Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 347257, 11 pages
doi:10.1155/2012/347257

Research Article

Distributional Similarity for Chinese: Exploiting Characters and Radicals

Peng Jin,[1] John Carroll,[2] Yunfang Wu,[3] and Diana McCarthy[4]

[1] *School of Computer Science, Leshan Normal University, 614004 Leshan, China*
[2] *Department of Informatics, Sussex University, Brighton BN1 9QJ, UK*
[3] *Institute of Computational Linguistics, Peking University, 100871 Beijing, China*
[4] *Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge CB3 9DB, UK*

Correspondence should be addressed to Peng Jin, jandp@pku.edu.cn

Received 10 April 2012; Accepted 1 June 2012

Academic Editor: Yuping Wang

Distributional Similarity has attracted considerable attention in the field of natural language processing as an automatic means of countering the ubiquitous problem of sparse data. As a logographic language, Chinese words consist of characters and each of them is composed of one or more radicals. The meanings of characters are usually highly related to the words which contain them. Likewise, radicals often make a predictable contribution to the meaning of a character: characters that have the same components tend to have similar or related meanings. In this paper, we utilize these properties of the Chinese language to improve Chinese word similarity computation. Given a content word, we first extract similar words based on a large corpus and a similarity score for ranking. This rank is then adjusted according to the characters and components shared between the similar word and the target word. Experiments on two gold standard datasets show that the adjusted rank is superior and closer to human judgments than the original rank. In addition to quantitative evaluation, we examine the reasons behind errors drawing on linguistic phenomena for our explanations.

1. Introduction

The computation of word similarity can make an important contribution to many language processing tasks and applications. Tasks include, for example, automatic creation of thesauruses [1], smoothing of statistical language models [2], and lexical expansion [3]. A recent more direct application has been for suggesting bid-terms for a commercial search engine [4, 5]. There are two main ways to compute word similarity: corpus-based and WordNet-based [6]. The first way avoids the requirement of a manually constructed lexical resource. Its disadvantage is that it is not always accurate, and some similar words are not found, particularly those with low frequency. The second method usually has better performance than the first one, but it requires a hand-crafted thesaurus and there will always be omissions in such a resource. Research is increasingly focusing on the former [7] or combining these two approaches [8, 9].

There are few previous studies on Chinese word similarity. Liu and Li [10] report work based on a manually created ontology named HowNet [11]. Rather than listing a target word's neighbors in order, they provide an interface to return the similarity score of any two terms. It is difficult to compare our ranking with their results because the program cannot be run in batch (noninteractive) mode.

The Chinese language has the feature that almost every individual Chinese character has some specific meaning. The meaning of a word is often highly related to the characters it is comprised of. For example, 汽车 (*automobile*) and 火车 (*train*) are kinds of 车 (*vehicle*). Moreover, each Chinese character is composed of one or more components. Usually, one component (the "radical") carries the meaning of a character to a certain degree and the other indicates the pronunciation. Characters sharing the same radicals have related or similar meanings. For example, the characters 吃 (*eat*) and 喝 (*drink*) have the same leftmost component 口 (*mouth*), the shared aspect of meaning being that both actions are performed by the mouth. It is easy to decompose a word into characters and to decompose a character into components according to their forms. Moreover, most characters have a single meaningful component.

Many researchers have noted that Chinese characters carry information about word meaning. This information has been used for predicting the semantic class of unknown words. Lu [12] proposed three knowledge-based models, out of which combining a character-category association model and a rule-based model performed best. A corpus-based model did not improve performance. Tseng [13] used a morphological analyzer to tag the syntactic categories of characters in a target word. Hsieh [14] predicted the sense of unknown two-character words with the help of a character ontology that was created manually. In contrast to these studies, our method does not require any special-purpose manually created resources. Veale and Chen [15] strived for valid decompositions of Chinese lexemes and divided them into three categories: metaphors, analogies, and blends. This information could be integrated into our method in further to improve the system's performance. As far as radicals were concerned, Liu and Lin [16] used structural information to identify similar Chinese characters. Wang et al. [17] integrated radical features into an SVM-based tool for part of speech tagging. Their experiment results showed that the precision was improved from 84.32% to 85.27% for out-of-vocabulary words.

In this paper, we improve the computation of word similarity in Chinese by taking into account both the characters in words and also the characters' components.

This paper is structured as follows. Section 2 describes a novel approach to rescoring the results from Lin's [1] corpus-based measure of word similarity taking advantage of structural features of the Chinese language. Section 3 presents an experimental evaluation. Section 4 discusses the results. Section 5 concludes and proposes directions for future work.

