Unsupervised Clustering of Morphologically Related Chinese Words

Chia-Ling Lee (r00922072@ntu.edu.tw)
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan

Ya-Ning Chang (yaningchang@gate.sinica.edu.tw)
Institute of Linguistics,
Academia Sinica, Taipei, Taiwan

Chao-Lin Liu (chaolin@nccu.edu.tw)
Department of Computer Science,
National Chengchi University, Taipei, Taiwan

Chia-Ying Lee (chiaying@gate.sinica.edu.tw)
Institute of Linguistics,
Academia Sinica, Taipei, Taiwan

Jane Yung-jen Hsu (yjhsu@csie.ntu.edu.tw)
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan

Abstract

Many linguists consider morphological awareness a major factor that affects children's reading development. A Chinese character embedded in different compound words may carry related but different meanings. For example, "商店(store)", "商品(commodity)", "商代(Shang Dynasty)", and "商朝(Shang Dynasty)" can form two clusters: {"商店", "商品"} and {"商代", "商朝"}. In this paper, we aim at unsupervised clustering of a given family of morphologically related Chinese words. Successfully differentiating these words can contribute to both computer assisted Chinese learning and natural language understanding. In Experiment 1, we employed linguistic factors at the word, syntactic, semantic, and contextual levels in aggregated computational linguistics methods to handle the clustering task. In Experiment 2, we recruited adults and children to perform the clustering task. Experimental results indicate that our computational model achieved the same level of performance as children.

Keywords: morphological awareness; human cognition; computational linguistics; Chinese character meaning

Introduction

Morphological awareness, defined as "children's conscious awareness of the morphemic structure of words and their ability to reflect on and manipulate that structure", is associated with children's reading ability and comprehension (Liu & McBride-Chang, 2010; Kirby et al., 2012; Ku & Anderson, 2003). It is thought by many linguists to strongly affect reading development in children (Liu & McBride-Chang, 2010).

A Chinese character embedded in different compound words may carry related but different meanings. For example, the meaning of the character "商/shang1/" in words"商店(store)" and "商品(commodity)" is commerce. In contrast, in "商代(Shang Dynasty)", "商" refers to a Chinese dynasty. Successful clustering of related Chinese words would make a contribution to Chinese learning. In addition, differentiating the character's meanings in such morphologically related words can facilitate Chinese word sense disambiguation and help improve Chinese word segmentation (Navigli, 2009).

In this research, we employ natural language processing and computational linguistics techniques to differentiate the meanings of a particular character that is embedded in different Chinese words. We apply different methods which take diverse factors into account, such as grammar, syntax, semantics, and context. We also aggregate all methods and build a better ensemble model. Furthermore, we conduct another experiment in which we asked adults and children to do the same clustering task. Experimental results indicate that our model can achieve the same level of performance as children in the clustering task.

There is previous work related to morphological awareness. Wang, Hsu, Tien, and Pomplun (2012) predicted raters' transparency judgments of Chinese morphological character based on latent semantic analysis (LSA) (Landauer, Foltz, & Laham, 1998). If a word is more similar to the primary meaning, it is more likely to be judged as semantically transparent, and opaque otherwise.

Galmar and Chen (2010) tried to identify different meanings of a Chinese character using LSA and semantic pattern matching in augmented minimum spanning tree. Galmar (2011) built a term-by-document matrix, and used the batch version of self-organizing maps (Kohonen, 2001) to visualize the interplay between morphology and semantics in Chinese words.

