1A3-IOS-3a-1

Unsupervised Sense Clustering of Related Chinese Words

Chia-Ling Lee[1] Yen-Ling Kuo[1] Chia-Mau Ni[1] Yu-Ju Chen[1] Chao-Lin Liu[2]
Jane Yung-jen Hsu[1]

[1] National Taiwan University [2] National Chengchi University

Chinese words which share the same character may carry related but different meanings, e.g., "花錢(spend)", "花費(expend)", "花園(garden)", "開花(bloom))". The semantics of these words form two clusters: {"花錢(spend)", "花費(expend)"} and {"花園(garden)", "開花(bloom)"}. In this paper, we aim at unsupervised clustering of a given set of such related Chinese words, where the quality of clustering results is to be judged based on the senses of the related words. Successfully differentiating these words not only represents an important step toward Chinese word sense disambiguation but also helps for computer assisted Chinese learning. Currently, we considered semantic context and tried to determine the clustering roughly. We employed two knowledge-based methods using ConceptNet5 and two corpus-based measures based on Chinese Wikipedia. Our experimental result shows it can be achieved 68.49% of Rand index in the best case using knowledge-based approach.

1. Introduction

Chinese words which share the same character may carry related but different meanings. Successfully differentiating these character senses in related words can facilitate Chinese word sense disambiguation (Navigli, 2009) significantly and even help improve Chinese word segmentation (Xue, 2003; Nakagawa, 2004). In addition to natural language processing, sense clustering of related Chinese words would give contribution to the field of linguistics. When children in early school years are learning to read Chinese, morphological awareness develops and grows with the increasing vocabulary. To be more specifically, we can find that most of the time, when we read a word that we have never read or heard before, based on our knowledge to word senses, we still can guess and infer its meaning. This skill is believed by many linguists to play an important role and strongly relates to reading ability (Liu & McBride-Chang, 2010; Ku & Anderson, 2003).

In this research, our goal is to use natural language processing techniques to differentiate sense of the same character embedded in different Chinese words. We believe that we could build software and offer an instrumental facility for computer assisted Chinese learning to help children and those dyslexic readers who are regarded as less aware of morphemes to construct morphological awareness (Shu, McBride-Chang, Wu, & Liu, 2006).

There are some related works, even though our goals are not identical but similar. Liou et al. (Liou, Cheng, Liou, & Liou, 2013) presented a method to train multi-code for the polysemous word by redesigning Elman network which has a context layer to find the hidden structure of sequential patterns. Some multi-meaning characters were encoded in this way and each meaning is represented as a code vector. Galmar (Galmar, 2011) built a term-by-document matrix, using such a matrix and batch self organizing maps (SOMs) to visualize the interplay between morphology and semantics in Chinese words sharing a common morpheme. The work of semantic clustering is designed for the study of morphological satiation in Chinese.

In our work, we used natural language processing approaches and took contextual information into consideration. Figure 1 presents an illustration of our framework. As we can see, from a given set of words sharing a common character, for each word, from our corpus we bagged the sentences containing the word into a set. The main idea behind this is that we used the concept of "bag of words" to capture the context of a word roughly. To compare the similarity between bags, we explored two well-known corpus-based and two knowledge-based methods. Ultimately, we applied a centroid clustering method to perform clustering.

Numerous methods discuss measures of similarity between two words or concepts using either knowledge-based or corpus-based approaches. For example, the well-known on-line lexical database, WordNet[1], is widely used to compute word-to-word semantic similarity (Pedersen, Patwardhan, & Michelizzi, 2004; Mihalcea, Corley, & Strapparava, 2006; Agirre et al., 2009). In addition to WordNet, in Chinese, some take HowNet[2] as their knowledge base (Dai, Liu, Xia, & Wu, 2008). For corpus-based approaches, perhaps the commonest one is the latent semantic analysis (LSA) proposed by Landauer et al.(Landauer, Foltz, & Laham, 1998). Additionally, for taking statistic or co-occurrence into account, Pointwise Mutual Information (PMI), Jaccard coefficient, Simpson coefficient, and Dice coefficient are measured (Manning & Schütze, 1999). In these corpus-based approaches, making use of web search engines and using web pages on the Internet as live big-scaled corpus is getting more and more popular recently(Turney, 2001; Bollegala, Matsuo, & Ishizuka, 2007; Iosif & Potamianos, 2010).

