Target Word Selection with Co-occurrence and Translation Information

Su Jian
Kent Ridge Digital Labs,
21 Heng Mui Keng Terrace,
Singapore 119613
sujian@krdl.org.sg

Guo Jin*
Motorola Lexicus Division,
3145 Porter Drive,
Palo Alto, CA 84304
guojin@lexicus.mot.com

Tong Loong Cheong
Kent Ridge Digital Labs,
21 Heng Mui Keng Terrace,
Singapore 119613
tongle@krdl.org.sg

Abstract

This paper presents a hybrid method for target word selection by applying a suitable statistical model to informative linguistic structures. The alternative interpretations of syntactic tuples in a source language sentence are mapped to the target language with bilingual lexicon. The target words are selected then according to the lexical co-occurrence probability in the target language. With the constraints propagating along the argument and modifier structure from modifier to predicate, all alternative target words are selected simultaneously using a tree-like Viterbi decoder. Although without a bilingual corpus, it is attempted to incorporate the translation probability information into the statistical model using some mathematical functions.

1 Introduction

It's well known that a given source language word can often be translated into different target language words depending on the context. The task of target word selection in MT is to decide which target language word is the most appropriate equivalent for a source language word. Consider the following sentence:

It's the most glaring restriction at the moment.

In our dictionary, '*glaring*' has 6 translation candidates (耀眼, 凶险, 明显, 闪耀, 眩目, 瞪视), while '*restriction*' has 3(限制, 拘束, 束缚). The correct translation of '*glaring*', 明显, is different in meaning from the other alternatives, while '*restriction*', 限制, differs mainly in usage.

In order to get the correct target word, different MT systems have different schemes[1]. Interlingua-based MT systems find the best rendering in the target language vocabulary from the interlingua derived from source language analysis. Transfer-based MT systems use lexical transfer rules mapping the source language words to target language words according to their syntactic/semantic features and contexts. To date, most of the efforts rely on syntactic information (part of speech and sub-categorization frame), some of these also use semantic information. Usage differences are mainly dealt with on an ad hoc basis. This usually results in unnatural translations.

Instead of using manually-constructed linguistic knowledge as in previous systems, the hybrid method proposed in this paper attempts to solve the problem with word co-occurrence statistics in the target language corpus. As one might expect, the correct translation pair, 明显限制 appears much more frequent in target language corpora than any other alternative pairs. It's these statistics that we use to select the most appropriate target words.

Word co-occurrence statistics has also been used by Ido Dagan[2] to carry out target word selection. Our method differs form Dagan's in several respects. One is the constraint propagation procedure that selects a consistent translation for sentences, where several ambiguous words appear in several tuples(two words and their syntactic relation — the relations between predicate and arguments, and also between head and modifier). Instead of selecting translations of syntactic tuples in descending order of co-occurrence probabilities(as in Dagan's method), we use dependency argument structure to consider all the constraints and their probabilities at same time. A tree-like Viterbi decoder then selects a consistent translation for all words with alternative translations. Thus more linguistic information from the output of source language parser are used, compared to Dagan's method. Furthermore, with a more sophisticated search strategy, N-best translations can be obtained.

Another difference lies in the use of the translation probabilities, where a mathematical function is used to mimic the effect of the translation probabilities without a bilingual corpus. Although this part still needs further refinement, our experiment has shown the benefits of this information.

* This work was done when the author was at Kent Ridge Digital Labs.

Experiments were carried out on an English news article randomly selected from the Editorial/Opinion pages of the International Herald Tribune, in which 163 target words were selected using our model. The result is very encouraging: we achieved 37% to 45% error reduction rate compared to the default translation which simply use the most frequent translation for each source word.

Section 2 of the paper describes several linguistic preprocessing procedures to obtain linguistic structure and data needed by the target word selection process, which employs a syntactic parser and a bilingual lexicon. Section 3 presents the statistical model, where a tree-like Viterbi search engine has been developed to take into account simultaneously all syntactic tuples in the sentences. The translation probability is also addressed in this section. Section 4 describes the experimental setting, results and analysis, with conclusions in Section 5.