To discriminate Chinese character meanings, in addition to LSA techniques, we consider diverse information from comprehensive aspects. There are numerous word-to-word semantic similarity or relatedness measures proposed in the past. In knowledge-based approaches, WordNet[1] was

[1] http://wordnet.princeton.edu

widely used (Pedersen, Patwardhan, & Michelizzi, 2004; Patwardhan & Pedersen, 2006; Mihalcea, Corley, & Strapparava, 2006). To compute word-to-word semantic similarity, syntactic dependency (Lin, 1997; Padó & Lapata, 2007), information content of the common subsumer of concepts (Resnik, 1995), and shortest path length between two concepts (Leacock & Chodorow, 1998) were used. In addition to WordNet, some adopted HowNet[2] as a knowledge base for Chinese (Dai, Liu, Xia, & Wu, 2008). For corpus-based approaches, perhaps the commonest one is the LSA. For recognizing synonyms, Turney (2001) used pointwise mutual information and information retrieval to measure the similarity of pairs of words. Additionally, for taking statistics and co-occurrence into account, Jaccard coefficient, Simpson coefficient, and Dice coefficient are measured (Manning & Schütze, 1999; Jackson, Somers, & Harvey, 1989)

Methods

Notations and Problem Definition

We first introduce terminologies and notations used in this paper. We denote a Chinese character by c, a word by w. $family(c)$ is a set of Chinese words in which all words contain a common character c. We call $family(c)$ a morphological family of a Chinese character c, each word in that set a target word, and the shared character c a target character. Target words within a morphological family sharing the same target character meaning compose a meaning group.

Given a set of Chinese words, $family(c)$, our goal is to differentiate meanings of the target character c in each word and to group them into clusters. Take the morphological family of "商" for example, the set {"商店", "商品", "商代", "商朝"} could be separated into two clusters: {"商店", "商品"} and {"商代", "商朝"}. This is because that the character "商" in both words within first cluster are related to commerce or business whereas the meaning of the same character in the second cluster is a dynasty in China history. In our work, we assume a character has only one meaning in each word.

Experiment 1

Framework Figure 1 illustrates the framework of Experiment 1. First, given a morphological family, we apply a couple of methods to calculate similarities between target words. In each method, we use different features to calculate similarity between words and get corresponding similarity matrices. Given these matrices, we are able to ensemble them, and get an ensemble matrix. The final step is to cluster target words by using a hierarchical clustering algorithm. In the end, we get a clustering result of morphologically related words.

Word-to-word Similarity Harris (1954) proposed a hypothesis that "words that occur in similar contexts tend to have similar meanings." In Chinese compounds, a constituent character provides some clues to the semantic of a compound.

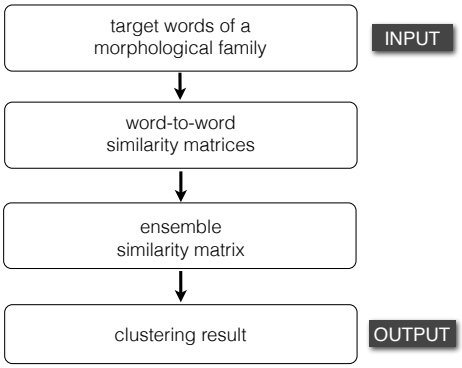

Figure 1: The framework of Experiment 1.

To begin with, we estimate word-to-word semantic similarities. In this paper, we apply LSA and propose three different methods, which are elaborated as following. In each method, a word w is represented by a feature vector $V(w)$. The similarity is determined by cosine similarity:

$$Sim_{vec}(w_i, w_j) = \cos(V(w_i), V(w_j)).$$

LSA: LSA has been widely used in natural language processing, text mining, and information retrieval. LSA assumes that words with closer meaning will occur in similar documents. From our corpus, we first construct a term-by-document matrix T. The value of a cell T_{ij} is the normalized frequency of word w_i shown in a document d_j. After preprocessing, truncated singular value decomposition (Golub & Reinsch, 1970) is used to reduce the dimensions of T. In our case, we reduce to 100 dimensions. Each value of the feature vector is mapped to a so-called latent topic. Thus, we get a latent semantic feature vector $V_{lsa}(w)$ for a target word w.

