating the parameters to fit the training data. Considering the nature of the current task, which requires a large corpus for training, we choose to use the Web as our corpus. In this setting, the parameters can be easily trained by submitting SC-only or TC-only terms to a search engine, and get the expected counts from the returned snippets by simple counting.

Intuitively, $P\left(A, \langle t_{-1}, t_0, t_1 \rangle, \langle s_{-1}, s_0, s_1 \rangle\right)$ will be high if the two context windows are highly "similar" (in probabilistic senses.) In the simplest model, one can simply assign a constant probability mass to the alignment probability when an aligned pair is an exact match. In other words, each match $s_j = t_{a_j}$ (j = -1, +1) will contribute a constant probability K to the alignment probability. And, the alignment probability will be proportional to the number of matched context word pairs. In other words,

$$P\left(A, \langle t_{-1}, t_0, t_1 \rangle, \langle s_{-1}, s_0, s_1 \rangle\right) = K \sum_{j=-1,1} \delta\left(s_j = t_{a_j}\right).$$

Under such circumstances,

$$\begin{aligned} P(t,s) &= \sum_{\langle l_t, r_t \rangle, \langle l_s, r_s \rangle} \sum_{A} P\left(A, \langle l_t, t, r_t \rangle, \langle l_s, s, r_s \rangle\right) \\ &= \sum_{\langle l_t, r_t \rangle, \langle l_s, r_s \rangle} \sum_{A} K \sum_{j} \delta\left(s_j = t_{a_j}\right) \\ &= K \sum_{\langle l_t, r_t \rangle, \langle l_s, r_s \rangle} \sum_{A} \sum_{j} \delta\left(s_j = t_{a_j}\right) \\ &= K \sum_{\langle l_t, r_t \rangle, \langle l_s, r_s \rangle} \begin{bmatrix} \delta(l_t = l_s) + \delta(r_t = r_s) \\ + \delta(l_t = r_s) + \delta(r_t = l_s) \end{bmatrix} \end{aligned} \quad (4)$$

which is proportional to the expected count defined as

$$\hat{c}(s,t) = \sum_{\langle l_t, r_t \rangle, \langle l_s, r_s \rangle} \sum_{A} \sum_{j} \delta\left(s_j = t_{a_j}\right). \quad (5)$$

The last equality assumes that, in a 3-gram window, the word alignment is restricted to 1-to-1 mapping. Therefore, the only allowed alignments, *A,* will be either <-1,0,1> or <1,0,-1>.

Global Translation Optimization

The term-wise optimization function, Equation (1), is normally applied independently of other SC-terms. Therefore, the same TC-term may act as the best translation for two or more SC-terms. Sometimes, this happens simply because a target word appears too often in every context such that it is ranked the best translation for almost all source words. When this happen, the top-1 candidate for most source words, except one, will be incorrect, resulting in very poor performance in finding

the correct translation. A loss of about 50% performance is actually observed! Such source words should actually choose the secondary best (or even lower-ranked) candidate as their best translation if some other source words has a higher *P(t|s)* for the best 't'.

To apply such a global optimization process, all the <s, t, *P(t|s)*> triples are sorted by the translation probability *P(t|s)* first. The translation pairs are then extracted starting from the most likely pair. Furthermore, an <s, t> pair will be extracted as a translation pair only when both of them were *not* seen in any previously extracted pair. The same global optimization steps can also be applied to all <t, s, *P(s|t)*> triples for TC-to-SC translation. Furthermore, the term-wise optimization function is unidirectional in nature. To optimize the translation process by considering both *P(t|s)* and *P(s|t)* scores, the above optimization steps can be applied to the <s, t, max{P(t|s), P(s|t)}> triples in the same manner.

The Web-based term alignment process can now be summarized as follows.

Step1 (Identify TC-specific terms and SC-specific terms): Acquire a TC dictionary and a SC dictionary. Identify TC-specific terms (D_T), SC-specific terms (D_S) and common vocabularies (D_{TS}) in the dictionaries. The terms in D_T and D_S will be the target for term alignment.

Step2 (Generate contextual windows): Submit TC-specific terms and SC-specific terms to a search engine, and use the returned search results as the training corpora. Apply a word segmentation algorithm (Chiang, 1992; Lin 1993) to the SC and TC search results, using the SC dictionary and TC dictionary. Then extract a sub-window of word tokens around SC-specific terms from the search results. Do the same for TC-specific terms in the same manner. Without loss of generality, the algorithm will assume that each sub-window consists of a word n-gram with the same number of word tokens to the left/right of a TC-specific or SC-specific term.

Step3 (Compute expected counts of s-t term pairs): Estimate the expected counts $\hat{c}(t,s)$ for each TC-specific term t, and SC-specific term s, using Equation (5).

Step4 (Estimate the alignment probabilities): The translation probability $\hat{P}(t \mid s)$ for each term pairs can be estimated by normalizing the expected counts.

Step5 (Global Translation Optimization): Sort the <s, t, *P(t|s)*> triples by the translation probability. Extract translation pairs starting from the most probable pairs if both *s* and *t* are never seen in previously extracted pairs.

