A Term Normalization Method for Better Performance of Terminology Construction

Myunggwon Hwang[1], Do-Heon Jeong[1,⋆], Hanmin Jung[1], Won-Kyoung Sung[1], Juhyun Shin[2], and Pankoo Kim[2]

[1] Korea Institute of Science and Technology Information (KISTI)
245 Daehak-ro, Yuseong-gu, Daejeon, South Korea
mg.hwang@gmail.com, {heon,jhm,wksung}@kisti.re.kr
[2] Chosun University
375 Seoseok-dong, Dong-gu, Gwangju, South Korea
{jhshinkr,pkkim}@chosun.ac.kr

Abstract. The importance of research on knowledge management is growing due to recent issues with big data. The most fundamental steps in knowledge management are the extraction and construction of terminologies. Terms are often expressed in various forms and the term variations play a negative role, becoming an obstacle which causes knowledge systems to extract unnecessary knowledge. To solve the problem, we propose a method of term normalization which finds a normalized form (original and standard form defined in dictionaries) of variant terms. The method employs a couple of characteristics of terms: one is appearance similarity, which measures how similar terms are, and the other is context similarity which measures how many clue words they share. Through experiment, we show its positive influence of both similarities in the term normalization.

Keywords: Term Normalization, Terminology, Appearance Similarity.

1 Introduction

Text resources are increasing explosively, due the large amount of data. Methods for efficient text resource management are of great importance and an area of recent concentration. The management begins with terminology construction, and term extraction is the most fundamental work in the construction. However, terms can be expressed in various forms in documents and this is big obstacle to quality terminology. Technical terms especially, which consist of two words at least, have more variations than general words. In the case of general words, the forms are differently expressed according to singular/plural types (ex. 'word' and 'words') and sometimes mistyping; the case of technical terms has additional expressions such as semantic replacement (ex. '**head** mounted display' and '**helmet** mounted display') and re-arrangement (ex. '**visna maedi** virus' and '**maedi visna** virus') of component word(s). These expressions cause a large

⋆ Corresponding author.

L. Rutkowski et al. (Eds.): ICAISC 2012, Part I, LNCS 7267, pp. 682–690, 2012.

quantity of unnecessary information, resulting in data sparseness of knowledge extraction [5], feature selection for machine learning [7], and knowledge integration/merging [8, 9], and so on. Even though such term variations need to be solved, research on normalization has been dealt with in only a few works [1, 2].

In order to cover the issue, we suggest a normalization method. First, we prepare a set of technical terms which are extracted from a huge corpus and divide the set into a normalized term set (NTS) and a variant term set (VTS). The method finds the variant terms original forms from NTS. We utilize Wikipedia[1] to collect the NTS and employ a couple of similarities, such as appearance similarity and context similarity. Through experimental evaluation, it is confirmed that both of the similarities can be positive factors for term normalization.

This paper is organized as follows: Section 2 describes our term normalization method. In Section 3, we evaluate the method through experiment. Finally, we summarize our research in the Fourth Section.

