On Detection of Synonyms Between Simplified Chinese of Mainland China and Traditional Chinese of Taiwan: A Semantic Similarity Method

Boli Wang and Xiaodong Shi(✉)

Department of Cognitive Science, Xiamen University, Xiamen 361005, China
me@bo-li.wang, mandel@xmu.edu.cn

Abstract. We present an approach for automatically detecting synonyms between simplified Chinese used in mainland China and traditional Chinese used in Taiwan from large scale corpus. After pre-processing step (including doing segmentation and POS tagging on our corpora), all words are classified into 3 categories according to their frequency: words exclusively used in mainland China, words exclusively used in Taiwan, and words commonly used in both sides. We use word vectors to represent meanings of words, calculate semantic similarities between words of both sides, and extract synonyms. The experiment shows that our approach can find synonyms that are not present in handcrafted dictionary.

Keywords: Synonyms between simplified and traditional chinese · Semantic representation · Semantic similarity · Automatic conversion between simplified and traditional chinese

1 Introduction

There is a lot of differences in language usage between simplified Chinese (SC) used in mainland China (MC) and traditional Chinese (TC) used in Taiwan (TW), including pronunciation, spelling system, punctuation marks, script, vocabulary, grammar and so on. In the aspect of vocabulary, some words have totally different meanings between MC and TW. For example, the word 土豆 [tu dou] means *potatoes* in MC but *peanuts* in TW. We call these words **homographs** between MC and TW. Meanwhile, some different words from two sides have the same meaning. For example, The English name *Obama* is translated into 奥巴马 [ao ba ma] in MC but 歐巴馬 [ou ba ma] in TW. We call these words **synonyms** between MC and TW.

Divergences in vocabulary confuse people from both MC and TW when they read articles from the other side. To handle this problem, some SC-TC automatic conversion systems integrate word conversion function using a MC-TW **synonyms table**. Typically, as reported in (Jinzhi Su, 1995) and (Xingjian Li, 2012), MC-TW synonyms table is constructed by handcrafted compiling, which is time-consuming and bounded by compilers' knowledge scope. Moreover, new words appear at all times and synonyms table thus should be renewed accordingly. In this paper, we present a

Q. Lu and H.H. Gao (Eds.): CLSW 2015, LNAI 9332, pp. 91–100, 2015.
DOI: 10.1007/978-3-319-27194-1_10

statistical approach to detect MC-TW synonyms from large scale corpus automatically. With our method, a MC-TW synonyms table can be constructed in shorter time and include some synonyms ignored by handcrafted work.

There are two types of words that should be listed in the synonym table. (i) Some words exclusively used in one side and incomprehensible to people from the other side. E.g. 离休 [li xiu] (*honored retirement of those engaged in the revolution before 1949*) is an **exclusive word** of MC and 博愛座 [bo ai zuo] (*priority seats for the weak*) is an exclusive word of TW. (ii) Some words are commonly used in both MC and TW but have different meanings in two sides, namely homographs, such as 土豆 [tu dou] we mentioned. These words are likely to be misunderstood and need to be translated into a synonym with the same meaning in word conversion.

Therefore, our task is to find correspondences of synonyms between MC and TW. However, some correspondences are not one-to-one. For example, in conversion from TC to SC, 土豆 [tu dou] should be translated into 花生 [hua sheng] (*peanuts*). But reversely, in conversion from SC to TC, 花生 [hua sheng] does not need to be translated into 土豆 [tu dou] because 花生 [hua sheng] also means *peanuts* in TW. Therefore, two different synonyms tables are needed to handle bidirectional conversions separately.

2 Related Work

When detecting synonyms, we need to calculate the semantic similarity between two words.

In the earlier researches, methods based on knowledge base have been commonly used on semantic similarity computing. (Richardson, 1994) measures semantic similarity of two words according to their relationship in WordNet, which is a knowledge base of English vocabulary. (Qun Liu, 2002) proposes a method to compute semantic similarity of Chinese words based on HowNet, which is a knowledge base of Chinese vocabulary. However, knowledge bases are handcrafted and always incomplete.