Experiments

To evaluate the above model, we have collected some well-known simplified-traditional Chinese term pairs, most of which are technical terms for information technology. The SC-terms and TC-terms are submitted to the Google search engine (http://www.google.com/) in order to collect their contextual windows. The Google search engine returns at most 1,000 search results for each query. All the snippets (i.e., summary of a search result) are word segmented using the mixture of the simplified Chinese vocabulary and the traditional Chinese vocabulary derived from the SIGHAN word-segmentation bakeoff corpora (Sproat, 2003; Emerson, 2005). Each contextual window consists of three terms, the central one being the source/target term to be translated/aligned with its counterpart. For each SC-term, its contextual windows are matched against that of the various target TC-terms. The expected co-occurrence counts are then estimated and normalized to compute *P(t|s)* or *P(s|t)*. Global optimization is then applied to prevent high-frequency target terms from being recognized as the best translation for almost all source terms.

Some measures are used to evaluate the performance of the system. The Top1 accuracy rate is the percentage of terms whose correct translation equivalent is ranked at the first place. Top10 including rate is the percentage of correct translations that fall within rank 10. The average reciprocal rank (ARR) is the average of '1/rank' for all the correct translations. The AR (average rank) is 1/ARR, which indicates an average rank for all the correct translations. The most important measure is the global accuracy rate (Acc), which is the percentage of term pairs that are correctly aligned after applying the global translation optimization step to resolve competitive source terms. We have considered three modes of optimization: one is to consider *P(t|s)* in the simplified-to-traditional (S2T) translation process, the other is to consider *P(s|t)* in the T2S process, and a third, labeled as ST+TS, is to jointly consider *P(t|s)* and *P(s|t)* in global optimization.

Size	Mode	Top1	Top10	ARR	AR	Acc
N=31	S2T	48%	77%	0.58	1.72	84%
	T2S	32%	90%	0.51	1.98	87%
	ST+TS					87%
N=50	S2T	34%	72%	0.45	2.20	82%
	T2S	30%	78%	0.43	2.32	82%
	ST+TS					86%

Table 4. Performance of C2C SMT Model

Table 4 summarizes the performance of this model for two different numbers of term pairs. Notice that, even with the highly simplified model for the alignment probability, the overall accuracy rate for finding the right translation is surprisingly high with the global optimization strategy applied. For SC-to-TC translation, an accuracy rate of 84% is observed, and 87% is achieved by the TC-to-SC translation with 31x31 alignment possibilities. When the number of translation pairs is

increased to 50x50 possible alignments, the accuracy drops a little due to the higher task perplexity. (The statistical significance for this difference could also be due to the small sample size.) Yet the global accuracy rates remain high. Also, the performance by joining the *P(t|s)* and *P(s|t)* scores achieves the best performance, which looks reasonable. On average, the correct candidates are ranked at about the second place (AR=2) if without considering global optimization.

It is surprising to notice that the Top1 accuracy rates are not absolutely high, yet the global accuracy rates are encouraging. The differences range from 36% to 55%. It is also noticeable that the global accuracy rate is not directly correlated to the Top1 performance. The reason is that a highly competitive target translation could be the best candidate for many source terms and therefore will introduce a huge yet unknown number of errors if the extraction process is based solely on the top1 performance. It is the *relative* translation probabilities with respective to other competitive pairs that matter in the translation equivalent extraction task. Failing to propose a global optimization strategy may therefore introduce a large searching error.

Table 5 in the Appendix shows the results of aligning 31x31 term pairs with bi-directional global translation optimization. The four errors (in shaded cells) come from mis-aligning 'software' to 'document' and 'chip' to 'interface'. These pairs actually share many common contexts and could be misjudged by non-native speakers too.

A few works related to translation equivalent mining from non-parallel corpora using special structure information, such as anchor texts, for different languages had been reported (Lu, 2002, 2003; Cheng, 2004). Their results for translation equivalent due to regional variations, however, cannot be directly compared with current work. Most such works place their emphases in improving the top1 accuracy rate or top-N including rate. Therefore, the absolute top1 performances are moderate but not really high since no global optimizations seem to be conducted.

Future Works

The major goal of the current study is to make possible large-scale term alignment between TC-only and SC-only terms. Therefore, the current simplified model has an EM variant behind its current form. With the small vocabulary size, the highly simplified alignment probability model has achieved some encouraging results. When the scale of the translation pairs is enlarged, it is expected that a better alignment probability model and training method are required. It is also found that some non-discriminative contexts (such as '的' de) may introduce high frequency competitors when ranking the candidates. Removal of such 'stop terms' from the context windows may be important to further polish the current model, and the context might need be extended to phrasal units of various window sizes.

Since the current task is directly applicable to finding synsets in a monolingual context, it might be possible to use such a model for ontology construction. Therefore, by properly extending the current model, some interesting applications or models might be possible. All such potential will be exploited in the future.

