Combining a Chinese Thesaurus with a Chinese Dictionary

Ji Donghong
Kent Ridge Digital Labs
21 Heng Mui Keng Terrace
Singapore, 119613
dhji@krdl.org.sg

Gong Junping
Department of Computer Science
Ohio State University
Columbus, OH
jgong@cis.ohio-state.edu

Huang Changning
Department of Computer Science
Tsinghua University
Beijing, 100084, P. R. China
hcn@mail.tsinghua.edu.cn

Abstract

In this paper, we study the problem of combining a Chinese thesaurus with a Chinese dictionary by linking the word entries in the thesaurus with the word senses in the dictionary, and propose a similar word strategy to solve the problem. The method is based on the definitions given in the dictionary, but without any syntactic parsing or sense disambiguation on them at all. As a result, their combination makes the thesaurus specify the similarity between senses which accounts for the similarity between words, produces a kind of semantic classification of the senses defined in the dictionary, and provides reliable information about the lexical items on which the resources don't conform with each other.

1. Introduction

Both ⟨⟨TongYiCi CiLin⟩⟩ (Mei. et al, 1983) and ⟨⟨XianDai HanYu CiDian⟩⟩ (1978) are important Chinese resources, and have been widely used in various Chinese processing systems (e.g., Zhang et al, 1995). As a thesaurus, ⟨⟨TongYiCi CiLin⟩⟩ defines semantic categories for words, however, it doesn't specify which sense of a polysemous word is involved in a semantic category. On the other hand, ⟨⟨XianDai HanYu CiDian⟩⟩ is an ordinary dictionary which provides definitions of senses while not giving any information about their semantic classification.

A manual effort has been made to build a resource for English, i.e., WordNet, which contains both definition and classification information (Miller et al., 1990), but such resources are not available for many other languages, e.g. Chinese. This paper presents an automatic method to combine the Chinese thesaurus with the Chinese dictionary into such a resource, by tagging the entries in the thesaurus with appropriate senses in the dictionary, meanwhile assigning appropriate semantic codes, which stand for semantic categories in the thesaurus, to the senses in the dictionary.

D.Yarowsky has considered a similar problem to link Roget's categories, an English thesaurus, with the senses in COBUILD, an English dictionary (Yarowsky, 1992). He treats the problem as a sense disambiguation one, with the definitions in the dictionary taken as a kind of *contexts* in which the headwords occur, and deals with it based on a statistical model of Roget's categories trained on large corpus. In our opinion, the method, for a specific word, neglects the difference between its definitions and the ordinary contexts: definitions generally contain its synonyms, hyponyms or hypernyms, etc., while ordinary contexts generally its collocations. So the trained model on ordinary contexts may be not appropriate for the disambiguation problem in definition contexts.

A seemingly reasonable method to the problem would be common word strategy, which has been extensively studied by many researchers (e.g., Knight, 1993; Lesk, 1986). The solution

would be, for a category, to select those senses whose definitions hold most number of common words among all those for its member words. But the words in a category in the Chinese thesaurus may be not similar in a strict way, although similar to some extend, so their definitions may only contain some similar words at most, rather than share many words. As a result, the common word strategy may be not appropriate for the problem we study here.

In this paper, we extend the idea of common word strategy further to a similar word method based on the intuition that definitions for similar senses generally contain similar words, if not the same ones. Now that the words in a category in the thesaurus are similar to some extent, some of their definitions should contain similar words. We see these words as *marks* of the category, then the correct sense of a word *involved* in the category could be identified by checking whether its definition contains such *marks*. So the key of the method is to determine the *marks* for a category. Since the *marks* may be different word tokens, it may be difficult to make them out only based on their frequencies. But since they are similar words, they would belong to the same category in the thesaurus, or hold the same semantic code, so we can locate them by checking their semantic codes.

In implementation, for any category, we first compute a *salience* value for each code with respect to it, which in fact provides the information about the *marks* of the category, then compute *distances* between the category and the senses of its member words, which reflect whether their definitions contain the *marks* and how many, finally select those senses as tags by checking whether their *distances* from the category fall within a threshold.