Finally, we achieved the best performance 68.49% measured by the Rand index using the knowledge-based method

Contact: Chia-Ling Lee, National Taiwan University, Taipei, Taiwan, 886-2-33664888, 886-2-23628167, spiderman0103@gmail.com

1 http://wordnet.princeton.edu
2 http://www.keenage.com

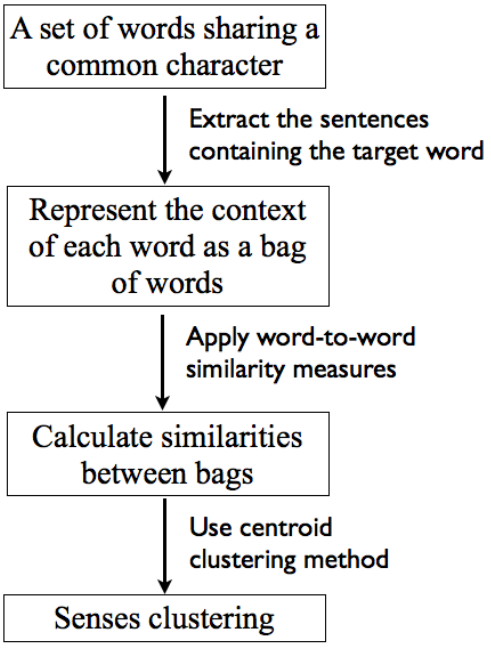


Figure 1: The framework of our approach

based on ConceptNet[3], which is a crowd-sourced commonsense knowledge database. At the current stage, we are simply able to differentiate senses roughly.

In Section 2, we explain the problem definition. Section 3 presents our methodology to do sense clustering. In Section 4, we explain how we obtain the corpora and test data. Section 5 shows our experiments. At last, summary and the future work are included in Section 6.

2. Problem Definition

Before describing the approaches we applied in this problem, we first define our notations used in this paper. We denote a Chinese character c, a word w and a document d. The notation D is a corpus containing $|D|$ documents. $B(w)$ represents the bag of words of a given word w, and $C(w)$ means the concepts of w in ConceptNet. $family(c)$ is a Chinese words set in which all words contain the common character c. We call $family(c)$ a morphological family of Chinese character c and words in the set "target words". At last, notice that "word" and "term" are interchangeable in this paper.

Given a set of Chinese words sharing a common Chinese character c, $family(c)$, our goal is to differentiate the sense of the character c in each word and to group them into clusters (three clusters in this paper). Take the morphological family of "花/hua1/" for example, the set {"花錢(spend)", "花費(expend)", "花園(garden)", "開花(bloom)"} could be separated into two clusters: {"花錢(spend)", "花費(expense)"} and {"花園(garden)", "開花(bloom)"}, since the character "花/hua1/" in both words within first cluster have the same meaning "expense" whereas the sense in the second cluster is "flower". That is, the character has two sense classes in this case. In our work, we assume a character has only one sense in each one of words.

3 http://conceptnet5.media.mit.edu

3. Methodology

3.1 Bag-to-Bag Similarity: Context-based

Since by observing a set of Chinese words that embedding the same character c, $family(c)$, we found that words in the same class, i.e., the words with similar senses of c, have similar semantic contexts. Take $family$(商/shang1/) for instance, context of "商店(store)" and "商品(commodity)" are related to commerce or business; "商代(Shang Dynasty)" and "商朝(Shang Dynasty)" are taking about a dynasty in Chinese history. For this reason, we tried to differentiate their semantic contexts.