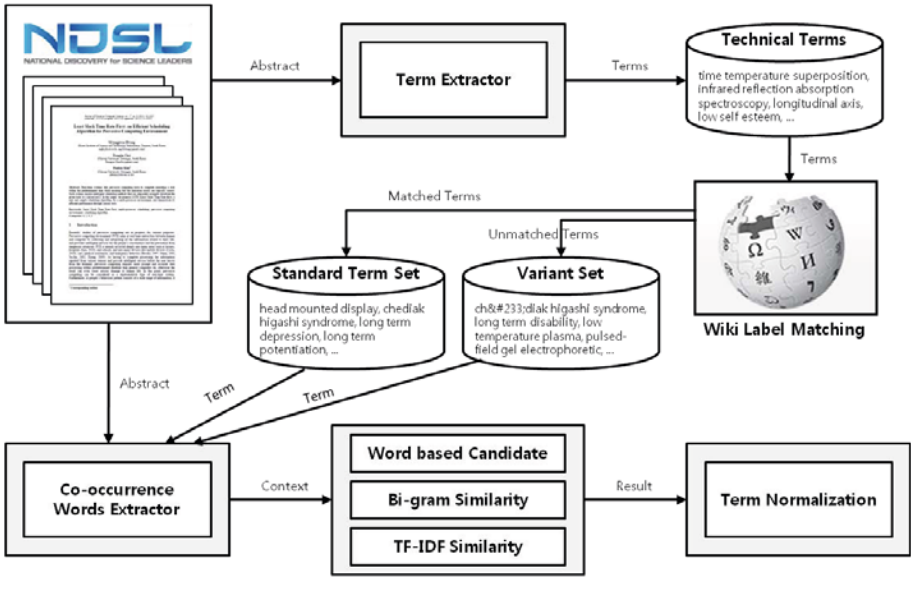

Fig. 1. System architecture for term normalization

2 Term Normalization Method

To construct terminology, many efforts are needed to extract and filter out terms. This current work is that kind of effort, and mainly consists of three parts for the term normalization as shown in Fig. 1. The first constructs a normalized term set (NTS) and a variant term set (VTS). The second step is to collect context

[1] Wikipedia, The Free Encyclopedia: `http://en.wikipedia.org/wiki/Main_Page`

information based on co-occurrence nouns of NT and VT. The last is to measure similarities by using the appearance feature and context information together. In this section, each step is described in detail with examples.

2.1 Construction of NTS and VTS

Terms can be expressed in various forms and a part of them is already generalized to the public and defined in dictionaries. The remaining part is considered to be variants or new words (in the research, the new word is not dealt with). The defined terms may be used more frequently than the other terms, but the undefined terms become an obstacle resulting in low performance in the works related to terminology. In order to resolve the obstacle, we divide terms into two groups: NTS, which is defined in a dictionary, and VTS, which is not defined. Through processing a huge paper abstract set of NDSL[2], 89,231 basic terms, which consist of only multi-words, have been prepared in advance (term extraction is out of range of the research so it is omitted). In addition, a label page of Wikipedia provided by DBPedia version 3.6[3] is employed for NTS. Through the step, 10,684 terms are collected into NTS. Table 1 shows a part of NTS and VTS.

Table 1. The examples of NTS and VTS

$Terms in NTS (with frequencies)$	$Terms in VTS (with frequencies)$
head mounted display (64),	long term disability (19),
chediak higashi syndrome (308),	chédiak higashi syndrome (80),
long term depression (2168),	low temperature plasma (26),
long term potentiation (8417),...	madine darby canine kidney (29),...

The elements of NTS can be a correct candidate of variant terms of VTS. From the following section, detail of the processes are described.

2.2 Construction of Context Information

In the previous step, NTS and VTS were prepared. To find the original form of VT, the work uses appearance similarity and context similarity together. For context similarity, co-occurrence nouns (clue words) are gathered as the context. After this step, each term of the NT and the VT has its co-occurrence nouns with their frequencies. In extracting nouns, the Stanford POStagger[4] is applied to tag part-of-speech to each word [4] and the Porter stemmer[5] is used for noun

[2] NDSL (National Discovery for Science Leaders): `http://www.ndsl.kr/index.do`
[3] DBPedia: `http://dbpedia.org/About`
[4] The Stanford Natural Language Processing Group: `http://nlp.stanford.edu/software/tagger.shtml`
[5] Porter Stemmer (The Porter Stemming Algorithm): `http://tartarus.org/martin/PorterStemmer/`

normalization [3][6]. Table 2 shows clue words and their frequencies as a part of terms in Table 1.