The remainder of this paper is organized as the following: in section 2, we give a formal setting of the problem and present the tagging procedure; in section 3, we explore the issue of threshold estimation for the distances between senses and categories based on an analysis of the distances between the senses and categories of univocal words; in section 4, we report our experiment results and their evaluation; in section 5, we present some discussions about our methodology; finally in section 6, we give some conclusions.

2. Problem Setting

The Chinese dictionary provides sense distinctions for 44,389 Chinese words, on the other hand, the Chinese thesaurus divides 64,500 word entries into 12 major, 94 medium and 1428 minor categories, which is in fact a kind of semantic classification of the words[1]. Intuitively, there should be a kind of correspondence between the senses and the entries. The main task of combining the two resources is to locate such kind of correspondence.

Suppose X is a category[2] in the thesaurus, for any word $w \in X$, let S_w be the set of its senses in the dictionary, and $S_X = \bigcup_{w \in X} S_w$, for any $s \in S_X$, let DW_s be the set of the definition words in its definition, $DW_w = \bigcup_{s \in S_w} DW_s$, and $DW_X = \bigcup_{w \in X} DW_w$, for any word w, let $CODE(w)$ be the set of its semantic codes that are given in the thesaurus[3], $CODE_s = \bigcup_{w \in DW_s} CODE(w)$, $CODE_w = \bigcup_{s \in S_w} CODE_s$, and $CODE_X = \bigcup_{s \in S_X} CODE_s$. For any $c \in CODE_X$, we

[1] The electronic versions of the two resources we use now only contain part of the words in them, see section 4.

[2] We generally use "category" to refer to minor categories in the following text, if no confusion is involved. Furthermore, we also use a semantic code to refer to a category.

[3] A category is given a semantic code, a word may belong to several categories, and hold several codes.

define its *definition salience* with respect to X in 1).

1) $Sal_1(c, X)=\frac{|\{w|w \in X, c \in CODE_w\}|}{|X|}$

For example, 2) lists a category Ea02 in the thesaurus, whose members are the synonyms or antonyms of word 高大(/gaoda/; high and big)[4].

2) 矮墩墩 矮小 参天 崔嵬 高大 高耸 岿然 魁岸 魁伟 魁梧 凌云 摩天 亭亭 突兀 突兀 万丈 巍然 巍巍 五短身材 细高挑儿 修长 屹立 屹然 崭然 颀长...

3) lists some semantic codes and their *definition salience* with respect to the category.

3) Ea02 (0.92), Ea03 (0.76), Dn01 (0.45), Eb04 (0.24), Dn04 (0.14).

To define a *distance* between a category X and a sense s, we first define a *distance* between any two categories according to the distribution of their member words in a corpus, which consists of 80 million Chinese characters.

For any category X, suppose its members are $w_1, w_2, ..., w_n$, for any w_i, we first compute its mutual information with each semantic code according to their co-occurrence in a corpus[5], then select 10 top semantic codes as its *environmental codes*[6], which hold the biggest mutual information with w_i. Let NC_i be the set of w_i's *environmental codes*, C_T be the set of all the semantic codes given in the thesaurus, for any $c \in C_T$, we define its *context salience* with respect to X in 4).

4) $Sal_2(c, X)=\frac{|\{w_i|c \in NC_i\}|}{n}$

We build a *context vector* for X in 5), where $k=|C_T|$.

5) $cv_X=<Sal_2(c_1, X), Sal_2(c_2, X), ..., Sal_2(c_k, X)>$

Given two categories X and Y, suppose cv_X and cv_Y are their *context vectors* respectively, we define their *distance* $dis(X, Y)$ as 6) based on the cosine of the two vectors.

6) $dis(X, Y)=1-cos(cv_X, cv_Y)$

Let $c \in CODE_X$, we define a *distance* between c and a sense s in 7).

7) $dis(c, s)=\underset{c' \in CODE_s}{Min} dis(c, c')$

Now we define a *distance* between a category X and a sense s in 8).

8) $dis(X, s)=\sum_{c \in CODE_X}(h_c \bullet dis(c, s))$

where $h_c=\frac{Sal_1(c, X)}{\sum_{c' \in CODE_X} Sal_1(c', X)}$.