2. Methods

2.1. Word Similarity Computation

Lin [1] defined the similarity between two objects as "the amount of information contained in the commonality between the objects divided by the amount of information in the descriptions of the objects." For a pair of words, w_1 and w_2, this similarity measure can be computed by the following equation:

$$\mathrm{sim}(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w)\in T(w_1)} I(w_1, r, w) + \sum_{(r,w)\in T(w_2)} I(w_2, r, w)}, \tag{2.1}$$

where $I(w, r, w')$ is the mutual information between w and w', which denotes the amount information in $\|w_1, r, w_2\| = c$. It is computed as follows:

$$I(w, r, w') = \log \frac{\|w_1, r, w_2\| \times \|*, r, *\|}{\|w_1, r, *\| \times \|*, r, w_2\|}, \tag{2.2}$$

where r is the relationship between w and w', and $T(w)$ is the set of pairs (r, w') such that $I(w, r, w')$ is positive.

In the experiments, Lin used a broad-coverage parser to extract triples from 64-million word corpus consisting of the Wall Street Journal (24 million words), San Jose Mercury (21 million words), and AP Newswire (19 million words).

For any given word w, (2.1) can be used to obtain any other word's similarity score with w and then ranking them in a descending order.

Lin's experimental results demonstrated that this method is better than other algorithms such as cosine and dice. It is widely used in the field of computational linguistics such as for word sense disambiguation [18] and semantic parsing [19].

2.2. Hierarchical Structure of a Chinese Word

As a logographic language, a Chinese word usually has some semantic relationship to all the characters it is comprised of while the character has some semantic relationship to one of the radicals that it is comprised of. In Figure 1, 艹 is the semantic radical of 草 (*grass*), a set of characters such as 蕨 (*fern*), 藤 (*rattan*), 葚 (mulberry), 花 (*flower*), and 萍 (duckweed), share the radical 艹 and are all related to herbaceous plant. Similar to the 草 (*grass*) example, 地 (*land*)'s semantic radical is 土 (*soil*). The radical 土 (*soil*) indicates the characters such as 坝 (*dam*), 坯 (*earthen brick*), 垡 (*turn up soil*), 墙 (*wall*), and 堡 (*small fort*) are related to soil.

It should be noticed that all subwords are easily obtained by straightforward morphological analysis. The semantic radicals are usually used as the index for words in a Chinese dictionary, and the Chinese government has published a national standard [20] to associate each character to a specific radical.

Most Chinese characters are composed of two radicals: one is the semantic radical, and the other is phonetic radical. Usually only the semantic radicals provide the important cue to measure the similarity for characters. Following Jin et al. [21], and in contrast to Yencken and Baldwin [22], in this paper we only use the semantic radical.

2.3. Integrating Character Similarity into Word Similarity

For each word, we find its top k neighbors with respect to Lin's similarity measure. The rank induced by (2.1) is the baseline. We then adjust the similarity score according to characters

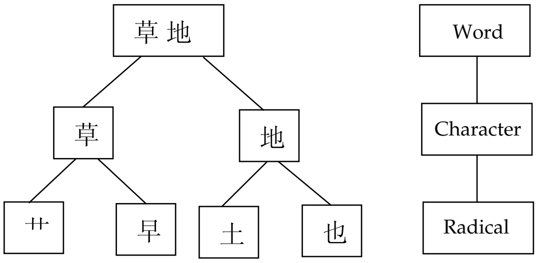


Figure 1: An example of the hierarchical structure of a Chinese word.

and radicals shared between the two words, based on a decomposition as illustrated in Figure 1, thus obtaining a reranking of the neighbors.

In broad terms, our approach works as follows. Firstly, if a neighbor of the target word contains a character that is also in the target word, the neighbor's score is increased. Secondly, for nonshared characters, if a radical of that character in a neighbor is the same as a radical of a character in the target, the neighbor's score is increased. Intuitively, there should be more weight on shared characters than shared radicals because characters are larger units and so in general carry a more substantial and specific meaning.

There are a number of ways in which the shared character and radical information can be used to adjust the distributional similarity score. The first approach we considered was to add constants that ensured shared characters were considered more important than shared radicals, which in turn were considered more strongly than the distributional similarity score. To do this we added a constant of 10 for each shared character in the target word and putative neighbor. A constant of 1 was applied for each shared radical because the original distributional similarity score is always below 1 (this constant was selected because no word is longer than 10 characters so it could never dominate the character-based weighting). This approach did not perform as well as an alternative approach which we present below, so we do not consider it further.