Concluding Remarks

Aligning the full set of terms that are TC-specific with terms that are SC-specific in all general domains has not been exploited so far. In the current paper, a monolingual statistical machine translation model is proposed to overcome such regional variations using a simple yet effective alignment model.

A global translation optimization method is also proposed to effectively utilize the translation probabilities by jointly considering all competitors in ranking the most probable target translation. With this optimization technique, the performances are boosted to more than 82% accuracy.

By its nature for mining synonymous terms in one language, application of the current context matching method might find its way for other applications such as synset discovery.

Acknowledgements

This work is partially supported by the National Science Council (NSC), Taiwan, Republic of China (ROC), under the contract NSC 95-2221-E-260-033-.

References

Brown, Peter F., J. Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. (1990). "A statistical approach to machine translation." *Computational Linguistics*, 16(2):79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. (1993). "The mathematics of statistical machine translation: Parameter estimation." *Computational Linguistics*, 19(2):263–311.

Chang, Jing-Shin and Keh-Yih Su. (1997). "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", International Journal of Computational Linguistics and Chinese Language Processing (CLCLP), vol. 2, no. 2, pp. 97-148, August, 1997.

Cheng, J., Y.-C. Pan, W.-H. Lu, L.-F. Chien. (2004). "Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora." In *Proc. of ACL 2004*, pp. 535-542.

Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su. (1992) "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of ROCLING-V*, pp. 123-146, Taipei, Taiwan, R.O.C., 1992.

Dempster, A. P., N. M. Laird, and D. B. Rubin. (1977). "Maximum Likelihood from Incomplete Data via the *EM* Algorithm", *Journal of the Royal Statistical Society*, 39 (b), pp. 1-38, 1977.

Emerson, Thomas. (2005). "The Second International Chinese Word Segmentation Bakeoff", In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Oct. 14-15, 2005, Jeju Island, Korea.

(http://www.sighan.org/bakeoff2005/)

Halpern, J. and Kerman J. (1999). "The Pitfalls and Complexities of Chinese to Chinese Conversion." *Proc. of the Fourteenth International Unicode Conference*, Cambridge, MA.

Lin, Ming-Yu, Tung-Hui Chiang and Keh-Yih Su. (1993). "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," Proceedings of ROCLING VI, pp. 119-142, 1993.

Lu, Wen-Hsiang, Lee-Feng Chien, Hsi-Jiann Lee. (2002). "Translation of Web Queries Using Anchor Text Mining," *ACM Transactions on Asian Language Information Processing*, Vol. 1, No. 2, Pages 159-172, June 2002.

Lu, Wen-Hsiang. (2003). *Term Translation Extraction Using Web Mining Techniques*, Ph.D. Doctoral thesis, Department of Computer Science and Information Engineering, National Chiao-Tung University, Taiwan, ROC., November, 2003.

Och, Franz J. and Hermann Ney. (2000a). "A comparison of alignment models for statistical machine translation." In COLING-2000: The 18th International Conference on Computational Linguistics, pages 1086–1090, Saarbrucken, Germany, August.

Och, Franz Josef and Hermann Ney. (2000b). "Improved statistical alignment models." In Proceedings of the 38th Annual Meeting of the ACL, pages 440–447.

Och, Franz Josef and Hermann Ney. (2004). "The alignment template approach to statistical machine translation." Computational Linguistics, 30:417–449.

Sproat, Richard and Thomas Emerson. (2003). "The First International Chinese Word Segmentation Bakeoff", In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, July 11-12, 2003, Sapporo, Japan.

(http://www.sighan.org/bakeoff2003/bakeoff_instr.html)

Appendix: Examples of Translation Pairs

位元組	字節	byte	-2.53121
支援	支持	support	-2.55842
數位	數碼	digital	-2.59346
螢幕	屏幕	screen	-2.73968
品質	質量	quality	-2.99494
印表機	打印機	printer	-3.1137
程式	程序	program	-3.20393
雷射	激光	laser	-3.5143
檔案	軟件	file (T) : software (S)	-3.52369
記憶體	內存	memory	-3.55996
滑鼠	鼠標	mouse	-3.64963
硬體	硬件	hardware	-3.70419
視窗	窗口	window	-3.71677
智慧	智能	intelligent	-3.8693
預設	默認	default	-3.89408
解析度	分辨率	resolution	-3.99121
搜尋	查找	search	-4.06546
晶片	接口	chip (T) : interface (S)	-4.09731
硬碟	硬盤	hard disk	-4.10083
通訊	通信	communication	-4.13117
軟體	文件	software (T) : document (S)	-4.20133
行銷	營銷	sale	-4.20861
資訊	信息	information	-4.31582
網際網路	互聯網	Internet	-4.38223
錄影	錄像	record	-4.4393
光碟	光盤	compact disc	-4.45002
介面	芯片	interface (T) : chip (S)	-4.64265
電腦	計算機	computer	-4.65538
行動	移動	mobile	-4.81808
連線	聯機	connection	-4.93523
網路	網絡	network	-5.03734

Table 5. Term Alignment with S2T+T2S Global Optimization (sorted by decreasing order of translation probabilities).