To take contexts into consideration, we would like to define the context of a target word w in a morphological family. We used a concept of "bag of words". From our corpus, we extracted all sentences which include the target word w into a bag. That is, we represented a context as a bag of words since we thought the neighbor words would contain some related information. When collecting "context" of each target word, we would filter out stop words to reduce some possible noises. Then, we introduced a scoring function proposed by Rada Mihalcea (Mihalcea et al., 2006) that originally calculated the similarity between two text segments. Nevertheless, in our case, we applied it to compare two contexts of words.

Let us denote the context of word w as $B(w)$, i.e., bag of words of word w. The similarity of two semantic contexts of words w_i and w_j is determined as

$$\begin{aligned} Sim_{context}(w_i, w_j) &= Sim(B(w_i), B(w_j)) \\ &= \frac{1}{2}\Big(\frac{\sum_{w \in B(w_i)} (maxSim(w, B(w_j)) * idf(w))}{\sum_{w \in B(w_i)} idf(w)} \\ &\quad + \frac{\sum_{w \in B(w_j)} (maxSim(w, B(w_i)) * idf(w))}{\sum_{w \in B(w_j)} idf(w)}\Big) \end{aligned} \quad (1)$$

For each word w in $B(w_i)$, we calculated the similarities to every words in $B(w_j)$ according to four word-to-word similarity measures to be described in the next section. We then picked the highest score among these, $maxSim(w, B(w_j))$ and weighted by the term inverse document frequency $idf(w)$ which reflects the word importance. To elaborate this step, we give an illustration as Figure 2. As we can see, the left part is the word "商店(store)", and the right part is some words from the bag of words of the word "商品(commodity)". We summed all up and normalized to avoid benefitting the bigger bag which contain larger number of words. In the other way, do the same process start from $B(w_j)$. Ultimately, the average of two scores is the contextual similarity.

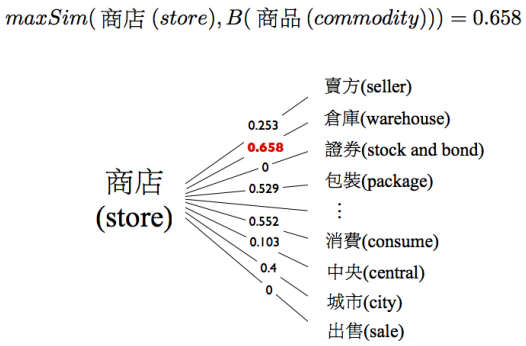


Figure 2: Computational procedure of $maxSim$(商店(store), B(商品(commodity)))

3.2 Word-to-Word Similarity

There are various measures for word-to-word semantic similarity in the literature. We chose two well-known corpus-based methods and two knowledge-based approaches for the comparison of semantic contexts explained in the previous subsection.

3.2.1 Corpus-based: Co-occurrence

We employ the concept of vector space models in information retrieval to define word similarity. The similarity between two words, w_i and w_j, is determined by the cosine value of their vectors, $v(w_i)$ and $v(w_j)$. The vector, $v(w_i)$, of a word, w_i, is defined by $(tf_{i1}, tf_{i2}, \ldots, tf_{i|D|})$, where a component tf_{ij} is defined as follows.

$$tf_{i,j} = \frac{occur(w_i, d_j)}{max\{occur(w, d_j) : w \in d_j\}}$$

$occur(w_i, d_j)$ represents the raw frequency of w_i in a document d_j in our corpus D. Term frequency is defined as the ratio of the frequency of w against the highest frequency of a certain word in the document. The cosine similarity of two vectors is defined below.

$$Sim_{TF}(w_i, w_j) = \frac{v_{TF}(w_i) \cdot v_{TF}(w_j)}{|v_{TF}(w_i)||v_{TF}(w_j)|}$$

The concept behind this is word co-occurrence. We assumed two words are more related if they appear in the same document more often.