In the alternative approach, we multiplicatively adjust the basic distributional similarity score according to the number of shared characters or radicals between the target word and neighbor whose similarity score we are adjusting. As (2.3) and (2.4) indicate, the original similarity score is multiplied by a parameter, to the power of the number of common characters:

$$\mathrm{sim}(w_1, w_2)_{CH} = \mathrm{sim}_{LIN} \times \alpha^{(\#\text{characters})}. \tag{2.3}$$

Analogous to the above equation, we consider a variant in which the character and radical counts are combined (*CH_RA*) as follows:

$$\mathrm{sim}(w_1, w_2)_{CH\_RA} = \mathrm{sim}_{LIN} \times \eta^{(\#\text{character})} \times \lambda^{(\#\text{radical})}. \tag{2.4}$$

We use held-out data from the CiLin evaluation described below to find optimal values for the parameters. In (2.3), α is set to 2.5. To (2.4), we first optimize η and next optimize λ, resulting in η equals to 2.5 and λ to 1.2.

3. Results

We develop a system to obtain the similar words from a large scale corpus based on the technique of Lin [1]. Then, given a target word, its neighbor words (the 50 most similar) are reranked by the method proposed in this paper. These ranks are evaluated on two gold standard datasets which are created by human.

3.1. Corpus

The corpus we use in our experiments comes from the Chinese Gigaword Second Edition (catalog number LDC2005T14). In this, we only use the Xinhua newswire material because it is Mandarin Chinese News Text, while the rest is not. The size of Xinhua newswire is 471,110 K Chinese characters, and there are in total 992,261 documents. This covers all news published by Xinhua News Agency (the largest news agency in China) from 1991 to 2004. In order to compute word similarity, we first segmented and part-of-speech tagged the corpus using the free downloadable system "ICTCLAS" (http://www.ictclas.org/), which is based on a Hierarchical Hidden Markov Model.

By analogy to Lin's [1] work on English word similarity which starts with dependency parsing, we use the Chinese version of the Stanford Parser to extract dependency triples from the Xinhua newswire corpus. The Stanford Parser also assigns part-of-speech labels using the PoS inventory of Xia [23], in which NN denotes noun, VV is verb, and VA is adjective. In total, there are nearly 850,000 word types in the corpus. Of these, 360,000 word types are tagged NN, VV, or VA, and 67,535 word types occur more than 100 times. Weeds et al. [24] proposed that "distributional similarity techniques have been shown to work well for words which occur more than 100 times in a given corpus." In fact, the threshold of 100 is adapted by many researchers [1, 25–27]. So, we also set 100 in our experiments. Finally, there are 45,667 word types in our experiment. Among them, there are 30,778 nouns; 13,164 verbs; 1,725 adjectives.

The Stanford Parser identifies 45 named grammatical relations; a 46th relation *dep* (dependent) is the default when the parser fails to identify the relation type. As in the analysis of Chang et al. [4, 5], we consider that the relations *dep*, *punct*, and *nummod* have nothing to do with word meaning (only being applicable for machine translation and similar applications), so they are removed from the parser results. In particular, Chinese sentences tend to contain more punctuation (*punct*) than English, which prefers to use conjunctions to link clauses.

There are in total 12,403,187 sentences after segmentation; of these, 1.16% sentences contain only one or two words and are removed before parsing. Since very long sentences can cause the parser to run out of memory or run slowly, we set the parameter "maxLength" to 300 bytes, so a further 6.3% sentences are removed before parsing. Although we chose the "factored parser" (xinhuaFactored.ser.gz), about 4.6% of sentences still failed to be parsed. In the end, 10,899,022 sentences were parsed, accounting for 87.9% of all 14 years of Xinhua newswire.

To decompose each Chinese character into its components, we use a publicly available table which lists each Chinese character and its radical. This list is a national standard published in 2009 [20]. Some radicals do not indicate the meaning of a character, so we filtered them out in our experiments.

Table 1: Evaluation on different parts of speech compared with the fourth level in CiLin extended.

