The Research of Word Sense Disambiguation Method Based on Co-occurrence Frequency of Hownet†

Erhong Yang, Guoqing Zhang, and Yongkui Zhang
Dept of Computer Science, Shanxi University,
TaiYuan 030006, P. R. China
Email: zyk@sxu.edu.cn

Abstract

Word sense disambiguation (WSD) is a difficult problem in natural language processing. In this paper, a sememe co-occurrence frequency based WSD method was introduced. In this method, Hownet was used as our information source , and a co-occurrence frequency database of sememes was constructed and then used for WSD. The experimental result showed that this method is successful.

Keywords

word sense disambiguation, Hownet, sememe, co-occurrence

1. Introduction

Word sense disambiguation (WSD) is one of the most difficult problems in NLP. It is helpful and in some instances required for such applications as machine translation, information retrieval, content and thematic analysis, hypertext navigation and so on. The problem of WSD was first put forward in 1949. And then in the following decades researchers adopted many methods to solve the problem of automatic word sense disambiguation, including:1) AI-based method, 2) knowledge-based method and 3) corpus-based method.[1] Although some useful results have been got, the problem of word sense disambiguation is far from being solved.

The difficult of WSD is as follow: 1) Evaluation of word sense disambiguation systems is not yet standardized. 2) The potential for WSD varies by task. 3) Adequately large sense-tagged data sets are difficult to obtain. 4) The field has narrowed down approaches, but only a little. [2]

In this paper, we use a statistical based method to solve the problem of automatic word sense disambiguation.[3] In this method, a new knowledge base------Hownet[4,5] was use as knowledge resources. And instead of words, the sememes which are defined in Hownet were used to get the statistical figure. By doing this, the problem of data sparseness was solved to a large degree.

2. A Brief Introduction Of Hownet

Hownet is a knowledge base which was released recently on Internet. In Hownet, the concept which were represented by Chinese or English words were described and the relations between concepts and the attributes of concepts were revealed. In this paper, we use Chinese knowledge base, which is an important part of Hownet, as the resource of our disambiguation. The format of this file is as follow:

W_X =word
E_X = some examples of this word
G_X= the pos of this word
DEF= the definition of this word

† This research project is supported by a grant from Shanxi Natural Science Foundation of China

A important concept used in Hownet that we must introduce is sememe. In Hownet, sememes refer to some basic unit of senses. They are used to describe all the entries in Hownet and there are more than 1,500 sememe all together.

3. Sense Co-occurrence Frequency Database

It is well known that some words tend to co-occur frequently with some words than with others[6]. Similarly, some meaning of words tend to co-occur more often with some meaning of words than with others. If we can got the relations of word meanings quantitatively, it would have some help on word sense disambiguation. In Hownet, all words are defined with limited sememes and the combination of sememes is fixed. If we make statistic on the co-occurrence frequency of sememe so as to reflect the co-occurrence of words, the problem of data sparseness would be solved to a large degree. Based on the above thought, we built a sense co-occurrence frequency database to disambiguate word senses.

3.1 The Preprocessing Of Hownet

The Hownet we downloaded from Internet is in the form of plain text. It is not convenient for computer to use and it must been converted into a database. In the database, each lexical entry is converted into a record. The formalization description of the records is as follow:

<lexical entry> ::= <NO.><morphology> <part-of-speech><definition>

Where NO. is the corresponding number of this lexical entry in Hownet. And the definition is composed of several sememes (short for SU) which were divided by comma. In addition, we have deleted the English sememes in order to saving space and speeding up the processing. Here are some examples after preprocessing:

NO.	Morphology	Part-of-speech	definition
21424	俭朴	ADJ	属性值,举止,俭,良
18888	坏	ADJ	属性值,好坏,坏,莠
18889	坏	V	损害
18887	坏	V	坏掉
18890	坏	N	念头,恶

3.2 The Creation Of Sememe Co-occurrence Frequency Database

The sememe co-occurrence frequency database is the basic of sense disambiguation. Now we will introduce it briefly.

The sememe co-occurrence frequency database is a table of two dimension. Each item corresponding to the co-occurrence frequency of a pair of sememes.

