Automatic Term Recognition based on the Statistical Differences of Relative Frequencies in Different Corpora

Junko Kubo, Keita Tsuji, Shigeo Sugimoto

Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
E-mail: kuboj35@slis.tsukuba.ac.jp, keita@slis.tsukuba.ac.jp, sugimoto@slis.tsukuba.ac.jp

Abstract

In this paper, we propose a method for automatic term recognition (ATR) which uses the statistical differences of relative frequencies of terms in target domain corpus and elsewhere. Generally, the target terms appear more frequently in target domain corpus than in other domain corpora. Utilizing such characteristics will lead to the improvement of extraction performance. Most of the ATR methods proposed so far only use the target domain corpus and do not take such characteristics into account. For the extraction experiment, we used the abstracts of a women's studies journal as a target domain corpus and those of academic journals of 39 domains as other domain corpora. The women's studies terms which were used for extraction evaluation were manually identified terms in the abstracts. The extraction performance was analyzed and we found that our method outperformed earlier methods. The previous methods were based on C-value, FLR and methods which were also used with other domain corpora.

1. Introduction

Automatic Term Recognition (ATR) is useful for updating terminology, dictionaries and various information retrieval. While many methods have been proposed in the past, most of them utilize only the target domain corpus (i.e., the corpus which contains the target terms). The features of the terms are: **"**(a) they appear frequently in documents of a target field, (b) they are not common words in the target fields and (c) they appear less frequently in the corpora of other fields**"** (Uchimoto et al., 2000). Based on this idea, we propose an ATR method which utilizes the statistical difference of relative frequencies of the academic terms in the target corpus and in other domain corpora. We compared our ATR results with others and showed our ATR measure outperformed the others.

The idea of using target domain corpus and other domain corpora for ATR is not new (Drouin, 2003; Harter, 1975; Nagao et al., 1976; Anselmo et al., 2001; Uchimoto et al., 2000). The differences between the preceding methods and ours are: (1) the measures proposed in some of the preceding studies are purely heuristic and do not have a statistical basis. Our method uses "the difference between two proportions" in the statistics which follows the normal distribution. In that sense, the statistical meaning of the value of our measure that represents the termhood is easy to understand, (2) we use academic texts not the general texts (e.g., newspapers) as other domain corpora. Some earlier methods used newspaper texts as other domain corpora. In that case, words that frequently appear in the academic texts but are not domain specific terms (e.g., "research" and "evaluation") are likely to be extracted. Our method does not have such drawbacks, (3) we compared our ATR measures which use target domain corpus and other domain corpora with the existing measures which use only the target domain corpus. By comparing ATR measures, we showed our ATR measure outperformed the existing measures.

2. Data

Below, we explain the corpus and the target terms used for our experiment.

2.1 Corpus

We used the following two corpora, i.e., target domain corpus from which we extract terms and other domain corpora from which we obtain words which are not the terms used in the target domain:

(1) Target corpus: We used the abstract texts of the Women's Studies International Forum [1] as the target domain corpus (henceforth "WSIF corpus"). The reason we chose women's studies as the target domain is that the first author is familiar with it and can identify the target terms (the terms which should be extracted) for the corpus. The number of abstracts was 1,212. They were published between 1982 and 2007. The number of tokens was approximately 180,000.

(2) Other corpora: We used the abstracts of the 39 journals [2] (e.g. "Applied Economics", "International Journal of Public Administration", "Urban Studies", etc. For details, see the Appendix A) for the other domain corpora (henceforth "OTHER corpora"). They were published between 1981 and 2007. We randomly extracted 1,212 abstracts from each journal and used them

[1] http://www.sciencedirect.com/science/journal/02775395
[2] http://www.taylorandfrancisgroup.com/

for the experiment. The reason we used the same number of abstracts as WSIF is that the relative frequencies of words slightly depends on the text size and we should eliminate such size-dependencies. The number of tokens was approximately seven million in total.

These texts were POS tagged using Brill tagger and word sequences whose POS patterns were as follows:

{(Adjective | Noun)*Noun +} or
{(Adjective)*(Noun)+(Preposition)*(Noun)+}.

There were 20,266 candidate terms in the WSIF corpus and 630,541 candidate terms in the OTHER corpora.

