PubTermVariants: biomedical term variants and their use for PubMed search

Lana Yeganova, Won Kim, Sun Kim, Rezarta Islamaj Doğan, Wanli Liu,
Donald C Comeau, Zhiyong Lu, W John Wilbur
National Center for Biotechnology Information, NLM, NIH, Bethesda, MD, USA
{yeganova, wonkim, sun.kim, islamaj, liu15,
comeau, luzh, wilbur}@mail.nih.gov

Abstract

Term normalization is frequently used in information retrieval task to reduce variant word forms to a common form. The most general term normalization technique used in practice is stemming, however it has been found to not be completely reliable. Here we present PubTermVariants, a high-quality data-driven resource of term variant pairs that can improve search results in PubMed. For a given pair, we consider two terms to be variants if they stem to the same form, pass the hypergeometric test, and pass the morpho-semantic test. We perform manual evaluation of a subset of PubTermVariants that confirms the high quality of the candidate pairs. We further present experiments that demonstrate their usefulness for PubMed search.

1 Introduction

Information retrieval, and biomedical text processing in general, profoundly depend on sensitive techniques for term normalization. Frequently, the link between a query and a document is not established because they use different forms of a term. These differences may be morphological (related by derivation or inflection) variations of a word, (e.g. autoimmune, autoimmunities, autoimmunity), synonyms (e.g. kidney disease and renal disease), abbreviations, etc. It is to the problem of morphological term variations that we wish to give attention here. Specifically, the goal of this study is to find pairs of string variants that have the same meaning and when used interchangeably benefit PubMed search.

Stemming (Porter 1980) is frequently used for the string normalization task to conflate different forms of a word that have the same meaning. This has been found useful in the task of information retrieval and has been shown to yield small improvements on typical test collections (Hull 1996, Hollink, Kamps et al. 2004, Manning, Raghavan et al. 2009, Moral, de Antonio et al. 2014). Since stemming is not completely reliable, different methods have been applied in an attempt to improve the final results of stemming such as limiting the results to forms found together in a lexicon (Krovetz 1993). This latter approach however is quite limiting and (Xu and Croft 1998) developed a method that uses a mutual information measure of the co-occurrence of two word forms to estimate how related they are and to put them in the same equivalence class if the information is above a threshold. This is done with the aim of improving the equivalence classes of forms with a common stem produced by the original application of the stemmer.

A study of morpho-semantic relationships in Medline (Wilbur and Smith 2013) identifies morphologically related tokens in Medline by using character n-grams as features and then computes the probability that two strings are related based on the context. This approach infers the morphological relatedness of two strings in a way more general than

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 141–145,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

stemming, but based on certain substrings of characters on which they match.

We take what we would describe as a more pragmatic approach. We define two terms to be variants of each other if they can be used interchangeably at the query stage. With the goal to obtain a reliable list of variants, we first find pairs of terms that stem to the same form and have common document in which they appear. We then apply the hypergeometric (HG) test (Larson 1982) to decide whether the observed co-occurrence for a particular pair of terms is above random. For the pairs that pass the HG test, we compute the morpho-semantic similarity score following (Wilbur and Smith 2013) and only retain the pairs that score above the threshold.

When all these conditions hold, we view the two forms as a good candidate term variant pair. We believe this is a less aggressive and clearly safer way to use stemming for query expansion that results in a conservative list of term variants.

In the next section we provide more details on how we generate term variants. We then present the results of manual evaluation of a randomly selected set of candidate term variant pairs. Further we describe two experiments that reveal how PubMed document retrieval is affected when term variants are used. In these experiments we consider both zero-result and nonzero-result queries.

2 Computing Term Variants

Stemming. To begin our processing, we first extracted space-separated tokens that appeared in ten or more PubMed articles. We stemmed every token with the Porter stemmer (Porter 1980) and collected pairs of tokens with the same stem. This process resulted in 201,219 unique pairs of tokens.

