Language, Information and Computation(PACLIC 11), 1996, 177-186

Finding a deficiency of a meaning in a Bunrui-goi-hyou entry by using corpora

Hiroyuki Shinnou
Ibaraki University Dept. of Systems Engineering
shinnou@lily.dse.ibaraki.ac.jp

Abstract

This paper presents a method to automatically find meanings which should be but are not entered in thesaurus. In this paper, we use Bunrui-goi-hyou as the thesaurus. To find the noun n with the meaning lacking in Bunrui-goi-hyou, we applies our presented method which extracts idioms from a corpus. We use the clue that many idioms with the noun n are extracted through that method. We have experimented with a corpus which consists of 5 years' worth of articles from a Japanese economic newspaper. As a result, we found 177 types of lacking meanings. Furthermore, our method can find not only a deficiency of a meaning but also meanings which are not used or are specifically used in the corpus domain.

1 Introduction

Bunrui-goi-hyou (The National Language Research Institute 1994) represents the Japanese thesaurus, and is used in many researches. Needless to say, it is important to expand and refine Bunrui-goi-hyou for the sake of Japanese natural language processing. This paper presents the method to automatically find meanings which must be entered in Bunrui-goi-hyou but not be entered. By this method, we can efficiently fill the deficiency of meanings in Bunrui-goi-hyou.

We should notice that our method automatically executes the following:

(1) to estimate that a certain noun n has the meaning g which is not entered as a meaning of this noun n in Bunrui-goi-hyou,

(2) to show some nouns which have the similar meaning to the meaning g,

however, our method does not trace what the meaning g is. This meaning is manually decided by observing nouns shown in (2).

For example, the Japanese noun "声" has meanings *"voice"* and *"opinion"*, but "声" in Bunrui-goi-hyou has only *"voice"* and does not have *"opinion*

". Our method points out that "声" in Bunrui-goi-hyou has a deficiency of a meaning, and shows some nouns such as "意見 (opinion)","見解 (view)","主張 (insistence)" etc. which have the similar meaning to *"opinion"*. By observing these nouns, we can manually decide that the lacking meaning of "声" is *"opinion"*.

The traditional research to acquire the unknown meaning was done by (Wilensky 1990). In his research, if a sense of a word is unknown, the sense is estimated from similar uses of other words to the use of the word. This approach, which estimates a feature of an unknown word from features of similar words, is taken to cope with the data sparseness problem of a corpus (Dagan 1993). In short, this approach is base on the idea that the unseen part can be estimated from seen similar parts. However, the simple use of this idea alone cannot find an unknown meaning of a word, because even similar words hardly have same polysemy to the word. For example, nouns such as "笑い声 (laughter)","喚声 (cheer)", "悲鳴 (scream)", "さえずり (song)" etc. are similar to the noun "声", but these nouns don't have the meaning *"opinion"*.

To estimate a lacking meaning, this paper applies the method presented by (Shinnou 1995), which extracts idioms from a corpus. In his research, first, the set of nouns which are co-occurred with a verb v is constructed from a corpus. Second, similar nouns to each other are removed from the above set. In this step, the similarity is measured by Bunrui-goi-hyou. Last, idioms are constructed from left nouns in the set and the verb v.

Suppose a noun n have multiple meanings and a meaning in them is not entered as a meaning of the noun n in Bunrui-goi-hyou. In above second step, the noun n tend to be left in the set. That is, we can estimate the noun n with the meaning lacking in Bunrui-goi-hyou by the clue that many idioms with the noun n are extracted through the above steps. Next our method extracts nouns which have a similar meaning to the lacking meaning of the noun n by mutual information. These nouns can be associated with the lacking meaning.

Our method can find not only a deficiency of a meaning but also meanings which are not used or are specifically used in the corpus domain.

We have experimented by the corpus which consists of Japanese economic newspaper 5 years articles with about 7.85 M sentences. We report the result of this experiment.