The next approach is pointwise mutual information, a measure motivated by information theory and intended to reflect statistical dependence between two words in the corpus. Given two target words w_i and w_j, their similarity is estimated as

$$Sim_{PMI}(w_i, w_j) = \frac{\Pr(w_i \& w_j)}{\Pr(w_i) * \Pr(w_j)}$$

The main idea of PMI is that, if w_i and w_j are independent statistically, the probability that they occur in the same document, $\Pr(w_i \& w_j)$, will equal to $\Pr(w_i) * \Pr(w_j)$. We therefore get their similarity $Sim_{PMI}(w_i, w_j)$=0. In contrast, $\Pr(w_i \& w_j)$ will be greater than $\Pr(w_i) * \Pr(w_j)$ if w_i and w_j are some extent dependent. Here, we define $\Pr(w)$ as

$$\Pr(w) = \frac{|\{d : w \in d\}|}{|D|}$$

It means the probability of a randomly selected document containing the word w.

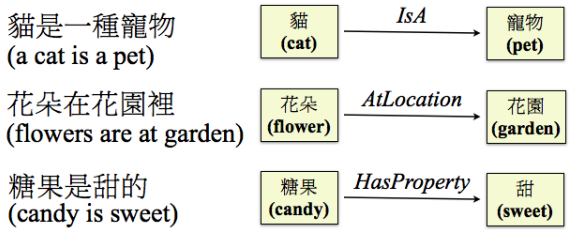


Figure 3: Some assertions and corresponding graph in Chinese ConceptNet

3.2.2 Knowledge-based: ConceptNet

In our knowledge-based metrics, we chose ConceptNet as knowledge database. ConceptNet is a commonsense knowledge base, and contains general human knowledge which are expressed in natural languages. We are using its Chinese version. ConceptNet is a large semantic graph in which nodes are concepts and edges are labeled with relations between the connected nodes. We call a pair of connected nodes an assertion. Figure 3 shows some assertions, e.g., in "candy is sweet", "candy" and "sweet" are concepts and "HasProperty" is their relation.

Compared with the other notable semantic knowledge base, WordNet, which is lexeme-based resource, ConceptNet excels at contextual concepts reasoning. This qualitative difference makes them suitable for different applications. In this paper, we explore ConceptNet as a knowledge source to capture contextual concepts at this moment and would get WordNet involved to improve our experiments in the future.

Based on ConceptNet, we explored two methods to estimate word similarity. First we considered the intersection of concepts related to the two target words as their similarity. We used the Jaccard index in this method simply. Second, since we thought both related concepts and their relations ought to be considered, we applied the AnalogySpace approach as well (Speer, Havasi, & Lieberman, 2008).

- Jaccard Index

  We assumed, intuitively, the more common concepts two words share, the more similar they are. We therefore applied Jaccard Index to ConceptNet. Given two target words w_i and w_j, their similarity score is simply defined as

  $$Sim_{Jaccard}(w_i, w_j) = \frac{|C(w_i) \cap C(w_j)|}{|C(w_i) \cup C(w_j)|}$$

  In this method, we ignored the relation type between two concepts

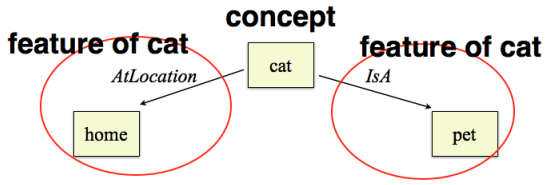


Figure 4: Example of feature for concept "cat"

Table 1: Part of AnalogySpace Matrix

	IsA (_pet)	*AtLocation (_home)*	*CapableOf (_fly)*	*PartOf (fur_)*
Cat	9	5	-3	6
Dog	12	3	-5	7
Airplain	-4	-3	7	0

- AnalogySpace

  AnalogySpace is a way of representing a concept in the knowledge base in a multidimensional feature vector. In ConceptNet, every node or concept is a word or a phrase. The feature of a concept is its neighbor and the relation type connecting them. For example, as shown in Figure 4, "is a pet" and "at home" are features of the concept "cat".