Table 2. The example of context information

Set	*Terms*	*ClueWords(ContextInformation)withFrequencies*
NTS	chediak higashi syndrome (308)	active (30), blood (12), cell (103), patient (45), rel (3), heterozyg (2), control (16), individu (2), enzym (7), protein (16), membran (19), gene (20), ...
VTS	chédiak higashi syndrome (80)	capac (8), neutrophil (18), monocyt (8), electron (5), microscopi (4), phagocytosi (2), leucocyt (1), aureu (1), vitro (1), abnorm (6), chemotaxi (2), ...

The context information of each term is utilized for context similarity between NTs and VTs.

2.3 Selecting Correct Candidates of VT

If the system tries to measure the similarity between all elements of NTS and VTS in order to find a NT as original forms of VTs, it wastes time and cost and causes low precision as well. Therefore the research selects correct candidates for all VTs. The terms the research deals with consist of multi-words, so if they share one word at least between a VT and a NT, the NT is added to the correct candidate set for the VT. Here, propositions, conjunctions and stop-words such as 'of,' 'for,' 'by,' 'an,' 'the,' 'and,' and 'or' are not involved in the selection process. Table 3 shows examples of candidate terms selected for VTs.

Table 3. Examples of lists of candidate terms

Variantterms	*Listsofcandidateterms*
vaso occlusive crises	hepatic veno **occlusive** disease, vaso **occlusive** crisis, **veno occlusive** disease, aorto iliac **occlusive** disease, **vaso** vagal syncope, **veno occlusive**
voltage operated ca2+ channels	**voltage** gated, l type calcium **channels**, **voltage** controlled filter, **voltage** sensitive calcium channel, plasma membrane **ca2+** atpase, voltage gated **ca2+** channel, store **operated** calcium channel, **voltage** dependent anion **channels**, ...

In order to find the correct NT (normalized term or original form), the work measures similarities between VT and each element of its candidate list. The next section explains the similarities.

[6] Term normalization and word normalization are different in a point of authors view. The motivation of both normalizations is to find original form but term normalization is more complex than that of the word. Please refer to the introduction of the paper for details.

2.4 Similarities for Term Normalization

The VTS may have two types of terms such as VTs and new words. For finding VTs which have a high possibility of variation we have two basic clues like the following.

- Decisive Clue 1. Similar appearance: The variations of multi-word terms could happen due to semantic replacement, re-arrangement, word inflection, and mistyping. If one term is originated from a NT, their term appearances are strongly similar.

- Decisive Clue 2. Similar context information: Even though two terms are similar on appearances, one is not always originated from the other, such as the case between 'vitamin c' and 'vitamin d'. Therefore the context similarity is additionally utilized.

Appearance Similarity. Terms contain a few words and each word consists of letter(s). The term variations occur due to additions, substitutions and removals of letter(s) or word(s) and thus a measure which inspects those specific changes is needed. The appearance similarity can grasp the changes and it measures bigram based word similarity. For the explanation, a set of words consisting of terms and a set of bigrams of each word are expressed by (1) and (2) respectively.

$$term_t = \{w_i, 1 \leq i \leq n\} \tag{1}$$

$$bigram_w = \{b_j, 1 \leq j \leq |w| - 1\} \tag{2}$$

where, $term_t$ is a term, w_i is i-th word of termt, n and b mean count of words and bigram of each word *(w_i)* individually. The count of bigrams of a word is different with word length by 1. For example, 'vaso occlusive crises' in Table 3 is expressed as: $term_{vasoocclusivecrises} = \{vaso, occlusive, crises\}$, $bigram_{vaso} = \{va, as, so\}$, $bigram_{occlusive} = \{oc, cc, cl, lu, us, si, iv, ve\}$, $bigram_{crises} = \{cr, ri, is, se, es\}$.