2 Estimation of nouns with the lacking meaning

Shinnou (Shinnou 1995) proposed the method to extract predicative idioms from a corpus by the lexical peculiarity for the noun in an idiom. His method firstly gathers cooccurrence data with the form $[noun, wo, verb]$

from the corpus. For example, from the sentence "昨夜ウイスキーを飲んだ (I drank whiskey yesterday)", cooccurrence data [ウイスキー, を, 飲む] ([*whiskey*, *obj*, *drink*]) is extracted. Second, the method chooses a verb v, and gathers nouns which can be an object of the verb v from cooccurrence data. Suppose the chosen verb is "飲む (drink)", we can gather "ウイスキー (whiskey)" and "要求 (request)" etc. from [ウイスキー, を, 飲む] and [要求, を, 飲む] etc. Next, similar nouns to each other are removed from these gathered nouns. Last, idioms are constructed by left nouns and the chosen verb. Figure 1 shows an example. In this example, "かたず - を - 飲む" and "息 - を - 飲む" are extracted as idioms. These extractions are correct.

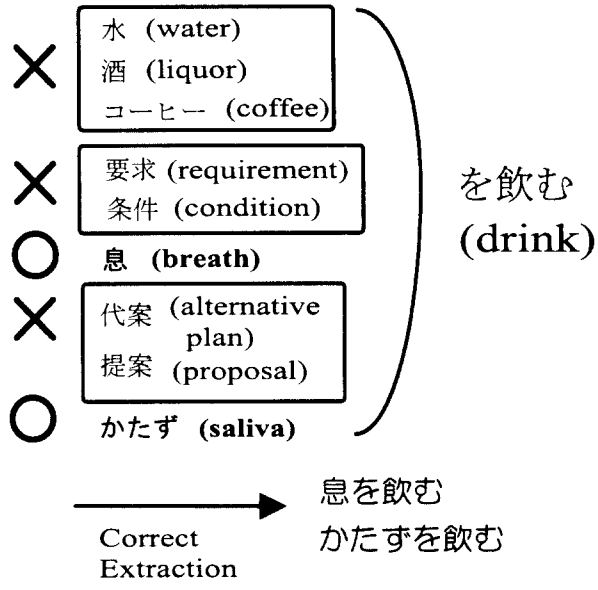

Figure.1

Suppose a noun n has multiple meanings and one meaning in these meanings is not entered as a meaning of that noun n in Bunrui-goi-hyou. Through above processes, expressions which comprise the noun n are extracted as idioms if the noun n in each expression is used as just the lacking meaning. These extractions are incorrect. For example, Figure 2 shows nouns co-occurred with the verb "異にする (distinguish)". In this example, the expression "立場 - を - 異にする" is extracted as an idiom, but it is incorrect.

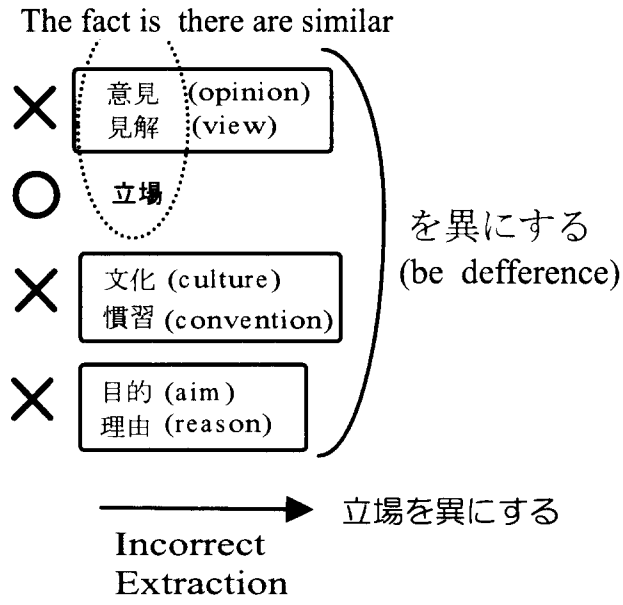

Figure.2

The noun "立場" has the meaning *"point of view"*, and "立場" is the similar noun to "意見 (opinion)" and "見解 (view)". Thus, the expression "立場 - を - 異にする" should not be extracted. This mistake occurs for lack of the meaning *"point of view"* of the noun "立場" in Bunrui-goi-hyou. As meanings of the noun "立場" in Bunrui-goi-hyou, only the meaning *"standing point"* is entered.

This mistake occurs not only in the case of the verb "異にする" but also in the case that the noun "立場" in the expression is used as the meaning *"point of view"*. For example, expressions such as "立場を明言する", "立場を明示する", "立場を力説する", "立場を無視する", "立場を要約する" are incorrect extractions.