The Hypergeometric Test. Here we used the hypergeometric distribution and the p-value test for every pair of words in a group. Let $\boldsymbol{N_s}$ and $\boldsymbol{N_t}$ be the number of documents in Medline that contain terms s and t respectively, let $\boldsymbol{N}$ be the size of Medline, and $\boldsymbol{N_{st}}$ be the number of documents in Medline containing both terms s and t. The random variable Y representing a number of documents containing both terms s and t is a hypergeometric random variable with parameters $\boldsymbol{N_s}$, $\boldsymbol{N_t}$ and N (Larson 1982) if s is randomly assigned to the $\boldsymbol{N_s}$ documents. The probability distribution of Y is:

$$\boldsymbol{P(y)} = \binom{\boldsymbol{N_t}}{\boldsymbol{y}}\binom{\boldsymbol{N - N_t}}{\boldsymbol{N_t - y}}\Big/\binom{\boldsymbol{N}}{\boldsymbol{N_s}}.$$

From N_{st} we compute the p-value, i.e. the probability of the observed N_{st} or a higher frequency arising by chance as follows:

$$\text{p-value} = \sum_{y=N_{st}}^{\min(N_s,N_t)} \boldsymbol{P(y)}.$$

The p-value reflects the significance of two words co-occurring in N_{st} documents given the frequencies of individual words and the size of the database. A low p-value indicates that the co-occurrence of the two words in not likely to be by chance, but because the words are closely related. By applying the HG test to these pairs at the 0.01 level, we obtain 124,548 word pairs that we will refer to as set StemHG.

The Morpho-Semantic Analysis. Finally, we make use of the study of morpho-semantic relationships in Medline (Wilbur and Smith 2013). For a candidate pair of strings the approach assigns a probability that the strings are semantically related. Using every candidate pair from the above 124,548 pairs, we retain only the pairs for which the probability of being related is 0.9 or higher. This analysis results in a collection of 82,216 term variant pairs that we will refer to as PubTermVariants. There are about 109,725K unique terms in PubTermVariants, since a term may be paired with multiple variants. Because of the HG and the morpho-semantic tests, the relationships between term variants are not transitive.

In the next section we confirm that this collection is of high quality by manual evaluation of a random sample and present experiments designed to demonstrate their usefulness.

3 Experimental Evaluation and Results

We evaluate the quality of PubTermVariants by performing manual evaluation of random pairs sampled from the collection. The manual evaluation reveals the high quality of the variants in the collection. Our further experiments are designed to prove their usefulness for PubMed search.

3.1 Manual Evaluation

Here we report a manual analysis of PubTermVariants with the goal to confirm that two variants are indeed word forms that carry the same meaning and can be safely interchanged in a query.

As mentioned earlier, PubTermVariants is a collection of 82,216 candidate term variant pairs. In addition to PubTermVariants we have 42,332 term pairs in the set StemHG\PubTermVariants that potentially may be enriched in term variants. We believe that the quality of term variants in PubTermVariants is attributable to the effect of independently applying different statistical methods.

We assessed the quality of proposed term variant pairs by manually evaluating 200 random pairs from PubTermVariants, as well as 200 random pairs from StemHG\PubTermVariants. The 400 pairs were shuffled and each pair presented to two annotators. Eight annotators reviewed 100 pairs each, so that each pair was evaluated by two different people. The annotators involved in the manual evaluation all have backgrounds in biomedical information retrieval.

A web-based tool was developed to carry out this evaluation process and this tool was designed to show the term pair, two PubMed abstracts that contained one variant but not the other, and one PubMed abstract that contained both word forms. The annotators were asked to judge whether the two word forms could be used interchangeably. This decision was made by judging the displayed abstracts and deciding whether all of them should be retrieved regardless of which term is being used.

At this round pairs of annotators agreed on 329 of 400 instances considered. All individual evaluations were compared and each pair of annotators met separately to discuss the discrepancies on the remaining 71 pairs. This was later followed by a meeting where all annotators were present, and all remaining cases were discussed. The annotation experiment found that 89% of pairs in PubTermVariants were true variants of the same concept, while only 81.5% of pairs in StemHG\PubTermVariants were true variants, presented in Table 1.