2.2 Target Terms

We used the following "target terms" manually identified in WSIF corpus. The number of terms was 2,104.

3. Methods

In this section, we first explain our measure MDP to extract terms from the corpora. Next, we explain the existing measures UC, C-value and FLR. Candidate terms are extracted as terms in descending order of their scores and the extraction performance is compared.

3.1 MDP and MDP$_\alpha$

We propose the following MDP (**M**inimum of the **D**ifference between **P**opulation Proportions) to use for ATR:

$$\mathrm{MDP(T)} = \min_{1 \le i \le N} \mathrm{D}$$

$$D_i = \frac{\frac{f_0(T)}{W_0} - \frac{f_i(T)}{W_i}}{\sqrt{\pi_i(T)(1-\pi_i(T))\left(\frac{1}{W_0}+\frac{1}{W_i}\right)}}$$

where $f_0(T)$ represents the frequency of candidate T in the corpora of the target domain and $f_i(T)$ represents those in the other domains ($1 \leqq i \leqq 39$). W_0 and W_i represent the number of candidate terms in the corpora of the target domain and other i-th domains, respectively. $\pi_i(T)$ is defined as follows:

$$\pi_i(T) = \frac{f_0(T) + f_i(T)}{W_0 + W_i}$$

D_i follows a normal distribution.

MDP focuses only on the difference of frequencies in the target corpus and in other domain corpora. However, if the words frequently occur in the OTHER corpora, they are not likely to be terms no matter how large the difference was. MDP does not consider the words' frequency in the OTHER corpora. Based on this idea, we propose the measure MDP$_\alpha$.
MDP$_\alpha$ is defined as:

$$\mathrm{MDP}_\alpha(T) = \begin{cases} \mathrm{MDP}(T) & \text{(if } T \text{ does not occur more than } \alpha \text{ times in each of the OTHER corpora)} \\ 0 & \text{(otherwise)} \end{cases}$$

We will show an example to illustrate the calculation of MDP. Suppose that we extract "feminist" as T from a target corpus (women's studies) to get W_0 and the other domain corpora to get W_i ($1 \leqq i \leqq 3$).

The items listed below are the number of candidate terms (e.g. "W_0 = 20000" and "W_1 = 21000") and the frequency of "feminist" in the corpora (e.g. "f_0 (feminist) = 100" and "f_1 (feminist) = 15").

- W_0 = 20000, f_0 (feminist) = 100
- W_1 = 21000, f_1 (feminist) = 15
- W_2 = 20000, f_2 (feminist) = 30
- W_3 = 19000, f_3 (feminist) = 0

i = 1: $$\frac{\frac{100}{20000} - \frac{15}{21000}}{\sqrt{\left(\frac{115}{41000}\right)\left(1-\frac{115}{41000}\right)\left(\frac{1}{20000}+\frac{1}{21000}\right)}} = 8.20$$

i = 2: $$\frac{\frac{100}{20000} - \frac{30}{20000}}{\sqrt{\left(\frac{130}{40000}\right)\left(1-\frac{130}{40000}\right)\left(\frac{1}{20000}+\frac{1}{20000}\right)}} = 6.15$$

i = 3: $$\frac{\frac{100}{20000} - \frac{0}{19000}}{\sqrt{\left(\frac{100}{39000}\right)\left(1-\frac{100}{39000}\right)\left(\frac{1}{20000}+\frac{1}{19000}\right)}} = 9.76$$

The result of the calculation of MDP is 6.15 which is the minimum value. Suppose that we extract "article" as T from the corpora in the same way as in the above example. The number of candidate terms (i.e., W_0 and W_i) is the same as above.

- f_0 (article) = 20
- f_1 (article) = 30
- f_2 (article) = 75
- f_3 (article) = 15

i = 1: $$\frac{\frac{20}{20000} - \frac{30}{21000}}{\sqrt{\left(\frac{50}{41000}\right)\left(1-\frac{50}{41000}\right)\left(\frac{1}{20000}+\frac{1}{21000}\right)}} = -1.24$$

i = 2: $$\frac{\frac{20}{20000} - \frac{75}{20000}}{\sqrt{\left(\frac{95}{40000}\right)\left(1-\frac{95}{40000}\right)\left(\frac{1}{20000}+\frac{1}{20000}\right)}} = -5.65$$

$i = 3$: $$\frac{\frac{20}{20000} - \frac{15}{19000}}{\sqrt{\left(\frac{35}{39000}\right)\left(1-\frac{35}{39000}\right)\left(\frac{1}{20000}+\frac{1}{19000}\right)}} = 0.69$$

The result of the calculation of MDP is -5.65 which is the minimum value. We can see from these examples that the MDP value of the term "feminist" is higher than "article". Thus, it seems that "feminist" has a high probability of being a women's studies terms.

