A Baseline System for Chinese Near-Synonym Choice

Liang-Chih Yu[1], Wei-Nan Chien[1,2] and Shih-Ting Chen[1]
[1]Department of Information Management, Yuan Ze University, Taiwan, R.O.C.
[2]Information Technology Center, National Taiwan Normal University, Taiwan, R.O.C.
Contact: lcyu@saturn.yzu.edu.tw

Abstract

Near-synonym sets represent groups of words with similar meaning, which are useful knowledge resources for many natural language applications such as query expansion for information retrieval (IR) and computer-assisted language learning. However, near-synonyms are not necessarily interchangeable in contexts due to their specific usage and syntactic constraints. Previous studies have developed various methods for near-synonym choice in English sentences. To our best knowledge, there is no such evaluation on Chinese sentences. Therefore, this paper implements two baseline systems: *pointwise mutual information (PMI)* and a *5-gram language model* that are widely used in previous work for Chinese near-synonym choice evaluation. Experimental results show that the 5-gram language model achieves higher accuracy than PMI.

1 Introduction

Lexical semantics plays an important role in many natural language applications. For instance, knowing that the word "arm" has (at least) the senses *weapon* and *bodypart* enables systems to perform word sense disambiguation. Knowing in addition the near-synonyms of a word can further improve the applications' effectiveness. For instance, knowing that the weapon sense of "arm" corresponds to the weapon sense of "weapon" and of "arsenal" means that systems can in addition perform term expansion for information retrieval (Moldovan and Mihalcea, 2000; Bhogal et al., 2007), (near-)duplicate detection for summarization, alternative word selection for writing support systems (Inkpen and Hirst, 2006; Inkpen, 2007; Wu et al., 2010), as well as computer-assisted language learning (Cheng et al., 2004; Ouyang et al., 2009).

Recent studies have shown that near-synonyms are not necessarily interchangeable in practical use due to their specific usage and collocational constraints, as shown in the following examples.

(1) {strong, powerful} coffee (Pearce, 2001)

(2) ghastly {error, mistake} (Inkpen, 2007)

(3) {bridge, overpass, tunnel} under the bay (Yu et al., 2010)

Example (1) and (2) present an example of collocational constraints in given contexts. In (1), the word "strong" in the near-synonym set {strong, powerful} is more suitable than "powerful" to fit the given context "coffee", since "powerful coffee" is an anti-collocation. Similarly, in (2), "mistake" is more suitable than "error" because "ghastly mistake" is a collocation and "ghastly error" is an anti-collocation. In (3), the near-synonym set {bridge, overpass, tunnel} represents the meaning of a physical structure that connects separate places by traversing an obstacle. Suppose that the original word in the given context "under the bay" is "tunnel". It can be found that the word "tunnel" cannot be substituted by the other words in the same set because all the substitutions are semantically implausible. The above examples indicate that not all words in a near-synonym set can be substituted with each other even though they share the same or similar meaning. Actually, some near-synonyms may produce inadequate substitutions, which may reduce the applications' effectiveness.

In order to verify whether near-synonyms do match the given contexts, previous studies have formulated the problem of near-synonym choice

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1366–1370,
Chiang Mai, Thailand, November 8 – 13, 2011.

English Sentence: This will make the _____ message easier to interpret.
Original word: error
Near-synonym set: {error, mistake, oversight}
Chinese Sentence: 這 將 使 這 _____ 訊息 容易 解釋
Original word: 錯誤
Near-synonym set: {錯誤, 錯, 差錯, 失察, 過失}

Figure 1. Example of the near-synonym choice evaluation for English and Chinese sentences.