Corpus-based approaches to semantic similarity computing have received much attention lately. Since the context surrounding a given word provides important information about its meaning, word meanings can be represented in distributional statistics of words' occurrence. (Chen, 2011) integrates lexical semantics consistency weight as a feature in SC-TC conversion model and semantic similarity is computed using cosine distance measure. (Shi Wang, 2013) and (Jing Shi, 2013) conduct a lot of experiments on different methods of semantic representation and similarity measure. Using deep neural network, (Mikolov, 2013) proposes two novel model architectures for computing continuous vector representations of words and achieves large improvements in accuracy of word similarity task with much lower computational cost.

3 Our Approach

The main purpose of this research is to detect synonyms between MC and TW automatically. The flowchart of our method is shown in Fig. 1.

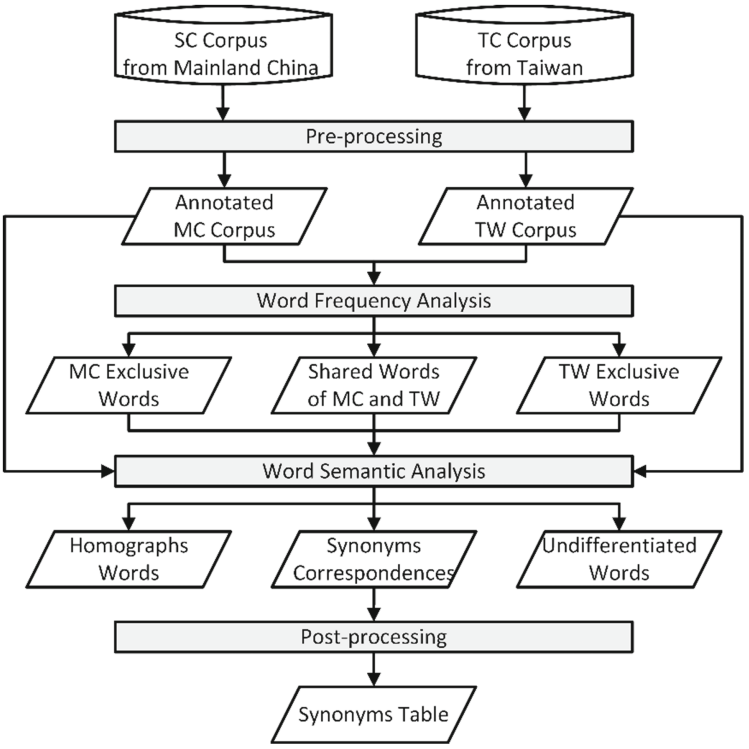

Fig. 1. General Flowchart of Synonyms Detection Method

3.1 Corpus and Pre-processing

A simplified Chinese corpus from MC and a traditional Chinese corpus from TW are needed. We choose MC and TW corpora from the same domain and at the same period, because meanings of words may differ considerably between different domains or periods. For example, in MC, 斑竹 [ban zhu] traditionally means *mottled bamboo*, but now it has a new meaning (*moderator*) in BBS.

Automatic Chinese segmentation with same granularity and POS tagging with same tag set is conducted on both two corpora[1] in pre-processing, which is shown in Fig. 2. We also convert all traditional Chinese characters in Taiwan TC corpus into SC after segmentation, because one word may have two different forms in SC and TC and TC-SC automatic conversion has a higher precision than SC-TC conversion.

[1] In our experiments, we use SEGTAG toolkit (http://cloudtranslation.cc/segtag.html) on both MC and TW corpora.