Metrics	Noun			Verb			Adjective		
	Lin	Char	Char/rad	Lin	Char	Char/rad	Lin	Char	Char/rad
P (1)	32.5%	42.7%	42.6%	30.4%	42.0%	42.0%	30.0%	42.0%	42.0%
P (5)	23.9%	33.0%	33.1%	17.7%	25.8%	25.9%	15.2%	23.0%	21.0%
P (10)	19.7%	26.5%	26.7%	12.6%	17.0%	17.1%	10.7%	14.1%	14.2%
INVR	0.865	1.100	1.101	0.608	0.816	0.817	0.519	0.755	0.757

3.2. Evaluation on CiLin Extended Version

The original version of the CiLin thesaurus was published by Mei et al. [28]. In that version, there are four levels in the taxonomy structure. Levels 1 to 3 are taxonomic categories, while Level 4 contains sets of near-synonym terms. There are 53,859 head terms listed. The thesaurus was extended in 2005 by the Information Retrieval Lab of Harbin Institute of Technology [29]. Compared with Mei et al.'s version, 14,706 head terms were removed because they were out-dated, and 38,244 new head terms were added. The extended CiLin consists of 12 large classes, 97 medium classes, 1,400 small classes (topics), and 17,817 small synonym sets which cover 77,343 head terms. In this extended version, there are five levels in the taxonomy structure. Levels 1 to 4 are taxonomic categories, and Level 5 consists of so-called *atom nodes*, which often contain a few words or only one word. To guarantee the quality of this thesaurus, all changes were done manually. This thesaurus can be downloaded freely (http://ir.hit.edu.cn/). In our experiments, we use the fourth level as the gold standard.

We used two evaluation metrics in our experiments: precision of the n top ranked neighbors ($P(n)$), and Inverse Rank ($INVR$), following Curran and Moens [30] (Since we only re-order word neighbours, the direct match metric (DIRECT) is not applicable). In our experiments, we set n to 1, 5, and 10. The $INVR$ score is the sum of the inverse rank of each match. For example, with 50 neighbors, if the first, the third, and the fiftieth neighbor words are present in the gold standard, then the inverse rank score is $1/1 + 1/3 + 1/50 = 1.353$. The ranking obtained using (2.1) serves as the baseline.

We evaluate all nouns, verbs and adjectives whose frequency is greater than 100. The results are shown in Table 1. "Lin" denotes the system's performance obtained by (2.1); "Char" is the score when character information is used; "Char/Rad" is the performance integrating both character and radical information.

There are in total 234,738 nouns in the corpus. Among them, 30,778 words occur more than 100 times. 13,065 of these do not occur in CiLin Extended. We therefore evaluate on the other 17,664 nouns. The $P(10)$ metric is improved by almost 45% compared with the original ranking.

Considering verbs, there are 100,863 in the corpus of which 13,164 occur more than 100 times. 4,355 do not occur in CiLin Extended. So, 8,809 verbs are evaluated. The $P(10)$ metric is improved by more than 31% over the baseline.

Considering adjectives, there are 24,889 in the corpus, with 1,725 occurring more than 100 times. 492 do not occur in CiLin Extended. So, 1,233 adjectives are evaluated. The $P(10)$ metric is improved by more than 35%.

In all, 27,706 target words are evaluated and $P(10)$ is improved by 39.4% over the baseline.

In every case, the improvement compared to the baseline is statistically significant (using pair-wise t-test at 95% confidence). It is unsurprising that the integration of radicals and characters makes only modest improvements compared to the results with characters alone because radicals are only a component of characters. It should be noted that with the evolution of words, the semantic relation between radicals and words becomes looser and looser (e.g., as words become to be used metaphorically).

3.3. *Evaluation on the Rank of Word Pairs*

Aside from the thesaurus style gold standard, another type of gold standard also ranks word pairs according to their similarity score. Finkelstein et al. [31] provided two language resources for the English language, carrying out a psycholinguistic experiment to create this dataset. First, they selected 353 word pairs. All the 30 noun pairs from Miller and Charles [32] were included in this set. Then, they asked a number of native speakers to assign a numerical similarity score between 0 and 10 (0 indicates that these two words are totally unrelated, 10 is very closely related). (Of the 353 word pairs, 153 were annotated by 13 people and the remaining 200 annotated by 16 persons. http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/). During the assignment, an annotator could look in a dictionary when he did not know the meaning of a particular word. But, he was required not to consult other people. Each word pair's similarity score was the mean of all annotators. Finally, the word pairs were ranked by their similarity scores. Following Finkelstein et al. [31], we created a new dataset for the Chinese language (This has been released as benchmark data by SemEval-2012 task 4. http://www.cs.york.ac.uk/semeval-2012/task4/). We first translate all word pairs from Finkelstein et al. [31] by two undergraduates who are Chinese native speakers with a high level of proficiency in English. If their translations are the same, this word pair is retained otherwise it is removed. (There are in fact 169 pairs completely same in literal. When an alias is used, e.g., the word "potato," we consider that the translation "土豆" and "马铃薯" are same. Another case is the preference of monosyllables and disyllable, e.g., the translation, "虎" and "老虎" for "tiger" are same). In the end, only five word pairs were discarded due to having different translations, and one pair was also removed because there were two identical word pairs. The remaining 347 word pairs are annotated by 12 undergraduates. The annotators are all native speakers and their major is linguistics. Each annotator annotated the whole set. In the SemEval-2012 campaign, 50 word pairs were used as trial data. So, the other 297 word pairs were used in this experiment. The Kendall's tau [33] is used as the evaluation metric. Given a rank, if its tau value is higher, it is closer to the gold standard rank.