as the "fill-in-the-blank" (FITB) task, and evaluated on English sentences (Edmonds, 1997; Inkpen, 2007; Gardiner and Dras, 2007, Islam and Inkpen, 2010). Figure 1 illustrates an example of FITB task on English and Chinese sentences. Given a near-synonym set and a sentence with one of the near-synonyms in it, the near-synonym is removed from the sentence to form a lexical gap. The goal is to predict an answer (best near-synonym) that can fill the gap from the given near-synonym set (including the original word). An evaluation can then be performed to examine whether the involved systems can restore the original word by filling the gap with the best near-synonym. To our best knowledge, there is no such evaluation for Chinese sentences. Therefore, this paper follows the FITB procedure to build a baseline system for Chinese near-synonym choice evaluation. Applications can benefit from such evaluation to provide more effective services. For instance, a writing support system can assist users, especially Chinese as Second Language (CSL) learners, to select a best alternative near-synonym when they have a need to avoid repeating the same word in composing a text.

In the following sections, we first present some previous work on near-synonym choice. Section 3 describes the two baseline systems: pointwise mutual information (PMI) and a 5-gram language model for Chinese near-synonym choice evaluation. Section 4 first introduces the Chinese near-synonym sets and test sets used in experiments, and then shows the evaluation results of the two baseline systems. Conclusions are finally drawn in Section 5.

2 Related Work

In the field of lexical semantics, the contextual information is useful for representing the meaning of words, phrases, as well as sentences (Mitchell and Lapata, 2008; Erk and Pado, 2008; Thater et al., 2010; Ó Séaghdha and Korhonen, 2011; Grefenstette et al., 2011). Therefore, the co-occurrences between a target word (the gap) and its context words have been commonly used in statistical approaches to measuring the substitutability of words. Edmonds (1997) built a lexical co-occurrence network from 1989 Wall Street Journal to determine the near-synonym that is most typical or expected in a given context. Inkpen (2007) used the PMI formula to select the best near-synonym that can fill the gap in a given context. The PMI scores for each candidate near-synonym are computed using a larger web corpus, the Waterloo terabyte corpus, which can alleviate the data sparseness problem encountered in Edmonds' approach. Following Inkpen's approach, Gardiner and Dras (2007) also used the PMI formula with a different corpus (the Web 1T 5-gram corpus) to explore whether near-synonyms differ in attitude.

Islam and Inkpen (2010) also used the Web 1T 5-gram corpus to build a 5-gram language model for near-synonym choice. Yu *et al*. (2010) presented a method to compute the substitution scores for each near-synonym based on n-gram frequencies obtained by querying Google. The dataset used in their experiments are derived from the OntoNotes copus (Hovy et al., 2006; Pradhan et al., 2007; Yu et al., 2008), where each near-synonym set corresponds to a *sense pool* in OntoNotes.

Besides the PMI and n-gram-based methods, another direction is to identify the senses of a target word and its near-synonyms using word sense disambiguation (WSD), comparing whether they were of the same sense (McCarthy, 2002; Dagan et al., 2006). Dagan *et al.* (2006) described that the use of WSD is an indirect approach since it requires the intermediate sense identification step, and thus presented a sense matching technique to address the task directly.

3 Baseline Systems

The baseline systems used for Chinese near-synonym choice are the PMI-based method (Inkpen, 2007; Gardiner and Dras, 2007) and the 5-gram language model (Islam and Inkpen, 2010). We choose these two methods because they are commonly used in previous work.

3.1 PMI-based method

The mutual information can measure the co-occurrence strength between a near-synonym and the words in a given context. A higher mutual information score indicates that the near-synonym fits well in the given context, thus is more likely to be the correct answer. The pointwise mutual information (Church and Hanks, 1991) between two words *x* and *y* is defined as