We further examined the quality of term variants as a function of token length, as shown in Figures 1 and 2. We find that tokens of length 3 are typically abbreviations and therefore not good term variant candidates in the absence of context. For example, a pair of terms *ohd* and *ohds* is labeled negative, because, while *ohd*/*ohds* could be used interchangeably as singular and plural forms of the abbreviation for "*occupational health departments*", *ohds* may also stand for "*hydroxylase deficiency syndrome*". Consequently, of 22 pairs from PubTermVariants that were labeled negative 12 pairs include 3 letter abbreviations. We also observe that the distribution of errors in StemHG\PubTermVariants is more uniform as a function of string length.

3.2 Effect of Term Variants on PubMed Search.

With the goal to understand the usefulness of these term pairs in a real-world setting, we examined the queries in PubMed logs and performed the following analyses:

1. We analyzed zero-result queries and identified real user queries that could have returned results by using PubTermVariants.
2. We analyzed a subset of result-producing queries and identified the difference in the result set had PubTermVariants been used.

	PubTerm-Variants	StemHG\Pub-TermVariants	Total
Positives	178(89%)	163(81.5%)	341
Negatives	22(11%)	37(18.5%)	59
Total	200	200	400

Table 1. Results of manual annotation of 200 random pairs from PubTermVariants and 200 pairs from StemHG\PubTermVariants. 89% of pairs in PubTermVariants and 81.5% of pairs in StemHG\PubTermVariants were found to be true variants.

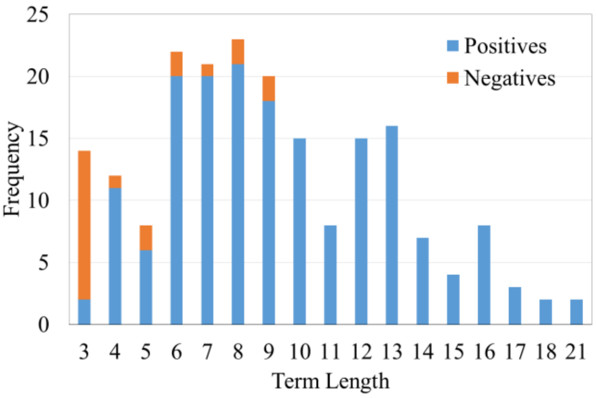

Figure 1. The quality of term variants as a function of token length in PubTermVariants.

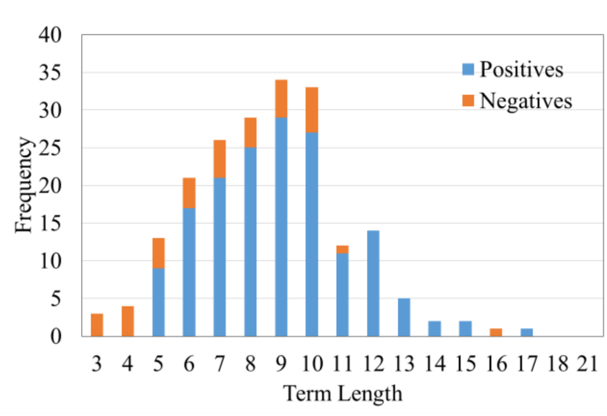

Figure 2. The quality of term variants as a function of token length in StemHG\PubTermVariants.

For these experiments we collected PubMed log data for 2015. We find that about 76% of PubMed queries contain a term from PubTermVariants.

Effect of using PubTermVariants on zero-result queries

Some PubMed queries do not produce a result. We call these queries zero-result queries. We ask whether using PubTermVariants could lead to results being retrieved for these queries.

In order to answer this question we chose a random day of PubMed logs in 2015. We preprocessed the data and kept only multi-term queries that contained alphanumeric characters, dashes and commas. We also removed queries that did not contain a term from PubTermVariants. For each query, we verified that if a term variant is removed, the query consisting of the remaining tokens retrieved a set of PubMed articles. This set contained 55,496 unique queries.

