Division of Example Sentences Based on the Meaning of a Target Word Using Semi-supervised Clustering

Hiroyuki Shinnou, Minoru Sasaki

Ibaraki University, 4-12-1 Nakanarusawa, Hitachi, Ibaraki, Japan 316-8511 {shinnou, msasaki}@mx.ibaraki.ac.jp

Abstract

In this paper, we describe a system that divides example sentences (data set) into clusters, based on the meaning of the target word, using a semi-supervised clustering technique. In this task, the estimation of the cluster number (the number of the meaning) is critical. Our system primarily concentrates on this aspect. First, a user assigns the system an initial cluster number for the target word. The system then performs general clustering on the data set to obtain small clusters. Next, using constraints given by the user, the system integrates these clusters to obtain the final clustering result. Our system performs this entire procedure with high precision and requiring only a few constraints. In the experiment, we tested the system for 12 Japanese nouns used in the SENSEVAL2 Japanese dictionary task. The experiment proved the effectiveness of our system. In the future, we will improve sentence similarity measurements.

1. Introduction

We perform the task of collecting example sentences from a corpus, based on the meaning of the target word. It is simple to extract example sentences that include the target word. The hurdle is to divide these sentences into clusters based on the meaning of the target word. This paper describes our system, which efficiently overcomes this hurdle.

Example sentences that are clustered based on the meaning of the target word, are useful for full-scale semantic analysis. For example, we can design a classifier to solve word sense disambiguation, using example sentences as training data during inductive learning (Masaki Murata and Masao Utiyama and Kiyotaka Uchimoto and Qing Ma and Hitoshi Isahara, 2001). We can easily construct the case slot of a verb (Philip, 1992) using example sentences clustered based on the meaning of the verb. A thesaurus can easily be designed (Hindle, 1990) using example sentences clustered based on the meaning of a noun.

Clustering of example sentences based on the meaning of a target word can be performed by distinguishing between the different senses of in an example sentence. In other words, this task is termed as word sense disambiguation. This task may be performed using a (semi-) supervised learning approach. However, this approach is not practical for two problems. The first problem is the cost of constructing the training data. Supervised learning requires a large amount of training data. If there are many target words, it is impossible to construct training data corresponding to each word. The second problem is the definition of the senses of a target word. When using a (semi-) supervised learning approach, it is necessary to define the senses of the target word in advance. It is difficult to maintain a uniform sense granularity, and a minor sense may easily be overlooked.

For this task, unsupervised learning, i.e., clustering, is available. However, several clustering methods, such as k-means, require the number of clusters, i.e., the number of meanings of the target word. These methods cannot be used for our task. Some clustering methods can estimate the number of clusters; however, this estimation is essentially impossible because it fixes the sense granularity, which depends on the target word.

Therefore, we use a semi-supervised clustering approach for this task. In this approach, a constraint on a pair of data, set by the user, is used for clustering (Cohn et al., 2003)(Basu et al., 2004)(Bilenko et al., 2004)(Klein et al., 2002). The system selects some pairs from the data set, and offers them to a user. The user enters the pair's constraint (must-link or cannot-link) into the system. Must-link indicates that the two data items must belong to the same cluster, while cannot-link indicates that they cannot belong to the same cluster. The cost to provide such constraints is lower than that of the cost to assign a class label to each data.

2. Division of example sentences by semi-supervised clustering

2.1. Algorithm

Figure 1 shows our system algorithm.