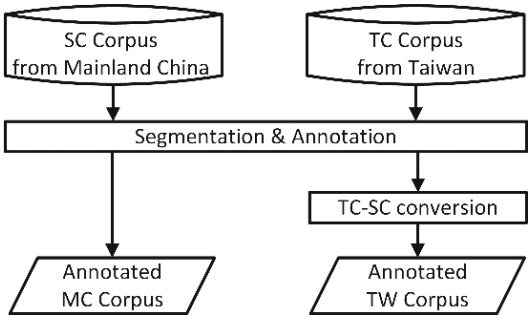

Fig. 2. Flowchart of Pre-processing

After these steps, we obtained an annotated MC corpus and an annotated TW corpus (in SC both).

3.2 Frequency Analysis

We use statistics of word frequency to find out words exclusively used in MC and TW respectively. The flowchart of word frequency analysis is shown in Fig. 3.

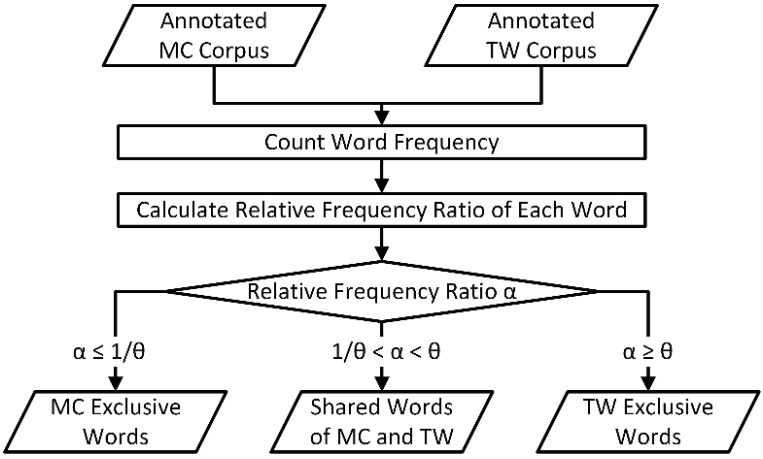

Fig. 3. Flowchart of Frequency Analysis

Let all words appear in the corpus be denoted by $W = \{w_1, w_2, \ldots, w_M\}$, where M is the size of W. The relative frequency ratios α of a word w_i between MC and TW is given by:

$$\alpha(w_i) = \frac{C_s(w_i)}{\sum_\omega C_s(\omega)} \Big/ \frac{C_t(w_i)}{\sum_\omega C_t(\omega)} \tag{1}$$

Here $C_s(w_i)$ is the number of times word w_i occurs in MC corpus and $C_t(w_i)$ is the number of times w_i occurs in TW corpus. For those words not occurring in the

corpus, we assign ε (smaller than 1) as their frequency to smooth data and avoid division by zero. In our experiment, we set $\varepsilon = 0.1$.

We set a threshold θ (larger than 1) and divide all words in W into three categories: (1) If $\alpha(w_i) \geq \theta$, word w_i is supposed to be an exclusive word of MC seldom used in TW. (2) If $\alpha(w_i) \leq 1/\theta$, word w_i is supposed to be an exclusive word of TW seldom used in MC. (3) If $1/\theta < \alpha(w_i) < \theta$, word w_i is supposed to be a **shared word** that is commonly used in both MC and TW.

3.3 Semantic Analysis

In semantic analysis, we extract synonyms between MC and TW by computing semantic similarity between words. The flowchart is shown in Fig. 4.