Intuitively, if $CODE_s$ contains the salient codes with respect to X, i.e., those with higher *salience* with respect to X, $dis(X, s)$ will be smaller due to the fact that the contribution of a semantic code to the *distance* increases with its *salience*, so s tends to be a correct sense tag of some word.

For any category X, let $w \in X$ and $s \in S_w$, if $dis(X, s) \leq T$, where T is some threshold, we will tag w by s, and assign the semantic code X to s.

3. Parameter Estimation

Now we consider the problem of estimating an appropriate threshold for $dis(X, s)$ to distinguish between the senses of the words in X. To do so, we first extract the words which hold only one code in the thesaurus, and have only one sense in the dictionary[7], then check the *distances* between these senses and categories. The number of such words is 22,028.

[4] "/gaoda/" is the Pinyin of the word, and "high and big" is its English translation.

[5] We see each occurrence of a word in the corpus as one occurrence of its codes. Each co-occurrence of a word and a code falls within a 5-word distance.

[6] The intuition behind the parameter selection (10) is that the words which can combined with a specific word to form collocations fall in at most 10 categories in the thesaurus.

[7] This means that the words are regarded as univocal ones by both resources.

Tab.1 lists the distribution of the words with respect to the *distance* in 5 intervals.

Intervals	Word num.	Percent(%)
[0.0, 0.2)	8,274	37.56
[0.2, 0.4)	10,655	48.37
[0.4, 0.6)	339	1.54
[0.6, 0.8)	1172	5.32
[0.8, 1.0]	1588	7.21
all	22,028	100

Tab. 1. The distribution of univocal words with respect to $dis(X, s)$

From Tab.1, we can see that for most univocal words, the *distance* between their senses and categories lies in [0, 0.4].

Let W_U be the set of the univocal words we consider here, for any univocal word $w \in W_U$, let s_w be its unique sense, and X_w be its univocal category, we call $DEN_{<t1, t2>}$ *point density* in interval $[t_1, t_2]$ as 9), where $0 \le t_1 < t_2 \le 1$.

9) $$DEN_{<t1, t2>} = \frac{\left|\{w | w \in W_U, t_1 \le dis(X_w, s_w) \le t_2\}\right|}{t_2 - t_1}$$

We define 10) as an *object function*, and take t^* which maximizes DEN_t as the threshold.

10) $DEN_t = DEN_{<0, t>} - DEN_{<t, 1>}$

The object function is built on the following inference. About the explanation of the words which are regarded as univocal by both Chinese resources, the two resources tend to be in accordance with each other. It means that for most univocal words, their senses should be the correct tags of their entries, or the distance between their categories and senses should be smaller, falling within the under-specified threshold. So it is reasonable to suppose that the intervals within the threshold hold a higher point density, furthermore that the difference between the point density in [0, t^*], and that in [t^*, 1] gets the biggest value.

With t falling in its value set $\{dis(X, s)\}$, we get t^* as 0.384, when for 18,653 (84.68%) univocal words, their unique entries are tagged with their unique senses, and for the other univocal words, their entries not tagged with their senses.

4. Results and Evaluation

There are altogether 29,679 words shared by the two resources, which hold 35,193 entries in the thesaurus and 36,426 senses in the dictionary. We now consider the 13,165 entries and 14,398 senses which are irrelevant with the 22,028 univocal words. Tab. 2 and 3 list the distribution of the entries with respect to the number of their sense tags, and the distribution of the senses with respect to the number of their code tags respectively.

Tag num.	Entry	Percent (%)
0	1625	12.34
1	9908	75.26
2	1349	10.25
≥3	283	2.15

Tab. 2. The distribution of entries with respect to their sense tags

Tag num.	Sense	Percent (%)
0	1461	10.15
1	10433	72.46
2	2334	16.21
≥3	170	1.18

Tab. 3. The distribution of senses with respect to their code tags

In order to evaluate the efficiency of our method, we define two measures, *accuracy rate* and *loss rate*, for a group of entries E as 11) and 12) respectively[8].

[8] We only give the evaluation on the results for entries, the evaluation on the results for senses can be done similarly.