In this experiment, we used the human annotated rank as a gold standard. The rank returned by (2.1) obtained a tau value of −0.018. The rank according to (2.4) obtained 0.023. While the tau values are not large and the difference is not significant, the reranking by our proposed method clearly moves the similarity calculations towards the human judgments which is encouraging. This result shows that the latter is closer to the gold standard than the original rank. We compared Liu and Li's [10] measure based on HowNet. The rank is obtained according to the word pairs' similarity score returned by the system which is developed by themselves with the default setting of parameters. The Kendall's tau of this system is 0.075, so it is better than our system. This could be expected because they used a thesaurus which was created manually by linguistic experts. It should be noted that there are only 69 word pairs considered for these results because the remaining 228 word pairs contain low-frequency words (less than 100) or where the similarity score of the two words in a pair is too small to be found as one's neighbour. These results highlight the need for more research to handle word similarity computation for low-frequency words and a better way of estimating similarity (dissimilarity) for distant words.

4. Discussion

4.1. Subword-Related Errors

Although our method can improve word similarity computation greatly, it does not always work well. There are in total 397 nouns, 208, verbs, and 32 adjectives whose performance goes down. We have conducted an error analysis to investigate the causes.

Firstly, some words' meanings are not related to the characters that compose the word. Nearly all the transliteration words' meaning have nothing to do with the characters they are composed of. For example, the noun "巴士" (*bus*) which comes from English, neither "巴" nor "士" has any relationship with *bus*. The widespread usage of metaphor decreases the performance when using subword information. The metaphorical mapping based on empirical basis and reason will move some words' meaning far from their original meaning. Another case is that some characters are only used for word construction and do not contribute a concrete meaning.

Secondly, performance could suffer from the ambiguity of characters and radicals. For instance, the noun 表 has at least four frequent meanings. They are *watch* (related to time) in word "手表" (*watch*), *tabulation* in word "表格" (*form*), *surface* in word "地表" (*the earth's surface*), and *piece of equipment for measuring* in "温度表" (*thermometer*). Similar issues apply to radicals. For example, there are two different meanings contributed by the radical 尸, as exemplified by *house* (屋) and *corpse* (屠). But in our study this difference is neglected; making allowances for such differences should also lead to improved performance. Although these factors are not problematic on the whole, because unrelated words do not tend to have high distributional similarity scores, there are cases where these factors result in incorrect rerankings.

4.2. The Performance on Different Word Frequencies

Weeds and Weir [26] demonstrated that word frequency can influence word similarity computation. We therefore selected out target terms according to their frequency, in increments of powers of 10: 2, 3, 4, 5, respectively (illustrated by Figures 2 and 3).

Our method works well both on high-frequency and low-frequency words when using the fourth level of CiLin Extended as the gold standard. For these experiments, we focused on nouns and verbs as there were more of these compared to adjectives (see Figures 2 and 3).

It should be noted that for verbs, the line climbs as frequency increases. However, the line for nouns climbs at the beginning but at "3" begins to drop. We intend to investigate this effect further in future work.

We note that gains over the baseline for low frequency words are larger than for high-frequency words. The former is 53.5% on average, and the latter is 42.6%. The original scores for low-frequency words are typically lower (in agreement with the results of Weeds and Weir [26]) and it is in general easier to make gains on a lower baseline; however, another reason for the difference may be that information from the characters and radicals is helpful in reducing the problems of sparse data which impact corpus-based approaches.