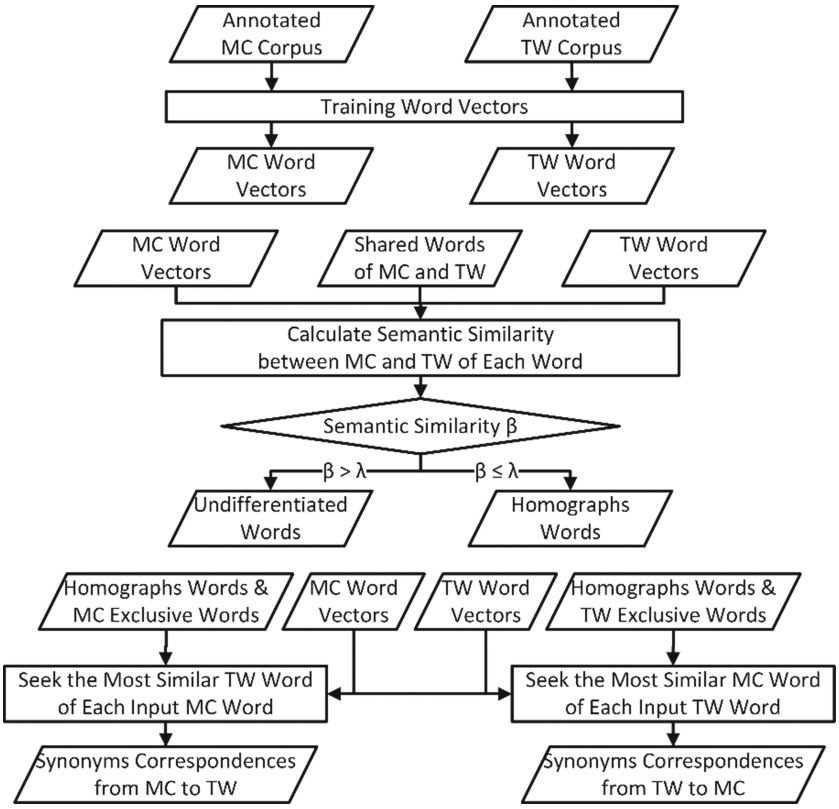

Fig. 4. Flowchart of Semantic Analysis

(i) We learn word vectors $R_s(w_i)$ and $R_t(w_i)$ as semantic representations of each word w_i from annotated MC and TW corpora respectively. Details of our learning algorithm will be described in Section 3.3.1.

(ii) Given a word w_i, which is a shared word, we calculate semantic similarity β between MC and TW as follows:

$$\beta(w_i) = Similarity(R_s(w_i), R_t(w_i)) \quad (2)$$

Details of similarity measure will be described in Session 3.3.2. We set a threshold λ and classify all shared words into two categories: (1) If $\beta(w_i) > \lambda$, word w_i is supposed to have same meaning between C and TC, namely an **undifferentiated word**. (2) If $\beta(w_i) \leq \lambda$, word w_i is supposed to have different meanings between MC and TW, namely a homograph word.

(iii) Given a word w_i, which is a homograph word or a MC exclusive word, we determine w_i's synonym $M_t(w_i)$ in TW by maximizing the semantic similarity as follows:

$$M_t(w_i) = argmax_\omega \, Similarity(R_s(w_i), R_t(\omega)) \quad (3)$$

If $w_i \neq M_t(w_i)$, we output w_i and $M_t(w_i)$ as a pair of synonyms from MC to TW.

(iv) Given a word w_i, which is a homograph word or a TW exclusive word, we determine w_i's synonym $M_s(w_i)$ in MC by maximizing the semantic similarity as follows:

$$M_s(w_i) = argmax_\omega \, Similarity(R_t(w_i), R_s(\omega)) \quad (4)$$

If $w_i \neq M_s(w_i)$, we output w_i and $M_s(w_i)$ as a pair of synonyms from TW to MC.

3.3.1 Lexical Semantic Representation

We represent meanings of words using vectors made of distributional statistics of words' co-occurrence. We set the size of context window as a constant N. Given a word w_i, we calculate the frequency of all words occurring in w_i's context window and obtain a probability distribution as the vector representation of w_i.

In this method, the bases of word vectors are words in vocabulary W. In word w_i's vector, the value on j-th dimension represents the probability that word w_j occurs in word w_i's context window. We assume that the number of homographs is few and all words in bases of vectors have consistent meanings between MC and TW[2]. Therefore, word vectors learned from MC corpus and TW corpus have the same meaning in every dimension and thus are comparable.

