Clustering acronyms in biomedical text for disambiguation

Naoaki Okazaki* **and Sophia Ananiadou**†‡

* Graduate School of Information Science and Technolog, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
okazaki@mi.ci.i.u-tokyo.ac.jp
† National Centre for Text Mining
‡ School of Informatics, University of Manchester
PO Box 88, Sackville Street, Manchester, M60 1QD, United Kingdom.
Sophia.Ananiadou@manchester.ac.uk

Abstract

Given the increasing number of neologisms in biomedicine (names of genes, diseases, molecules, etc.), the rate of acronyms used in literature also increases. Existing acronym dictionaries cannot keep up with the rate of new creations. Thus, discovering and disambiguating acronyms and their expanded forms are essential aspects of text mining and terminology management. We present a method for clustering long forms identified by an acronym recognition method. Applying the acronym recognition method to MEDLINE abstracts, we obtained a list of short/long forms. The recognized short/long forms were classified by a biologist to construct an evaluation set for clustering sets of similar long forms. We observed five types of term variation in the evaluation set and defined four similarity measures to gathers the similar long forms (i.e., orthographic, morphological, syntactic, lexico semantic variants, nested abbreviations). The complete-link clustering with the four similarity measures achieved 87.5% precision and 84.9% recall on the evaluation set.

1. Introduction

Given the increasing number of neologisms in biomedicine (names of genes, diseases, molecules, etc.), the rate of acronyms used in literature also increases. It has been reported that around 64,000 new acronyms have been introduced in 2004 (Chang and Schütze, 2006). Existing acronym dictionaries cannot keep up with the rate of new creations. The number of acronyms existing currently only in MEDLINE demands automated methods for their identification, disambiguation and management.

Acronyms are compressed forms of terms, and are used as substitutes of the fully expanded termforms. An acronym is also referred as a short form (e.g. *HMM*) having a long or expanded form, also called its definition (e.g. *hidden markov model*). A recent study (Wren et al., 2005) reported that only 25% of documents relevant to the concept *c-jun N-terminal kinase* could be retrieved by using the full form, as in more than 33% of the documents the concept is referred to by using its acronym *JNK*. In this way, search engines using acronyms rather than just full forms achieve better performance.

Thus, discovering acronyms and relating them to their expanded forms is an essential aspect of text mining and terminology management. Acronym identification deals with extracting pairs of short and long forms occurring in text. Research has been devoted into the construction of acronym databases and their expanded forms. However it is almost impossible to ensure completeness in coverage given that not all biomedical texts are publicly available. Most research has focused on methods for the recognition of acronyms and their expanded forms (or definitions) from running text (Adar, 2004; Pustejovsky et al., 2001; Schwartz and Hearst, 2003). Another approach used machine learning techniques to generate automatically acronyms from expanded forms (Tsuruoka et al., 2005) to overcome the problem of acronym-definition databases.

Acronyms are ambiguous, i.e. the same acronym may refer to different concepts (*GR* is an abbreviation for both *glucocorticoid receptor* and *glutathione reductase*). In order to deal with ambiguity, automatic merging of long forms using n-gram and context similarities has been proposed (Gaudan et al., 2005). Acronyms also have variant forms, i.e. the same term may have several acronyms (e.g. *NF kappa B*, *NF kB*). Both phenomena present substantial challenges for terminology management and for text mining.

In this paper, we present a method for clustering long forms identified by an acronym-recognition method. Figure 1 shows the outline of the method. Applying an acronym-recognition method to biomedical documents (e.g., MEDLINE abstracts), we obtain a list of the short/long forms identified in the text. Given a list of long forms for an acronym, the proposed method gathers similar long forms (e.g., orthographic, morphological variants, synonyms, etc.) long forms into a cluster.