Our method removes correct extractions, that is real idioms, from expressions extracted by the above method. Next, our method classifies left expressions according to the noun in the expression. If the class for a noun is big, we can estimate that the noun has a lacking meaning.

To decide whether an expression is an idiom or not, we use the idiom dictionary (Inoue 1994). The expression is decided as an idiom if it is an entry in that dictionary.

3 Decision of a lacking meaning

Now, we have cooccurrence data $[n, wo, v_i]$ corresponding to the expression "n - を - v_i" whose noun n is estimated to have a lacking meaning. Thus,

by mutual information (Hindle 1990), we can extract nouns, which have the similar meaning to the lacking meaning, from cooccurrence data. In this paper, we extract five nouns by this procedure. The set of five nouns is expressed by the following U_n.

$$U_n = \{n_{ur_0}, n_{ur_1}, n_{ur_2}, n_{ur_3}, n_{ur_4}\}$$

We have also cooccurrence data $[n, wo, v_j]$ corresponding to the not extracted expression "n - を - v_j" such as $v_j \neq v_i$. So, by the same procedure we can extract nouns, which have the similar meaning to the meaning of the noun n in Bunrui-goi-hyou, from cooccurrence data. The set of these nouns is expressed by the following R_n.

$$R_n = \{n_{r_0}, n_{r_1}, n_{r_2}, \cdots, n_{r_i}\}$$

By comparing U_n and R_n, we decide whether the noun n has a lacking meaning or not. If the noun n_{ur_0} is equal to a noun n_{r_k} in R_n and there is some similarity between n_{r_k} and n, the noun n is decided not to have a lacking meaning.

By the above processes, we can show

1. a noun n in Bunrui-goi-hyou lacks a meaning,
2. nouns which have the similar meaning to the lacking meaning are elements of U_n.

In final step, we must decide what the lacking meaning is. This decision are manually done by observation of U_n.

For example, suppose the noun "チェーン (chain)" is estimated to have a lacking meaning, and the $U_{チェーン}$ are shown as following.

$U_{チェーン}$ = {アミューズメント施設 (*a place of amusement*),
ホテル (*hotel*), レストラン (*restaurant*),
レジャー施設 (*leisure facilities*), 食品スーパー (*supermarket*)}

The noun with meaning *"a number of shops or hotels etc. owned by the same company"*, which is the lacking meaning of the noun "チェーン", is not in $U_{チェーン}$. However, by observing $U_{チェーン}$, we can easily decide that the lacking meaning is its meaning, i.e. *"a number of shops ... "*.

4 Experiment and evaluation

To evaluate our method, we have experimented by the corpus which consists of Japanese economic newspaper 5 years articles. The corpus has about 7.85 M sentences, and the average number of characters in one sentence was about 49.

From the corpus, we gathered about 4.41 M bits of cooccurrence data (about 1.48 M types) whose postpositional particle was "*wo*". From them, we removed the cooccurrence data whose frequency was 1, or whose verb does not appear more than 20 times. In all, we obtained about 3.27 M bits of cooccurrence data, which consisted of about 0.24 M types. Next, we extracted 46,930 types of expressions as idioms from cooccurrence data by the lexical peculiarity for the noun in an idiom, and removed real idioms from these expressions by using the idiom dictionary (Inoue 1994). Next, we classified left expressions by their nouns, and picked out 2,142 types of nouns whose frequency was more than 4. These 2,142 types of nouns were candidates of nouns with lacking meanings.

Next, we made U_n and R_n for each candidate noun n. By comparing U_n and R_n, 1,110 types of nouns were estimated to have a lacking meaning, and U_n for these nouns were shown. These shown U_n are evaluated by Kouji-en (Shinmura edition 1993) which is Japanese standard dictionary. Suppose the shown U_n is associated in our mind with a meaning g. If the meaning g is not entered as a meaning of the noun n in Bunrui-goi-hyou, but is entered in Kouji-en, then we evaluate that the U_n is effective. For example, there was the noun "スキー (ski)" in 1,110 types of extracted nouns, and $U_{スキー}$ is as following.

$$U_{スキー} = \{靴下\,(sock),\ 長靴\,(rubber\ \ boot),\ 靴\,(shoe),\ 運動靴\,(sport\ \ shoe),\ ワラジ\,(ancient\ \ Japanese\ \ straw\ \ sandals)\}$$

By observing nouns in $U_{スキー}$, we can estimate that the noun "スキー" has the meaning *"ski"*. And Kouji-en shows two meanings for the noun "スキー" as following.