First, the user enters a target word w, and the rough number k of meanings of w^{-1} . The system then collects the sentences, including w, from a corpus. The set of these sentences is the data set D. Next, the system divides D into k clusters with a general clustering tool.

$$C = \{C_1, C_2, \cdots, C_k\}$$

The set A in Figure 1 is the final clustering result, i.e., a set of clusters. The system first sets $A = \{C_1\}$, then sequentially picks up C_i from C, and evaluates whether the C_i must be added into A, or merged into a cluster in A. This is based on the user feedback. The system shows the central sentence of C_i and the central sentence of a cluster of C_j in A to the user. The user judges whether the meanings of the target words w in each sentence are identical. If they are identical, the user enters "must-link," and is merged into the cluster. If not ("cannot-link"), the procedure is repeated

¹It is about from five to ten times of the estimated number.

for the next cluster of A. If the user judgments are "not identical" for every central sentence of A, the C_i is added into A. Note that the central sentence does not change on merging the two clusters, i.e., one central sentence is chosen and then continuously used. Therefore, the system output is unique, and the maximum number of user judgments is k(k-1)/2.

Input W and initial cluster number k construct $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_N\}$ cluster D to k clusters $C = \{C_1, C_2, \dots, C_k\}$ $A = \{C_1\}$, $C = \{C_2, C_3, \cdots, C_k\}$ for(i =2; i < k+1; i++) { z is the center of C_i foreach C in A { x is the center of C give (x,z) to user get Const from user if (Const = must-link) { $C \leftarrow C \cup C_i$; break } } if (Const = cannot-link) { $A \leftarrow A \cup \{C_i\}$ } }





2.2. System description

Figure 2 \sim 7 show our system. The system is implemented using Perl CGI. First, we input the target word (Figure 2).

ØSentence Search – Windows Inter	net Explor
- R http://localhost/cgi-bi	n/index.html
😭 🍄 🏾 🏈 Sentence Search	
keywords: 核 Search Reset	

Figure 2: Input of a target word

software, we use HyperEstraier provided in the following site.

http://hyperestraier.sourceforge.net/
index.html

- Search Results - Windows Internet Explorer の提供元	Yahoo! JAPAN		للعر	(1
* Ø http://kcehost/cgi-bin/searchcgi		💌 fy 🐹 Öccele	8	p.
谷 適根 - Search Results		∆ •⊙ • ∂ • ∂ <-	9C • @ 7-NO	•
例検索		種の検索	紀果1-1000 幷	i
sters: 20				
Run clutering				
1. 核物質の防護に関する条約は、核物質の国際轄	送中に一定の核物質防護権	措置が採られることを確保する	らこと、このような	£.
措置が採られる保証のない核物質の輸出入を許 物とする。	可しないこと及び核物質に多	除る一定の行為を犯罪とし処	向すること等を内)
 	3燃料安全基準專門部会(6	第163年11月設置)において	行われており	
これまでに枢燃料施設に共通の安全審査の基本 委員会決定)及び同業実際計を数まえたウランカ	3D考え方として、「核燃料施 IT協協国有の安全審査指導	設安全審査基本指計」(昭和 計として、「ウラン加工協設安	55年2月安全 全審査提詳 (昭	2
和55年12月安全委員会決定)並びに、核燃料制	着設の万一の事故に関連し	て、その立地条件の適否を判	断するために用	,
いるブルトニウムのめやす線量として、「核燃料系 和Eo年4日中央委員会は使いがそれぞれ留空な	総の立地評価上必要なブル わプいる	レトニウムに関するめやす線	動について」(昭	
3. なお、核兵器の製造・保有は、原子力基本法の則	地定の上からも禁止されてい	るところであるが、さらに、わ	6個は、昭和61	
年6月、核兵器の不拡散に関する条約を批准し、 ネ	非核兵器国として核兵器を	製造しない、取得しないなどの)鉄稿を負ってい	ŀ
 で、 この炉は、在来型炉がウラン235などの核分裂! 	によって生じる高速中性子を	軽水などにより減速するのに	対して、滅速過	
程を経ないで核分裂反応を続けるもので、有効に	利用できる中性子が多いた	め、核分裂を起こさないウラ	ン238を核分裂	
15000000000000000000000000000000000000	ために約40010000000000000000000000000000000000	院電の進展に対応した核燃料	物質等の利用	
の本格化に対処して、原子炉施設以外の核燃料	サイクル施設や核燃料物質	の輸送等に係る安全性に関	、調査審議する	ò
常設の「核燃料安全専門審査会」を設置した。	51年4月,征用6月時逃窜肥	設女主 御近 明11102129580	5月に満年に月して、	
該兵器国は、核兵器をを他国に移譲せず、また、	その製造について非核兵業	国を振動しないこと。		
 「核物質防護」とは。核物質の盗難寺の不法は移 行為に対する防護をいう。 	転に対する防護及び原子力	施設又は核物質の構造に対	する防害,奴壊	
8. 最近、核実験により我が国への放射能の影響が	みられるのは中国の大気圏	内核実験のみであるが、昭和	054年度は中国	l.
の秘実験は行われなかった。				