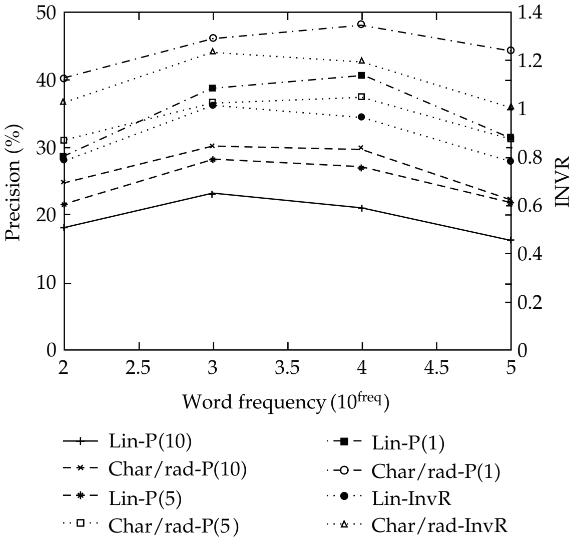


Figure 2: The effect of noun frequency compared with the fourth level in CiLin Extended.

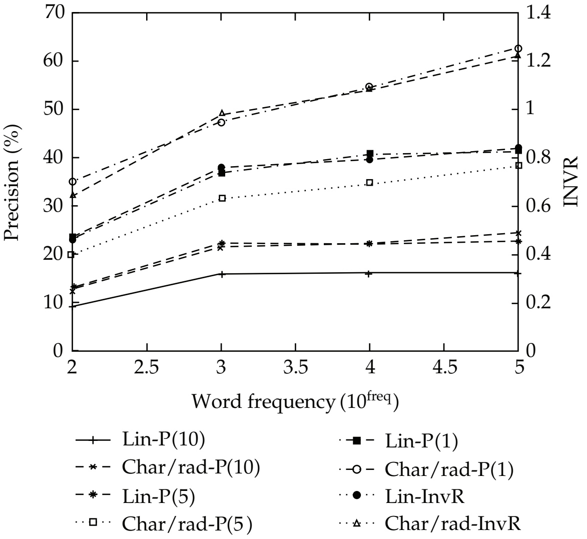


Figure 3: The effect of verb frequency compared with the fourth level in CiLin Extended.

5. Conclusion

We have described a conceptually simple but effective way to improve the computation of Chinese word similarity. The method leverages the hierarchical nature of the Chinese language: the meaning of a word is partly related to the characters it contains. Furthermore, radicals provide clues to the meaning of characters. Evaluated on two gold standard datasets, our method outperforms the state-of-the-art algorithm. We demonstrate that the effectiveness of our method spans different parts of speech (noun, verb, and adjective) and word frequencies.

However, our method makes some simplifying assumptions and takes no account of known special cases. In particular, certain characters such as 阿 should be disregarded because its purpose is for constructing a word and it does not have any concrete meaning. Conversely, some distinct characters contribute similar meanings, so should be taken account of in the similarity scoring. The work of Yencken and Baldwin [34] could be extended to Chinese to capture similarities between pairs of characters.

Our measure does not take the length of words into consideration. Shorter words will have fewer characters and therefore perhaps be ranked lower than they should. We need to investigate this in future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (no. 61003206, 60703063, and 61103089).

References

[1] D. Lin, “Automatic retrieval and clustering of similar words,” in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL/COLING '98)*, pp. 768–774, 1998.

[2] L. Lee, “Measures of distributional similarity,” in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)*, pp. 25–32, 1999.

[3] L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet, “Directional distributional similarity for lexical expansion,” *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (ACLShort '09)*, pp. 69–72, 2009.

[4] P. C. Chang, H. Tseng, D. Jurafsky, and C. D. Manning, “Discriminative reordering with Chinese grammatical relations features,” in *Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation*, pp. 51–59, 2009.

[5] W. Chang, P. Pantel, A. M. Popescu, and E. Gabrilovich, “Towards intent-driven bidterm suggestion,” in *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, pp. 1093–1094, 2009.

[6] A. Budanitsky and G. Hirst, “Evaluating wordnet-based measures of lexical semantic relatedness,” *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[7] G. Dinu and M. Lapata, “Measuring distributional similarity in context,” in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP '10)*, pp. 1162–1172, 2010.

[8] A. Fujii, T. Hasegawa, and T. Tokunaga, “Integration of hand-crafted and statistical resources in measuring word similarity,” in *Proceedings of Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 45–51, 1997.

[9] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, “A study on similarity and relatedness using distributional and WordNet-based approaches,” in *Proceedings of the of Human Language Technology—North America Annual Computational Linguistics*, pp. 19–27, 2009.

