AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

Term Evolution: Use of Biomedical Terminologies

Gintare Grigonyte, Fabio Rinaldi and Martin Volk
Institute of Computational Linguistics, University of Zurich
Binzmuehlestr. 14, 8057 Zurich, Switzerland

Abstract

This extended abstract presents a work in progress of using terminological resources from the biomedical domain to systematically study the change of domain terminology over time. In particular we investigate term replacement. In order to study term replacement over time, semantic knowledge like conceptual granularity of a term is necessary. We analyze three popular biomedical terminology resources (UMLS, CTD, SNOMED CT) and show how information provided there can be used to extract lexically distinctive synonym sets that exclude variants. We use the entire PubMed dataset to chronologically study occurrences of extracted synonyms. Our experiments on the disease subsets of three terminologies reveal that the phenomenon of term replacement can be observed in around 60% of the extracted synonym sets.

Terminology change phenomenon

Domain terminology evolves over time: some terms disappear and others get introduced, terms can get more specific, or become more abstract, also the meaning of terms change. In this article we focus on term replacement. Term replacement occurs due to the competition of terms describing the same phenomenon, unless the denoted object disappears (Cowie 1998; Norri 2004). Term replacement can be observed among synonymous terms. For instance, new terms can be adopted to describe the same phenomenon, or in some cases one or several synonymous terms cease to being used.

Term replacement can be modeled upon a set of synonyms referring to the same concept. Let us assume a concept C containing a synonym set S where each member s_i of the synonym set S is a term and refers to the concept C.

C = S$(s_1, s_2, ..., s_n)$

For instance, C1=S(disease, illness, sickness): S is the synonym set containing 3 synonyms, they all refer to concept C1.

The change of usage of terms over time can be explained by replacement between synonyms. Hypothetically, if we observe the decreased usage of the term 'sickness' over time, but terms 'disease' and 'illness' - increase, we call it a replacement of a term.

Synonym set extraction from biomedical terminologies

Biomedical terminologies like UMLS provide conceptual classifications for terms, but do not differentiate between synonyms and morphological or orthographical variants. Consider the following example:

concept ID	term ID	term
C2697452	S16752029	HAIR MORPHOLOGY 2 (disorder)
C2697452	S11971678	Curly hair
C2697452	S16616027	CURLY HAIR
C2697452	S16608447	HAIR MORPHOLOGY 2
C2697452	S16612260	HAIR CURVATURE, VARIATION IN
C2697452	S16619873	HRM2

From this data one can infer the following:

Concept	C2697452
Terms	curly hair; hair morphology 2; hair curvature, variation in
Lexical variants	{HAIR MORPHOLOGY 2; HRM2; HAIR MORPHOLOGY 2 (disorder)}, {Curly hair; CURLY HAIR}

This granularity of concept-term-variant helps to separate between very different type of information: synonyms of a term and variants of a term.

In order to derive lexically distinct synonym sets from biomedical terminologies we proceed with the following processing steps:

1. removing lexically identical terms, or so-called duplicates;
2. normalization of capitalization and punctuation;
3. normalization of term variants that differ in word order, e.g. 'dentin dysplasia' and 'dysplasia dentin';
4. reducing lexically similar strings, e.g. 'coronary ostium stenosis' and coronary ostial stenosis'.

The latter processing step is based on calculating Levenshtein's distance among strings on a character basis and allows a difference less than 10% (i.e., strings are modified by sorting characters in each string, Levenshtein's distance among two modified strings is set to an empirically chosen threshold of 0.9). For instance terms 'acinetobacter infections' and 'acinetobacter infection' are too similar as the distance between these two strings is 0.9787.

Acquired synonym sets

We have used three terminological resources from which we have derived terminologies of diseases: UMLS (2012, January release; diseases tagged with the t047 semantic type), CTD (2012, January release, acquired from www.ctdbase.org) and SNOMED CT (20120131 release, acquired from www.ihtsdo.org).

Table 1 summarizes results of extracting synonym sets from three terminologies.