To capture the sequence information of words, we actually use the concatenation of two vectors to represent the meaning of one word: one represents the distribution in the preceding context window and the other one represents the distribution in the succeeding context window. Therefore, the whole word vector of w_i is given by:

$$R(w_i) = [R_l(w_i), R_r(w_i)] \quad (5)$$

[2] For example, when word 土豆 [tu dou] served as the basis of word vectors, we ignore the truth that it means *potatoes* in MC but *peanuts* in TW.

Here, $R_l(w_i)$ represents the distribution of word occurrence in the left window of w_i and $R_r(w_i)$ represents that in the right window of w_i. The length of $R_l(w_i)$ and $R_r(w_i)$ are all equal to M, the size of W.

3.3.2 Similarity of Lexical Semantic

Considering that word vectors are probability distributions, we use histogram intersection distance (HID) to measure the semantic similarity between words.

With the inspiration of tf-idf, we set different weights to different dimensions of word vector when computing HID. The weight of the j-th dimension is set to be the reciprocal of frequency that word w_j occurs in the whole corpora. Thus, when computing similarity, dimensions of infrequent words play a more important part than those of frequent words. The enhanced HID can be described by the following formula:

$$H(R_1, R_2) = \frac{\sum_k \frac{\min(R_1[k], R_2[k])}{C(w_k)}}{\sum_k \frac{\max(R_1[k], R_2[k])}{C(w_k)}} \tag{6}$$

Here, $H(R_1, R_2)$ is the similarity of word vectors R_1 and R_2. $R[k]$ is the value on the k-th dimension of word vector R. $C(w_k)$ is the frequency that word w_k occurs in the whole corpora.

Since representation of a word consists of two probability distribution, when computing the similarity of two words R_s and R_t, as described in (7), we firstly compute the similarity between R_{sl} and R_{tl} and similarity between R_{sr} and R_{tr} separately, and then set the average of two similarities as the similarity between R_s and R_t.

$$Similarity(R_s, R_t) = \frac{H(R_{sl}, R_{tl}) + H(R_{sr}, R_{tr})}{2} \tag{7}$$

3.4 Post-processing

After synonyms detection, we verify the result manually[3] to improve the accuracy and convert TW words into TC.

4 Experiment

4.1 Experiment Setup

To evaluate our approach, we use a MC corpus from Xinhuanet with 189 million words[4] and a TW corpus from MSN with 183 million words[5] to detect synonyms

[3] According to the result of manual verification, we can also update the value of thresholds in our detection algorithm to achieve better performance.

[4] The Xinhuanet corpus is available in Superfection Simplified Chinese Corpus (http://cloudtranslation.cc/corpus_sc.html).

[5] The Taiwan MSN corpus is available in Superfection Traditional Chinese Corpus (http://cloudtranslation.cc/corpus_tc.html).

from MC to TW. Contents of both two corpora are news published on Internet from 2011 to 2014. To narrow the searching space, we only focused on words with 2 or 3 Chinese characters. We also filtered out MC words with frequency lower than 2000 and TW words with frequency lower than 10. We set our program to output 5 best synonym candidates for each MC word.

4.2 Experiment Result

Our program output 421 synonyms pairs. We verified all the results manually and the precision is shown in Table 1. 148 pairs are correct synonyms and 76 results list the correct synonyms in the 2nd to 5th candidates. Some correct synonyms are shown in Table 2.

Table 1. Precision of the Experiment Result

	p@1*	p@3*	p@5*
Precision	35.15%	49.64%	53.21%

* p@N means precision considering N-best results.

Table 2. Synonyms Detected in the Experiment (Partial)*

Word in mainland China		Word in Taiwan		Meaning
美联储	[mei lian chu]	**聯準會**	[lian zhun hui]	U.S. Federal Reserve
短信	[duan xin]	簡訊	[jian xun]	SMS; short messages
硬件	[ying jian]	硬體	[ying ti]	Hardware
网点	[wang dian]	**據點**	[ju dian]	Branches; outlets
网络	[wang luo]	網路	[wang lu]	Network
群体	[qun ti]	**族群**	[zu quan]	Social groups
概率	[gai lv]	機率	[ji lv]	Probability
民警	[min jing]	**員警**	[yuan jing]	Policemen
入市	[ru shi]	**進場**	[jin chang]	Entering the stock market
芯片	[xin pian]	晶片	[jing pian]	Silicon chips
出租车	[chu zu che]	計程車	[ji cheng che]	Taxies
欺诈	[qi zha]	**詐欺**	[zha qi]	Fraud

* Pairs in bold are those not annotated as synonyms in [3].

