A Method of Automatic Translation of Words of Multiple Affixes In Scientific Literature

Lei Wang[1,2], Baobao Chang[1], Janet Harkness[3]
[1]Key Lab of Computational Linguistics of Ministry of Education, Peking University, Beijing 100871, China
[2]Department of English, Peking University, Beijing 100871, China
[3]UNL Gallup Research Center, Nebraska 68588, Lincoln, USA
wangleics@pku.edu.cn, chbb@pku.edu.cn, jharkness2@unl.edu

Abstract

Inflection and derivation have been the main ways of creating new words in English. With the development of science and technology, words as such appear faster than ever in scientific literature. Influenced by English, Chinese words with multiple affixes are also becoming a major way of new word creation in scientific literature. By studying the similarities of their original sources, this paper employs a head transduction model in an attempt to automatically translate these words from English to Chinese. With this method we hope to solve the problem of such words usually as unknown words in machine translation systems and build a bilingual lexicon with a richer content.

Keywords

scientific literature; words of multiple affixes; automatic translation

1 Introduction

An affix is a morpheme that is attached to a word stem to form a new word. Words with multiple affixes (WMAs) such as "translatability", "postmodernism" and "surrealism" are very common in English. But in Chinese most WMAs are borrowed from Japanese which were also influenced by English. In recent years, research such as in Pan et al. (2004) and Shen (1995) has been conducted to explore the new phenomenon of adding affixes to Chinese words to form new words in scientific literature, which are mainly translated from English or other western languages. Although these words have an English origin, they show their own characteristics. To distinguish them from their counterparts in English, Chinese linguists call them "quasi-affixes". The concept was first proposed by Lu Shuxiang in *The Analysis of Chinese Grammar* in 1978, which turned a new leaf of study Chinese affixes. For this Xu(1997) also remarked: "In Chinese-Tibetan languages, the derivation that plays an important role in new word creation is not those affixes whose senses are fading and that only serve as formal markers, but those quasi-affixes that retain their certain senses". Because quasi-affixes emerge very fast, many new words have been created in this way, especially in scientific literature. As we mentioned above, many Chinese compound words were borrowed from Japanese (more than 20,000) in late Qing Dynasty by introducing Japanese textbooks. In the process, WMAs became a part of Chinese lexicon, e.g. words with common affixes like "-性(-ity)" and "-度(-dom)".

Prefixes	Examples	Suffixes	Examples
软—(soft--)	软着陆(soft-landing)	--化(--ize)	绿化(greenize)
自—(self--)	自尊(self-respect)	--度(--ity)	灵敏度(sensitivity)
类—(quasi--)	类词缀(quasi-affix)	--门(--gate)	伊朗门(Iran gate)
后—(post--)	后现代(post-modern)	--性(--ity)	灵活性(flexibility)

Table 1. Common quasi-affixes in Chinese words

By comparing the WMAs in both English and Chinese, their features that can be used for automatic translation are summarized into the following three categories: First, in English the part-of-speech a word is generally shown by derivation and many multiple affixes are used to indicate their syntactic functions. But this is not true for Chinese. Therefore, when an English WMA is translated, the stem word is usually kept while the affixes will be taken off. For instance, the structure of the word "modernization" is "stem word+suffix+suffix". Its Chinese translation is "现代化(modernization)" and the last suffix has to be omitted and corresponds to the empty character ε.

In contrast, English words with a single affix may be translated to Chinese WMAs. For instance, when a word like "usable" is translated, the Chinese prefix "可-(-able)" has to be added and its translation becomes "可利用的(usable)" (with the structure "prefix+stem word+suffix"). When analyzing this type of words, new characters need to be generated. Some WMAs correspond very well in both languages, e. g., word "nongovernmental" with the structure "prefix+stem word+suffix" and "非政府的(nongovernmental)" with the structure "prefix+stem word +suffix".

The work in this paper considers only verbs and adjectives, the majority of WMAs. Ordinary dictionaries usually cannot collect all the WMAs, especially in scientific literature. They often become "unknown words" in machine translation and cannot be translated successfully. Section 5 provides a few examples for such words.

2 Weighted Head Transducer

This paper aims to employ weighted head transducer (Alshawi, 2000) to translate WMAs automatically. Formally, a weighted head transducer is a 5-tuple: an alphabet W of input symbols; an alphabet V of output symbols; a finite set Q of states $q_0 \ldots q_s$; a set of final states F⊆Q; and a finite set T of state transitions. A transition from state q to state q' has the form where w is a member of W or is the empty string ε; v is a member of V or ε; the integer α is the input position; the integer β is the output position; and the real number c is the weight or cost of the transition. A transition in which $\alpha = 0$ and $\beta = 0$ is called a head transition.

$$< q, q', w, v, \alpha, \beta, c >$$

The interpretation of q, q', w, and v in transitions is similar to left-to-right transducers, i.e., in transitioning from state q to state q', the transducer "reads" input symbol w and "writes" output symbol v, and as usual if w (or v) is ε then no read (respectively write)

takes place for the transition. The difference lies in the interpretation of the read position α and the write position β. To interpret the transition positions as transducer actions, we consider notional input and output tapes divided into squares. On such a tape, one square is numbered 0, and the other squares are numbered 1, 2 . . . rightwards from square 0, and -1,-2 . . . leftwards from square 0 (Fig. 1).

A transition with input position α and output position β is interpreted as reading *w* from square α on the input tape and writing *v* to square β of the output tape; if square β is already occupied, then *v* is written to the next empty square to the left of β if $\beta < 0$, or to the right of β if $\beta > 0$, and similarly, if input was already read from position α, *w* is taken from the next unread square to the left of α if $\alpha < 0$ or to the right of α if $\alpha > 0$.

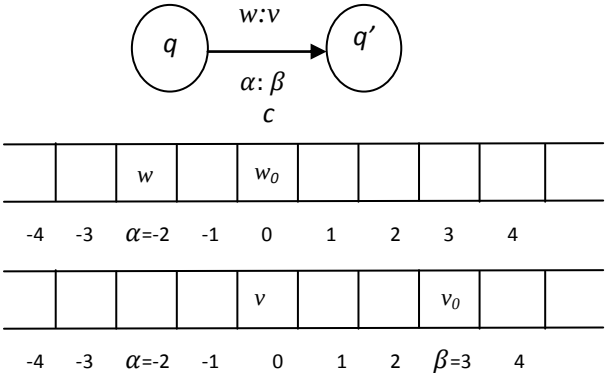

Figure 1. Transition symbols and positions

By head transducers, we can translate headword *w* and its dependent nodes as its source string into the target word *v* and dependent nodes as its target string. We can use conditional probability as the weight for transition and the probability for head words *w* and *v* and their dependents *w'* and *v'* can be as:

$$p(q', w', v', \alpha, \beta | w, v, q)$$

Here *q* and *q'* are the from-state and to-state for the transition and α and β are the source and target positions, as before. We also need parameters $p(q_0, q_1|w,v,)$ for the probability of choosing a head transition given this pair of headwords. To start the derivation, we need parameters $p(roots(w_0, v_0))$ for the probability of choosing w_0 and v_0 as the root nodes of the two trees[1].

3 Automatic Translation of WMAs by Head Transducer

Alshawi(1997) uses head transducer to translate languages L_1 and L_2 from two directions based on two vocabulary sets V_1 *and* V_2. Compared with standard finite state automata, the input string and the output string need not correspond on the word level but can have a certain distance. For WMAs, a translation model with head transducer mainly consists of

[1] In this paper, a stem word lexicon is employed and the head is the verb or the adjective. So the value of this probability is 1.

the following components (Alshawi, 1997):

- A set of head transducers for WMAs;
- Lexicons with WMAs and affixes with lists of probabilities for transduction;
- Search algorithm for computing the probability of input word or affix.

As to a simple grammatical structure as in Fig. 2, a WMA can be analyzed into a tree by rules applied; in the meanwhile a dependency tree of a Chinese WMA will be generated to show the structure of the affixes that are attached to the word stem.

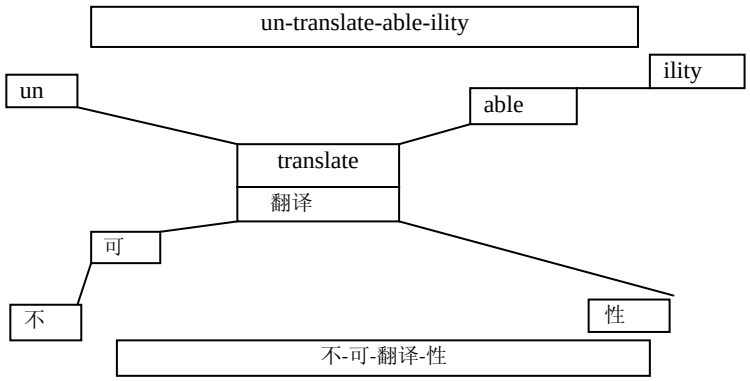

Figure 2. Synchronized dependency tree

4 Experiments to Testify Effectiveness

To do the above lexical analysis on WMA constituents and the translation, we need to construct three lexicons: a bilingual stem word lexicon, a bilingual prefix lexicon and a bilingual suffix lexicon. Statistical test is applied on the following aspects: One is to select sense for the stem word, and Chang (2003) elaborates how corresponding words can be extracted from an aligned parallel corpus. For the lack of training data from scientific literature, many WMAs have data sparseness and are themselves "unknown words". In this paper, we do not assign a probability to a word for selection, but number them to distinguish their senses in the order of the lexicon provided by Language Data Consortium (LDC).

Prefix lexicon		**Suffix lexicon**		**Stem word lexicon**		
prefix	sense/probability	suffix	sense/probability	stem word	common prefixes	common suffixes
un-	未/0.32 不/0.54 没有/0.14	-ity	性/0.95 度/0.04 程度/0.01	translate	un- auto-	-tion -able
ir-	多元/0.95	-ale -ible	可…的/1.00	reverse	ir-	-able -sibilty
…	…	…	…	…	…	…

Table 2. Three lexicons for automatic translation of WMAs

In the above stem word lexicon, the word "translate" as a verb has two common prefixes to be attached and two common suffixes to be attached. The other task to be completed is the selection of affixes, for instance, the prefix for negation in Chinese can be "非-, 不-, 难以-(un-, in-, non-)", while the suffixes that have similar senses are "性, 度, 程度(-ity, -dom, -ility)". We separate the options by "/" along with their probabilities, which are obtained from (1) the PKU news parallel corpus, (2) the LDC lexicons and (3) the PKU People's Daily (1998) corpus and by computing their co-occurrence. (1) and (2) are used to compute the matching probabilities of an affix and its corresponding senses, while (3) is used to obtain the probability of a stem word with the affixes that go with it.

To translate WMAs, our first task is to identify them in scientific literature. As most WMAs have many an affix, the number of affixes of a word is a good indicator to judge whether it is a WMA. We design a stemming algorithm with affixes attached to separate the word stem and its affixes and to identify WMAs by calculating the number of affixes deprived from a word as well. The algorithm is in Fig. 3.

In our work, present participle "-ing" or past participle "-ed" are not considered as suffixes. Words like "predecessor" and "unable", although they start with prefix such as "pre-" and suffix such as "-able", once deprived of their affixes, the rest of the word is not a word in the lexicon, therefore we will not consider it as a WMA.

Input: an English word
Output: word stem and its affixes
Algorithm:

1 Input an English word *W*.
2 Testify whether it has prefixes. If yes, count the number of prefixes C_p and go to 4; if no, go to 3.
3 Testify whether it has suffixes. If yes, count the number of suffixes C_s and go to 4; if no, go to 4.
4 $C_i=C_p+C_s$, if C_i >=2, *W* is a WMA, search the dictionary first; if not found, implement the WMA Translate module. If C_i <2, input next word.
5 Exit.

Figure 3. Stemming algorithm with affixes separated

Upon obtaining a WMA, what we need to do next is to translate it by the lexicons in TABLE II. Since both the lexicon of prefixes and the lexicon of suffixes contain over a couple of hundred items, if we compute the probability of each of them that goes with each item in the word stem lexicon that includes a few thousand items , it will be a time-consuming task. Thus we only select those affixes that are commonly attached to a word stem and neglect those that will never appear before or after a word stem. Another problem is that a word stem itself may contain multiple meanings, especially when it appears in a specific domain it may adopt a specific meaning like "conduct" in physics in Table IV. Unable to do word sense disambiguation well in our work, we have to assume that the user of the translated text is able to select a meaning within a specific domain himself or herself. This job actually can be done once the process of actually translating WMAs starts since many dictionaries provide senses of a word in specific domains with special markers.

To testify the method, two experiments are conducted. Experiment 1 aims to testify its

generality, i. e. the translation of WMAs from different domains. We select 10 sentences with WMAs from www.engkoo.com. The process is as follows: First WMAs are identified from scientific literature by using the affix lexicons and then the stemming algorithm with affixes separated is used to extract its stem word (head) and its affixes (See the "stem word" column in Table III). Finally head transducer is used to translate the WMAs. We compare the translation results with the explanations provided by www.engkoo.com to find the acceptable ones. With judgment from an expert translator, we select those acceptable results from our WMA Translate module based on head transducers and list the results in Table III with the original translated sentences pairs on the left for comparison.

We also input these words into the automatic translation system in Google Translate (Google). We find that the problem with Google is that the translations cannot reflect the changes after affixes are added to a stem word, especially for adjective suffix "-able" and its derivative "-ability". Most of them are translated into the structure of "prefix + stem word". The word "nonconductive" is even translated as "导电(conductive)", its opposite meaning.