2 Linguistic Scheme

Before the statistical model takes effect, there are several steps in the procedure to obtain the necessary information. This includes a) getting the argument structure of the input sentence, b) locating the word translation ambiguity, c) mapping the syntactic tuple to the target language, d) calculating the target word co-occurrence frequency.

a) Getting the argument structure of the input sentence

Tapestry English parser[3] outputs a dependency argument structure for each English sentence. It use a 3-slot valency that corresponds to deep subject (ARG0), deep object(ARG1), indirect object(ARG2), and other modifiers(CIRC), while the heads of phrases and their modifiers or adjuncts are also connected as ATG(attributes of the governor). These syntactic tuples, are hence connected to each other in the output structure of each sentence.

b) Locating word translation ambiguity

All possible translations for each content word in the source language sentence are found with a bilingual MT dictionary. Some translations will be eliminated by syntactic rules in the dictionary, according to the syntactic feature or context of that word in the source language sentence. The remaining ambiguous translations will be solved by the selectional constraints existing in corresponding target language syntactic tuples, which is represented by syntactic tuple probabilities.

This basic concept is based on our understanding that words in direct syntactic relation have strong dependencies. Chomsky defines such dependencies as selectional relations[4]. Subject and verb, for example, have a selectional relation, and so do verb and object. Subject and object, on the other hand, are assumed to be chosen independently of one another. It should be noted that this independence is only an approximation. (Such second-degree effects are considered indirectly through the dependency structure in this model.)

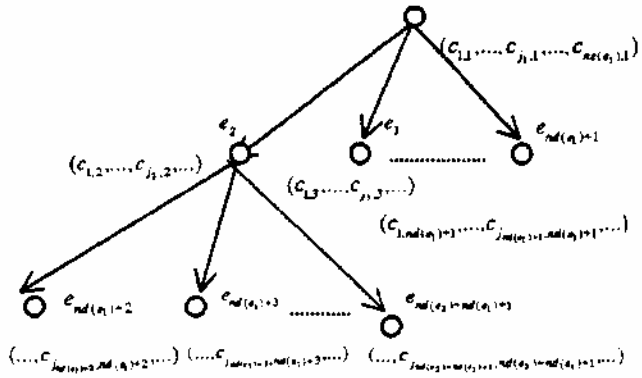
c) Mapping syntactic tuples to the target language

At this point, English syntactic tuples are mapped to corresponding Chinese syntactic tuples. Theoretically, the mapping is somehow problematic. There may not be one-to-one mapping from source language relations to target ones. In some cases the mapping depends on the words of the syntactic tuple.

In practice, this is less of a problem. In most cases, the source language relation has a direct equivalent in the target language. In other cases, lexicon or syntactic transformation rules can be used to make accurate mappings.

d) Calculating the target word co-occurrence probability

Since a Chinese parser is not available in our experiment, each Chinese syntactic tuple is approximated by two Chinese words which co-occur in the most frequent word order for that syntactic relation (e.g., an adjective precedes the noun it modifies). The word co-occurrence probability of given Chinese translations is calculated within a short distance in the sentences from the Chinese corpus. These word co-occurrence frequencies are then used, along with lexical translation information, to solve the remaining ambiguities in the statistical model.



(a)

Fig. English content words connected in the dependency tree structure.

3 Statistical Model

In general, a source sentence may contain several ambiguous words that appear in several tuples, while different relations put different constraints on an ambiguous word, and may favour different translations. As seen in the previous section, these ambiguous source language syntactic tuples are connected in the dependency structure, with predicate/phrase head at root/parent node, arguments/modifiers at child/leaf node (Figure 1(a)).* This structural information is used in our statistical model for considering these constrains at the same time and a tree-like Viterbi decoder is devised for selecting a

* In practice, the whole dependency are separated by words without ambiguous, which means instead of one dependency tree, there may be several dependency tree in a sentence.

consistent translation for all ambiguous words in the structure.

3.1 Statistical selection model

To describe the statistical model, the following notation is used:

$e = \{e_1, e_2, ...\}$, the set of English content words in the tree structure, where e_1 is the word in the root node;

$c = \{c_1, c_2, ...\}$, the set of Chinese translation for e;

$nd(e_i) / nd(c_i)$, number of child node of node e_i / c_i;

$e_{sdn}(e_i) / c_{sdn}(c_i)$, the start child node of node e_i / c_i;

$nc(e_i)$, Number of translation candidates for English word e_i;

$c(e_i) = \{c_{i,1}, c_{i,2}, ..., c_{i,j}, ...\}$, the set of Chinese translation candidates for the English word e_i, where $c_{i,j}$ is the jth candidates;

$A = \{a_{j,j_k} = \Pr(c_{j_k,i} / c_{j,i})\}$, the co-occurance probability of Chinese terms $c_{j,i}$ and $c_{j_k,i}$, where c_j is the father nod of c_{j_k};

$B = \{b_{j,i}(e_i) | b_{j,i}(e_i) = \Pr(c_{j,i} / e_i)\}$, the lexical translation probability;

$\Pi = \{\pi_{j,i} | \pi_{j,i} = \Pr(c_{j,i})\}$, the a priori probability of the Chinese term $c_{j,i}$.