  The whole ConceptNet can be converted to an AnalogySpace matrix. The rows of AnalogySpace matrix are concepts; the columns are their features. The weight of an entry came from numbers of matched assertion or sentences collected. Table 1 is part of the AnalogySpace matrix. Since there are about 300 thousand concepts in total in Chinese ConceptNet, the scale of AnalogySpace matrix will be very large and sparse. We applied truncated singular value decomposition (truncated SVD) to get smaller and less noise matrix (Golub & Reinsch, 1970). A concept was then transformed to a smaller dimensional vector. Given the lowered dimensions matrix, we mapped word w_i and w_j directly to the concepts in ConceptNet and denote them as $v_{AS}(w_i)$ and $v_{AS}(w_j)$ here. We could simply compute the cosine similarity between two concepts and denote it as $Sim_{AnalogySpace}(w_i, w_j)$.

$$Sim_{AnalogySpace}(w_i, w_j) = \frac{v_{AS}(w_i) \cdot v_{AS}(w_j)}{|v_{AS}(w_i)||v_{AS}(w_j)|}$$

3.3 Clustering

We used centroid clustering of hierarchical clustering as our clustering method. Although in this paper the words should be clustered into three groups, in the future work, the number of clusters is uncertain. Thus, we used hierarchical instead of flat clustering method which require a prespecified number of clusters as input.

To begin with, we assign the similarity scores obtained from the measures just mentioned in previous subsections to form a symmetric similarity matrix. Next, we initialize the clusters by viewing each target word as a singleton cluster first. In the following iteration, two most similar clusters are merged into the same group. We run iteratively until the given number of clusters are formed (three clusters in this paper) and return the outcome.

In the standard centroid clustering (Manning, Raghavan, & Schütze, 2008), the similarity of two clusters is defined as the similarity of their centroids. In our application, centroid similarity is equivalent to average similarity of all pairs of words from different clusters. That is, in the main iteration of the algorithm, it computes the centroid similarity of the merged cluster of all pair of clusters. Next, the two clusters with maximum centroid similarity score are merged into the same cluster.

3.4 Example

To help comprehensively understand, we give an example as Figure 5. Given $family$(花/hua1/), we aim to differentiate their senses and cluster them into different groups. Note that for illustration, here we focus on only one of the four similarity measures $Sim_{AnalogySpace}$. In our experiments, we applied four approaches.

First, step (1), we extract all sentences containing target words and pack them to bags. Next, we computed pairwise similarities of contexts, i.e., bags of words, using equation (1) based on $Sim_{AnalogySpace}$ as seen in step (2). Through this calculation, we would acquire the matrix of pairwise similarity of these Chinese words. In the end, separate them to three clusters based on centroid clustering as seen at step (4).

4. Dataset

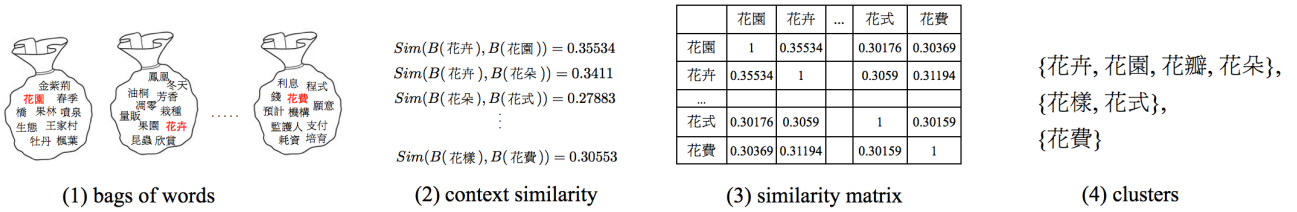
4.1 Corpus

We used traditional Chinese version of Wikipedia webpages crawled during August 2012 as our corpus. The original data format is HTML. For the current study, we removed the HTML tags (e.g., <a>,) and extracted the main textual content using the tool boilerpipe[4]. After this process, we obtained 444,838 valid plain text files. In order to employ the simplified Chinese version of Stanford Word Segmenter[5], we first made use of Open Chinese Convert[6] to convert traditional Chinese to simplified Chinese. Finally, we obtained 361,712,495 words. Among them, we found 187,265 types that appeared at least 30 times in the corpus (see Table 2).

