A REFINING METHOD OF OBTAINED ATTRIBUTES TO CHARACTERIZE UNDEFINED CONCEPTS USING SEARCH ENGINE

Noriyuki Okumura[1] and Yuto Hatakoshi[2]
Department of Electronics and Computer Science, Nagano National College of Technology, 381-8550 Nagano, Japan
Faculty of Culture and Information Science, Doshisha University, Kyo-tanabe, 610-0394 Kyoto, Japan

Keywords: Concept-base, Machine learning, Search engine, Time series, Obtaining attributes, Undefined concept.

Abstract: In this paper, we propose a method to resolve problems of the attributes obtaining method using WWW search engine characterizing undefined concepts, which do not exist in Concept-base. Concept-base is an key database constituting Word Association Mechanism to perform commonsense judgment. Concept-base was constructed automatically by electronic dictionaries and newspapers. Therefore Concept-base has about 120,000 statically defined concepts. Nevertheless, it has no effective learning system. This paper proposes a method to make Concept-base to learn concepts dynamically using Auto Feedback system. In addition, a removal method of noise at-tributes is also proposed. We present attributes refinement method that paid attention to changing with time of the Internet. Furthermore, we inspect the effectiveness by an evaluation experiment.

1 INTRODUCTION

The objective of this research is to construct an automatically learning method using WWW for existing Concept-base(Okumura et al., 2007). Concept-base is a large-scale Knowledge-base constructed by electronic dictionaries and newspapers. Nevertheless, it has no effective learning system. Therefore, we need a machine learning system for existing Concept-base.

Concept-base has a large number of concepts, which have some Attribute-Weight pairs. It is difficult to deal with new concepts such as new words, proper nouns, and etc. generated momently because Concept-base was constructed by static data. We proposed Auto-Feedback method(Tsuzi et al., 2004) using a search engine[1] to resolve the problems. However, this method obtained Attribute-Weight pairs in retrieved point. As were shown in earlier reports(Gulla et al., 2007; Gordon et al., 2010), words gathering systems were proposed. These methods were not suitable for our Concept-base because these systems did not work in the long period. The method which paid attention to time series(Horiuchi and Uchida, 2011), but this method did not refine attributes. Consequently, this method had the problem that different results were obtained whenever we retrieved new concepts.

This paper proposes a method to resolve abovementioned problems. Proposed method statistically refines Attribute-Weight pairs, which were obtained for long periods. By evaluating experiments, we showed that the proposed method was superior to the method in the past from the standpoint of obtaining Attribute-Weight pairs.

[1]http://www.google.co.jp/

2 METHOD

In the following, we present a method to refine Attribute-Weight pairs obtained by search engine. First, our Concept-base, Auto Feedback and Revision of Morphological Analysis are briefly depicted. Second, our proposal method was described. Finally, we presented the evaluation method.

2.1 Concept-base

Concept-base(Okumura et al., 2007) is a large-scale Knowledge-base constructed by electronic dictionar-

Okumura N. and Hatakoshi Y..
A REFINING METHOD OF OBTAINED ATTRIBUTES TO CHARACTERIZE UNDEFINED CONCEPTS USING SEARCH ENGINE.
DOI: 10.5220/0003697804930497
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (SSW-2011), pages 493-497
ISBN: 978-989-8425-80-5

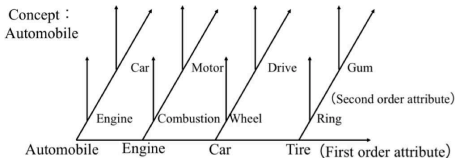

Figure 1: Concept (automobile) is extended to second order attribute. This figure shows association words for the concept (Automobile).

ies and newspapers. Headwords in dictionaries were assumed to be concepts and content words in explanation sentences were assumed to be attributes for headwords (concepts). A concept (A) consists of pairs of attributes (a_i) which characterizing the concept (A) and weights (w_i) which mean the importance of each attributes (eif is a natural number for each concepts, '$znum$' is a number of attributes)(1).

$$A = (a_i, w_i)|0 < i < znum + 1. \quad (1)$$

Attributes for each concept were also defined in Concept-base as concepts. Therefore, one concept was defined as attributes chain model of n-th-order dimension. In this paper, Concept-base has about 120,000 concepts, and each concept has 30 attributes on average. Fig.1 shows the example of concept (automobile). eAutomobilef has attributes (engine, car, tire, etc.). eEnginef, eCarf, and eTiref are also defined in Concept-base. Thus, eEnginef has attributes (Combustion, Motor, etc.).