4.3 Analysis

From Table 1 we see that our method reports a satisfactory precision. Moreover, Table 2 shows that a number of synonyms, that are hard to be found in handcrafted compilation, can be found in our method.

However, according to the experiment result, our method still has some limitations.

(i) Ambiguous words cannot be handled in our method. For example, our program output TW word 類股 [lei gu] (*sectors of stocks*) as the synonym of MC word 板块 [ban kuai] (*plates or sectors*). This correspondence is correct in text about stock market but incorrect in geology context.

(ii) Synonym phrases cannot be detected by our method. For example, TW word 殘障人士 [can zhang ren shi] (*disabled people*) is corresponding to MC word 残疾人 [can ji ren] (*disabled people*). But this correspondence cannot be found by our method because 殘障人士 [can zhang ren shi] is segmented as two words 殘障 [can zhang] and 人士 [ren shi].

5 Conclusion and Future Work

We have presented an automatic approach to detect synonyms between MC and TW, and showed that it can extract synonyms missing by traditional handcrafted compilation. Thus, our method can be also applied to computer aided dictionary compilation.

Our future work includes:

- An attempt to integrate continuous vector representations of words using deep neural network;
- An investigation into an appropriate method of synonym phrases detection;
- An attempt to apply our method to machine translation and try to extract translation rules from comparable corpora.

Acknowledgments. The work described in this paper is supported by the Special Fund Project of Ministry of Education of China (Conversion System from Simplified to Traditional Chinese Characters), National High-Tech R&D Program of China (No. 2012BAH14F03), the National Natural Science Foundation of China (Nos. 61005052 and 61303082) and the Research Fund for the Doctoral Program of Higher Education of China (No. 20130121110040).

References

1. Su, J.: Research on Homographs across the Straits. Studies of the Chinese Language **1995**(2), 107–117 (1995). (苏金智: 海峡两岸同形异义词研究. 中国语文. 1995(2), 107–117 (1995)). (in Chinese)
2. Li, X., Qiu, Z.: Determination and Treatment of Diverse Words in the Cross-Straits Dictionary. Applied Linguistics **2012**(4), 74–81 (2012). (in Chinese)
3. The Common Words Dictionary of the Cross-Straits. http://www.zhonghuayuwen.org/PageInfo.aspx?Id=375. (in Chinese)
4. Richardson, R., Smeaton, A., Murphy, J.: Using WordNet as a knowledge base for measuring semantic similarity between words. In: Proceedings of AICS Conference (1994)
5. Liu, Q., Li, S.: Word Similarity Computing Based on How-net. Computational Linguistics and Chinese Language Processing **7**(2), 59–76 (2002). (in Chinese)
6. Chen, Y., Shi, X., Zhou, C.: A simplified-traditional chinese character conversion model based on log-linear models. In: Proceedings of International Conference on Asian Language Processing (2011)
7. Wang, S., Cao, C., Pei, Y., Xia, F.: A Collocation-based Method for Semantic Similarity Measure for Chinese Words. Journal of Chinese Information Processing. **27**(1), 7–14 (2013). (in Chinese)

8. Shi, J., Wu, Y., Qiu, L., Lv, X.: Chinese Lexical Semantic Similarity Computing Based on Large-scale Corpus. Journal of Chinese Information Processing **27**(1), 1–6+80 (2013). (in Chinese)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR (2013)
10. Swain, M.J., Ballard, D.H.: Color Indexing. IJCV **7**(1), 11–32 (1991)
11. Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. Information Processing & Management **24**(5), 513–523 (1988)