2. Methodology

Most acronym-recognition methods make use of letter matching of the expressions appearing near parentheses to identify short/long form candidates. Letter matching methods would extract an acronym-definition pair ⟨HMM, hidden markov model⟩ from a text such as, "We used hidden markov model (HMM) to capture the patterns of acronym generation," by searching for letters 'h', 'm', and 'm' before the expression '(HMM)'. Schwartz and Hearst (2003) proposed an algorithm for identifying acronyms by using parenthetical expressions as a marker of a short form. Long form candidates were extracted by estimating the maximum number of words for a short form. Then the algorithm applied a character matching technique, i.e., all letters and digits in a short form had to appear in the corresponding long form in the same order. Even though the core algorithm was very simple, the authors reported

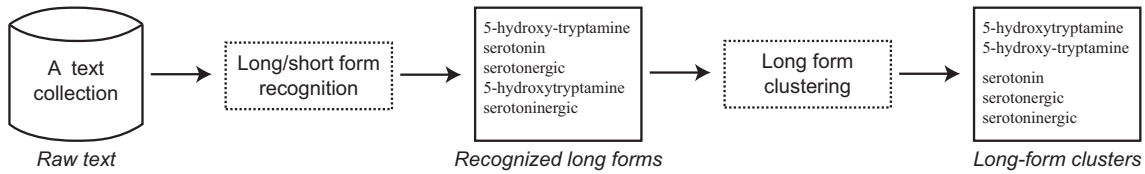

Figure 1: Acronym recognition and long-form clustering

99% precision and 84% recall on the Medstract gold standard[1].

However, the letter-matching approach cannot recognize a short/long form pair that has a synonymous relation expressed by parenthesis, e.g., *serotonin (5-HT)*. Hisamitsu and Niwa (2001) proposed a method for extracting useful parenthetical expressions from Japanese newspaper articles. Their method measured the co-occurrence strength between the inner and outer phrases of a parenthetical expression by using statistical measures such as mutual information, χ^2 test with Yate's correction, Dice coefficient, log-likelihood ratio, etc. Although their method did not focus on the acronym-definition relation, it dealt with generic parenthetical expressions (e.g., abbreviation, non abbreviation paraphrase, supplementary comments, etc.). In short, the statistical approach can extract a short/long form pair that has, for example, the synonymous relation expressed by parenthesis as long as it co-occurs frequently in the source text.

Since the aim of this study is to identify clusters from a set of long forms corresponding to an acronym, it is preferable that the evaluation data contains long forms recognized by unidentified relations (e.g., acronym-definition, synonym, paraphrase, etc.). For our experiments we have applied the C-value method (Frantzi and Ananiadou, 1999) to a set of expressions appearing before a specific short form, to extract short/long form pairs. The C-value method is a domain-independent method for automatic term recognition based on linguistic and statistical information. The linguistic analysis enumerates all potential terms appearing before parenthesis expressions of a specific acronym in a given text. The statistical analysis assigns a termhood (likelihood to be a long form) to a candidate long-form by using the following features: the frequency of occurrence of the candidate term; the frequency of the candidate term as part of other longer candidate terms; the number of these longer candidate terms; and the length of the candidate term.

Given a long-form candidate, w, the C-value termhood function $\mathrm{CV}(w)$ is defined in the following formula,

$$\mathrm{CV}(w) = \log_2 [\mathrm{len}(w)] \cdot \mathrm{freq}(w) - \frac{\sum_{t \in T_w} \mathrm{freq}(t)}{|T_w|}. \quad (1)$$

Therein: w is a candidate term; $\mathrm{freq}(x)$ denotes the frequency of occurrence of term x; $\mathrm{len}(x)$ denotes the length (number of words) of term x; T_w is a set of candidate terms which contain term w (*nested terms*); and $|T_w|$ represents the number of such candidate terms T_w. The C-value approach is characterized by the extraction of nested terms which gives preference to terms appearing frequently in a given text but not as a part of specific longer terms. This is a desirable feature for acronym recognition to identify long-form candidates in contextual sentences.

[1] `http://www.medstract.org/`

Acronym	# long-forms	# clusters
CD	29	29
PC	28	25
CI	21	16
RT	19	17
PG	18	18
MAP	18	15
CT	15	9
LH	13	9
HR	13	12
TPA	12	2
SD	12	12
BP	12	9
CAT	11	6
ER	10	7
...	..	..