Figure 3: Example sentences

Then, the system conducts clustering for example sentences (Figure 4). As the clustering engine software, we use CLUTO provided in the following site.

```
http://glaros.dtc.umn.edu/gkhome/
views/cluto
```

Clustering Result - Windows Internet Explorer の歴色元: Yahoo! JAF	AN	12
 • Image: http://localhost/cgi-bin/cluster.cgi?8B38C8 	💌 fg 🗙 Coople 🖉	
Clustering Result		• •
Clustering Result		100
Typical sentences		
Cluster 1		
・54 核戦力については、ソ連の全戦略ミサイルの約30%に当 ・178 核戦力では、この地域における核抑止力を維持するため ば漸次戦略任務から外され、米本土に基地をおくトライデント原 りついる。	たるICBMやSLBM等の核戦力が配備されていると推定される。 にボラリス原子力潜水艦等が配備されているが、ボラリス潜水艦 子力潜水艦に代替することで核抑止力を強化する計画が進めら	
297 核戦力は, ICBM若干, IRBM65~85基, MRBM約5 327 戦略振戦力は, ICBM若干, IRBM50~70基, MRBM 517 戦域振戦力としては, SS-20やTU-22Mバックファイア)基及び書-6(TU-16)約90機である。 40~60基, TU-16約90機である。 が取州への配備開始に引き続いて, この地域にも配備されつつ	
95。 889 第2次大戦後、ソ連は、通常戦力の分野で圧倒的優位で 主義諸国は、通常戦力の面で量的に対抗するのが困難である/ 5基本としてきている。	ある強大な軍事力を維持してきたのに対して、米国などの自由 こめ、核戦力と通常戦力の組み合わせにより戦争を抑止すること	
Cluster 2		
9 核都市と、その核都市への通勤者が600人以上、または、	その核都市への通勤者を在住雇用者で除した値が0.05を超え	
る市町村を加えて都市圏とする。 • 12 そのため、これら核都市内において幹線道路等の整備や 有機的に結合する交通網の整備を推測し、これら核都市の育成	都市再開発等を実施する(まか、東京外郭環状道路等級都市を 2. 勢備に怒める必要がある。	
 17 平成3年度においては、千葉県の業務応都市基本構想() 備事業の活用により、幕張新都心の拠点協設等に対して日本長 57 東京大都市圏における核都市の育成整備を図るため、63 	(更津業務該都市基本構成)を承認するとともに、核都市拠点整 発銀行からの低利融資を行い、事業の促進を図った。 ・年度は、核都市拠点地区の整備に係る日本開発銀行の出融資	
制度の拡充を図る。 • 61 このため、昭和63年に制定した多種分散型国土形成(現) 基本方針を平成元年3月に策定した。	法に業務該都市制度が盛り込まれ、同法に基づく業務該都市	

Figure 4: Initial clustering result

The system retrieves example sentences including the target word from a corpus (Figure 3). As the search engine The number of clusters is given by the user. Then, the system selects a typical sentence from each cluster (Figure 5).