1. one of a pair of long narrow strip of wood etc. fixed under the feet for traveling over snow
2. sport on snow with ski

First meaning shows that the noun "スキー" has the meaning *"ski"* in practice. However, the noun "スキー" in Bunrui-goi-hyou has only the above second meaning. That is, we can decide the lacking meaning of the noun "スキー" is *"ski"*. Thus, the $U_{スキー}$ is evaluated to be effective.

This evaluation found 177 types of effective U_n. The appendix table shows a part of lacking meanings found by our method. For information, the appendix table shows the meaning entered in Bunrui-goi-hyou besides the found lacking meaning.

5 Remarks

It is difficult to extract all knowledge from only a corpus because of incomplete analysis and data sparseness. And it is nonsense to create new one from scratch. It can be preferable to adjust or expand an existing one to meet an application's needs.

Such efforts have been done by many researchers. Hearst decomposed the Word Net into a set of categories, and adjusted and expanded the set by the Word Space created from corpora in order to better serve a text labeling task (Hearst 1996). Kaneda pointed out that we cannot gather enough examples, which is used to automatically acquire rules to select a verbal meaning, from only a corpus, and tried to adjust the handmade rules by the corpus (Kaneda 1995). Shinnou pointed out that the general thesaurus cannot be constructed from only a corpus, and used the existing thesaurus to cover the sparseness of the corpus (Shinou 1996). Our method is a kind of these approaches.

However, only addition of the lacking knowledge to existing one is not so useful. It is preferable to find the specific knowledge in the needed task, and to modify to suit the used corpus domain. Our method can not only find lacking meanings but also modify existing meanings.

Our method extracts nouns which have the similar meaning to the meaning entered in Bunrui-goi-hyou. But there are some nouns which don't have these similar nouns. Meanings entered in Bunrui-goi-hyou for these nouns are estimated not to be used in the corpus domain. Thus, we can lower the priority of these meanings. The noun "在庫 (stock)" is the such example. The meaning of "在庫" in Bunrui-goi-hyou is *"something exist in a warehouse"*. In the corpus, this meaning is not used and only the meaning *"merchandise kept in a warehouse"*, which is not entered in Bunrui-goi-hyou, is used.

Further, our method can find also the incorrect meaning in Bunrui-goi-hyou, when deciding the lacking meaning. For example, $U_{\text{離脱}}$ is shown as following.

$$U_{\text{離脱}} = \{\text{停止}(to\ stop),\ \text{脱退}(to\ leave),\ \text{解除}(to\ cancel),\ \text{解散}(to\ break\ up),\ \text{制裁}(sanctions)\}$$

We can estimate that "離脱 (to leave)" has the meaning *"to leave a party etc."*, but can notice strangeness that "離脱" is not similar to "脱退" in $U_{離脱}$. This is because the position of the noun "離脱" on the thesaurus in Bunrui-goi-hyou is wrong.

Moreover, when deciding the lacking meaning, we can find also meanings which are not entered in Kouji-en. This is because these meanings are conventionally used in the used corpus domain. For example, the $U_{トップ}$ for the noun "トップ (top)" is as following.

$$U_{トップ} = \{\text{社長}(president),\ \text{首相}(prime\ minister),\ \text{大統領}(President),\ \text{幹部}(leading\ member),\ \text{代表}(representative)\}$$

The meaning of the noun "トップ" is obviously *"person with the highest rank or degree"*, but this meaning is not entered in Bunrui-goi-hyou and Kouji-en. That is, the meaning is specifically or conventionally used in the used corpus. As other examples, there are the meaning *"gang"* for the noun "組 (class)", and the meaning *"character"* for the noun "活字 (type)".

The problem of our method is that many mistake estimations make the decision of the lacking meaning hard work. To overcome this problem, in this paper we use similarities to the meaning entered in Bunrui-goi-hyou to reduce mistake estimations. In the future, we will try other ways.

The scale of the used corpus is another problem. The large corpus is favorable for our method. However, the usable corpus is generally small. Therefore, in the future, we must consider how to effectively apply our method to a small corpus.