Document: In this method, we would like to capture document-category level context of a word. That is to say, we view the category of the document where a target word occurs as a feature. We construct a matrix D where a row dimension represents a target word and a column dimension is a type of a document. The value of a cell is the normalized frequency of a target word occurring in the corresponding document type. In our corpus, there is a total of 90 genres, styles, and topics. A feature vector of the target word w_i, $V_{doc}(w_i)$, is denoted by $(D_{i,1}, D_{i,2}, ..., D_{i,90})$.

Relation: To the sentence level, we would also like to know the relation between words in a sentence. Through the tool of Stanford Parser[3], which provides the grammatical relations between words, a sentence could be parsed and represented as several Stanford typed dependencies. The tool supports Chinese as well. A typed dependency is a triplet: name of the relation, governor and dependent. Name of the relation is what we focus on now. Take the sentence, "I love you.", for example, one of dependencies is *nsubj(love, I)*, which means

[2]http://www.keenage.com

[3]http://nlp.stanford.edu/software/lex-parser.shtml

that the dependent "I" is a nominal subject of the governor "love" and their relation is "nsubj". We count how frequently a target word plays a role of a dependent in each kind of relation. A grammatical feature vector $V_{rel}(w)$ is generated.

POS: Additionally, we are interested in another syntactic feature: part-of-speech (POS), such as noun, verb, adjective, etc. From segmented texts with POS tags, we construct a matrix where column dimensions are POS tags and each row i is mapped to a target word w_i. Likewise, the value of a matrix cell is the normalized frequency that a word is tagged by the corresponding POS tag in a text. A syntactic feature vector of a target word w_i, $V_{pos}(w_i)$, reflects a distribution of POS tags of the word. We expect that two similar target words have similar distributions of POS tags in a corpus.

Ensemble Until now, we have taken various aspects into consideration, including semantic, topical, grammatical, and syntactic. For each method, we generate one word-to-word similarity matrix. Next step is to integrate them.

Source of ensemble approach comes from these m similarity matrices. Here, we denote a similarity matrix by M. We then aggregate the similarity matrices by accumulating them with different weights. The idea of this approach is that if two words are similar in many aspects, there should be more matrices giving this pair a high similarity score, and the resultant score will be higher comparatively. Through weighted accumulation, scores of similar and dissimilar words will be distanced. The aggregated matrix is defined mathematically as:

$$M_{ensemble} = \frac{1}{m}\sum_{i=1}^{m}(\omega_i \times M_i)$$

where ω_i is the weight of the word-to-word similarity of method i.

Clustering The objective of clustering is to group similar words together and separate dissimilar words into different clusters. Clustering algorithms are classified into three main categories: hierarchical, partitional, and hybrid. Although a partitional clustering method (e.g., K-means) runs faster since it does not need to compare all pairs of items, the number of clusters is usually required to be given. In our case, the number of clusters is not determined in advance, we thus employ a hierarchical agglomerative clustering (HAC) algorithm in this work.

When HAC starts, each target word is viewed as a singleton cluster. At each iteration, two most similar clusters are merged. We run iteratively until clusters similarity score reaches a predefined threshold θ. We set θ to 0.15 based on some pilot trials. If a similarity of two clusters is greater than the threshold, they will be merged and become a new cluster. When no new cluster could be produced, the outcome is generated. Figure 2 illustrates a cluster dendrogram of $family$("商"). It demonstrates how HAC works on the target words of "商".

Average link is adopted to estimate similarity between two clusters. Given two clusters, C_a and C_b, the similarity between clusters is defined as below.

$$sim_{cluster}(C_a, C_b) = \sum_{w_i \in C_a, w_j \in C_b} \frac{sim(w_i, w_j)}{|C_a||C_b|}$$

We finally apply HAC on the matrix $M_{ensemble}$ and get a clustering result.