3.2 UC, C-value and FLR

We used the following C-value and FLR as the representative measures which use only the target domain corpus. In addition, we chose the measure by Uchimoto et al. (2000) (henceforth "UC") as representative one which uses not only the target domain corpus but also the other domain corpora.

UC (Uchimoto et al., 2000) is defined as:

$$UC(T) = TF \times \frac{TF}{DF} \times \left(\frac{1}{FF_i}\right)^3$$

where *TF* is the number of occurrences of *T* in the corpus of the target domain, and *DF* is the number of documents in the corpus of the target domain which contains *T*, and FF_i is the number of fields[3] which contain *T*.

We will show an example to illustrate the calculation of UC. We used the *T* and these frequencies are the same in the MDP example. We set DF of the target domain for UC example.

- *DF* (feminist) = 40
- *DF* (article) = 10

UC (feminist) $= 100 \cdot \frac{100}{40} \cdot \left(\frac{1}{3}\right)^3 = 9.26$

UC (article) $= 20 \cdot \frac{20}{10} \cdot \left(\frac{1}{4}\right)^3 = 0.63$

We can see from the example that the UC value of the term "feminist" is higher than "article". Therefore, it seems that "feminist" has a higher probability of being a women's studies term.

C-value (Frantzi et al. 2000) is defined as:

$$C\text{-}value(T) = \log_2 |T| \left(f(T) - \frac{t(T)}{c(T)} \right)$$

where $|T|$ is the number of constituent words of the term candidate *T*, *f(T)* is the number of times the term candidate *T* appeared in the corpus. *t(T)* is the frequency of occurrence of *T* in longer (already extracted as the above word formation) candidate terms, and *c(T)* is the number of those candidate terms.

We will show an example to illustrate the calculation of C-value. Suppose that we extract "feminist" and "feminist theory" as *T* from the target corpus. The items listed below are the candidate terms and these frequencies.

- socialist feminist theory (7)
- contemporary feminist theory (6)
- feminist theory book (2)
- feminist theory (30)
- feminist movement (20)
- feminist organization (6)
- feminist group (4)
- socialist feminist (10)
- radical feminist (15)
- feminist (100)

C-value (feminist) $= \log_2 1\left(200 - \frac{100}{9}\right) = 0$

C-value (feminist theory) $= \log_2 2\left(45 - \frac{15}{3}\right) = 40$

As known from the formula and the example, a single-word term's C-value comes to be 0. We can see from the example that the C-value of the term "feminist theory" is higher than "feminist". Therefore, it seems that "feminist theory" has a high probability of being a women's studies term.

FLR (Nakagawa et al. 2003) is defined as:

$$FLR(T) = f'(T)\left(\prod_{i=1}^{|T|}(FL(t_i)+1)(FR(t_i)+1)\right)^{\frac{1}{2|T|}}$$

where $f'(T)$ is the number of times the term candidate *T* appeared in the corpus as an independent phrase or compound (in other words, it is not included in a longer phrase or compound). t_i is the *i*-th constituent word of *T*, $FL(t_i)$ is the total number of adjoining nouns on the left side of *t*, and $FR(t_i)$ is the total number of adjoining nouns on the right side of *t*.

We will show an example to illustrate the calculation of FLR. We used the candidate terms and these frequencies are the same in the C-value example.

FLR (feminist) $= 100 \cdot ((38+1)\cdot(75+1))^{\frac{1}{2}} = 5444.26$

FLR (feminist theory) $= 30 \cdot ((38+1)\cdot(75+1)\cdot(45+1)\cdot(2+1))^{\frac{1}{4}}$
$= 758.68$

[3] The NACSIS database was partitioned into 59 field corpora according to the names of their academic societies.