4.2 Test Data

Our test data and ground truth were provided by Chia-Ying Lee, a psycholinguistics researcher of the Institute of Linguistics, Academia Sinica. We have ten morphologic families, and they include hundreds of candidate words for our experiments.

4 http://code.google.com/p/boilerpipe/
5 http://nlp.stanford.edu/software/segmenter.shtml
6 http://code.google.com/p/opencc/



	花園	花卉	...	花式	花費
花園	1	0.35534		0.30176	0.30369
花卉	0.35534	1		0.3059	0.31194
...					
花式	0.30176	0.3059		1	0.30159
花費	0.30369	0.31194		0.30159	1

Figure 5: A walk-through example $family$(花/hua1/) using $Sim_{AnalogySpace}$ word-to-word similarity

Table 2: Statistics of the corpus - traditional Chinese version Wikipedia

# of webpages	# of words	# of types
444,838	361,712,495	187,265

Table 3: Two of test examples: the morphological families of Chinese character "花/hua1/" and "商/shang1/"

Family	Rough meaning in English	Selected Words
花/hua1/	指植物 (things related to plants)	花卉, 花園, 花瓣, 花朵
	形容炫耀或技巧華麗的 (patterns or styles)	花式, 花樣
	消耗後述時間或能量 (expenditure or costs)	花費
商/shang1/	以營利爲目的的事業 (things related to commerce)	商業, 商品, 商店
	討論解決問題的辦法 (negotiation or discussion)	商量, 商議, 商討
	專有年代 (a Chinese dynasty)	商代, 商朝

We selected sense classes in common use and high frequency words. From each family, we chose top three most common classes of senses, i.e., the senses have more related words, and then selected up to four most frequent words from each class as our test data. As a result, each family contains 7 to 9 words sharing common Chinese character and total 84 words. We give two families for examples as below (see Table 3)

5. Experimental Result

5.1 Baseline

As a baseline for comparison, when calculating bag-to-bag similarity scores, we assigned a number between 0 and 1 from a uniform random distribution. The formula is defined as:

$$Sim_{random}(B(w_i), B(w_j)) = rand(0, 1)$$

where rand is a random function. Next, we applied centroid clustering in the same way.

5.2 Evaluation

In order to evaluate the performance of the proposed similarity scoring methods in unsupervised senses clustering, in this section, we will introduce the criteria of clustering quality we employed: Rand index and F-measure.

In the evaluation of clustering, recall and precision are two common criteria. Clustering can be viewed as a series of decisions. That is, during the process, a clustering algorithm decides to gather two words into the same cluster or not.

Thus, we can evaluate the performance by the Rand index which is the percentage of correct decision made by our algorithm. We denote true positives, true negatives, false positives, and false negatives by TP, TN, FP, and FN, respectively. Rand index is defined as

$$Rand\ index = \frac{TP + TN}{TP + TN + FP + FN}$$

We also used F-measure and it is defined as

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F = \frac{2 * Precision * Recall}{Precision + Recall}$$

5.3 Result

Given 10 morphological families as test data, we applied our approaches and evaluated their performances as described above. In order to estimate pairwise context similarities using equation (1), we used four scoring methods to measure word-to-word similarity: AnalogySpaec, Jaccard, TF, and PMI. The final performance of each method was obtained by averaging the scores of 10 families. Table 4 shows experimental results in the Rand index and F-measure scores.

As seen in table 4, it can be found that the best performance was achieved by AnalogySpace. Its performance in the Rand index and F-measure can be achieved at 68.49% and 56.87%, respectively. The Rand index of the baseline is 57.50% and F-measure is 31.44%. AnalogySpace outperforms the baseline 10.99% of Rand index and 25.43% of F-measure. The worst performances of our approaches are still better than those of the baseline. We will talk about more detail in next subsection soon.

Table 4: Performances of Random, TF, PMI, Jaccard, and AnalogySpace evaluated by the Rand index and F-measure

Method	Rand index	F-measure
Random	57.50%	31.44%
TF	63.93%	48.56%
PMI	65.32%	53.96%
Jaccard	63.93%	49.00%
AnalogySpace	**68.49%**	**56.87%**

5.4 Discussion

In addition to showing the performances, in this subsection, we are also trying to make a discussion, look into and analyze the test data.