6 Conclusions

In this paper, we proposed the method to find a deficiency of a meaning in Bunrui-goi-hyou by a corpus. The proposed method first extracts idioms from the corpus by the lexical peculiarity for the noun in an idiom. By incorrect extractions, the noun with a lacking meaning is estimated. Next, nouns which have the similar meaning to the lacking meaning are shown. By observing these nouns, the lacking meaning is manually decided. We experimented by the corpus which consists of Japanese economic newspaper 5 years articles. As the result, we could find 177 types of lacking meanings.

This method is a kind of approaches to expand and modify the existing knowledge by using a corpus. The acquisition of knowledge from a corpus is an important research, but the completely automatic method is not practice. The approach to expand and modify the existing knowledge by a corpus is practical and effective. A problem of the proposed method is many mistaken estimations. The solution is our future work.

Acknowledgments

The corpus used in our experiment was extracted from CD-ROMs '90–'94 sold by the Nihon Keizai Shinbun Company. We deeply appreciate the work of everybody involved in securing permission from the Nihon Keizai Shinbun Company to use this corpus.

References

Dagan,I., Marcus,S., and Markovitch,S. : "Contextual word similarity and estimation from sparse data", Proc. of ACL-93, pp. 164–171 (1993).

Hearst,M. and Schütze,H. : "Customizing a Lexicon to Better Suit a Computational Task ", Corpus Processing for Lexical Acquisition, MIT press, pp.77–96 (1996).

Hindle,D. : "Noun classification from predicate-argument structures ". Proc. of ACL-90, pp.268–275(1990).

Inoue, M : "Reikai kanyouku jiten (in Japanese)", Soutaku Publishing (1994).

Kaneda,S., Akiba,Y., and Ishii,M. : "Jireini motozuku eigodousi sentakuruuru no syuuseigata gakusyuuhou (in Japanese)", Proceedings of the first annual meeting of the Association for Natural Language Processing, pp. 333–336(1995).

Shinmura,I edition: "Kouji-en Fourth printing (in Japanese)", Iwanami Publishing (1993).

Shinnou,H : "Redefining similarity in a thesaurus by using corpora ", Proc. of COLING-96, pp.1131–1135 (1996).

Shinnou,H and Isahara,H : "Automatic Acquisition of Idioms on Lexical Peculiarity (in Japanese)", Journal of Information Processing, Vol. 36, No. 8, pp. 1845–1854 (1995).

The National Language Research Institute : "Bunrui-goi-hyou (in Japanese)", Shuuei Publishing (1994).

Wilensky,R. : "Extending the Lexicon by Exploiting Subregularities", Proc. of COLING-90, Vol. 2, pp. 407–412 (1990).

Appendix

noun	the meaning in Bunrui-goi-hyou	the lacking meaning
キー	key to lock the door	keys on keyboard
紙面	surface of the paper	article
定期	a fixed period	abbreviation for time deposit
仏	Buddha	abbreviation for French
ゴール	goal	getting score in soccer games etc
ネット	net	abbreviation for network
覚悟	to give up	metal attitude
大勢	rough situation	many persons
両国	both countries	Ryougoku (place-name)
パンチ	to beat	a ticket punch
レバー	liver	lever
手間	work with wages	effort
出血	bleeding	sacrifice
助け	helping	something to need
騒動	making a noise	a state of emergency
大手	arm	abbreviation for major companies
膜	film to wrap and separate an organ	skin or film
クラブ	society	club to hit a ball in golf
ハンドル	handle	steering wheel
獲物	something to take by force	fishes and animals to get by fishing and shooting
頭脳	brain	main person
マイナス	mark of minus	lack or loss
火種	small fire to cause big fire	cause of an event
近く	near place	state almost reaching a value
待遇	service	pay
ダイヤ	abbreviation for diamond	abbreviation for diagram
ボタン	button	push-button
電源	power supply	electric outlet
土俵	straw bag to be filled with earth	place to carry put an event
無理	to be un-reasonable	to force a person to do
涙	tear	sympathy
潮流	ocean current	movement of times
機	machine	chance
悲劇	play with tragedy ending	tragedy event
チップ	tip	microchip
シート	seat	sheet
腕	arm	skill
ブレーキ	something to hinder a plan and a movement	brake
土壌	soil	environment to cause an event
反響	reflection of sound wave	response
便	letter	convenience of transportation
いす	chair	position in a company
口	mouth	talk
頭	head	thought
柱	pillar	person to support an organization
壁	wall	obstacle
立場	standing place	view or opinion
夢	dream	ambition
声	voice	opinion