We can see from the example that the FLR of the term "feminist" is higher than "feminist theory". Therefore, it seems that "feminist" has a higher probability of being a women's studies term.

4. Results and Discussions

We calculated the precision and recall of extracting terms based on the four ATR measures and MDP_{α}.

The precision is defined as (the number of extracted terms which were listed in target terms)/(the number of extracted words). The recall is defined as (the number of extracted terms which were listed in target terms)/(the number of terms which were listed in target terms and existed in the corpus).

4.1 MDP and other existing ATR measures

The precision and recall of extracting terms based on the three ATR measures (MDP, C-value and FLR) are shown in Figure 1. In Figure 1, we compare the existing measures which use only the target domain corpus with MDP which uses not only the target domain corpus but also the other domain corpora. In Figure 2, the precision and recall of extracting terms are based on the two ATR measures (MDP and UC) which use both the target domain corpus and the other domain corpora.

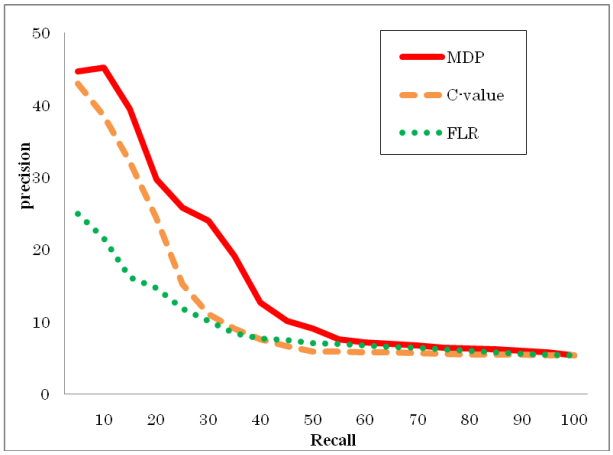

Figure 1: Precision and Recall for MDP, C-value and FLR

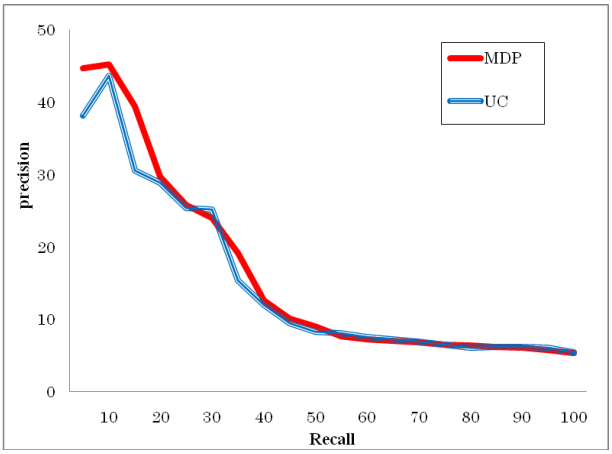

Figure 2: Precision and Recall for MDP and UC

Figure 1 shows that the precision of extraction based on MDP is much higher than that based on C-value and FLR. As we previously mentioned, MDP uses other domain corpora while C-value and FLR use target corpus only. From this result, we can say that using other domain corpora is effective for ATR. Figure 2 shows that the precision of extraction based on MDP is higher than that based on UC especially at the low recall. From these results, we showed MDP outperformed the earlier three measures.

Now, let us look at extraction errors. Table 2 (in Appendix B) shows the top twenty candidate terms extracted based on the MDP, UC, C-value and FLR. Two types of errors were observed in the terms which were extracted based on MDP and other measures. The first type of error is the wrongly extracted words which are common in the academic field but are not the domain specific terms. This type of error was frequently observed among the words which were extracted based on C-value ("case study", etc) and FLR ("paper", etc). On the other hand, these kinds of errors were rare among the words which were extracted based on MDP and UC. The reason is that MDP and UC assign high scores only to candidate terms which rarely occur in the OTHER corpora.

The second type of error is wrongly extracted common words such as "work" by C-value and "self" by MDP. The main reason is that MDP focuses only on the difference of frequencies in the target corpus and in other domain corpora and it does not take into consideration how often the words occur in the OTHER corpora, i.e., MDP does not consider the words frequency in the OTHER corpora. Based on this idea, we propose another measure, MDP_{α} which will be described in the next section.