WMA	Stem word	Sentence pairs	Acceptable sentences	Percentage
irreversibility	reverse	10	9	90%
incompatibility	compatible	10	8	80%
irregularity	regular	10	5	50%
unavailability	avail	10	7	70%
unsuitability	suit	9	6	67%
nonconductive	conduct	7	5	71%
unpredictability	predict	10	8	80%
decomposable	decompose	10	7	70%
transformable	transform	10	9	90%
interoperability	operate	10	8	80%

Table 3. Comparison of Translations with www.engkoo.com

The purpose of Experiment 2 is to translate WMAs in specific domains. We choose the verb "conduct" and its two senses in physics – "传导(to conduct)" and "导电(to conduct electricity)". From LDC, we find 38 WMAs with the stem word "conduct" and construct its affix lexicons. Then we use head transducers to translate them and compare the results with the senses from LDC and Google's translations as in Table IV.

In Experiment 2 we find that among the 38 words, Google fails on 13 that account for 34% of the total and succeeds on 14 that account for 34%. The rest one-third is the same as in Experiment 1, i.e., the translation is the same as the stem word even though affixes have been added. For instance, five words with the prefix "photo-" are all translated as "光电导(photoconductive)", but actually they mean "光电导体(photoconductor)", "光传导率(photoconductivity)", etc. respectively. In this sense, the errors of Google can account for two-thirds of all the translations. The method in this paper fails on 4 and its main reason is that the lexicons built do not have the affixes needed. If the senses in LDC are used as the standards, the translation errors are 3. The problem with our method is if we abide by the

rules designed strictly but the term itself changes, especially when the term becomes so specialized as to its register is different, our method will fail. For instance, in LDC the sense of the word "photoconductivity" changes to be "光导增益(increment of photoconductivity)" in a particular field of physics.

MWA	**LDC's sense**	**Our translation**	**Google's translation**
conductibility	传导性	传导性 导电性	传导
conductimetry	电导分析法	传导分析法 导电分析法	×
conductivity	导电率 传导率 传导度	传导性 导电性	电导率
conductograph	传导仪	传导仪 导电仪	×
conductometer	电导计 热导计	传导计 导电计	电导(×)
…	…	…	…
multiconductor	多触点	多传导体 多导电体(×)	多触点
paraconductivity	顺电导	×	×
photoconduction	光电导	光传导 光导电	光电导
…	…	…	…
thermoconductivity	热传导率	热传导性 热导电性	×

Table 4. Comparison of Translations in Specific Domains

5 Concluding Remarks

Chao(1968) points out: "In modern Chinese, many words have become disyllabic or polysyllabic. Many monosyllabic morphemes – words, have become inflections in compounds. Also, a few inflect ions in compounds have lost their senses as stem words and become affixes that symbolize their roles in words and form various derivations." The similarities of WMAs in both English and Chinese enable us to formalize them and construct templates for automatic translation. This paper proposes a method for automatic translation of WMAs in both English and Chinese based on head transducers, and conducts a couple of experiments to testify its effectiveness. Our research will provide a solution to WMAs as unknown words in machine translation systems and help to build lexicons with a richer content.

6 Acknowledgment

Our work is supported by a grant from the 973 National Basic Research Program of China (No. 2004CB318102).

7 References

Pan, Wenguo; Ye, Buqing; Han, Yang. Research on Chinese Word-formation. Shanghai: Huadong Normal University, 2004.

Shen, Mengying. The New Trend of Chinese Suffixition. In Proceedings of The First National Conference of Lexical Study of Chinese. Beijing: Language and Literature Press, pp. 32-36, 1995.

Xu, Shixuan. The analysis of word creation by derivation in Chinese-Tibetan languages," in National Languages, vol. 4, pp. 23-31, 1999.

Alshawi, H.; Bangalore, S.; Douglas, S.. Learning Dependency Translation Models as Collections of Finite-State Head Transducers, in Computational Linguistics, vol. 26(1), pp. 46-60, 2000.

Alshawi, H.; Buchsbaum, A. L.; Xia, F.. A Comparison of Head Transducers and Transfer for a Limited Domain Translation Application, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 360-365, 1997.

Alshawi, H.; Xia, F.. English-to-Mandarin Speech Translation with Head Transducers, Proceedings of the Workshop on Spoken Language Translation, pp. 54-60, 1997.

Baobao Chang, Research on translating equivalent word pairs based on statistical models, in Journal of Computer, vol. 26(5), pp. 616-621, 2003.

Chao, Yuen Ren, A Grammar of Spoken Chinese, Berkeley: University of California Press, pp. 24, 1968.