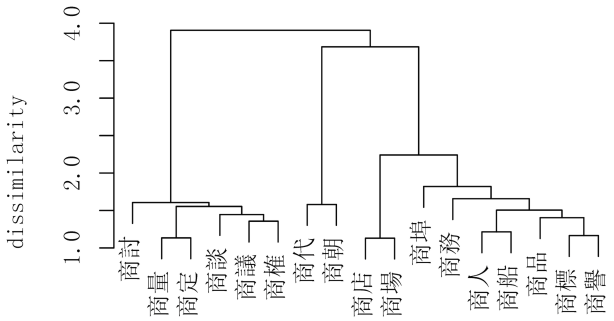

Figure 2: Cluster dendrogram of $family$("商")

Experiment 2

Participants What about human performance on our clustering task and how does it compare with our computational results? In Experiment 2, we asked two groups of participants, 14 adults and 9 children, to perform the clustering task. 14 adults were graduate students recruited from National Taiwan University and Academia Sinica. All were native speakers of Chinese. Child group consists of 9 children from fifth to sixth grade of primary school students.

Materials and procedure In adult group, a task of clustering of morphologically related Chinese words was completed by using a questionnaire. The questionnaire included 11 morphological families. In each family, all target words were listed and participants were asked to group morphologically related words into clusters. The number of clusters were not limited.

In child group, instead of a questionnaire, we gave them word cards to reduce task difficulty. For each target word in all 11 families, we made a word card with notional phonetic alphabets. Since our task of clustering is not very easy for most children, we separated the task into two stages. In each stage, a kid was asked to finish 4 to 7 families. Moreover, during the task, if they did not know a target word, we would ask them to guess or put it to another group named "I do not know." When evaluating, we would view each target word in that group as a single cluster. (Among 285 target words, 8% words were not recognized on average per child.)

Results

Corpus and Dataset

The corpus we used is the Academia Sinica Balanced Corpus (ASBC) version 3 (Huang & Chen, 1998). ASBC is a

Table 1: An example of easy-degree morphological families, *family*(“商”). High-frequency target words are shown in boldface.

Target character	Target Words	Meaning of target character
商/shang1/	**商標, 商務, 商品, 商店, 商人, 商場**, 商譽, 商埠, 商船	things related to commerce or business
	商議, 商量, 商討	negotiation or discussion
	商代, 商朝	a China dynasty
花/hua1/	**花卉, 花園, 花草, 花香, 花瓣**, 花店, 花農, 花季, 花海, 花蜜, 花苞	things related to plants
	火花, 浪花, 雪花, 花樣, 花紋, 花邊, 花布, 油花, 水花, 花式, 花招	patterns or styles
	花費, 花錢, 花用, 花掉, 花光	expenditure or costs
	花蓮	a place name

balanced Chinese corpus with part-of-speech tagging. Each article in the corpus is classified and marked according to five criteria: genre, style, mode, topic, and medium. The corpus is also segmented according to the word segmentation standard proposed by Huang, Chen, and Chang (1996). ASBC contains more than 9 thousands articles, near 5 millions words, and 144 thousands unique words.

Our input data and ground truth were provided by psycholinguistic researchers of the Institute of Linguistics, Academia Sinica. There are 11 morphological families, including 285 target words. To test word-frequency effects, we separated all words into two groups based on word frequency, a threshold of 20. The high-frequency group and low-frequency group contain 139 and 146 target words respectively. We expected that the high-frequency group would have better performance than the low-frequency one. In addition, to test difficulty effects, the psycholinguists also annotated these morphological families with two degrees of difficulty: hard and easy. Hard-degree means that it is more difficult to discriminate the target character's meanings. Inversely, a target character in a easy-degree family can be differentiated easier. These 11 families are divided into 6 hard-degree and 5 easy-degree ones.

Two examples of morphological families, *family*(“商”) and *family*(“花”), are shown in Table 1. Each row presents a meaning group and high frequency target words are printed in boldface.