11) $$\frac{|RT_E \cap CT_E|}{|RT_E|}$$

12) $$\frac{|CT_E - (RT_E \cap CT_E)|}{|CT_E|}$$

where RT_E is a set of the sense tags for the entries in E produced by the tagging procedure, and CT_E is a set of the sense tags for the entries in E, which are regarded as correct ones somehow.

What we expect for the tagging procedure is to select the appropriate sense tags for the entries in the thesaurus, if they really exist in the dictionary. To evaluate the procedure directly proves to be difficult. We turn to deal with it in an indirect way, in particular, we explore the efficiency of the procedure of tagging the entries, when their appropriate sense tags don't exist in the dictionary. This indirect evaluation, on the one hand, can be carried out automatically in a large scale, on the other hand, can suggest what the direct evaluation entails in some way because that none appropriate tags can be seen as a *special tag* for the entries, say *None*[9].

In the first experiment, let's consider the 18,653 univocal words again which are selected in parameter estimation stage. For each of them, we create a new entry in the thesaurus which is different from its original one. Based on the analysis in section 3, the senses for theses words should only be the correct tags for their corresponding entries, the newly created ones have to take *None* as their correct tags.

When creating new entries, we adopt the following 3 different kinds of constraints:

i) the new entry belongs to the same medium category with the original one;

ii) the new entry belongs to the same major category with the original one;

iii) no constraints;

With each constraint, we select 5 groups of new entries respectively, and carry out the experiment for each group. Tab. 4 lists average accuracy rates and loss rates under different constraints.

Constraint	Aver. accuracy(%)	Aver. loss (%)
i)	88.39	11.61
ii)	94.75	5.25
iii)	95.26	4.74

Tab. 4. Average accuracy, loss rates under different constraints

From Tab. 4, we can see that the accuracy rate under constraint i) is a bit less than that under constraint ii) or iii), the reason is that with the created new entries belonging to the same medium category with the original ones, it may be a bit more likely for them to be tagged with the original senses. On the other hand, notice that the accuracy rates and loss rates in Tab.4 are complementary with each other, the reason is that $|RT_E|$ equals $|CT_E|$ in such cases.

In another experiment, we select 5 groups of 0-tag, 1-tag and 2-tag entries respectively, and each group consists of 20~30 entries. We check their accuracy rates and loss rates manually. Tab. 5 lists the results.

Tag num.	Aver. accuracy(%)	Aver. loss(%)
0	94.6	7.3
1	90.1	5.2
2	87.6	2.1

Tab. 5. Average accuracy and loss rates under different number of tags

Notice that the accuracy rates and loss rates in Tab.5 are not complementary, the reason is that $|RT_E|$ doesn't equal $|CT_E|$ in such cases.

In order to explore the main factors affecting accuracy and loss rates, we extract the entries which are not correctly tagged with the senses, and check relevant definitions and semantic codes.

[9] A default sense tag for the entries.

The main reasons are:

i) No salient codes exist with respect to a category, or the determined are not the expected. This may be attributed to the fact that the words in a category may be not strict synonyms, or that a category may contain too less words, etc.

ii) The information provided for a word by the resources may be incomplete. For example, word 全数(/quanshu/, all) holds one semantic code Ka06 in the thesaurus, its definition in the dictionary is:

全数: 全部[Eb02]
/quanshu/ /quanbu/
all

The correct tag for the entry should be the sense listed above, but in fact, it is tagged with *None* in the experiment. The reason is that word 全部 (/quanbu/, all) can be an adverb or an adjective, and should hold two semantic codes, Ka06 and Eb02, corresponding with its adverb and adjective usage respectively, but the thesaurus neglects its adverb usage. If Ka06 is added as a semantic code of word 全部 (/quanbu/, all), the entry will be successfully tagged with the expected sense.

iii) The distance defined between a sense and a category fails to capture the information carried by the order of salient codes, more generally, the information carried by syntactic structures involved. As an example, consider word 谣传 (/yaochuan/), which has two definitions listed in the following.