To measure bigram based word similarity, Dice's coefficient is used and (3) shows the equation.

$$s(w_k, w_l) = \frac{2 \times P(bigram_{w_k} \bigcap bigram_{w_l})}{|w_k| + |w_l| - 2}, w_k \in VT, w_l \in NT \tag{3}$$

To inspect word re-arrangement, the order of words can be ignored in (3). Its reason is indicated with Table 4 later. The results by (3) are used for appearance similarity of (4).

$$s(term_{VT}, term_{NT}) = \frac{2 \times \Sigma arg\ max(s(w_k, w_l))}{|term_{VT}| + |term_{NT}|} \tag{4}$$

Where term_{VT} and term_{NT} are a variant term and candidate term respectively. Tables 4 and 5 show the examples of the appearance similarities, and bold typed result means the maximum.

Table 4 is an example of the word re-arrangement, and Table 5 is about the substitution or the mistyping. For the word re-arrangement case, we should

Table 4. An appearance similarity of re-arrangement case

terms	{visna, maedi, virus}	{maedi, visna, virus}
Bigram	$bigram_{visna}$ = {vi,is,sn,na}	$bigram_{maedi}$ = {ma,ae,ed,di}
	$bigram_{maedi}$ = {ma,ae,ed,di}	$bigram_{visna}$ = {vi,is,sn,na}
	$bigram_{virus}$ = {vi,ir,ru,us}	$bigram_{virus}$ = {vi,ir,ru,us}
$s(w_k, w_l)$	$s(w_{visna},w_{maedi})$=0, **$s(\mathbf{w}_{visna}, \mathbf{w}_{visna})$=1 {vi,is,sn,na}**,	
	$s(w_{visna},w_{virus})$=0.25 {vi}, **$s(\mathbf{w}_{maedi}, \mathbf{w}_{maedi})$ = 1 {ma,ae,ed,di}**,	
	$s(w_{maedi},w_{visna})$=0, $s(w_{maedi}, w_{virus})$ = 0, $s(w_{virus},w_{maedi})$=0,	
	$s(w_{virus},w_{visna})$=0.25 {vi}, **$s(\mathbf{w}_{virus}, \mathbf{w}_{virus})$ = 1 {vi,ir,ru,us}**.	
$s(term_p, term_q)$	(2*(1+1+1))/(3+3) = 1.	

Table 5. An appearance similarity of mistyping or substitution case

terms	{chediak, higashi, syndrome}	{chédiak, higashi, syndrome}
Bigram	$bigram_{chediak}$ = {ch,he,...,ak}	$bigram_{ch\&\#233;diak}$ = {ch,h&,...,ak}
	$bigram_{higashi}$ = {hi,ig,...,hi}	$bigram_{higashi}$ = {hi,ig,...,hi}
	$bigram_{syndrome}$ = {sy,yn,...,me}	$bigram_{syndrome}$ = {sy,yn,...,me}
$s(w_k, w_l)$	**$s(\mathbf{w}_{chediak},\mathbf{w}_{ch\&\#233;diak})$=0.471 {ch,di,ia,ak}**,	
	$s(\mathbf{w}_{higashi},\mathbf{w}_{higashi})$=1 {hi,ig,ga,as,sh,hi},	
	$s(\mathbf{w}_{syndrome},\mathbf{w}_{syndrome})$=1 {sy,yn,nd,dr,ro,om,me}.	
$s(term_p, term_q)$	(2*(0.471+1+1))/(3+3) = 0.824.	

follow that the word order is not important. The appearance similarity is utilized as one factor for the normalization.

Context Similarity. As described previously, the appearance similarity is not sufficient to find an original form. As a supplement, the context similarity is additionally considered. From the section 2.2, the context information of each term has been collected. This section measures context similarity between NT and VT. For the similarity, clue (co-occurrence noun) weights of NT are calculated by TF-IDF (Term Frequency-Inverse Document Frequency). Table 6 is an example (chediak higashi syndrome) of NT.

The weights about all clues of every NT are applied to context similarity which measures how many clues they share.

$$Context\_Similarity\,(term_{VT}, term_{NT}) = \Sigma weight\,(matched\_clue) \quad (5)$$

where *matched_clue* is a clue which appears with $term_{VT}$ and $term_{NT}$ together. Table 7 shows an example of context similarity measure of a VT and each element of its candidate set.