Evaluation

To evaluate our performance, we use two metrics: F1 and *normalized mutual information* (NMI) (Manning & Schütze, 1999). F1 describes how correctly when we make a decision of assigning two words to the same cluster. Mutual information measures the amount of information by which our knowledge about the classes increases when we are told what the clusters are. Normalization is required to penalize large cardinalities which will cause high mutual information.

However, we found a trend that we did not expect. In terms of F1, when the threshold θ of HAC increased, meaning that a larger number of clusters, F1 became worse. In contrast, when evaluated with NMI, the performance improved as the number of clusters increased. Actually the tendencies may not be difficult to understand when we look deeper into the definitions of F1 and MNI. When the number of clusters becomes larger, there are less and less pairs of words within a cluster, even one-word clusters. That is to say, to a certain extent, we lose chances to gain F1. However, maximum mutual information can be gathered when clusters are further subdivided into smaller clusters. These trends are more obvious especially in clustering on small data sets.

To prevent the threshold dominating our performance, in this paper, we propose a new metric named F-NMI to address the issue particularly. We define F-NMI as $\alpha \times \text{F1} + (1-\alpha) \times \text{NMI}$ and set α to 0.5 in our experiments.

Experimental Results

We averaged performances across 11 families of each method and each human group. Table 2 summarizes the results of Experiments 1 and 2.

Table 2: F1, NMI, and F-NMI of Experiment 1 (computational methods) Experiment 2 (human clustering result)

Method	NMI	F1	F-NMI	Frequency Effects	Difficulty Effects
Experiment 1					
Random	8.34%	42.54%	25.44%	5.18%	-1.29%
LSA	27.51%	45.86%	36.68%	8.85%	7.05%
Document	7.95%	50.62%	29.28%	7.86%	-2.43%
Relation	19.07%	50.01%	34.54%	6.26%	0.59%
POS	40.85%	54.05%	47.45%	-1.97%	0.96%
Ensemble	40.85%	60.83%	**50.84%**	4.66%	6.73%
Experiment 2					
Adult	77.49%	76.12%	76.80%	3.22%	7.90%
Child	56.46%	54.58%	55.52%	8.08%	2.84%

As expected, the ensemble method worked best in general since they considered various linguistic factors. It achieved the best performances compared with other computational methods across all three metrics: 40.85% of NMI, 60.83% of F1, and 50.84% of F-NMI. In addition to the ensemble method, the POS method had prominent performances as well. Although other methods did not have distinguished performances, they all outperformed a random similarity in terms of F-NMI. In Experiment 2, adult group achieved 77.49% of NMI, 76.12% of F1, and 76.80% of F-NMI in average. It shows a high agreement with our gold standard. The child group reached 56.46% of NMI, 54.58% of F1, and 55.52% of F-NMI. It is obvious that the performance of chil-

dren is far from the adult group. Moreover, although there was some distance away from the adult group, the ensemble method (F-NMI = 50.84%, $s.d.$ = 20.41%) reached the same performance level with child group (F-NMI = 55.52%, $s.d.$ = 10.56%) in terms of F-NMI.

Both word-frequency effects and difficulty effects are also what we are interested in. Figure 3 illustrates frequency effects. The child group had frequency effects ($N = 9$, one-tailed, $p = 0.035$, $t = 1.95$) where the F-NMI score in high-frequency words was higher than in low-frequency. However, probably due to small samples (11 families, 14 adults), the difference between high and low-frequency words was not statistically significant in the adult group and ensemble method.

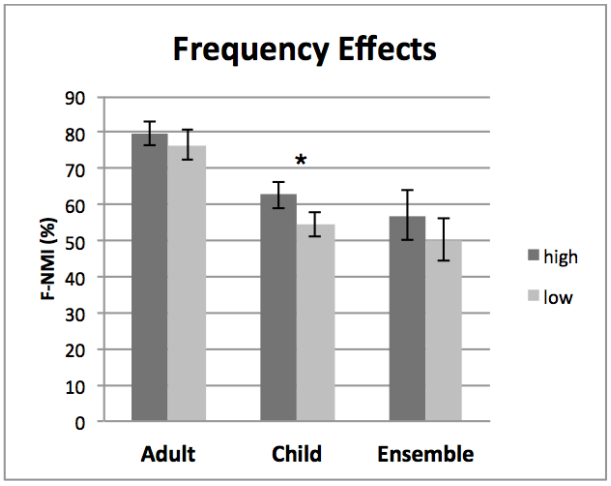

Figure 3: Frequency effects of adult group, child group, and the ensemble method.