Table 1: Results of the long forms in the evaluation set

After having selected 50 acronyms used frequently in MEDLINE abstracts, we applied the C-value method to extract terms which co-occur frequently with the parenthetical expression of the acronyms. Terms with their termhood score higher than 10 were chosen as the long forms. After having revised manually the recognition result to erase the misrecognized long-forms, we obtained 50 sets of long forms each of which corresponds to an acronym. Finally, we asked a biologist to classify the long forms for each acronym. Table 1 shows the example of acronyms, the number of unique long forms, and the number of clusters identified by the biologist. Figure 2 shows the long forms for the acronyms *CAT* and *LPS*.

3. Clustering long forms

Nenadic et al. (2004) described different types of term variations in the context of term normalization as an integral part of the automatic term recognition. They classified the types of term variations into four: *orthographical*, *morphological*, *syntactic*, *lexico-semantic* variations. We observed the following five types of term variation in the evaluation set.

1. *Orthographic* variation includes optional usage of hyphens (e.g., *5-hydroxy-tryptamine* and *5-hydroxytryptamine*), different Latin/Greek transcriptions (e.g., *oestrogen receptor* and *estrogen receptor*),

CAT

- computed axial tomography; computerized axial tomographic; computerised axial tomography; computerized axial tomography; computer assisted tomography
- chloramphenicol acetyltransferase; chloramphenicol acetyl transferase
- catalase
- carnitine acetyltransferase
- choline acetyltransferase
- total calcium

LPS

- lipopolysaccharide; endotoxin
- late potentials

Figure 2: The except of the evaluation corpus

and different spelling usages (e.g., *computerised tomography* and *computerized tomography*).

2. *Morphological* variation includes the usage of plural or singular nouns (e.g., *body mass index* and *body mass indices*) and the different usage of adjectives and nouns (e.g., *computerized tomographic* and *computerized tomography*).

3. *Syntactic* variation includes structural differences such as use of possessives and nouns (e.g., *amygdaloid central nucleus* and *central nucleus of the amygdala*) and ordinal differences of words (e.g., *human immunodeficiency virus type 1* and *type 1 human immunodeficiency virus*).

4. *Nested abbreviations* are found even in long forms for acronyms, e.g., *angiotensin-converting enzyme* and *ang i-converting enzyme*); *human immunodeficiency virus type 1* and *hiv type 1*; *systemic lupus erythematosis* and *systemic le*.

5. *Lexico semantic* variation includes the use of synonyms in the process of assigning names to concepts, e.g., *lipopolysaccharide* and *endotoxin*; *5-hydroxytryptamine* and *serotonin*; *arterial pressure*, *blood pressure* and *response time*, *reaction time*.

An approach dealing with variation types 1 and 2 is to write a set of conversion rules such as: "remove spaces and hyphens;" "assume letter *z* as *s*;" and "replace letters *oe* with *e*." In this paper, we use the cosine similarity between two strings t_i and t_j to capture variation types 1 and 2. Letter n-gram similarity $\mathrm{sim}_{\mathrm{ch}}(t_i, t_j)$ between two terms, t_i and t_j, is calculated as follows:

$$\mathrm{sim}_{\mathrm{ch}}(t_i, t_j) = \frac{1}{k}\sum_{n=1}^{k} \frac{|\text{n-gram}_{\mathrm{ch}}(t_i) \cap \text{n-gram}_{\mathrm{ch}}(t_j)|}{\sqrt{|\text{n-gram}_{\mathrm{ch}}(t_i)||\text{n-gram}_{\mathrm{ch}}(t_j)|}}. \quad (2)$$

Therein: $\text{n-gram}_{\mathrm{ch}}(t_i)$ is a set of letter n-grams generated from term t_i; and k is a parameter to determine the maximum order of n-gram calculation (i.e., uni-gram, bi-gram, tri-gram, ..., k-gram). Similarity $\mathrm{sim}_{\mathrm{ch}}(t_i, t_j)$ assesses the concordance of letters or letter sequences in the two terms.
An approach dealing with variation type 3 is to write a set of conversion rules such as "X of Y" → "Y X". Since it is difficult to define a comprehensive set of rules, we measure the concordance of words or word sequences in the two terms. We use the cosine similarity between two strings t_i and t_j in word n-grams,

$$\mathrm{sim}_{\mathrm{wd}}(t_i, t_j) = \frac{1}{k}\sum_{n=1}^{k} \frac{|\text{n-gram}_{\mathrm{wd}}(t_i) \cap \text{n-gram}_{\mathrm{wd}}(t_j)|}{\sqrt{|\text{n-gram}_{\mathrm{wd}}(t_i)||\text{n-gram}_{\mathrm{wd}}(t_j)|}}. \quad (3)$$