[10] Q. Liu and S. Li, “Word similarity computing based on How-Net,” *Computational Linguistics and Chinese Language Processing*, vol. 7, no. 2, pp. 59–76, 2002.

[11] Z. Dong and Q. Dong, *HowNet and the Computation of Meaning*, World Scientific, Singapore, 2006.

[12] X. Lu, “Hybrid models for semantic classification of Chinese unknown words,” in *Proceedings of the Human Language Technologies 2007 and the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL/HLT '07)*, pp. 188–195, 2007.

[13] H. Tseng, “Semantic classification of Chinese unknown words,” in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03)*, vol. 2, pp. 72–79, 2003.

[14] S. K. Hsieh, “Word meaning inducing via character ontology: a survey on the semantic prediction of Chinese two-character words,” in *Proceedings of the 4th SIGHAN Workshop*, pp. 56–63, 2005.

[15] T. Veale and S. Chen, “Unlocking the latent creativity of orthographic structure,” in *Proceedings of the ECAI-06 Workshop on Computational Creativity*, 2006.

[16] C. L. Liu and J. H. Lin, "Using structural information for identifying similar Chinese characters," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL-HLT '08)*, pp. 93–96, 2008.

[17] L. Wang, W. Che, and T. Liu, "A SVMTool-based Chinese POS tagger," *Journal of Chinese Information Processing*, vol. 23, no. 4, pp. 16–21, 2009.

[18] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll, "Unsupervised acquisition of predominant word senses," *Computational Linguistics*, vol. 33, no. 4, pp. 553–590, 2007.

[19] S. Pradhan, W. Ward, K. Hacioglu, J. H. Martin, and D. Jurafsky, "Shallow semantic parsing using support vector machines," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '04)*, pp. 233–240, 2004.

[20] Y. Chen, Y. Lin, J. Chen, and Y. Song, *Specification for Identifying Indexing Components of GB, 13000.1 Chinese Characters Set*, Language and Literature Press, Beijing, China, 2009.

[21] P. Jin, J. Carroll, Y. Wu, and D. McCarthy, "Improved word similarity computation for Chinese using sub-word information," in *Proceedings of the International Conference on Computational Intelligence and Security*, pp. 459–462, 2011.

[22] L. Yencken and T. Baldwin, "Modelling the orthographic neighbourhood for Japanese Kanji," in *Proceedings of the 21st International Conference on the Computer Processing of Oriental Languages*, Singapore, 2006.

[23] F. Xia, "The part-of-speech guidelines for the Penn Chinese Treebank project," Technical Report IRCS 00-06, University of Pennsylvania, 2000.

[24] J. Weeds, D. Weir, and B. Keller, "The distributional similarity of sub-parses," in *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp. 7–12, 2005.

[25] K. Lindén and J. Piitulainen, "Discovering synonyms and other related words," in *Proceedings of the 3rd International Workshop on Computational Terminology*, pp. 63–70, 2004.

[26] J. Weeds and D. Weir, "Co-occurrence retrieval: a flexible framework for lexical distributional similarity," *Computational Linguistics*, vol. 31, no. 4, pp. 439–475, 2006.

[27] L. Sarmento, P. Carvalho, and E. Oliveira, "Exploring the vector space model for finding verb synonyms in Portuguese," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP '09)*, pp. 393–398, 2009.

[28] J. Mei, Y. Zheng, Y. Gao, and H. Yin, *TongYiCi CiLin*, Commercial Press, Shanghai, China, 1984.

[29] Z. Lu, H. Wang, J. Yao, T. Liu, and S. Li, "An equivalent pseudoword solution to Chinese word sense disambiguation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL/COLING '06)*, pp. 457–464, 2006.

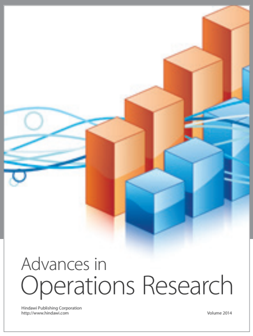
[30] J. Curran and M. Moens, "Scaling context space," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pp. 231–238, Philadelphia, Pa, USA, July 2002.

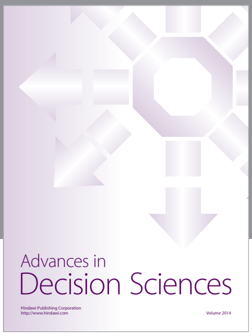
[31] L. Finkelstein, E. Gabrilovich, Y. Matias et al., "Placing search in context: the concept revisited," *ACM Transactions on Information Systems*, vol. 20, no. 1, pp. 116–131, 2002.