3 Experimental Evaluation

For evaluation of the normalization method, we chose 171 variant terms *(vt)* randomly and selected their normalized terms *(nt)* which have the term similarity

Table 6. An example of target term

Term(Occ.)	*Clues*	*Occ.*	*TF*	*IDF*	*TF − IDF*
Chediak higashi syndrome (308)	mk	9	0.029	2.478	0.072
	cell	103	0.334	0.686	0.229
	enzym	7	0.023	1.151	0.026
	protein	16	0.052	0.836	0.043
	studi	29	0.094	0.439	0.041
	blood	12	0.039	0.981	0.038
	clone	4	0.013	1.469	0.019
	phosphatas	4	0.013	1.722	0.022
	fetus	3	0.010	2.023	0.020
	cytotox	5	0.016	1.741	0.028
	...	...	...	...	...

Table 7. An example of target term

Variantterm	*Candidateterms*	*Contextsimilarity*
Chédiak	Chediak higashi syndrome	5.5815
higashi syndrome	hermansky pudlak syndrome	3.2683
	wiskott aldrich syndrome	2.4093
	naevoid basal cell carcinoma syndrome	2.9550
	...	...

(ts), by multiplying appearance similarity *(as)* and context similarity *(cs)*. In other words, we have prepared 171 pairs *(p)* which are expressed to *p(vt, nt, as, cs, ts)*. By manual evaluation, the result shows that 40 pairs (about 23.4%) are correctly normalized. As described in Section 2.1 for construction of NTS and VTS, all elements in VTS are not kinds of variant. In order to check the effectiveness of the *as* and the *cs*, we assign threshold values to each similarity from 0.1 to 0.9 and evaluate each result, totalling 81 results.

In the research, we deal with term normalization and it should be performed carefully like knowledge enrichment [6] because the result influences its application area. In other words, these kinds of research guarantee that the resulted data should be pure. Therefore, precision is the most important factor among evaluation methods. We evaluate our method and Table 8 shows the result in detail from this point of view.

In the table, *c_p* means count of pairs which remain after applying TV_AS. Table 8 summarizes the cases on 0.5 and 0.7 of TV_AS which attain the best performances on F1 and precision. In the case of F1, it could reach 76.7(%) when 0.5 and 0.1 are given to TV_AS and TV_CS respectively. However, the research was not designed to have a wrong result, but pursues perfect precision with the maximum count of right pairs *(vt, nt)*. Accordingly we could find the result having 19 correct pairs with 100(%) at TV_AS 0.7 and TV_CS 0.2. Through the evaluation, we could confirm that the method proposed in the paper has positive normalization. In future, by concentrating on terminology construction when processing large amounts of data, it could help extract a high quality of

Table 8. Performance evaluations on precision, recall, and F1 rates (TV_AS: threshold value of appearance similarity, TV_CS: threshold value of context similarity)

	TV_AS=0.5						TV_AS=0.7					
TV_CS	c_p	O	X	Pr.(%)	Re.(%)	F1(%)	c_p	O	X	Pr.(%)	Re.(%)	F1(%)
0.0	63	39	24	61.9	100	76.5	34	29	5	85.2	100	92.1
0.1	47	33	14	70.2	84.6	**76.7**	27	24	3	88.9	61.5	72.7
0.2	34	27	7	79.4	69.2	74.0	19	**19**	0	**100**	48.7	65.5
0.3	24	19	5	79.2	48.7	60.3	12	12	0	100	30.8	47.1
0.4	20	17	3	85.0	43.6	57.6	10	10	0	100	25.6	40.8
0.5	18	17	1	94.4	43.6	59.6	10	10	0	100	25.6	40.8
0.6	13	13	0	100	33.3	50.0	7	7	0	100	17.9	30.4
0.7	13	13	0	100	33.3	50.0	7	7	0	100	17.9	30.4
0.8	12	12	0	100	30.8	47.1	6	6	0	100	15.4	26.7
0.9	11	11	0	100	28.2	44.0	6	6	0	100	15.4	26.7

knowledge, because the normalization helps to prevent unnecessary information extraction. However the research depends on the term appearance, rather than the context information or its semantics. We will prepare another method for better performance.