Figure 4 depicts difficulty effects. The adult group had difficulty effects ($N = 14$, one-tailed, $p = 0.023$, $t = 2.09$) where the performance in easy-degree families was better than hard-degree ones. Because we only have 5 hard-degree and 6 easy-degree families, and the variation among 11 families are considerably large, in the child group and our ensemble method, the differences between hard and easy-degree families were not significant.

Discussion

In this paper, we aim at differentiating the meanings of the character shared by different target words. We propose several computational methods to calculate word-to-word similarities. Not only latent semantics but also various factors, including document category, dependency, and POS tags, are taken into account. Through aggregating an array of methods, the ensemble method achieved the same level of performance with the child group.

It seems that the ensemble method could provide more comprehensive information to us for discriminating the meanings of Chinese characters, compared with other non-ensemble methods. The results of Experiment 1 suggest

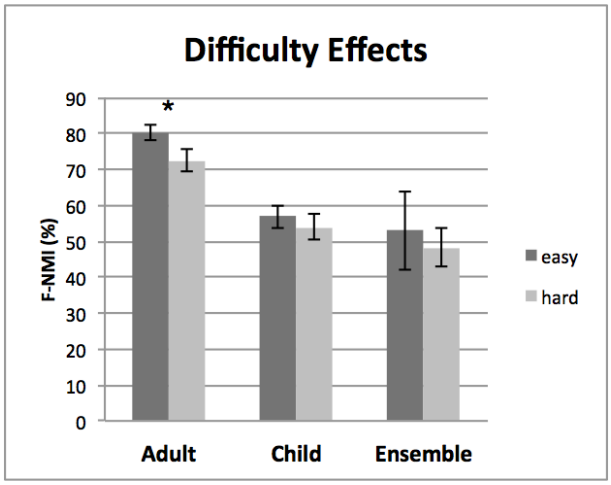

Figure 4: Difficulty effects of adult group, child group, and the ensemble method.

that it is hard to discriminate a Chinese character's meanings by concerning only one factor. For example, both "商店(store)" and "商品(commodity)" often appeared in articles about commerce or business. However, some other words which are assigned into the same meaning group because their target character provides common implicit senses to the words and further impact the roles of the words in a sentence. Take another morphological family of "生/sheng1/" for instance. Target words "醫生(doctor)", "女生(girl; woman)", "出生(born)", and "誕生(born)", to some degree, often occur in medicine or gynecology-related articles. Nonetheless, these words can be divided into two groups: {"醫生", "女生"} and {"出生", "誕生"}. In the former group, "生" means "person", and it is a noun. The latter means "to give birth to" and it is a verb. Thus, POS tag distribution complements the factor of document category. To summarize, even though each single method did not have distinguished performances and could be improved further, all of these methods were essential to achieve the best system performance.

We observe that methods that consider internal structures of the words captured the meaning of the shared character more precisely. In addition to the ensemble method, among other four computational methods, the POS method worked the best. Specifically, the POS method had the best performance; LSA was second to POS; next one was the Relation method, and the Document method was the worst. A POS tag provides a clue to how a word functions in a sentence. In contrast, document category was too abstract to help us differentiate the characters' meanings. The relation of a dependency provided information at the sentence level and its performance was between the document category method and the POS method.