Figure 5: Typical sentences

The next stage is the semi-supervised clustering. The system shows a pair of typical sentences to the user, and the user judges whether target words in two sentences are used in same meaning, or not (Figure 6).

Figure 6: User judgment



@Seni-12	supervised Clustering - Windows Internet Explorer の提供元: Y	ahoo! JAPAN	_ D ×
GO.	[g] http://localhost/cgi-bin/semicls1.cgi3803008	¥ 49 🗙 Google	ρ·
\$ 61	Semi-supervised Clustering		· ③ 7-14Q · *
Final cl	lutering result		4
Cluster Cluster	rumber : 4 r1		
 1 核特 な措置が 	物質の防護に関する条約は、核物質の国際輸送中に一 が採られる保証のない核物質の輸出入を許可しないこと	定の核物質防護措置が採られることを確保する。 及び核物質に係る一定の行為を犯罪とし処罰す	こと、このよう ること等を内
 12 核別 12 核別 14 長子 15 5年11 14 55年11 	5. 燃料施設の指計等の検討は、安全委員会の総燃料安全 までに該燃料施設に共通の安全審査の基本的考え方と 決定)及び同基本指針を設まえたつラン加工施設置有の 2月安全委員会決定)並じに、総料料施設の万一の事基 ニウムのめやず練量として、「総燃料施設の立地評価上点	基準専門部会(昭和53年11月設置)において行 して、「総燃料施設安全審査基本特許」(昭和56 安全審査指計として、「ウンノ加工施設安全審査 なブルトニウムに関するのやす縁量について	うわれてお 3年2月安全 計録計」(昭和 30に用いる 」(昭和58年
 4月女当 5またの本格(制を整) 「核燃料 	主要員会の定力が不住でも認定されている。 こ、原子力施設の安全審査については、原子力委員会は 化に対処して、原子内施設以外の核燃料サイクル施設で 値することが逸航であると判断し、昭和61年4月、従前の 半安全専門審査会」を設置した。	1、今後の原子力発電の進展に対応した総燃料 対応料物質の輸送等に係る安全性に関し、調整 の再処理施設安全審査専門部会を発展的に解消	物質等の利用 転譲する体 して、常設の
 7 18 壊行為(8 最) 国の総防 	時物質防護」とは、総物質の盗難等の不法な移転に対する に対する防護をいう。 近、該実験により我が国への放射能の影響がみられるの 実験は行われなかった。	5防護及び原子力施設又は核物質の執影に対す)は中国の大気圏内核実験のみであるが、昭和6	る妨害, 岐 54年度(3中
 9 核都 超える市 10 19 	都市と、その核都市への通勤者が500人以上、または、 市町村を加えて都市圏とする。 992年9月の核物質防護に関する条約の再検討会議に	その核都市への通動者を在住雇用者で除した信 おいて、廃棄物中の核物質に関する核物質防護	助10.05を の在り方等
の 検討(・ 11 こ) で 試験語 を目標()	のため、カイトラインの見直し会合の開催要求が出され の基本計画では、核融合動力炉実現の前提となる臨界 装置(JT-60)」(プラズマ温度数千万度から1億度程度。 装置を開発することを中ふとして、非円形断面トーラス総構要	こ。 プラズマ条件の達成に重点を置き。トカマク型の1 プラズマ密度と閉じ込め時間の積2~6×1013 編の研究間段、高ペータブラズマに関する研究	臨界ブラズ Seec/cm3 開発、ブラズ ・

Figure 7: Final clustering result

2.3. Similarity between sentences

A measure of the similarity of the sentences is essential in our system.

The system converts a sentence into a feature list. Using this feature list, the system measures similarities between sentences. Our system uses the following features used in the paper (Shinnou and Sasaki, 2003). Suppose the target word is $w = w_i$, which is the *i*-th word in the sentence.