Therein, $\text{n-gram}_{\mathrm{wd}}(t_i)$ is a set of word n-grams generated from term t_i.
A straightforward approach dealing with variation type 4 is to use an acronym dictionary for generating potential abbreviations, e.g., *iotensin* → *i* and *human immunodeficiency virus* → *hiv*. However, it is difficult to prepare such a comprehensive dictionary in advance. Moreover, we need to estimate the probability that two strings have the acronym-definition relation (e.g., Tsuruoka et al., 2005) to determine the position of the expressions to be abbreviated. For simplicity, we use the overlap coefficient (Manning and Schütze, 1999) between two strings t_i and t_j (in letters),

$$\mathrm{sim}_{\mathrm{ov}}(t_i, t_j) = \frac{|t_i \cap t_j|}{\min(|t_i|, |t_j|)}. \quad (4)$$

Therein, $t_i \cap t_j$ represents the number of letters appearing in the terms t_i and t_j; and $|t_i|$ denotes the number of letters in term t_i.
We cannot deal with variation type 5 only by examining letters in two terms. Hence, we define contextual similarity $\mathrm{sim}_{\mathrm{cont}}(t_i, t_j)$ between two terms, t_i and t_j, which measures how terms appearing around the two terms are similar in text. We define *context sentence* for a short/long pair as the exact sentence in which the pair appears. We collect context sentences for each short/long pair and apply the C-value method to enumerate multi-word terms in the context. Thus, we create a context vector w_i that consists of context terms and their weights calculated by the C-value method. We define contextual similarity $\mathrm{sim}_{\mathrm{cont}}(t_i, t_j)$ as the cosine coefficient of context vectors w_i and w_j:

$$\mathrm{sim}_{\mathrm{cont}}(t_i, t_j) = \frac{w_i \cdot w_j}{|w_i||w_j|} \quad (5)$$

If terms appearing around the two terms are similar in the text, $\mathrm{sim}_{\mathrm{cont}}(t_i, t_j)$ will be greater.
Finally we combine four similarity metrics as a liner combination:

$$\mathrm{d}(t_i, t_j) \quad = \quad 1 - \mathrm{sim}(t_i, t_j) \quad (6)$$

α	β	γ	θ	δ	F-measure
0.7	0.2	0.0	0.1	0.3	0.862
0.7	0.1	0.1	0.1	0.3	0.851
0.7	0.2	0.1	0.0	0.3	0.847
....	....	....	....	....	....

Table 2: The optimal combination of parameters.

$$\begin{aligned} \mathrm{sim}(t_i, t_j) &= \alpha \mathrm{sim}_{\mathrm{ch}}(t_i, t_j) + \beta \mathrm{sim}_{\mathrm{word}}(t_i, t_j) \\ &\quad + \gamma \mathrm{sim}_{\mathrm{ov}}(t_i, t_j) + \delta \mathrm{sim}_{\mathrm{cont}}(t_i, t_j) \quad (7) \\ &\quad \alpha + \beta + \gamma + \delta = 1 \quad (8) \end{aligned}$$

Therein: $\mathrm{d}(t_i, t_j)$ is the distance (dissimilarity) of two terms t_i and t_j; and α, β, γ, and δ are positive values satisfying Equation 8. Using the distance function $\mathrm{d}(t_i, t_j)$, we applied complete link clustering with a given threshold θ, since the number of clusters is unknown in advance. Parameter n (the maximum order of n-gram) was set to three, i.e., uni-gram, bi-gram, and tri-gram are used for calculating $\mathrm{sim}_{\mathrm{ch}}(t_i, t_j)$ and $\mathrm{sim}_{\mathrm{word}}(t_i, t_j)$.