5.4.1 Four word-to-word similarity measures and their performances

In terms of performance, the reason why AnalogySpace is prominent might be the inferences based on ConceptNet could catch the concepts and differentiate semantic context directly and easier. Besides, we can notice that PMI and AnalogySpace outperform both TF and Jaccard index methods. We think this is because both AnalogySpace and PMI measure take dependencies into account. Moreover, the weighting scheme of term frequency maybe too simple and Jaccard index ignored relation types in which contain a few latent semantic information, resulting bad performance.

5.4.2 Is "bag of words" a good way to catch the context?

The basic assumption we made is that character senses of morphological family words could be differentiated by considering their semantic contexts. Thus, we represented a context as a bag of words by collecting all the sentences where contain the target words from our corpus. In this way, we can catch semantic information nearby. The experimental result shows that currently we could discriminate their senses roughly.

However, owing to the window size we chose are too wide, taking the concept of "bag of words" as context may have bad effect. On one hand, even though after filtering out stop words, it would still contain too many noisy words to represent a context precisely. On the other hand, the more words the bag contains the more possible to gain high similarity score it is. It will benefit the words with high term frequency since they appear in more sentences.

Moreover, how to catch the word context is a critical problem. Bagging the segmented words into a set destroyed and ignored the structure of syntactic for sure. The structure contains very rich potential and useful information for our task. Therefore, in order to enhance our approach and improve the performance, we should try other approaches to capture the context. For example, considering the structure of syntactic grammar, parsing trees, and the relations between words in term of part-of-speech tags, etc.

5.4.3 Context distances between sense classes

According to our idea, if two words have similar semantic contexts, we grouped them into the same cluster, and assigned them into different clusters if dissimilar. So the performance is highly conntected to the "distances" between different sense classes. That is to say, for those morphological families whose distances between classed are larger, they are easier to cluster and the performances are better.

Table 5: Two examples to show the relation between extent of variation of senses and performance

Family	Ground Truth	Average Rand index	Average F-measure
花/hua1/	{花卉, 花園, 花瓣, 花朵}, {花式, 花樣}, {花費}	85.71%	75.00%
格/ge2/	{格律, 格言}, {格式, 格局, 格子}, {資格, 格調, 合格}	57.14%	42.93%

Our approach worked well especially when semantic contexts of sense classes are clearly different from other sense class contexts. Such as families containing a sense class of dynasty, e.g., the family("商/shang1/") and the family("明/ming2/"). Because their contexts are prominent from other senses classes, they will form a firm and inseparable group. In these cases, the performances will be leveraged up no matter using corpus-based or knowledge-based approaches. Moreover, as seen in table 5, the senses in the morphological family of Chines character "花/hua1/" can be easily discriminated by semantic context and has higher performance because their contexts are totally distinct. (For comparison, the performances in the table are average accuracies of four measures.) Yet the words in family("格/ge2/") are not far away enough to each other and difficult to differentiate. Actually, even most of our team members in this work had incorrect clustering results against ground truth provided by linguists of the family("格/ge2/").

Nevertheless, there still exist some vague situations. For example, words belonging to different sense classes may share similar or overlapping semantic context, or the same sense class words have unrelated context. These cases will lower the performance. Besides, we should notice that some senses classes are derived from another sense class, resulting in that the distances between classes are smaller and difficult to differentiate. Furthermore, the common usage of some words have been changed or extended by we human along with development of the language. For example, "光明(bright)" is talking about luminosity of something originally, but nowadays we use this word when describing somebody's future or life attitude as well. These cases will bring about different meanings or contexts from origin and reduce the effectiveness of our method in some way.

6. Summary and Future Work

In this work, our goal is to use natural language processing skill to differentiate senses of the same Chinese character embedded in different words and do clustering. Since it can

be found that most of words in the same class tend towards sharing more similar semantic contexts, we proposed a semantic context-aware approach to solve this problem first roughly. We tried to capture the context of a target word by the "bag of words" approach and compared the similarity between them. To estimate the similarity bag of words, we employed two corpus-based and two knowledge-based word-to-word similarity measures. If two bags are talking about related topic or having similar semantic context, we assigned them into the same cluster.