Given the set of English content words e, the statistical model is to find the best Chinese translation c' with the maximum likelihood, i.e. with maximum $\Pr(c|e)$.

$$c' = \arg\max_c \Pr(c|e) \tag{1}$$

$$= \arg\max_c \frac{\Pr(c,e)}{\Pr(e)} \tag{2}$$

$$= \arg\max_c \Pr(c,e) \tag{3}$$

Given the dependency tree structure T_1 and the mapping of syntactic tuples to target language, we have,

$$\Pr(c,e) = \Pr(c_1)\Pr(e_1|c_1)\prod_{i=0}^{nd(c_1)-1}\Pr(c_{2,i}|c_1,e_1)\Pr(e_{2,i}|c_1,c_{2,i},e_1)$$
$$\prod_{m=0}^{nd(c_{2,i})-1}\Pr(c_{sdn(c_{2,i})+m}|c_1,c_{2,i},e_1,e_{2,i})\Pr(e_{sdn(e_{2,i})+m}|c_1,c_{2,i},c_{sdn(c_{2,i})+m},e_1,e_{2,i})... \tag{4}$$

Following the first-order approximations:

$$\Pr(c_{sdn(c_{2,i})+m}|c_1,c_{2,i},e_1,e_{2,i}) \approx \Pr(c_{sdn(c_{2,i})+m}|c_{2,i}) \tag{5}$$

$$\Pr(e_{sdn(e_{2,i})+m}|c_1,c_{2,i},c_{sdn(c_{2,i})+m},e_1,e_{2,i}) \approx \Pr(e_{sdn(e_{2,i})+m}|c_{sdn(c_{2,i})+m}) \tag{6}$$

That is the current translation c_i is only dependent on its parent node c_f, but not any earlier history. In equation (6), we assume that given the current Chinese translation c_i, it is sufficient to observe the English word e_i. With these Markov assumptions, the original equation (4) simplifies tremendously:

$$\Pr(c,e) = \Pr(c_1)\Pr(e_1|c_1)\prod_{i=0}^{nd(c_1)-1}\Pr(c_{2,i}|c_1)\Pr(e_{2,i}|c_{2,i})$$
$$\prod_{m=0}^{nd(c_{2,i})-1}\Pr(c_{sdn(c_{2,i})+m}|c_{2,i})\Pr(e_{sdn(e_{2,i})+m}|c_{sdn(c_{2,i})+m})... \tag{7}$$

Intuitively, here we take the view that the sentence was actually composed by a native Chinese speaker in his mind, but was somehow "distorted" in the way of communication. That is, what a listener heard become an English sentence. Our model is devised to find the original Chinese sentence for the listener.

To the listener, the English sentence is "generated" from an unknown Chinese sentence. The generation pro-

$$\Pr(e_i|c_i) = \frac{\Pr(e_i)\Pr(c_i|e_i)}{\Pr(c_i)}$$

cess takes two steps: $\Pr(c_i|c_f)$ models the production of dependent c_i given its governor c_f in Chinese; and $\Pr(e_i|c_i)$ models the mapping from Chinese word c_i to its real presence e_i in English. Considering that

Equation(9) is the same with equation(7) for choosing c, since e is constant for all c.

$$\Pr(c,e) = \Pr(c_1|c_0)\frac{\Pr(c_1|e_1)}{\Pr(c_1)}\prod_{i=0}^{nd(c_1)-1}\Pr(c_{2,i}|c_1)\frac{\Pr(c_{2,i}|e_{2,i})}{\Pr(c_{2,i})}$$
$$\prod_{m=0}^{nd(c_{2,i})-1}\Pr(c_{sdn(c_{2,i})+m}|c_{2,i})\frac{\Pr(c_{sdn(c_{2,i})+m}|e_{sdn(e_{2,i})+m})}{\Pr(c_{sdn(c_{2,i})+m})}... \tag{8}$$

$$\Pr(c_1|c_0) = \Pr(c_1) \tag{9}$$

Equation (9) shows that 3 factors are considered in this model: target language word co-occurrence probability, lexical translation probability and a priori probability.

An effective and efficient sub-optimal search for the Chinese string that maximizes equation (9) are devised (see appendix for the details). Considering the fact that some times more than one translation is correct, a more sophisticated search strategy can be further developed to obtain N-best translations.

3.2 Lexical Translation Probability

From the above description, it's obvious that we've employed the target language statistics and lexical translation probability in our model. The translation probability has also been used by Brown et al[5] in their target word selection, where the translation probability is derived from a bilingual corpus. However a sufficient large bilingual corpus is usually not available for most domains and most language pairs. Hence Dagan's method incorporates only target language probabilities and ignores any notion of translation probabilities.