Terminology	UMLS	CTD	SNOMED CT
Number of concepts	9,986	9,657	94,147
Number of terms	530,149	69,494	376,062
Processing step-1 (number of terms, %)	288,988 54.51%	68,919 99.17%	269,897 71.77%
Processing step-2 (number of terms, %)	258,290 48.72%	68,410 98.44%	179,289 47.67%
Processing step-3 (number of terms, %)	220,546 41.6%	46,965 67.58%	178,266 47.40%
Processing step-4 (number of terms, %)	155,038 29.24%	31,223 44.92%	132,456 35.22%
Number of ambiguous terms	267	1,134	326
Number of concepts containing synonym sets	28,643	7,009	20,923

Table 1: Summary of the extraction of synonym sets from UMLS, CTD and SNOMED CT terminologies.

For the purpose of tracking term replacement we need to consider concepts containing two or more synonyms. From the UMLS terminology we have derived 28,643 such concepts containing 91,695 lexically different terms (synonyms) which is on average 3.2 synonyms per concept. From the CTD terminology we have extracted 7,009 concepts - 28,543 terms, i.e. 4.0 synonyms per concept. From the SNOMED CT we have extracted 20,923 concepts - 59,232 terms, i.e. 2.8 synonyms per concept.

Discovering term replacement

In order to track term replacement over time we need a substantial reference corpus of the domain. For this purpose the citation database PubMed storing over 22 million citation references for biomedical literature was used as a chronological corpus containing documents between 1881 and 2012. Over 11 million of documents (consisting of titles and abstracts) have been indexed with the Indri IR (www.lemurproject.org) system.

We propose to capture term replacement by dividing the chronological reference corpus into time periods and using the simple linear regression model to analyze tendencies of occurrence for each synonym over time.

Figure 1 illustrates a typical term replacement situation. The synonym set extracted from the SNOMED CT disease terminology contains 8 lexically distinct synonyms, i.e., ”lung fluke disease”, ”pulmonary distomatosis”, ”pulmonary paragonimiasis”, ”endemic oriental haemoptysis”, ”paragonimiasis”, ”oriental lung fluke disease”, ”infection by paragonimus” and ”lung fluke infection”. Terms like ”pulmonary distomatosis”, ”oriental lung fluke disease”, and ”endemic oriental haemoptysis” show a clear tendency of being replaced with far more frequently used ”paragonimiasis” and ”pulmonary paragonimiasis”.

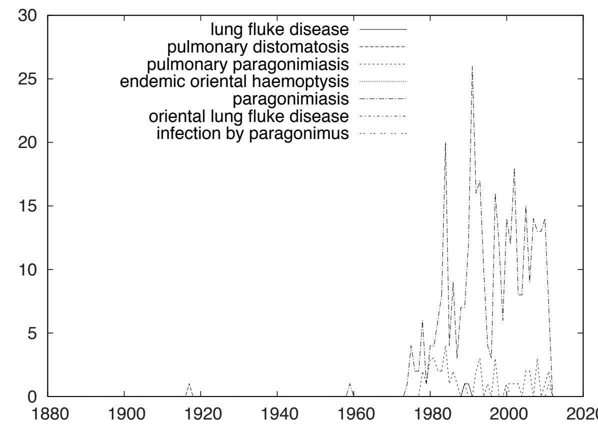

Figure 1: Term replacement example: analysis of synonym occurence.

The poster presentation of this paper will include the presentation of the approach for capturing term replacement and experimental results revealing how pervasive term replacement is in the biomedical domain.

Acknowledgments.

This research is funded by the Sciex NMS-CH programme of the Rector's Conference of the Swiss Universities (CRUS). Project Code 11.002.

References

Cowie, C. 1998. The Discourse Motivations for Neologising: Action Nominalization in the History of English. In: Coleman, J., Christian J. eds. *Lexicology, Semantics and Lexicography.* Selected Papers from the Fourth G. L. Brook Symposium, Manchester, 179–206.

Norri, J. 2004. Entrances and exits in English medical vocabulary, 1400–1550. In: Taavitsainen, I., Pahta, P. eds. *Medical and Scientific Writing in Late Medieval English.* Cambridge University Press, 100–143.