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}, \quad (1)$$

where $P(x,y) = C(x,y)/N$ denotes the probability that *x* and *y* co-occur; $C(x,y)$ is the number of times *x* and *y* co-occur in the corpus, and *N* is the total number of words in the corpus. Similarly, $P(x) = C(x)/N$, where $C(x)$ is the number of times *x* occurs in the corpus, and $P(y) = C(y)/N$, where $C(y)$ is the number of times *y* occurs in the corpus. Therefore, (1) can be re-written as

$$PMI(x,y) = \log_2 \frac{C(x,y) \cdot N}{C(x) \cdot C(y)}. \quad (2)$$

Inkpen (2007) computed the PMI scores for each near-synonym using the Waterloo terabyte corpus and a context window of size $2k$ (k=2). Given a sentence *s* with a gap, $s = ...w_1...w_k \_\_\_\_ w_{k+1}...w_{2k}...$, the PMI score for a near-synonym NS_i to fill the gap is defined as

$$PMI(NS_j, s) = \sum_{i=1}^{2k} PMI(NS_j, w_i). \quad (3)$$

The near-synonym with the highest score is considered as the answer. In this paper, we use the Chinese Web 5-gram corpus to compute PMI scores. The frequency counts $C(\cdot)$ are retrieved from this corpus in the same manner within the 5-gram boundary.

3.2 5-gram language model

The n-grams can capture contiguous word associations in given contexts. Given a sentence $s = ...w_{i-4}w_{i-3}w_{i-2}w_{i-1}w_i w_{i+1}w_{i-2}w_{i-3}w_{i+4}...$, where w_i represents a near-synonym in a set. In computing the 5-gram scores for each near-synonym, Islam and Inkpen (2010) considers only the five product items $P(w_i | w_{i-4}^{i-1})$, $P(w_{i+1} | w_{i-3}^{i})$, $P(w_{i+2} | w_{i-2}^{i+1})$, $P(w_{i+3} | w_{i-1}^{i+2})$, and $P(w_{i+4} | w_{i}^{i+3})$. The other items are excluded because they do not contain the near-synonym and thus will have the same values. Accordingly, the 5-gram language model (n=5) with a smoothing method can be defined as

$$\begin{aligned} P(s) &= \prod_{i=0}^{5} P(w_i | w_{i-n+1}^{i-1}) \\ &= \prod_{i=0}^{5} \frac{C(w_{i-n+1}^{i}) + (1+\alpha_n) M(w_{i-n+1}^{i-1}) P(w_i | w_{i-n+2}^{i-1})}{C(w_{i-n+1}^{i-1}) + \alpha_n M(w_{i-n+1}^{i-1})} \end{aligned} \quad (4)$$

where $M(w_{i-n+1}^{i-1})$ denotes a missing count used in the smooth method, defined as

$$M(w_{i-n+1}^{i-1}) = C(w_{i-n+1}^{i-1}) - \sum_{w_i} C(w_{i-n+1}^{i}) \quad (5)$$

where $C(\cdot)$ denotes an n-gram frequency, which can be retrieved from the Chinese Web 5-gram corpus. Additionally, the 5-gram language model is implemented as a back-off model. That is, if the frequency of a higher-order n-gram is zero, then its lower-order n-grams will be considered. Conversely, if the frequency of a higher-order n-gram is not zero, then the lower-order n-grams will not be included in computation.

4 Experimental Results

4.1 Experiment setup

1) Chinese near-synonym sets: Since there is no standard dataset for Chinese near-synonym sets, we created seven Chinese near-synonym sets based on the seven English near-synonym sets used as the standard dataset in the previous studies (Edmonds, 1997; Inkpen, 2007; Gardiner and Dras, 2007, Islam and Inkpen, 2010). For each English near-synonym set, we first identified its corresponding senses (entries) in the Chinese WordNet (CWN) (Huang et al., 2008). Each corresponding Chinese near-synonym set