4.2 MDP_{α}

We calculated the precision and recall of extracting terms based on the MDP_{α}. The results are shown in Figure 3.

Figure 3 shows that the precision of extraction based on MDP_5 is the highest especially at the low recall. Table 1 shows the top 10 candidate terms extracted based on the MDP and MDP_5. Table 3 (in Appendix B) shows the top twenty candidate terms extracted based on the MDP_{α}

We can see in Table 1 that MDP_5 failed to extract the core terms in the women's studies such as "woman" and "gender", while MDP could do that. However, note that we do not have to extract these core terms when terminological dictionaries for that domain already exist and these core terms are included in them. Therefore, we should choose between MDP and MDP_5 depending on the purpose, and whether we would like to extract both the core terms and newly-coined terms or just the latter.

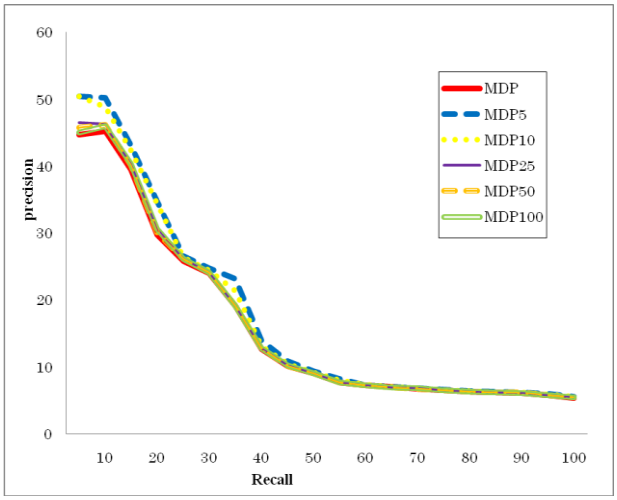

Figure 3: Precision and Recall for MDP_{α}

rank	***MDP***	***MDP$_5$***
1	woman	feminist
2	feminist	feminism
3	gender	feminist theory
4	feminism	struggle
5	self	motherhood
6	movement	femininity
7	article	rape
8	feminist theory	patriarchy
9	way	masculinity
10	lesbian	domestic violence

Table1: Top ten candidate terms extracted based on the MDP and MDP_5

5. Conclusion

In this paper, we showed that our ATR measure which uses MDP or MDP_5 outperformed the traditional measures based on UC, C-value and FLR. Unlike C-value and FLR, MDP (or MDP_5) needs other domain corpora in addition to the target domain corpus. The cost of obtaining such corpora is the drawback of MDP. However, the academic texts are now prevailing and it is getting easier to obtain corpora of various domains. We think that this problem is now being ameliorated.

Further work is needed to clarify the amount of the target corpus and the other domain corpora that is sufficient for extracting terms.

References

Anselmo, P., Felisa, V. and Julio, G. (2001). Corpus-based Terminology Extraction Applied to Information Access, In *Proceedings of the Corpus Linguistics 2001*, Lancaster, pp.458--465.

Drouin, P. (2003). Term Extraction using Non-technical Corpora as a Point of Leverage. *Terminology*, 9(1), pp.99--115.

Frantzi, K., Ananiadou, S. and Mima H. (2000). Automatic Recognition of Multi-word Terms: the C-value/NC-value Method, *International Journal on Digital Libraries*, 3(2), pp.115--130.

Harter, S. P. (1975). A Probabilistic Approach to Automatic Keyword Indexing – Part I. On the Distribution of Specialty Words in a Technical Literature, *Journal of the American Society for Information Science*, 26(4), pp.197--206.

Harter, S. P. (1975). A Probabilistic Approach to Automatic Keyword Indexing – Part II. An Algorithm for Probabilistic Indexing, *Journal of the American Society for Information Science*, 26(5), pp.280--289.

Nagao, M., Mizutani, M. and Ikeda, H. (1976). An Automatic Method of the Extraction of Important Words from Japanese Scientific Documents, *Journal of Information Processing Society of Japan*, 17(2), pp.110--117.