Before introducing the sememe co-occurrence frequency database, we gave the following definition:

Definiton: *suppose word W has m sense items in hownet, and the corresponding definition of each sense item is:* $y_{11}, y_{12}, \ldots, y_{1(n1)}$; $y_{21}, y_{22}, \ldots, y_{2(n2)}$; …; $y_{m1}, y_{m2}, \ldots, y_{m(nm)}$ *respectively. We call* $\{y_{i1}, y_{i2}, \ldots, y_{i(ni)}\}$ *a sememe set of W(short for SS), and call* $\{\{ y_{11}, y_{12}, \ldots, y_{1(n1)}\}, \{ y_{21}, y_{22}, \ldots, y_{2(n2)}\}, \ldots, \{ y_{m1}, y_{m2}, \ldots, y_{m(nm)}\}\}$ *the sememe expansion of W (short for SE).*

For example, in the above mentioned example, the word “俭朴” has only one sense item. The corresponding sememe set of this

sense item is 〈属性值,举止,俭,良〉 and the sememe expansion of "俭朴" is 〈〈属性值,举止,俭,良〉〉. The word "坏" has four sense items, and the corresponding sememe set of each item is〈属性值,好坏,坏,莠〉,〈损害〉,〈坏掉〉 and 〈念头,恶〉 respectively. The sememe expansion of word "坏" is 〈〈属性值,好坏,坏,莠〉,〈损害〉,〈坏掉〉,〈念头,恶〉〉。

When building the sememe co-occurrence frequency database, the corpus is segmented first and each word is tagged with its sememe expansion in Hownet. Then for each unique pair of words co-occurred in a sentence (here a sentence is a string of characters delimited by punctuations.), the co-occurrence data of sememes which belong to the definition of each words respectively were collect. When collecting co0occurrence data, we adopt a principle that every pair of word which co-occurred in a sentence should have equal contribution to the sememe co-occurrence data regardless of the number of sense items of this word and the length of the definition. Moreover, the contribution of a word should be evenly distributed between all the senses of a word and the contribution of a sense should been evenly distributed between all the sememe in a sense. The algorithm is as follow:

1.Initial each cell in the sememe co-occurrence frequency database(short for SCFD) with 0.

2.For each sentence S in training corpus, do 3-7.

3.For each word in sentense S, tag the sememe expansion to it.

4.For each unique pair of sememe expansion (SE_i, SE_j), do 5-7.

5.For each sememe SU_{imp} in each sememe set SS_{im} in SE_i, do 6-7.

6.For each sememe SU_{jnq} in each sememe set SS_{jn} in SE_j, do 7.

7.Increase the value of cell SCFD(SU_{imp}, SU_{jnq}) and SCFD(SU_{jnq},SU_{imp}) by the product of w(SU_{imp}) and w(SU_{jnq}). Where w(SU_{xyz}) is weight of SU_{xyz} given by

$$W(SU_{xyz}) = \frac{1}{|SE_x| \times |SS_{xy}|}$$

It can be concluded from the above algorithm that the SCFD are symmetrical. In order to saving space and speeding up the processing, we only save those cells （SU_i,SU_j） that satisfying $SU_i \leqslant SU_j$.

3.3 The Sememe Co-occurrence Frequency Database Based Disambiguation Method

3.3.1 The Sememe Co-occurrence Frequency Based Scoring Method

When disambiguate a polysemous word, we given the following equation as the score of a sense item of the polysemous word and the context containing this polysemous word. The context of the word is the sentence containing this word.

$$score(S,C) = score(SS,C') - score(SS,GlobalSS) \quad (1)$$

Where S is a sense item of polysemouse word W, C is the context containing W, SS is the corresponding sememe set of S, C' is the set of sememe expansion of words in C and GlobalSS is the sememe set that containing all of the sememe defined in Hownet.

$$score(SS,C') = \sum_{\forall SE \in C'} score(SS,SE') \Big/ |C'| \quad (2)$$

for any sememe set SS and sememe expansion set C'.

$$score(SS,SE') = \max_{SS' \in SE'} score(SS,SS') \quad (3)$$

for any sememe set SS and sememe expansion SE'.

$$score(SS,SS') = \sum_{\forall SU' \in SS'} score(SS,SU') \Big/ |SS'| \quad (4)$$

for any sememe set SS and SS'.

$$score(SS, SU') = \sum_{\forall SU \in SS} score(SU, SU') \Big/ |SS| \quad (5)$$

for any sememe set SS and sememe SU'.

$$score(SU, SU') = I(SU, SU') \quad (6)$$

for any sememe SU and SU'.

$$I(SU, SU') = \log_2 \frac{f(SU, SU') \cdot N^2}{g(SU) \cdot g(SU')} \quad (7)$$

Where f(SU,SU') is the co-occurrence frequency corresponding to sememe pair （SU，SU'） in SCFD. And for g(SU) and N, we have the following equation:

$$g(SU) = \sum_{\forall SU'} f(SU, SU') \quad (8)$$

$$N = \sum_{\forall SU, \forall SU'} f(SU, SU') \Big/ 2 \quad (9)$$

In equation (7), the mutual-information-like measure deviated from the stardard mutual-information measure by multiple a extra multiplicative factor N, this is because that the scale of the corpus is not large enough that the mutual-information of some sememes pairs would be negtive if it was not normalized by a extra multiplicative factor N. In equation (9), the sum of f(SU,SU') was divided by 2, this is because for each pair of sememes, $\sum_{\forall SU, \forall SU'} f(SU, SU')$ is increase by 2.