Consequently, we reached near 68.5% of Rand index using the knowledge-based approach based on ConceptNet for it could capture semantic concepts easily to certain degree. The performance shows we could do unsupervised sense clustering roughly. We found that our approach worked well especially in those morphological families whose distances between classes are larger, namely the characters have distinct senses.

However, not all character senses are easily to be differentiated due to their contexts are not such clearly different from others. Some words from different sense classes may have overlapping context. For those families that have bad performances, we contributed them to the following reasons. First, catching the semantic context by extracting all sentences brought about many noisy words. Second, using the concept of "bag of words" ignored structure of syntax which contains rich information. Moreover, some of word-to-word similarity measures we used maybe too naive to estimate the context similarity precisely.

In the future, we will try not only using "bag of words" concept to capture semantic context but also take other features widely used in natural language processing into account. Such as part-of-speech tagging, syntactic grammar, etc. Also, in addition to context level, we would explore lower levels. For example, in word level or go deep into Chinese character level. We even could decompose Chinese character into many components or radicals for those are thought to provide rich and valuable information related to sense of the character.

Back to the motivation of this work, we hope to use computing linguistic technique to help model the morphological awareness. By views of many linguists, this skill is regarded as key ability to Chinese reading achievement. Even though currently we are simply able to differentiate senses roughly, it' s not only a good start but also a fundamental and preliminary work for the future.

Acknowledgments

This work was supported in part by the grants NSC 101-2221-E-004-018, NSC 99-2410-H-001-041, NSC 99-2221-E-002-139-MY3, and NSC 101-2627-E-002-002 from the National Science Council, Taiwan.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional andWordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27).

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring Semantic Similarity between Words Using Web Search Engines. In *Proceedings of WWW* (Vol. 766).

Dai, L., Liu, B., Xia, Y., & Wu, S. (2008). Measuring Semantic Similarity between Words Using HowNet. In *Computer Science and Information Technology, ICCSIT'08.* (pp. 601–605).

Galmar, B. (2011). Using Kohonen Maps of Chinese Morphological Families to Visualize the Interplay of Morphology and Semantics in Chinese. In *ROCLING 2011 poster papers* (pp. 240–251).

Golub, G. H., & Reinsch, C. (1970). Singular Value Decomposition and Least Squares Solutions. *Numerische Mathematik*, *14*(5), 403–420.

Iosif, E., & Potamianos, A. (2010). Unsupervised Semantic Similarity Computation between Terms Using Web Documents. *Knowledge and Data Engineering, IEEE Transactions on*, *22*(11), 1637–1647.

Ku, Y.-M., & Anderson, R. C. (2003). Development of Morphological Awareness in Chinese and English. *Reading and Writing*, *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.

Liou, C.-Y., Cheng, C.-W., Liou, J.-W., & Liou, D.-R. (2013). Autoencoder for Polysemous Word. In *Intelligent Science and Intelligent Data Engineering* (pp. 458–465). Springer.

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.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (Vol. 1). Cambridge University Press Cambridge.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, p. 775).

Nakagawa, T. (2004). Chinese and Japanese Word Segmentation Using Word-level and Character-level Information. In *Proceedings of the 20th International Conference on Computational Linguistics* (p. 466).

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet:: Similarity: Measuring the Relatedness of Concepts. In *Demonstration Papers at HLT-NAACL 2004* (pp. 38–41).

Shu, H., McBride-Chang, C., Wu, S., & Liu, H. (2006). Understanding Chinese Developmental Dyslexia: Morphological Awareness as A Core Cognitive Construct.

Journal of Educational Psychology, *98*(1), 122.

Speer, R., Havasi, C., & Lieberman, H. (2008). AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. In *Proceedings of AAAI.*

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL.

Xue, N. (2003). Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, *8*(1), 29–48.