In this paper, we aim to construct an automatically learning method for Concept-base using search engine.

2.2 Auto Feedback

An undefined concept in Concept-base was input, and the documents that were de-scribed about the undefined concept, were obtained from the retrieval result pages of search engine. The words included in the retrieval result pages were attributes of undefined concepts. The weight of each attribute was granted by tf and idf. tf was the frequency that undefined concepts appear in the retrieval result pages. idf was calculated from the number of the retrieval pages and the number of all pages of search engine. Table.1 showed examples of the obtained attributes of undefined concepts.

In this research, we obtained 100 candidate attributes descending in weight order by Auto Feedback. The Auto Feedback got attributes at the point in time when I retrieved undefined words. Therefore retrieval results were influenced by a temporary topic, and it was considered that Auto Feedback was not able to obtain attributes definitely.

Table 1: The attributes of undefined concepts gHarrison Fordh and gFinePixh.

Harrison Ford		FinePix	
attributes	weights	attributes	weights
movie	225.16	digital	331.21
actor	120.77	camera	326.95
appearance	87.46	pixel	301.11

2.3 Revision of Morphological Analysis

This paper used MeCab(Kudo et al., 2004) as a Morphological Analyzer. Japanese have no custom leaving a space between words like English. A problem to divide sentences needlessly too much happened when we used a Morphological Analyzer. It unnecessarily divid-ed a sentence into words by the default MeCabfs setting. It had an original revision rule for this problem. However, we set a simple rule without using its rule.

1. Connecting words and phrases in the parenthesis.
2. Connecting if nouns were next to each other.

For example, in the case of a sentence gJiEVJh, uJiEVJiNAUSICAA/of Valley of the Windjvwas divided withuv,uJv,uiEVJv before reviewing setting, and the title of the movie is divided needlessly. We united nouns to be adjacent by uv after the setting changed, we can extract uJiEVJv(Table 2).

2.4 Proposal Method

Auto Feedback was a method to learn undefined concepts on the spot. Consequently, the method paid no attention to changing with time of Internet. Proposal method re-peats the Auto Feedback trial many times and refines attributes and weights of undefined concepts statistically (Fig. 2).

2.5 Evaluation Method

In this section, we explain the evaluation method of our work.

Evaluation Method. Three subjects evaluated these acquired all attributes (about 20,000 words). We adopted attributes which two or three subjects answered suitable as correct words. In all of Auto Feedback trials, we calculated precision (Eq.3), recall (Eq.4), and F-measure (Eq.4).

About Recall, there may be correct attributes other than acquired attributes using Auto Feedback. However, it was difficult to collect attributes that human beings thought suitable by a questionnaire. In this research, we adopted correct attributes evaluated by

Table 2: uJiEVJv for leaving a space between words.

Before revision		After revision	
Acquired attributes	Evaluation	Acquired attributes	Evaluation
	Right		Right
	Wrong	JiEVJ	Right
J	Wrong	-	-
iEVJ	Right	-	-

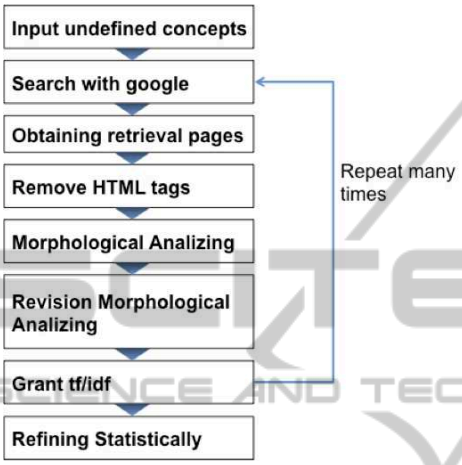

Figure 2: The flowchart of proposal method.

three subjects to a numerator of Recall. Therefore Recall takes 1.0 value when we output all acquired attributes. We evaluate the method in the past by averaging all trials.

$$precision = \frac{1}{n}\sum_{i=1}^{n}\frac{correct\ attributes}{all\ obtained\ attributes} \quad (2)$$

$$recall = \frac{1}{n}\sum_{i=1}^{n}\frac{correct\ attributes\ in\ sellected\ attributes}{correct\ attributes} \quad (3)$$

$$F-measure = \frac{1}{n}\sum_{i=1}^{n}\frac{2\times precision\times recall}{precision+recall} \quad (4)$$

Evaluation Data. We use 49 undefined concepts in Concept-base as evaluation data.

3 RESULTS

We obtained attributes for 49 undefined concepts by Auto Feedback for one month. The number of obtained attributes except the repetition was 302 on average. For each undefined concepts, we sorted the attributes for the threshold at the number of times which attributes was obtained every ten times in the experiment period. The horizontal axis of Fig. 3 shows the number of times that was not obtained as attributes in entire experiment period. About evaluation data,

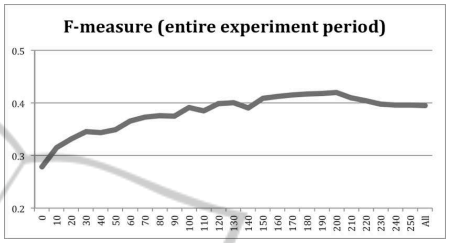

Figure 3: Changing of F-measure in entire experiment period.

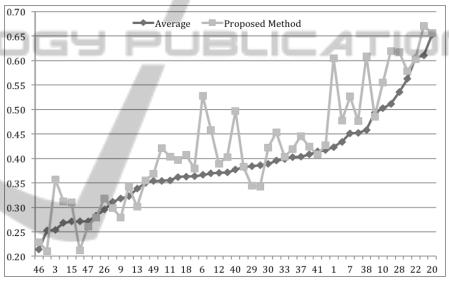

Figure 4: Comparison proposal method with average of all Auto Feedback trials.

we calculated F-measure with each threshold. Fig. 3 shows a change of the average.

Numerical value of the horizontal axis of Fig. 4 means a number appropriated to undefined concepts in Table 3. The horizontal axis of Fig. 4 is sorted in order of the average of all Auto Feedback trials using the method in the past (250 trials).

Table 4 shows samples of the retrieval result (undefined concept: Google) of Auto Feedback once trials and the result of proposed method that refined Attribute-Weight pairs based on the number of appearance.

4 DISCUSSION

When number of times that was not obtained as attributes in Fig. 3 is smaller than 200, F-measure takes the maximum. In other words, when number of times that was attributes were obtained is greater than 50, F-measure takes 0.42 of the maximum. We adopted

Table 3: Evaluation data (written in English notation for explanation).

Undefined concepts (Evaluation Data)					
1	Ichiro	18	Sorting works	35	relation between Japan & China
2	Influenza	19	Kyoto Sanga FC	36	Hokkaido Nippon-Ham Fighters
3	Barack Obama	20	Keihan Electric Railway	37	confront Waseda and Keio
4	cartel	21	All Japan University Road relay	38	outflow of pictures
5	Google	22	the Northen Territories	39	Hideki Matsui
6	smartphone	23	Chiba Lotte Marines	40	Yokohama APEC
7	Sony Ericsson	24	Jun Natsukawa	41	Hiro Mizushima
8	Yu Darvish	25	astronaut	42	Tomomi Kasai
9	Saeko Darvish	26	Yuko Ogura	43	Kawori Manabe
10	mine of Chile	27	Senkaku	44	Yutaka Takenouchi
11	Domino's Pizza	28	Senkaku's video	45	Singing contest of red and white
12	Myanmar	29	Naoko Yamazaki	46	news of Marriage
13	Your Party	30	University of Yamanashi	47	Shin-ichi Hatori
14	rare earth element	31	Hisashi Iwakuma	48	self-defense
15	LAWSON	32	new year letter	49	group infection
16	Yonaguni Island	33	a great war of Warring period	-	-
17	World Volley	34	Yuki Saito	-	-

Table 4: Comparison of obtained Attribute-Weight pairs of Proposed Method and Auto Feedback in the past.

Auto Feedback in the past		Proposed Method	
Attributes	Weights	Attributes	Weights
Retrieval Result	907.821	USA	219.071
Many	664.259	Firewall	177.136
High	442.840	How to use	133.609
China Government	404.549	Company	131.618
New	309.988	Many	128.388
China	280.353	Site	127.669
Search Engine	258.543	Access Analyzing	119.792
Detailed	243.562	Search	104.181
USA	231.078	All over the world	97.671
Facebook	221.420	Google Map	90.177

this value (200) for the threshold and sorted the attributes. As a comparison experiment, we calculated precision, recall, and F-measure for all of Auto Feedback trials. Fig. 4 shows relations of F-measure in all Auto Feedback trials and sorting by the proposal method.

When number of trials that attributes were obtained is greater than 50, F-measure takes maximum value in Fig. 3. We guessed that the refinement method of the attributes works effectively. We calculated F-measure whenever Auto Feedback was carried out and found average about 250 trials.

5 CONCLUSIONS

In this research, we proposed refining method of obtained attributes for undefined concepts using Auto Feedback. Proposed method refined Auto Feedback attributes based on the number of appearance statistically.

In addition, we showed that higher F-measure score was provided than Auto Feed-back in the past. It is necessary to examine weighting method for refined attributes. We showed that it is effective to limit weights using dispersion by a precedence study(Hatakoshi, 2010). We compound with a precedence study and proposal method to gain high performance.

ACKNOWLEDGEMENTS

This work was supported by Grant-in-Aid for Young Scientists (B) (No. 23720222)

REFERENCES

Gordon, J., Van Durme, B. D., and Schubert, L. K. (2010). Learning from the web: Extracting general world knowledge from noisy text. In *Proceedings of the AAAI 2010 Workshop on Collaboratively-built Knowledge Sources and Artificial Intelligence*.

Gulla, J. A., Borch, H. O., and Ingvaldsen, J. E. (2007). Ontology learning for search applications. In *Proceedings of the 2007 OTM Confederated international conference on On the move to meaningful internet systems: CoopIS, DOA, ODBASE, GADA, and IS - Volume Part I*, OTM'07, pages 1050–1062, Berlin, Heidelberg. Springer-Verlag.

Hatakoshi, Y. (2010). An acquisition method of attributes for unknown words considering time factor. In *The 73rd National Convention of Information Processing Society of Japan*.

Horiuchi, Y. and Uchida, O. (2011). Extraction of unsteadiness of concept attributes using weblog articles. In

Proceedings of the 10th WSEAS international conference on Applied computer and applied computational science, ACACOS'11, pages 137–141, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to japanese morphological analysis. In *In Proc. of EMNLP*, pages 230–237.

Okumura, N., Yoshimura, E., Watabe, H., and Kawaoka, T. (2007). An association method using conceptbase. In *Proceedings of the 11th international conference, KES 2007 and XVII Italian workshop on neural networks conference on Knowledge-based intelligent information and engineering systems: Part I*, KES'07/WIRN'07, pages 604–611, Berlin, Heidelberg. Springer-Verlag.

Tsuzi, Y., Hirokazu, W., and Tsukasa, K. (2004). The method of acquisition of the new concept and its attribute using the world wide web. In *The 18th Annual Conference of the Japanese Society for Artificial Intelligence*.