In our approach, with the target language probabilities playing the main role, we've tried to incorporate this information without bilingual corpus by some mathe-

$$b_{j,i}(e_i) = \frac{1}{j_i^n} \tag{11}$$

matical functions. Based on the fact that usually a bilingual dictionary are ordered according to their translation frequency, with the most frequent translation in the default first position, we define the following function,

Adjusting the parameter in the experiment, as a result $n = 2$. From the experiment, we found that such a function do improve the precision, by about 5%.

More elaborate work can be done here. In fact, $v_{j_i}(e_i)$ is not only related to the order j_i, but also the number of translation candidates $nc(e_i)$, and the translation candidates themselves. Although we cannot give a different function for different target word in practice, the work on word class is still possible. The word class can be generated from automatic clustering routine[6] [7], or use the predefined class in the thesaurus, such as <<Cilin >>[8], WordNet[9]. Instead of function (11), a three-layer neural network that is well known for realizing accurately any continuous function, having been used as a post processor to deal with the mismatch problem between acoustic model and language model in speech recognition,[10] can be also used here.

3.3 Data Sparseness

As revealed in equation (9), the seleciton of target word depends on three kinds of probability. Given a sub-billion word corpus as was employed in this study, it is straightforward to achieve a satisfactory estimation for target word a priori probability.

Without employing huge bilingual corpus that is almost impossible to collect for open/dynamic domain, it is believed that to achieve significant improvement over what is proposed in the section above for estimating lexical translation probability is almost impossible.

Target word co-occurrence probability is basically a kind of bigram probability. Numerous studies have been conducted on its proper estimation in the context of statistical language modeling, especially in association with its application in speech recognition.[11] [12]For the experiments reported here, we used the backoff approach. In the future work, we'll do experiments with other approaches.

4 Experiment

To evaluate the proposed target word selection method, we implemented and tested the method on an article that is randomly selected from the Editorial/Opinion pages of the International Herald Tribune.

After the preprocessing stages described in section 2, content English words are connected in the argument dependency structure, and some translations are eliminated by the lexical transfer rules for syntactic reasons. Those translations that are missing in the dictionary are added into translation candidates manually. Ignoring words with parsing errors (words are not connected in the right syntactic tuples), words that need sentential translations (for example, idioms.) and words having only one translation (most of them are proper nouns), the selection of the remaining 163 words are evaluated. For each English word we had an average of 4.37 alternative translations and an average of 1.2 correct translations.

The Statistical data are acquired from Chinese corpus consisting of People Daily newspaper articles, with about 350 million Chinese characters. The word co-occurrence probabilities are calculated on the words next to each other. The word unigrams are also calculated.

After acquiring all the relevant data, the algorithm in section 3 is executed for each of the test sentences. The precision of the statistical model was 80%~83%, whereas relying just on Word Frequencies, which always selects the most frequent target word, yields 69%. This result means that compare with Word Frequency procedure, the error reduction rate of our method is 37%~45%. The error reduction is defined as:

$$\text{error reduction} \stackrel{\text{def}}{=} \left(1 - \frac{\text{error on the statistical method}}{\text{error on Word Frequency Procedure}}\right) \times 100\% \quad (12)$$

After analysis of the result, two main errors within this statistical model were found. One is the mismatching problem between the bilingual dictionary and the lexicon list used to segment Chinese corpus. Some translations are not in the segmentation lexicon list, which cause the corresponding statistics unavailable. This problem can be easily solved by putting all the translation words into the segmentation lexicon list, and then re-segmenting Chinese corpus and recalculating the statistics.

Another problem is insufficient data and noise data produced in automatic acquisition of the data. As described above, word co-occurrence probabilities are calculated using the word next to each other, and it will surely miss out target syntactic tuples with the two content words separated from each other, but collecting words that have no syntactic relations. Further experiment can be done with word co-occurrence in a window. Intuitively, smaller window size would introduce less noise but produce higher data sparseness, while larger window size does the opposite. More accurate solution is to have Chinese parser, even partial parser that can produce the syntactic tuples needed in our model.

Even without the above potential improvements, the experiment result has shown the robustness and usefulness of our model.

5 Conclusion

In this paper, we proposed a hybrid method for target word selection in MT by applying a suitable statistical model to highly informative linguistic structures. Unlike pure statistical models that ignore linguistic information, a syntactic parser is used to get linguistically meaningful data and structures on which the statistical algorithm can operate. At the same time, lexical co-occurrences found in a large corpus reflect a huge amount of semantic/usage information. This quantitative information has been used for our resolution of lexical ambiguity, which is traditionally constructed manually on an ad hoc basis or ignored altogether.