- e1: the word w_{i-1}
- **e2:** the word w_{i+1} **e3:** two content words in front of w_i

e4: two content words behind w_i

e5: thesaurus ID number of e3 and e4

For example, let's consider the following sentence² in which the target word is '*kiroku*(記録)'³.

kako/saikou/wo/kiroku/suru/ta/.

The system generates the following feature list from the above example sentence.

```
{ e1=wo, e2=suru, e3=saikou, e3=kako,
 e4=suru, e4=., e5=3192, e5=31920,
 e5=1164, e5=11642 }
```

Because of space limitations, we sketch out the similarity measurement sim(A, B) between two feature lists, A and B.

The sim(A, B) is the sum of three kinds of similarities.

sim(A,B) = 1/3 * {
 (Similarity of the feature e1)
 + (Similarity of the feature e2)
 + (Similarity of the feature e3,e4 and e5)}

The similarity is 1 if the feature has the same value, and 0 if it does not. In addition, the similarity is adjusted using some ad-hoc rules.

²A sentence is segmented into words, and each word is transformed into its original form by morphological analysis.

³The word 'kiroku(記録)' has at least two meanings: "memo" and "record".

2.4. Central sentence of cluster

We define the central sentence x_c of cluster

$$C = \{x_1, x_2, \cdots, x_n\}$$

as follows:

$$c = \arg \max_{i \in 1:n} \sum_{j \in 1:n} sim(x_i, x_j).$$

This indicates that the average of the similarities between x_c and sentences in C is the largest.

3. Experiment

We tested our system for 12 nouns (shown in Table 1) used in the SENSEVAL2 Japanese dictionary task (Shirai, 2003). The data set was constructed from the training and test data provided by SENSEVAL2. It gives 50 nouns, from which we picked only the words that produced 300 or more example sentences were picked. This procedure gave 12 nouns, which were then set as target words.

Table 1: Data sets			
word	# of example	# of	
	sentences	meanings	
mono(もの)	754	10	
mondai(問題)	636	4	
daihyou(代表)	466	3	
mae(前)	426	4	
kankei(関係)	414	3	
gogo(午後)	396	3	
jibun(自分)	362	2	
jidai(時代)	360	4	
kodomo(子供)	354	2	
genzai(現在)	341	2	
syakai(社会)	339	6	
ima(今)	329	4	
average	431.42	3.91	

First, the user must enter the initial cluster number to the system by overestimating the number of possible meanings of the target word. In this experiment, the number was fixed at 20.

The system then converts a sentence into a feature list, and constructs the similarity matrix using the similarity measure, sim(A, B). To perform clustering for the similarity matrix, we used the clustering tool kit CLUTO. We performed clustering using CLUTO with the cluster number set to 20, and without any optional parameters⁴.

Next, through the user feedback described in the previous section, the system generates the final clustering result from the above 20 clusters. Table 2 shows the result.

The "# of clusters" in Table 2 lists the number of clusters generated by our system. The number in parentheses is the true number given by SENSEVAL2 using a dictionary. The "# of constraints" lists the number of constraints entered by the user. The "semi-supervised" shows the system accuracy, and "un-supervised" shows only the clustering accuracy.

Table 2: Result of experiment

target	# of	# of	semi-super	unsuper
word	clusters	constrains	-vised	-vised
mono	4 (10)	66	0.391	0.309
mondai	1 (4)	19	0.969	0.969
daihyou	3 (3)	35	0.858	0.667
mae	3 (4)	24	0.855	0.371
kankei	2 (3)	30	0.785	0.848
gogo	3 (3)	27	0.634	0.444
jibun	1 (2)	22	0.942	0.942
jidai	2 (4)	28	0.653	0.550
kodomo	2 (2)	26	0.588	0.480
genzai	2 (2)	26	0.974	0.707
syakai	4 (6)	22	0.755	0.395
ima	3 (4)	23	0.687	0.423
Average	3.25 (3.91)	24.6	0.757	0.592

Table 2 shows that the performance using the semisupervised approach is better than when using the unsupervised, and the number of constraints is small. Therefore, the system is efficient.