	Near-Synonym sets	Sinica Corpus		News Corpus		All	
		PMI	5GRAM	PMI	5GRAM	PMI	5GRAM
1	難懂的, 困難的, 艱難的, 艱苦的 (difficult, hard, tough)	67.14%	**70.32%**	71.51%	71.51%	68.83%	**70.78%**
2	錯誤, 錯, 差錯, 失察, 過失 (error, mistake, oversight)	**67.25%**	52.64%	**60.13%**	54.22%	**63.58%**	53.46%
3	任務, 工作, 義務 (job, task, duty)	65.08%	**76.43%**	67.62%	**70.91%**	66.54%	**73.26%**
4	責任, 職責, 職務, 約定 (responsibility, burden, obligation, commitment)	50.03%	**68.05%**	**56.79%**	56.67%	54.41%	**60.67%**
5	質料, 物質, 材料 (material, stuff, substance)	**72.98%**	59.84%	**75.90%**	65.32%	**74.09%**	61.92%
6	給, 給予, 給與, 供應, 供給 (give, provide, offer)	**69.39%**	65.89%	51.88%	**58.11%**	60.59%	**61.98%**
7	決定, 定奪, 終結, 確定 (settle, resolve)	61.43%	**71.97%**	60.59%	**72.35%**	60.85%	**72.23%**
	Average	65.20%	**70.05%***	61.38%	**66.62%***	62.99%	**68.07%***
	Number of test cases	26,504		36,427		62,931	

***** Statistically significant ($p < 0.05$) using *Binomial Exact Text*

Table 1. Accuracy of PMI and 5GRAM on different test sets.

can then be created by selecting the Chinese translations of the identified entries in the Chinese WordNet. Table 1 shows the seven Chinese near-synonym sets.

2) Test set: The test sentences containing the near-synonyms were selected from two corpora: the Sinica Corpus and Chinese News Corpus, released by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). If a test sentence contained two (or more) near-synonyms, then this sentence was divided into two (or more) test examples. The near-synonyms were then removed from the test examples for FITB evaluation.

3) Implementation of baseline systems: The two baseline system, PMI and 5GRAM, were implemented using the (3) and (4), respectively. For PMI, the size of the context window k was set to 2. For 5GRAM, only the 5-gram with a near-synonym in the middle position of each test example was selected the s for testing.

4) Evaluation metric: The answers proposed by PMI and 5GRAM are the near-synonyms with the highest score. The correct answers are the near-synonyms originally in the gap of the test examples. The performance is measure by the accuracy, which is defined as the number of correct answers made by each baseline system, divided by the total number of test examples.

4.2 Evaluation results

Table 1 shows the evaluation results of Chinese near-synonym choice using PMI and 5GRAM. The results show that 5GRAM achieved better performance than PMI on both test corpora. In comparison with the results on the seven English near-synonym sets, previous studies reported that the accuracy of PMI and the 5-gram language model were 66.0 (Inkpen, 2007) and 69.9 (Islam and Inkpen, 2010), respectively, which was greater than 62.99 and 68.07 reported in Table 1. One possible reason is that the total number of near-synonyms in the seven Chinese near-synonyms sets is 28 and the average is 4 for each set, and that in the seven English near-synonyms sets is 21 and the average is 3 for each set. More near-synonyms in a set may decrease systems' ability to discriminate among near-synonyms.

5 Conclusion

This work has presented the use of the PMI and 5-gram language model for Chinese near-synonym choice. Additionally, this work has also created seven Chinese near-synonym sets based on the standard dataset of the seven English near-synonym sets. Experimental results show that the 5-gram language model that can capture contiguous word associations in given contexts achieved higher accuracy than PMI.

Acknowledgement

This work was supported by National Science Council, Taiwan, R.O.C (NSC99-2221-E-155-036-MY3 and NSC100-2632-S-155-001), and Aim for the Top University Plan, Ministry of Education, Taiwan, R.O.C. The authors would like to thank the anonymous reviewers and the area chairs for their constructive comments.

References

J. Bhogal, A. Macfarlane, and P. Smith. 2007. A Review of Ontology based Query Expansion. *Information Processing & Management*, 43(4):866-886.

C. C. Cheng. 2004. Word-Focused Extensive Reading with Guidance. In *Proc. of the 13th International Symposium on English Teaching*, pages 24-32.