4. Evaluation

Using the evaluation set as a gold standard, we measured the quality of clustered descriptions obtained by the proposed method. Calculate F-measure scores from precision and recall (Anquetil et al., 1999), we maximize the F-measure score by searching for the optimal combination of parameters, α, β, γ, and θ (note that $\delta = 1 - \alpha - \beta - \gamma$). Table 2 shows F-measure scores over the different parameters. The proposed method with,

$$(\alpha, \beta, \gamma, \delta, \theta) = (0.7, 0.2, 0.0, 0.1), \quad (9)$$

yields the best clustering result (F-measure = 0.862; precision = 0.875; recall = 0.849). This result indicated that the cosine similarity between two terms in letter n-grams was dominant in the clustering process. The overlap coefficient between two strings γ could not contribute to the clustering process due to the small number of instances (five instances) corresponding to variation type 4. The result also showed that the effect of the context similarity for disambiguation of long forms.

5. Conclusion

In this paper, we presented a method for clustering long forms identified by an acronym-recognition method. We observed five types of term variation in the evaluation set and defined four similarity measures to gathers long forms (e.g., orthographic, morphological variants, synonyms, etc.) into a cluster. The proposed method achieved 87.5% precision and 84.9% recall on our evaluation set.

A future direction of this study would be to distinguish orthographic and morphological (types 1 and 2) variations from other types of variations. Although this study dealt with the different types of variations in the liner combination of the similarity measures, we found that variation types 1 and 2 were dominant in the long forms identified by the acronym-recognition method. We need to explore a methodology for each type of variation independently and to construct its evaluation set.

Acknowledgements

The U.K. National Centre for Text Mining is funded by the Joint Information Systems Committee, the Biotechnology and Biological Sciences Research Council, and the Engineering and Physical Sciences Research Council.

6. References

Eytan Adar. 2004. SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics*, 20(4):527–533.

Nicolas Anquetil, Cedric Fourrier, and Timothy C. Lethbridge. 1999. Experiments with hierarchical clustering algorithms as software remodularization methods. In *Proceedings of the International Workshop on Program Comprehension*, pages 235–255.

Jeffrey T. Chang and Hinrich Schütze. 2006. Abbreviations in biomedical text. In S. Ananiadou and J. McNaught, editors, *Text Mining for Biology and Biomedicine*, pages 99–119. Artech House, Inc.

Katerina T. Frantzi and Sophia Ananiadou. 1999. The C-value / NC-value domain independent method for multiword term extraction. *Journal of Natural Language Processing*, 6(3):145–179.

Sylvain Gaudan, Harald Kirsch, and Dietrich Rebholz-Schuhmann. 2005. Resolving abbreviations to their senses in medline. *Bioinformatics*, 21(18):3658–3664.

Toru Hisamitsu and Yoshiki Niwa. 2001. Extracting useful terms from parenthetical expression by combining simple rules and statistical measures: A comparative evaluation of bigram statistics. In Didier Bourigault, Christian Jacquemin, and Marie-C L'Homme, editors, *Recent Advances in Computational Terminology*, pages 209–224. John Benjamins.

Christopher. D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, London, England.

Goran Nenadic, Sophia Ananiadou, and John McNaught. 2004. Enhancing automatic term recognition through term variation. In *Proceedings of 20th International Conference on Computational Linguistics (COLING 2004)*, pages 604–610.

James Pustejovsky, José Castaño, Brent Cochran, Maciej Kotecki, and Michael Morrell. 2001. Automatic extraction of acronym meaning pairs from MEDLINE databases. *MEDINFO 2001*, pages 371–375.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing (PSB 2003)*, number 8, pages 451–462.

Yoshimasa Tsuruoka, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. A machine learning approach to acronym generation. In *BioLink 2005, Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, ACL/ISMB*, pages 25–31.

Jonathan D. Wren, Jeffrey T. Chang, James Pustejovsky, Eytan Adar, Harold R. Garner, and Russ B. Altman. 2005. Biomedical term mapping databases. *Database Issue*, 33:D289–D293.