4. Discussion

4.1. Initial number of clusters

In our system, the user must provide the initial number of clusters for the target word; while in the experiment, the number was fixed at 20.

We varied the number of clusters from 10 to 100, in steps of 5. The result is shown in Figure 8. In Figure 8, the xaxis is the initial number of clusters entered by the user, the y-axis in the upper figure of Figure 8 is the average of 12 accuracies, and the y-axis in the lower portion of Figure 8 is the average number of constraints entered by the user.

It can be concluded, that the accuracy is higher when the number of initial clusters is large. However, a large number of clusters requires more constraints.

The initial number of clusters depends on the target word. If the target word is estimated to have many meanings, the initial number should be large. If not, we should set the initial number small. One of the advantages of our system is that the initial number of clusters is not fixed, and is instead entered by the user.

Accuracy primarily depends on the measurement of similarity. We plan to develop methods to further improve the accuracy of these measurements. We will need to construct a thesaurus for this task..

4.2. Similarity measurement

Precsion of our system depends on similarity measurement between example sentences. Our defined similarity measure is ad-hoc, so is needed to be improved. At the present, our system handles only noun words. For verb words, we must use case slots and syntactic information to measure similarity.

⁴The program used is "scluster," and its input is a similarity matrix.



Figure 8: Constrains and accuracy for the initial cluster numbers

To measure similarity, use of a thesaurus is essential. We use Bunrui-goi-hyou⁵ as the thesaurus. We can improve similarity measurement by using more powerful thesaurus. Improvement of similarity measurement is our future work. To do it, we must construct the thesaurus suitable for our task.

5. Conclusion

In this paper, we have described a system that divides example sentences based on the meaning of the target word. In this task, the estimation of the number of possible meanings is essential. Therefore, our system uses semi-supervised clustering. First, a data set is divided into many small clusters by using an initial clusters number entered by the user. Next, the system merges these clusters by asking the user for feedback. An experiment using 12 nouns demonstrated the high accuracy of our system, using a small number of constraints given by the user. In the future, we will improve the accuracy of the similarity measurements to obtain more accurate clustering results.

Acknowledgements

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas, "Japanese Corpus", 19011001, 2007.

6. References

- Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. pages 333–344.
- Mikhail Bilenko, Sugato Basu, and Raymond J. Mooney. 2004. Integrating Constraints and Metric Learning in Semi-Supervised Clustering. In *ICML-2004*, pages 81–88.
- David Cohn, Rich Caruana, and Andrew McCallum. 2003. Semi-supervised Clustering with User Feedback. Technical Report TR2003–1892, Cornell University.
- Donald Hindle. 1990. Noun classification from predicate argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics* (ACL-90), pages 268–275.
- Dan Klein, Sepandar D. Kamvar, and Christopher D. Manning. 2002. From Instance-level Constraints to Spacelevel Constraints: Making the Most of Prior Knowledge in Data Clustering. In *ICML-2002*, pages 307–314.
- Masaki Murata and Masao Utiyama and Kiyotaka Uchimoto and Qing Ma and Hitoshi Isahara. 2001. Japanese word sense disambiguation using the simple Bayes and support vector machine methods. In *Proceedings of the SENSEVAL-2*, pages 135–138.
- Resnik Philip. 1992. WordNet and Distributional Analysis: A Class-based Approach to Lexical Discovery. In *Proceedings of AAAI-92 Workshop on Statistically-Based NLP Techniques*, pages 48–56.
- Hiroyuki Shinnou and Minoru Sasaki. 2003. Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the EM algorithm. In Proceedings of Seventh Conference on Natural Language Learning (CoNLL-2003), pages 41–48.
- Kiyoaki Shirai. 2003. SENSEVAL-2 Japanese Dictionary Task (in Japanese). Journal of Natural Language Processing, 10(3):3–24.

⁵Japanese standard thesaurus