K. Church and P. Hanks. 1991. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16(1):22-29.

I. Dagan, O. Glickman, A. Gliozzo, E. Marmorshtein, and C. Strapparava. 2006. Direct Word Sense Matching for Lexical Substitution. In *Proc. of COLING/ACL-06*, pages 449-456.

P. Edmonds. 1997. Choosing the Word Most Typical in Context Using a Lexical Co-occurrence Network. In *Proc. of ACL-97*, pages 507-509.

K. Erk and S. Pad´o. 2008. A Structured Vector Space Model for Word Meaning in Context. In *Proc. of EMNLP-08*, pages 897-906.

M. Gardiner and M. Dras. 2007. Exploring Approaches to Discriminating among Near-Synonyms, In *Proc. of the Australasian Technology Workshop*, pages 31-39.

E. Grefenstette, M. Sadrzadeh, S. Clark, B. Coecke, and S. Pulman. 2011. Concrete Sentence Spaces for Compositional Distributional Models of Meaning. In *Proc. of IWCS-11*, pages 125-134.

E. H. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. In *Proc. of HLT/NAACL-06*, pages 57–60.

C. R. Huang, S. K. Hsieh, J. F. Hong, Y. Z. Chen, I. L. Su, Y. X. Chen and S. W. Huang. 2008. Chinese Wordnet: Design, Implementation, and Application of an Infrastructure for Cross-lingual Knowledge Processing. In *Proc. of the 9th Chinese Lexical Semantics Workshop*.

D. Inkpen. 2007. A Statistical Model of Near-Synonym Choice. *ACM Trans. Speech and Language Processing*, 4(1):1-17.

D. Inkpen and G. Hirst. 2006. Building and Using a Lexical Knowledge-base of Near-Synonym Differences. *Computational Linguistics*, 32(2):1-39.

A. Islam and D. Inkpen. 2010. Near-Synonym Choice using a 5-gram Language Model. *Research in Computing Science: Special issue on Natural Language Processing and its Applications*, Alexander Gelbukh (ed.), 46: 41-52.

D. McCarthy. 2002. Lexical Substitution as a Task for WSD Evaluation. In *Proc. of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation at ACL-02*, pages 109-115.

J. Mitchell and M. Lapata. 2008. Vector-based Models of Semantic Composition. In *Proc. of ACL-08*, pages 234-244.

D. Moldovan and R. Mihalcea. 2000. Using Wordnet and Lexical Operators to Improve Internet Searches. *IEEE Internet Computing*, 4(1):34-43.

D. Ó Séaghdha and A. Korhonen. 2011. Probabilistic Models of Similarity in Syntactic Context. In *Proc. of EMNLP-11*, pages 1047-1057.

S. Ouyang, H. H. Gao, and S. N. Koh. 2009. Developing a Computer-Facilitated Tool for Acquiring Near-Synonyms in Chinese and English. In *Proc. of IWCS-09*, pages 316-319.

D. Pearce. 2001. Synonymy in Collocation Extraction. In *Proc. of the Workshop on WordNet and Other Lexical Resources at NAACL-01*.

S. Pradhan, E. H. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. In *Proc. of ICSC-07*, pages 517-524.

S. Thater, H. Fürstenau, and M. Pinkal. 2010. Contextualizing Semantic Representations Using Syntactically Enriched Vector Models. In *Proc. of ACL-10*, pages 948-957.

C. H. Wu, C. H. Liu, H. Matthew, and L. C. Yu. 2010. Sentence Correction Incorporating Relative Position and Parse Template Language Models. *IEEE Trans. Audio, Speech and Language Processing*, 18(6):1170-1181.

L. C. Yu, C. H. Wu, R. Y. Chang, C. H. Liu, and E. H. Hovy. 2010. Annotation and Verification of Sense Pools in OntoNotes. *Information Processing & Management*, 46(4):436-447.

L. C. Yu, C. H. Wu, E. H. Hovy. 2008. OntoNotes: Corpus Cleanup of Mistaken Agreement Using Word Sense Disambiguation. In *Proc. of COLING-08*, pages 1057-1064.