When disambiguation, we tag the sememe T that satisfying the following equation to polysemous word W.

$$T = \arg\max_{S} score(S, C) \quad (10)$$

3.3.2 The Creation Of Mutual Information Database

We have created a mutual information database according to (7),(8) and(9) Here is some examples:

The examples in table 1 have a high mutual information. The sememe pairs in this table have certain semantic relations. While the examples in table 2 have a low mutual information. And the sememe pairs in this table have no patency semantic relations.

Table 1: example of sememe pairs which have a high mutual information

Sememe 1	Sememe 2	Mutual-Information	Sememe 1	Sememe 2	Mutual-Information
赌博	寻欢	33.811057	表情	羞愧	27.418417
鼓吹	夸大	29.441937	昏迷	醒	27.234630
光洁度	摸	28.024560	味道	香	27.093292
跑	气喘	28.023521	慢待	漠	26.984521
使净	整理	27.571478	低植	蔬菜	26.710478

Table 2: example of sememe pairs which have a low mutual information

Sememe 1	Sememe 2	Mutual-Information	Sememe 1	Sememe 2	Mutual-Information
食品	政	8.693242	合作	末	9.171023
交往	医	8.754611	侧	液	9.357734
车	圆	8.793914	驱赶	正误	9.448947
合作	疾病	9.121846	程度	交换	9.528801
机构	疾病	9.150412	禽	主次	9.599495

It can been concluded from table 1 and table 2 that the mutual information can reflect

the tightness of semantic relations.

4. Experiment And Analysis

We did the experiment on a corpus of 10,000 characters from People's Dialy.

Firstly, the corpus is segmented, and then the sememe co-occurrence frequecny database and mutual information database is created. In the mutual-information database, there is 709,496 data items corresponding to different sememes pairs. In order to speeding up the processing, the mutual-information database was sorted and indexed according to the first two bytes of each sememe pair. At last the experiment of disambiguation of some polysemous words was done. Here is two examples:

Example 1：全|省|两万四千|多|名|党政|干部|累计|处理|信访|案|十万|余|件|。

Example 2：这|是|香港|海关|今年|破获|的|第|一|宗|来自|内地|的|文物|走私|案。

We use the following euqation to access the accuracy ratio of disambiguation:

$$accuracy\quad ratio = \frac{the\ number\ of\ correctly\ tagged\ examples}{the\ total\ number\ of\ examples\ in\ testing\ set} \tag{11}$$

the experimental result is shown in table 4.

Table 3: Two examples that disambiguate using sememe co-occurrence frequency database

The definition of word "案"	The score of sense items and the context of word "案" in example 1	The score of sense items and the context of word "案" in example 2
文书	14.459068	8.659968
事情	9.817648	10.817648
事情 警	7.415986	12.415986
家具 放置	-0.134779	-0.134779
语文 提出 商讨 辩论	-0.818518	-0.818518
最大同现频率	14.459068	12.415986
排歧结果	文书	事情 警

Table 4: the experiment result

	Total number of testing examples	The number of correctly tagged examples	Accurracy ratio
Close test	100	75	75%
Open test	100	71	71%

The disambiguation method introduced above have the following charatristics:

(1) The problem of data spraseness is solved in a large degree.

(2) This disambiguation method avoids the laborious hand tagging of training corpus.

(3) This method can been easily applied to other kind of corpus.

Reference

[1]. Nancy Ide, Jean Veronis, Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, Computational Linguistics, 1998, Volume 24, number 1, pp 1-40

[2]. Philip Resnik, David Yarowsky, A Perspective on Word Sense Disambiguation Methods and their

Evaluation, http://www.cs.jhu.edu/~yarowsky/pubs.html

[3]. Alpha K. Luk, Statistical Sense Disambiguation with Relatively Small Corpus Using Dictionary Definitions, 33rd Annual Meeting of the Association for Computational Linguistics,26-30 June, 1995, Massachusetts Institute of Technology, Cambridge, Massachusetts. USA, pp.181-188

[4]. 董振东, 语义关系的表达和知识系统的建造, 语言文字应用, 1998 年第 3 期, 总第 27 期, pp.76-82

[5]. 董振东, 知网，http://www.how-net.com.

[6]. Kenneth Ward Church, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, 1990,Volume 16, Number 1, pp.22-29