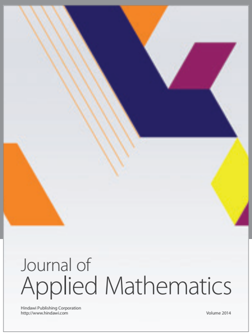
[32] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.

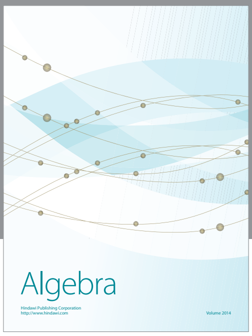
[33] M. Lapata, "Automatic evaluation of information ordering: Kendall's Tau," *Computational Linguistics*, vol. 32, no. 4, pp. 471–484, 2006.

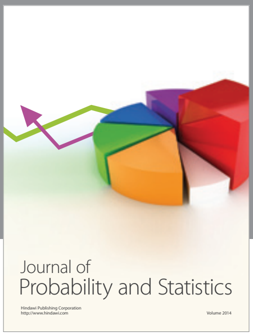
[34] L. Yencken and T. Baldwin, "Measuring and predicting orthographic associations: modeling the similarity of Japanese Kanji," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, pp. 1041–1048, 2008.


Advances in
Operations Research

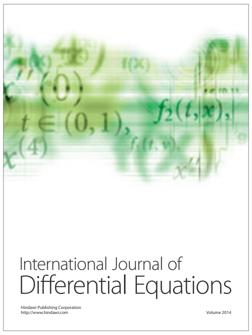

Advances in
Decision Sciences

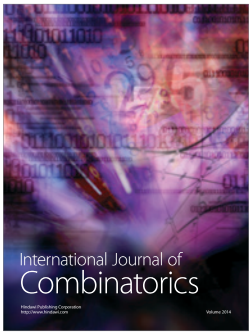

Journal of
Applied Mathematics

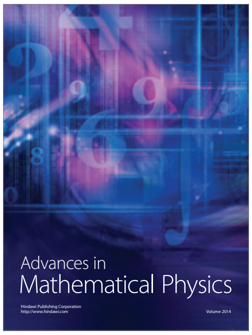

Algebra

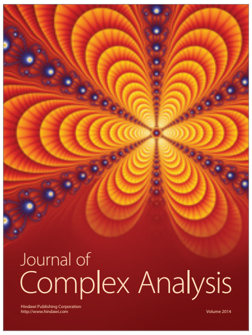

Journal of
Probability and Statistics

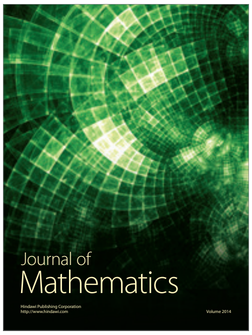
The Scientific
World Journal

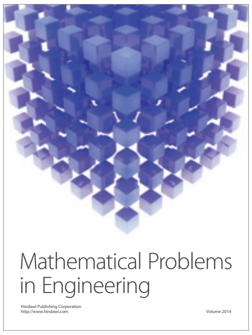

International Journal of
Differential Equations

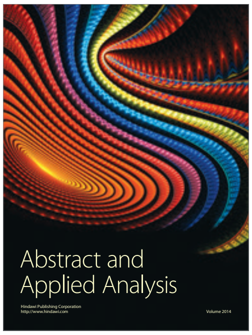

International Journal of
Combinatorics

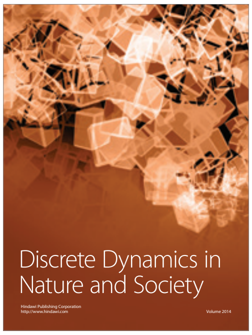

Advances in
Mathematical Physics

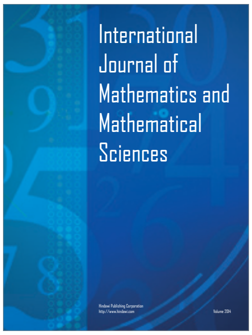

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

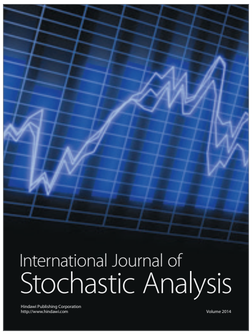

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization