Grasping Related Words of Unknown Word for Automatic Extension of Lexical Dictionary

Myunggwon Hwang
Intelligent Computing Lab.
Gwangju, Korea
mghwang@chosun.ac.kr

Jongan Park
Dept. of Information & Communication
Gwangju, Korea
japark@chosun.ac.kr

Sunkyoung Baek
Intelligent Computing Lab.
Gwangju, Korea
Zamilla100@chosun.ac.kr

Pankoo Kim
Dept. of Computer Engineering
Gwangju, Korea
pkkim@chosun.ac.kr

Junho Choi
Intelligent Computing Lab.
Gwangju, Korea
spica@chosun.ac.kr

Abstract

An aim of this research is to grasp related words of unknown word. Currently, several lexical dictionaries have been developed for semantic retrieval such as WordNet and FrameNet. However, more new words are created in every day because of new trends, new paradigm, new technology, etc. And, it is impossible to contain all of these new words. The existing methods, which grasp the meaning of unknown word, have a limitation that is not exact. To solve this limitation, we have studied the way how to make relations between known words and unknown word. As a result, we found a noble method using co-occurrence, WordNet and Bayesian probability. The method could find what words are related with unknown word and how much weight other words relate with unknown word.

1. Introduction

This paper deals with a research for real semantic information retrieval with understanding all of words in the document. Many lexical dictionaries have been developed to overcome the limitation of current retrieval methods which use a simple keyword matching using probability, but it is impossible to contain words generated everyday by new trends, new items and new technologies. So, to guess the meaning of these unknown words, which are not defined in lexical dictionaries, many researchers have studied using several ways such as co-occurrence, TF-iDF or learning methods. These researches for unknown words show much advanced result but, have a limitation that could not find related words with unknown word.

This paper addresses research to grasp related words of unknown word for automatic extension of lexical dictionary with getting over the limitation. Known words occurring in a document may have only one meaning, but, generally have many meanings and relations between other words. Unknown words also can have many meanings and relations with known words which are defined in lexical dictionary. However, existing researches have concentrated on methods to define a meaning of unknown word or to find a base class (super-class) of unknown word, which is considered as instance. So, in this paper, we have studied what words are related with unknown word and how much weight is related between unknown word and known word. To do this, we have used the Word-Net that is lexical dictionary, co-occurrence and Bayesian probability. We implemented

E-FORENSICS 2008, January 21-23, Adelaide, Australia

DOI 10.4108/wkdd.2008.2634

a system depending on this research. The results showed that this research is very valuable and noble.

This paper organized as follows: Section 2 introduces our main discourse. Section 3 shows experimental result and evaluation. Finally, we summarize this study in Section 4.

2. Proposed Method

When we are reading, we meet UWs (unknown word) often. We can guess the meaning or related words easily without referring to a dictionary. Because the explanation or the definition of UW occurs in the same sentence using related words. In other words, meaning of UW can be found generally through co-occurrence words [5]. Especially, a noun word has much meaning than other part of speeches, so, plays core role in understanding document [3]. In this way, we can guess UWs using co-occurred noun words. However, a computer has many limitations to grasp the meaning of UW exactly. So, we focused on finding related words with UW before grasping the meaning of UW, which is an obstacle in NLP research area.

We define an UW as word not contained in WordNet v2.0 as English lexical dictionary. And we have researched based on following assumptions.

Assumptions:

- The semantically nearest words from UW occur in the same sentence.
- UW can have many meanings and many relations with known words those are defined in a dictionary.

In this chapter, we present our proposed method through several steps. Table 1 shows each step and summary. And, the following sub-chapters explain the detail contents about table 1.

Table 1. Overall steps of proposed method

Step	Contents
Pre-processing	- PoS Tagging - Extract Nouns from Sentence - Stemming
UW Extraction	- If word which is not contained in WordNet occurs then that is UW.
Creation of Noun Set	- Make a noun set using noun words occurring with UW in same sentence
Weight Measure of Co-occurrence Word	- To give weight to co-occurrence word using Bayesian probability.
Making a Related Noun Group	- Creating a group of related nouns - Calculating Bayesian probability of the group
Normalization	- Normalizing the Bayesian probability of group

2.1. Pre-Processing

To extract noun words from sentences, first we need PoS tagging. We employed the PoSTagger [1] that is possible to tag PoS quite exactly. It just tags the PoS of each word, so, we need to grasp compound noun, which occurs using sequential nouns in a sentence. For example, if a noun 'football' and a noun 'player' occur sequentially, we generate 'football', 'player', and, 'football_player' to match those into WordNet. If WordNet contains a compound noun 'football_player' then, 'football' and 'player' will exclude in a part to measure a weight. Chapter 2.4 deals with this problem. After these steps, a stemming, which is a process to find an original form of inflected word, is added. Table 2 shows the result of this pre-processing. The sentence in table 2 is a part of a document about soccer player 'Zidane' from WIKIPEDIA [2].

Table 2. The result of pre-processing

Sentence	Zinedine Yazid Zidane, popularly nicknamed Zizou, is a French football player of Algerian descent, famous for leading France to winning the 1998 World Cup.
PoS Tagging	Zinedine/NNP Yazid/NNP Zidane/NNP, popularly/RB nicknamed/VBN Zizou,/NNP is/VBZ a/DT French/JJ football/NN player/NN of/IN Algerian/NNP descent,/NNP famous/JJ for/IN leading/VBG France/NNP to/TO winning/VBG the/DT 1998/CD World/NNP Cup/NNP
Extracted Nouns	Zinedine, Yazid, Zidane, Zinedine_Yazid, Yazid_Zidane, Zizou, football, player, football_player, Algerian, descent, Algerian_descent, France, World, Cup, World_Cup

2.2. UW Extraction

Noun words extracted in the pre-processing step are mapped into a noun part of WordNet to make a decision which word is UW. If there is no agreement word in the WordNet, this word is considered as UW. Table 3 is pseudo-code to extract UW. UW is a set of UW that will get growing by processing many sentences. The UWs extracted from a sentence in Table 2 are presented below of table 3.

Table 3. Pseudo-code to extract UW

Pseudo Code	N_i : i-th noun extracted from document;
	UW : Unknown Word Set;
	WN : Noun part of WordNet;
	$if(N_i \notin WN)$
	{
	$UW = UW + N_i$;
	}

Extracted UW	Zinedine, Yazid, Zidane, Zinedine_Yazid, Yazid_Zidane, Zizou, Algerian_descent

2.3. Creation of Noun Set

In this step, we make a noun set occurring with UW in the same sentence because co-occurring words have a possibility to relate with UW. Table 4 is pseudo-code to create a noun set.

Table 4. Pseudo-code to extract UW

$UW = \{UW_i \mid 1 \le i \le m\}$: Unknown word(UW_i) set; m : Total count of unknown word(UW_i); $K_i = \{k_{ij} \mid 1 \le j \le n_i\}$: Collection of noun(k_{ij}) set occurring with UW_i; n_i : Total count of noun occurring with UW_i (total count of noun set of UW_i);

Many UWs are extracted and each UW has many co-occurring noun words through processing the whole document using previous steps. Table 5 shows the just 3 results of processing the document about 'Zidane'.

Table 5. Extracted UW and noun set

i	UW_i	Co-occurred Noun Set	n_i
1	Zidane	Algerian, Cup, descent, football, France, June, player, World, World_Cup, attention, Brazil, country, Europe, fame, goal, playmaker, Year, April, football_player	19
2	Juventus	club, Madrid, Real, attention, Brazil, country, Cup, Europe, fame, goal, playmaker, World, championship, Euro, Italy, level, Spain, victory, World_Cup	19
3	FIFA	attention, Brazil, country, Cup, Europe, fame, goal, play-maker, World, World_Cup, Player, Year, April, football	14

2.4. Weight-Measure of Co-occurring Words

This step measures the weight of co-occurring words with UW using Bayesian probability because frequency is important factor to grasp related degree between UW and known words. In the case of compound noun, it has 2 more nouns and we generated both each single noun and compound noun in chapter 2.1. In here, if the WordNet contains the compound noun, all of single noun words contained in compound noun should be excluded from weight-measure.

For example, if the WordNet contains a compound noun 'World Cup' and 'World Cup' occurs twice, single noun 'World' occurs once then, these are calculated 'World' 3 times occurrence, 'Cup' twice and 'World Cup' twice in the previous step. These make an affection of duplication to occurring count of noun words because only single word 'World' occurs once independently. Therefore, if a compound noun which is consisted of more than 2 words sequentially is detected, we find a pure value of Bayesian probability of single nouns by extracting Bayesian probability of compound noun. Table 6 presents the process to measure the weight of co-occurrence words. And, table 7 shows the weight of each co-occurrence word with UWs in table 5.

Table 6. Weight-measure of co-occurring words

$$Bayesian = \frac{P(B \mid A)P(A)}{P(B)}$$

$oc(UW_i)$: Frequency of UW_i, $oc(k_{ij})$: Frequency of k_{ij}

$$Bayesian = \frac{P(oc(UW_i) \mid oc(k_{ij}))P(oc(k_{ij}))}{P(oc(UW_i))}$$

If, $k_{ij} = \{k_{i(j-q)} \mid 1 \le q \le r\}$ is compound noun,

r : count of single nouns composed in compound noun

$P(oc(k_{iq'}) \mid oc(UW_i)) = P(oc(k_{iq}) \mid oc(UW_i)) - P(oc(k_{ij}) \mid oc(UW_i))$

q' : independent occurrence of single noun

$$P(oc(k_{iq'}) \mid oc(UW_i)) = \frac{P(oc(UW_i) \mid oc(k_{iq}))P(oc(k_{iq}))}{P(oc(UW_i))} - \frac{P(oc(UW_i) \mid oc(k_{ij}))P(oc(k_{ij}))}{P(oc(UW_i))}$$

Table 7. Weight of co-occurring words

UW_i	Noun(k_{ij})	W	Noun(k_{ij})	W
Zidane	Algerian	0.2	April	0.2
	football	0.2	football_player	0.2
	June	0.2	player	0.2
	World_Cup	0.6	attention	0.2
	country	0.2	Europe	0.2
	goal	0.2	playmaker	0.2
	descent	0.2	Brazil	0.2
	France	0.2	fame	0.2
	World	0.2	Year	0.2
Juventus	club	0.6667	Madrid	0.6667
	attention	0.3333	Brazil	0.3333
	victory	0.3333	Europe	0.3333
	goal	0.3333	playmaker	0.3333
	World_Cup	0.3333	championship	0.3333
	Italy	0.3333	level	0.3333
	Real	0.6667	Spain	0.3333
	country	0.3333	Euro	0.3333
	fame	0.3333		
FIFA	attention	0.2	Brazil	0.2
	football	0.2	Europe	0.2
	goal	0.2	playmaker	0.2
	World_Cup	0.6	Player	0.2
	country	0.2	Year	0.2
	fame	0.2	April	0.2
	World	0.2		

W : Weight of Co-occurring Word

2.5. Making a Related Noun Group

The words occurring with UW can have relations with each other. These related words can be grasped by matching into WordNet. If some words are related, we make a group using these related words because UW has much more related probability with the group of co-occurring words [4, 6]. The step in this chapter makes a group through co-occurring words, and measures the related probability between UW and the groups using sum of Bayesian probability of each word. Table 8 shows the result of group that is consisted of related nouns and Bayesian probability of group.

Table 8. Related noun group and sum of Bayesian

UW_i	Group of Related Nouns	Sum
Zidane	09951469(football_player)-10283858(player) 10283858(player)-10284756(playmaker)	0.6
	08802093(France)-09142657(Europe) 08802093(France)-08060674(Europe)	0.4
Juventus	08895440(Madrid)-08894294(Spain)	1.0
FIFA	10284756(playmaker)-10283858(Player)	0.4

2.6. Normalization

An aim of this paper is to find related words of UW. To do this, we measured Bayesian probability of each co-occurring word and made related noun groups through previous steps. However, that is not probability of relation between UW and group. Moreover, to reflect both probability and relation, it needs normalization. Table 9 and table 10 show way to normalize and results.

Table 9. Normalization of group

$G_i = \{G_{il} \mid 1 \le l \le s\}$: Relate Related Group of UW_i (This group is consisted of co-occurrence words with UW_i through matching into WordNet.)

s : count of groups related with UW_i

G_{il} : Related words in the co-occurrence set k_i

$$Normalization = \frac{\sum_{l=1}^{s} Bayesian(G_{il} \mid UW_i)}{\sum_{j=1}^{n_i} Bayesian(k_{ij} \mid UW_i)}$$

Table 10. Result of normalization

UW_i	Group	Group of Related Nouns	N
Zidane	$G_{1,1}$	Algerian	0.056
	$G_{1,2}$	football	0.056
	$G_{1,3}$	football_player, player, playmaker	0.167
	$G_{1,4}$	France, Europe	0.111
	$G_{1,5}$	World	0.056
	$G_{1,6}$	World_Cup	0.167
	$G_{1,7}$	Brazil	0.056
	$G_{1,8}$	Goal	0.056
	…	…	…

UW_i	Group	Group of Related Nouns	N
Juventus	$G_{2,1}$	Club	0.100
	$G_{2,2}$	Madrid, Spain	0.145
	$G_{2,3}$	victory	0.050
	$G_{2,4}$	Goal	0.050
	$G_{2,5}$	World_Cup	0.050
	$G_{2,6}$	Europe	0.050
	$G_{2,7}$	football	0.050
	…	…	…
FIFA	$G_{3,1}$	Goal	0.071
	$G_{3,2}$	World_Cup	0.214
	$G_{3,3}$	Europe	0.071
	$G_{3,4}$	football_player, player	0.143
	…	…	…

N :Normalization

In Table X, the $UW_{Zidande}$ has many groups and especially two groups ('football_player', 'player', 'playmaker'), ('World_Cup') show the most related value as 0.167. And, the second is the group of 'France' and 'Europe'. In case of $UW_{Juventus}$, the group of 'Madrid' and 'Spain' is the most related. This result is obtained from a short document that is just consisted of 301 words. Though, the result is very meaningful. Through many experiments, we obtained that the longer document is, the better we get result.

3. Experimental Result

We collected documents which are consisted of more than 1000 words from Wikipedia and gather 10 (e) evaluators for this experiment. And, we showed the extracted unknown word to evaluators and the related 10 (g) noun groups depending on the normalization value and, we asked them to judge whether each word has relation with UW or not. The amount of UW is 100(u). And then, the relevancy of this result is calculated by (1).

$$\text{Relevancy}(\%) = \frac{\sum_{i=1}^{e}\sum_{j=1}^{u}\sum_{k=1}^{g} R\_UW_{ijk}}{e \times u \times g} \times 100 \qquad (1)$$

Where, R_UW_{ijk} has a value 1 or 0 depending on being relation between UW and related noun groups. For example, if all of evaluators ($e=10$) judge the values of every R_UW_{ijk} ($g=10$) to be 1 then the value of $\sum_{k=1}^{g} R\_UW_{ijk}$ is 10. So, the relevancy of this experiment is 100(%). Through this experiment and evaluation, we found that this research is very meaningful and excellent as 95.88(%).

4. Conclusions

An ultimate aim of this research is to extend lexical dictionary through addition of unknown words automatically. To accomplish this aim, the meaning, related words and relation such as hyponym, hypernym, synonym, antonym, etc of unknown word have to be grasped. In this paper, to grasp related words of unknown word, we applied co-occurrence, Bayesian probability and WordNet as lexical dictionary. Through experiment and evaluation, we got semantic and excellent result. Based on this research, grasping the meaning of unknown words and relation is future works for auto-extension of lexical dictionary.

Acknowledgment

This study was supported by Ministry of Culture & Tourism and Culture & Content Agency in Republic of Korea.

Reference

[1] http://nlp.stanford.edu/software/tagger.shtml
[2] http://en.wikipedia.org/wiki/Zidane
[3] Hyo-Jung Oh, Sung-Hyoun Myaeng, "A Hypertext Categorization Method using Incrementally Computable Class Link Information", Korean Institute of Information Scientists and Engineering, no. 07, vol. 29, pp.498-509, August, 2002.
[4] Hyunjang Kong, Myunggwon Hwang, PanKoo Kim: The Method for the Unknown Word Classification, PKAW2006, pp.207-215, August, 2006.
[5] A. Jobbins and L. Evett, "text Segmentation Using Reiteration and Collocation," Proceedings of the COLING-ACL'98, pp. 614-618, August 1998.
[6] Hyunjang Kong, Myunggwon Hwang, Gwangsu Hwang, Jaehong Shim, Pankoo Kim, "Topic Selection of Web documents using Specific Domain Ontology", MICAI2006 : Advances in Artificial Intelligence, LNAI 4293, pp. 1047-1056, November 2006.