4 Conclusion

This paper proposed a normalization method of term variations which is necessary in constructing knowledge from large amounts of data. To do this, we divided technical terms into a normalized term set (NTS) and a variant term set (VTS) through Wikipedia concept matching, constituted context information for each term, prepared candidate terms for original forms, and finally found normalized terms based on the appearance similarity *(as)* and the context similarity *(cs)*. In the experimental evaluation, we could have the maximum count of correct pairs of *vt* and *nt* under the condition of 0.7 and 0.2 of threshold values for *as* and *cs* respectively.

In automatic knowledge construction, term normalization is a significant requirement, to avoid generating unnecessary information. To this end, this research is expected to contribute to diverse fields of knowledge mining. However it still has a limitation, which is that it cannot find 'Vitamin C' as an original form from 'L ascorbate' or 'L ascorbic acid' because the work depends with more weight on the *as*. We will continue study for the solution which is based on semantics.

References

1. Dowdal, J., Rinaldi, F., Ibekwe-SanJuan, F., SanJuan, E.: Complex Structuring of Term Variants for Question Answering. In: Proc. of the ACM Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, vol. 18, pp. 1–8 (2003)

2. Ibekwe-Sanjuan, F.: Terminological Variation, a Means of Identifying Research Topics from Texts. In: Proc. of Intl. Conf. on Computational Linguistics, vol. 1, pp. 564–570 (1998)
3. Porter, M.F.: An algorithm for suffix stripping. J. of Program 14(3), 130–137 (1980)
4. Toutanova, K., Manning, C.: Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In: Proc. Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 63–70 (2000)
5. Hwang, M., Kim, P.: A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary. Intl. J. on Semantic Web and Information Systems (IJSWIS) 5(1), 48–64 (2009)
6. Hwang, M., Choi, C., Kim, P.: Automatic Enrichment of Semantic Relation Networks and its Application to Word Sense Disambiguation. IEEE Transactions on Knowledge and Data Engineering 23(6), 845–858 (2011)
7. Brank, J., Mladenic, D., Grobelnik, M., Milic-Frayling, N.: Feature Selection for the Classification of Large Document Collections. Journal of Universal Computer Science 14(10), 1562–1596 (2008)
8. Duong, T.H., Jo, G., Jung, J.J., Nguyen, N.T.: Complexity Analysis of Ontology Integration Methodologies: A Comparative Study. Journal of Universal Computer Science 15(4), 877–897 (2009)
9. Jung, J.J.: Semantic business process integration based on ontology alignment. Expert Systems with Applications 36(8), 11013–11020 (2009)
10. Hwang, M., Choi, D., Choi, J., Kim, H., Kim, P.: Similarity Measure for Semantic Document Interconnections. Information-An International Interdisciplinary Journal 13(2), 253–267 (2010)
11. Hwang, M., Choi, D., Kim, P.: A Method for Knowledge Base Enrichment using Wikipedia Document Information. Information-An International Interdisciplinary Journal 13(5), 1599–1612 (2010)
12. Bawakid, A., Oussalah, M.: Using features extracted from Wikipedia for the task of Word Sense Disambiguation. In: Proc. of IEEE Intl. Conf. on Cybernetic Intelligent Systems, pp. 1–6 (2010)
13. Fogarolli, A.: Word Sense Disambiguation Based on Wikipedia Link Structure. In: Proceedings of IEEE Intl. Conf. on Semantic Computing, pp. 77–82 (2009)