For any MT system that needs syntactic parsing, a syntactic parser and bilingual lexicon are available. Hence, only a sufficiently large target language corpus is

needed by our method. With current availability of texts in electronic form, such a corpus is feasible in many domains and for many languages. We believe that our method is feasible and cost effective for target word selection in practical MT systems. It also has potential applications in the area of cross language information retrieval and text mining, where information recorded in other languages is needed.

Based on the linguistic phenomenon that mapping between words and word senses among different language is different, this method has another apparent usage, namely word sense disambiguation. It solves the circularity problem in acquiring sense disambiguation data. Using a bilingual lexicon (which maps each sense separately into its possible translation) and a monolingual corpus of the target language, we can acquire statistics on word senses automatically, without manual tagging. In cases where different senses of a source language word map to the same target word, a third language can be used to distinguish among these senses.

References

[1] Leo Wanner(1996). "Lexical Choice in Text Generation and Machine Translation", Machine Translation, Volume 11, Nos. 1-3.

[2] Ido Dagan, Alon Itai(1994). "Word Sense Disambiguation Using a Second Language Monolingual Corpus", Computational Linguistics.

[3] Wang Tongsheng, Tong Loong Cheong, Tan Chew Lim,(1996). "An Example-Based Approach to Prepositional Phrase Attachment in NLP", ICCC'96.

[4] Chomsky(1965). "Aspects of the theory of syntax". MIT press.

[5] Peter F. Brown et al, "Word sense disambiguation using statistical methods", proceeding, Annual Meeting of the Association for Computational Linguistics 264-270, 1991.

[6] Peter F. Brown et al, "Class-based n-gram Models of Natural Language", Computational Linguistics, Volume 18, number 4.

[7] Haizhou Li, Shuanghu Bai,ICASSP'97.

[8] Mei Jiaju, 1983. Chinese thesaurus <<Tongyici Cilin>>, Shanghai thesaurus Press, Shanghai, China.

[9] George A. Miller, "Introduction to WordNet: An Online Lexical Database", 1993

[10] Guo Jin, Lui Ho Chung, "A Multilayer Perceptron PostProcessor to Hidden Markov Modeling for Speech Recognition", ICASSP'93.

[11] Jelinek F. 1997, Statistical Methods for Speech Recognition, The MIT Press.

[12] Ney H. et al 1997, Statistical Language Modeling Using Leaving One-Out, In Young S. and Bloothooft G. (eds.), Corpus-Based Methods in Language and Speech Processing, Kluwer Academic Publishers.

Appendix

Search Algorithm

Step 1 Bottom-Up. From leaf node to root node,

Initialization. e_i is the leaf node, and e_f is the father node,

$$\delta_{1,j_1} = a_{1,j_1} b_{j_1}(e_1)/\pi_{j_1}$$

$$\psi_{j_1}(i) = \arg\max_{j_1}[\delta_{1,j_1}]$$

Recursion. e_i is an intermediate/root node, $e_d, e_{d+1}, \ldots, e_{d+nd(e_i)-1}$ are daughter nodes, and e_f is the father node,

$$\delta_{i,j_i} = (a_{j_i,j_i} b_{j_i}(e_i)/\pi_{j_i}) \prod_{k=0}^{nd(e_i)-1} \max_{j_{d+k}} \delta_{j_{d+k},d+k}$$

$$\psi_{j_i}(i) = \arg\max_{j_i}[\delta_{i,j_i}]$$

Termination. e_1 is the root node, thus $i=1$, $d=2$, (* indicates the optimized results).

$$P^* = \max_{j_1}[\delta_{n,j_1}]$$

$$j_1^* = \arg\max_{j_1}[\delta_{n,j_1}]$$

Step2 Top-down. From e_2 to leaf node

$$j_i^* = \psi_{j_f}(i)$$

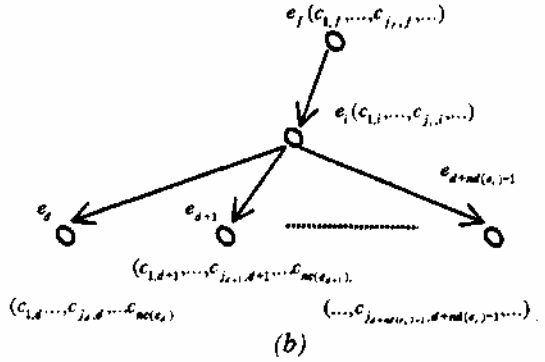


Fig. The tree structure of the English content words