For each of these queries, we replaced a term with its variant from the PubTermVariants list. Since some terms have multiple variants, this process resulted in 110,103 queries which are now used to query PubMed and report results. We call the use of the term variants successful if at least one of the variant queries resulted in a successful search. For example, term *monoubiquitination* is mapped in our collection to terms *monoubiquitin*, *monoubiquitinate*, *monoubiquitinated*, and *monoubiquitinates*. For a given query *hdm2 monoubiquitination*, one substitution to *hdm2 monoubiquitinated* is found to be successful, and so we call that substitution successful. Articles were retrieved for 8.83% (4,902 queries) of the original 55,496 queries. This percentage, however, represents a lower bound of success because for every query only one term was considered for replacement. Queries can have several candidate terms for replacement.

Effect of using PubTermVariants on result-producing queries

Since PubMed is quite successful at producing relevant results for most queries, we wanted to examine the effect of the variants in PubTermVariants on these searches. For this experiment we randomly selected 1,000 user queries that contained a term listed in PubTermVariants and where the number of PubMed results for each of these queries was between 1 and 20. For each of these queries we produced only one variant query so that the original term variant was replaced with one of its paired variants in the PubTermVariants list, randomly selected. The resulting variant queries were used to query PubMed and we retrieved results for 480 of these queries. We compared the results set for the original user queries with their variant queries and found that the average number of results for the original queries was 6.8, however, if we combine the results with the results of the variant queries this number increases to 8.5. Furthermore, 38% variant queries retrieve additional relevant PubMed articles without overwhelming the search results. Similar to the zero result case, this percentage represents a lower bound.

4 Conclusions

We presented a high-quality list of biomedical term variants which we call PubTermVariants. The PubTermVariants resource is generated in a data-driven way by applying two statistical tests to pairs of tokens that stem to the same form. Both, the hypergeometric and the morpho-semantic tests, provide a useful tool for deciding whether terms in the pair are related or not.

PubTermVariants provides a clean and reliable high-quality collection of terms that can be used interchangeably in PubMed queries. The manual examination revealed that 89% of the pairs are true variants, and removing three letter tokens results in higher quality. Our experiments on PubMed log data demonstrated that some zero-result queries that contain a term variant can return results by applying a substitution from PubTermVariants. Our other experiments revealed that when a term variant is applied to create a variant query, in 38% of the cases the result set was enriched with articles which were not present in the initial request, thus increasing recall.

PubTermVariants is available for other applications of biomedical term variants from ftp://ftp.ncbi.nlm.nih.gov/pub/wilbur/PubTermVariants/pairs.txt.gz.

5 Acknowledgements

The authors thank Grigory Balasanov for his help in preparing the web-based tool for the manual annotation task.

Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

Church, K. and P. Hanks (1989). Word association norms, mutual information, and lexicography. Proceedings of the 27th ACL Meeting: 76-83.

Hollink, V., et al. (2004). Monolingual Document Retrieval for European Languages. Information Retrieval 7(1).

Hull, D. (1996). Stemming Algorithms - A Case Study for Detailed Evaluation. Journal of the American Society for Information Science 47(1): 70-84.

Krovetz, R. (1993). Viewing Morphology as an Inference Process. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.

Larson, H. J. (1982). Introduction to Probability Theory and Statistical Inference. New York, John Wiley & Sons.

Manning, C., et al. (2009). Introduction to Information Retrieval. Cambridge, England, Cambridge University Press.

Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems.

Moral, C., et al. (2014). A survey of stemming algorithms in information retrieval. Information Research 19(1).

Porter, M. F. (1980). An algorithm for suffix stripping. Program 14(3): 130-137.

Wilbur, W. J. and L. Smith (2013). A Study of the Morpho-Semantic Relationship in Medline. Open Inf Syst J 6.

Xu, J. and W. B. Croft (1998). Corpus-based stemming using cooccurrence of word variants. ACM Transactions on Information Systems (TOIS) 16(1): 61-81.