Finally, although the ensemble method could achieve similar performances as children, it is still not as good as adults. Some internal structure of the word may contain more useful information and should be further explored. In the future,

Chinese character-based computational techniques can be investigated, such as character-level syntax tree and character-based dependency (Zhang, Zhang, Che, & Liu, 2013; Zhao, 2009). Moreover, we hope to build an e-learning platform and apply our methods to assist teachers in teaching students to learn Chinese.

Acknowledgments

This work was supported in part by the grants NSC 99-2221-E-002-139-MY3, NSC 101-2627-E-002-002, NSC101-2221-E-004-018, and NSC 102-2420-H-001-006-MY2 from the National Science Council, Taiwan, and NTU 102R890864 from National Taiwan University, Taiwan.

References

Dai, L., Liu, B., Xia, Y., & Wu, S. (2008). Measuring semantic similarity between words using HowNet. In *Proceedings of the international conference on computer science and information technology* (pp. 601–605).

Galmar, B. (2011). Using Kohonen maps of Chinese morphological families to visualize the interplay of morphology and semantics in Chinese. In *Proceedings of the twenty third conference on computational linguistics and speech processing* (pp. 240–251). Taipei, Taiwan.

Galmar, B., & Chen, J.-Y. (2010). Identifying different meanings of a Chinese morpheme through semantic pattern matching in augmented minimum spanning trees. *International Journal of Computational Linguistics and Applications*(1-2), 153–168.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, *14*(5), 403–420.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 146–162.

Huang, C.-R., & Chen, K.-J. (1998). Academia Sinica balanced corpus (version 3). *Taipei, Taiwan: Academia Sinica*.

Huang, C.-R., Chen, K.-J., & Chang, L.-L. (1996). Segmentation standard for Chinese natural language processing. In *Proceedings of the sixteenth conference on computational linguistics* (pp. 1045–1048).

Jackson, D. A., Somers, K. M., & Harvey, H. H. (1989). Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? *American Naturalist*, *133*(3), 436–453.

Kirby, J. R., Deacon, S. H., Bowers, P. N., Izenberg, L., Wade-Woolley, L., & Parrila, R. (2012). Children's morphological awareness and reading ability. *Reading and Writing*, *25*(2), 389–410.

Kohonen, T. (2001). *Self-organizing maps*. Verlog, Berlin: Springer.

Ku, Y.-M., & Anderson, R. C. (2003). Development of morphological awareness in Chinese and English. *Reading and Writing: An Interdisciplinary Journal*, *16*(5), 399–422.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In *WordNet: An electronic lexical database.* MIT Press.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the thirty-fifth annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (pp. 64–71). Madrid, Spain.

Liu, P. D., & McBride-Chang, C. (2010). What is morphological awareness? Tapping lexical compounding awareness in Chinese third graders. *Journal of Educational Psychology*, *102*(1), 62–73.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the twenty-first conference on artificial intelligence* (Vol. 21, p. 775). Boston, MA.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, *41*(2), 1–69.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL workshop on making sense of sense: Bringing psycholinguistics and computational linguistics together* (pp. 1–8). Trento, Italy.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::similarity - measuring the relatedness of concepts. In *Proceedings of the nineteenth national conference on artificial intelligence* (pp. 38–41). San Jose, CA.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In (pp. 448–453). Montreal, Canada.

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the twelfth european conference on machine learning* (pp. 491–502). Freiburg, Germany.

Wang, H.-C., Hsu, L.-C., Tien, Y.-M., & Pomplun, M. (2012). Estimating semantic transparency of constituents of English compounds and two-character Chinese words using latent semantic analysis. In *Proceedings of annual meeting of the cognitive science society.* Sapporo, Japan.

Zhang, M., Zhang, Y., Che, W., & Liu, T. (2013). Chinese parsing exploiting characters. In *51st annual meeting of the association for computational linguistics.*

Zhao, H. (2009). Character-level dependencies in Chinese: Usefulness and learning. In *Proceedings of the 12th conference of the european chapter of the association for computational linguistics* (pp. 879–887).