谣传 1) 谣言[Da19] 传播[Ie01].
/yaochuan/ /yaoyan/ /chuanbo/
hearsay spread
the hearsay spreads.
2) 传播[Ie01] 的 谣言[Da19]
/chuanbo/ /de/ /yaoyan/
spread of hearsay
the hearsay which spreads

The two definitions contain the same content words, the difference between them lies in the order of the content words, more generally, lies in the syntactic structures involved in the definitions: the former presents a sub-obj structure, while the latter with a "的(/de/,of)" structure. To distinguish such definitions needs to give more consideration on word order or syntactic structures.

5. Discussions

In the tagging procedure, we don't try to carry out any sense disambiguation on definitions due to its known difficulty. Undoubtedly, when the noisy semantic codes taken by some definition words exactly cover the salient ones of a category, they will affect the tagging accuracy. But the probability for such cases may be lower, especially when more than one salient code exists with respect to a category.

The distance between two categories is defined according to the distribution of their member words in a corpus. A natural alternative is based on the shortest path from one category to another in the thesaurus (e.g., Lee at al., 1993; Rada et al., 1989), but it is known that the method suffers from the problem of neglecting the wide variability in what a link in the thesaurus entails. Another choice may be *information content* method (Resnik, 1995), although it can avoid the difficulty faced by shortest path methods, it will make the minor categories within a medium one get a same distance between each other, because the distance is defined in terms of the information content carried by the medium category. What we concern here is to evaluate the dissimilarity between different categories, including those within one medium category, so we make use of semantic code based vectors to define their dissimilarity, which is motivated by Shuetze's word frequency based vectors (Shuetze, 1993).

In order to determine appropriate sense tags

for a word entry in one category, we estimate a threshold for the distance between a sense and a category. Another natural choice may be to select the sense holding the smallest distance from the category as the correct tag for the entry. But this choice, although avoiding estimation issues, will fail to directly demonstrate the inconsistency between the two resources, and the similarity between two senses with respect to a category.

6. Conclusions

In this paper, we propose an automatic method to combine a Chinese thesaurus with a Chinese dictionary. Their combination establishes the correspondence between the entries in the thesaurus and the senses in the dictionary, and provides reliable information about the lexical items on which the two resources are not in accordance with each other. The method uses no language-specific knowledge, and can be applied to other languages.

The combination of the two resources can be seen as improvement on both of them. On the one hand, it makes the thesaurus specify the similarity between word senses behind that between words, on the other hand, it produces a semantic classification for the word senses in the dictionary.

The method is in fact appropriate for a more general problem: given a set of similar words, how to identify the senses, among all, which account for their similarity. In the problem we consider here, the words fall within a category in the Chinese thesaurus, with similarity to some extent between each other. The work suggests that if the set contains more words, and they are more similar with each other, the result will be more sound.

References

Knight K. (1993) *Building a Large Ontology for Machine Translation.* In "Proceedings of DARPA Human Language Conference", Princeton, USA, 185-190.

Lesk M. (1986) *Automated Word Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone.* In "Proceedings of the ACM SIGDOC Conference", Toronto Ontario.

Lee J. H., Kim M. H., and Lee Y. J. (1993). *Information retrieval based on concept distance in IS-A hierarchies.* Journal of Documentation, 49/2.

Mei J.J. et al. (1983) *TongYiCi CiLin(A Chinese Thesaurus).* Shanghai Cishu press, Shanghai.

Miller G.A., Backwith R., Fellbaum C., Gross D. and Miller K. J. (1990) *Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography*, 3(4) (Special Issue).

Rada R. and Bicknell E (1989) *Ranking documents with a thesaurus.* JASIS, 40(5), pp. 304-310.

Resnik P. (1995) *Using Information Content to Evaluate the similarity in a Taxonomy.* In "Proceedings of the 14th International Joint Conference on Artificial Intelligence".

Schutze H. (1993) *Part-of-speech induction from scratch.* In "Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics", Columbus, OH.

XianDai HanYu CiDian(A modern Chinese Dictionary) (1978), Shangwu press, Beijing.

Yarowsky D. (1992) *Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora.* In "Proceedings of COLING'92", Nantas, France, pp. 454-460.

Zhang J, Huang C. N. Yang E. H. (1994) *Construction a Machine Tractable Dictionary from a Machine Readable Dictionary.* Communications of Chinese and Oriental Language Information Processing Society, 4(2), pp. 123-130.