Nakagawa, H., Mori, T. and Yumoto, H. (2003). Term Extraction Based on Occurrence and Concatenation Frequency. *Journal of Natural Language Processing*, 10(1), pp.27--45.

Uchimoto, K., Sekine, S., Murata, M., Ozaku, H. and Isahara, H. (2000) Term Recognition Using Corpora from Different Fields, *Terminology*, 6(2), pp.233--256.

Appendix A

The list of journals used as "OTHER corpora" is shown below.

- Aerosol Science and Technology
- Applied Economics
- Australian Journal of Earth Sciences
- Avian Pathology
- British Poultry Science
- Clinical Toxicology
- Current Eye Research
- Early Child Development and Care
- Educational Gerontology
- Electric Power Components and Systems
- Energy Sources, Part A: Recovery, Utilization, and Environmental Effects
- Ergonomics
- Ferroelectrics
- International Journal of Computer Mathematics
- International Journal of Control
- International Journal of Electronics
- International Journal of Mathematical Education in Science and Technology
- International Journal of Neuroscience
- International Journal of Polymeric Materials
- International Journal of Public Administration
- International Journal of Radiation Biology
- International Journal of Remote Sensing
- Journal of Applied Statistics
- Journal of Clinical and Experimental Neuropsychology
- Journal of Macromolecular Science, Part A
- Journal of Modern Optics
- Journal of Natural History
- Journal of Statistical Computation and Simulation
- Medical Mycology
- Medical Teacher
- Molecular Physics
- Pathology
- Petroleum Science and Technology
- Phase Transitions
- Radiation Effects and Defects in Solids
- Systems Biology in Reproductive Medicine
- The Journal of Adhesion
- Urban Studies
- Xenobiotica

Appendix B

rank	***MDP***	***UC***	***C-value***	***FLR***
1	woman	feminist	feminist theory	woman
2	feminist	hijab	violence against woman	feminist
3	gender	motherhood	United States	study
4	feminism	masculinity	higher education	gender
5	self	rape	feminist research	article
6	movement	patriarchy	young woman	research
7	article	Mari	human right	work
8	feminist theory	BFM	black woman	paper
9	way	feminist perspective	group of woman	state
10	lesbian	feminist research	domestic violence	experience
11	struggle	peace movement	feminist perspective	group of woman
12	young woman	violence against woman	South Africa	movement
13	motherhood	feminist objectivity	world war	role of woman
14	violence	meaning of home	American woman	world
15	identity	cosmetic surgery	case study	class
16	femininity	migrant woman	number of woman	family
17	rape	study student	same time	experience of woman
18	body	symbolic violence	role of woman	young woman
19	patriarchy	gender justice	division of labour	American woman
20	masculinity	virginity	Muslim woman	status of woman

Table 2: Top twenty candidate terms extracted based on the MDP, UC, C-value and FLR

rank	MDP_5	MDP_{10}	MDP_{25}	MDP_{50}	MDP_{100}
1	feminist	feminist	feminist	feminist	woman
2	feminism	feminism	feminism	gender	feminist
3	feminist theory	feminist theory	self	feminism	gender
4	struggle	lesbian	feminist theory	self	feminism
5	motherhood	struggle	lesbian	feminist theory	self
6	femininity	young woman	struggle	lesbian	movement
7	rape	motherhood	young woman	struggle	feminist theory
8	patriarchy	violence	motherhood	young woman	way
9	masculinity	femininity	violence	motherhood	lesbian
10	domestic violence	rape	identity	violence	struggle
11	feminist perspective	patriarchy	femininity	identity	young woman
12	Irish woman	masculinity	rape	femininity	motherhood
13	American woman	domestic violence	patriarchy	rape	violence
14	academy	feminist perspective	masculinity	body	identity
15	oppression	Irish woman	domestic violence	patriarchy	femininity
16	feminist research	American woman	feminist perspective	masculinity	rape
17	post	academy	Irish woman	domestic violence	body
18	diary	oppression	American woman	feminist perspective	patriarchy
19	black woman	feminist research	academy	Irish woman	masculinity
20	nationalism	post	oppression	American woman	domestic violence

Table 3: Top twenty candidate terms extracted based on the MDP_5, MDP_{10}, MDP_{25}, MDP_{50} and MDP_{100}