How preferred are preferred terms?

Gintare Grigonyte[1], Simon Clematide[2], Fabio Rinaldi[2]

[1]Computational Linguistics Group, Department of Linguistics, Stockholm University
Universitetsvagen 10 C SE-106 91 Stockholm, Sweden
[2]Institute of Computational Linguistics, University of Zurich
Binzmuhlestrasse 14, CH-8050 Zurich, Switzerland
E-mail: gintare@ling.su.se, siclemat@cl.uzh.ch, rinaldi@cl.uzh.ch

Abstract

We present a novel approach for synonymous term preference detection that relies on chronological text analysis. Our approach analyses the use of synonymous term entries in a chronological reference corpus. As a result of preference evaluation, a ranking of preference between all the synonymous term entries belonging to the same concept is established.

Keywords: automatic terminology curation; synonymous terms; term preference; chronological corpus.

1. Introduction

This article discusses the problem of automatically determining preferred terms in terminological databases. The notion of a preferred term becomes important for automatic domain text processing. We have experimented with biomedical terminology; however the approach presented in this paper can be extended to other domains and terminologies.

Terminological entries in databases like Unified Medical Language System (UMLS) contain manually assigned tags denoting which synonym among all listed synonyms is the preferred one.

To illustrate the impact of the UMLS, consider the largest database of biomedical domain literature PubMed. PubMed publishes more than 500,000 documents each year and its publications are indexed with UMLS terms.

The UMLS (Bodenreider, 2004) is a human-expert curated terminological resource that has the following micro-structure:

ConceptID

Synonym 1

Synonym 2... PreferredTerm

Synonym n

The conceptID is a conceptual identifier for all subsumed terms. The conceptual identifier is similar to a synset identifier in WordNet. Just like a synset contains synonymous interchangeable expressions, so a concept in the UMLS also has synonymous terms. The preferred term tag is reviewed periodically and assigned manually by domain experts who curate terminological entries.

Domain terminology is extremely responsive to changes and new developments inside the respective domain, which motivates the development of automatic approaches for terminology maintenance. We view term preference in domain texts as a usage-based, and thus dynamic, phenomenon. An automatic preference detection is important if we want to take into account how terms are actually used in domain literature.

2. Data and tools

We used a subset of the UMLS terminology covering the topic of diseases. This subset contains over 90,000 concepts. The total number of terms is over 500,000. As a chronological reference corpus to study the usage of domain terms, we used all publications of the PubMed[1] January 2012 release. The 2012 PubMed dataset release contains over 22 million documents consisting of titles and some abstracts between 1881 and 2012.

In order to consistently detect occurrence of terminology in the PubMed2012 corpus we have used a specialized tool MetaMap[2], developed by the National Library of Medicine, which identifies biomedical concepts from unstructured texts and maps them into concepts from the UMLS (Pratt and Yetisgen-Yildiz, 2003).

3. Possible approaches

A terminological concept in UMLS contains multiple synonyms expressing the same concept and one of those synonyms is marked as a *preferred term*. For instance, the *C0008049* concept in UMLS has 16 synonyms, of which one is marked as preferred: ‘*varicella infection*’.

This paper proposes a corpus-based approach for automatically detecting preference among synonymous terms in terminologies such as UMLS. We see term preference as a usage related, dynamic phenomenon. The simplest way of automatically measuring term preference is counting the number of occurrences in a reference corpus:

[1] http://www.ncbi.nlm.nih.gov/pubmed

[2] http://metamap.nlm.nih.gov

chickenpox varicella	11
varicella infection	346
varicella	3820
chicken pox	1767

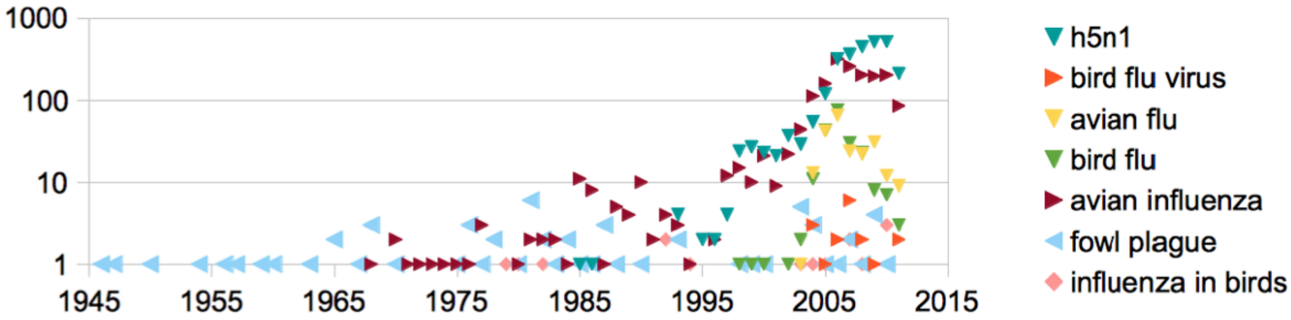


Figure 1: Chronological occurrences of synonyms of the concept 'C0016627'.

However, in case of recently emerged and topical terms, like *'h5n1'* in the concept *C0016627* (see Table 1), we find that their frequency is overwhelming and that this criteria for determining term preference might be inadequate. Thus, chronological information such as a time interval between the first and last occurrences of a term (see column 3, Table 1) or the total number of years for which a term is used in a reference corpus (see column 4, Table 1) might also constitute informative criteria of a term usage.

Taking into account time dimension alone is also insufficient, particularly if term occurrence is sparse. Besides, analyzing frequency and time data separately creates a biased view of term preference. Consider, for instance, synonyms of the concept *C0016627* (Table 1, Figure 1): *'h5n1'* is the most frequent; *'fowl plague'* is the most chronologically prominent.

Synonymous terms	# occurrences	year interval	# years
h5n1	**2722**	26	20
bird flu virus	20	9	7
avian flu	219	9	8
bird flu	206	13	13
avian influenza	1737	43	40
fowl plague	65	**64**	**42**
influenza in birds	15	31	11

Table 1: Analysis of synonyms of the concept *C0016627*.

In this paper we argue that in order to determine the preference of a term among its synonyms, time and frequency criteria should be used in combination. The simplest model that considers both dimensions is a linear regression.

4. Method

We model the series of data of the occurrence of a term over time as a simple linear regression, where α and β are unknown parameters, and ε corresponds to noise:

$$\alpha + x_i * \beta + \varepsilon_i \qquad (1)$$

The fitted line is equal to the correlation between term occurrence (y_i) and time (x_i) corrected by the ratio of standard deviations of these variables. The unknown parameter β corresponds to the steepness of the slope. We use an ordinary least squares method for estimating unknown parameters α and β.

$$\hat{\beta} = \frac{Cov[x,y]}{Var[x]} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \qquad (2)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (3)$$

Chronological data sparseness is a major obstacle if we want to compare all synonyms and estimate their parameters for linear regression. From Figure 1 we see that some terms occur rather consistently throughout the years, e.g. ‘fowl plague’, while other occur very rarely, e.g. ‘influenza in birds’. In order to obtain the same number of data points we included all years when at least one of the synonyms has occurred; also, in cases when a synonym has not occurred though other synonyms from the group have occurred during that year, we set the basic value for a non-occurring synonym to 0.1[3].

We use relative frequency of occurrences normalized by the total number of occurrences within the set of synonyms occurring during a specific year.

The final ranking of term preferences is based on parameter β multiplied by two constants: 1) the total number of years that a synonym has occurred divided by the maximum number of years available from the set of synonyms; and 2) the total number of occurrences of a synonym divided by the total number of occurrences

[3] Arbitrarily chosen in order to differentiate between situations: a) 0, none of the synonyms of a concept have occurred that year; and b) 0.1, a synonym has not occurred, but other synonyms from the concept have.

within the synset.

The estimated parameter β from the linear regression model based on term occurrence and time enables a ranking of different synonyms of the same concept. For instance, the term preference ranking over time for the concept *C0016627* in the PubMed corpus is:

avian influenza	0.00478
h5n1	0.00345
fowl plague	0.00204
bird flu	0.00079
avian flu	0.00024
avian flu virus	0.00019
influenza in birds	0.00018

This approach can be used for several interpretations of term evolution. The first interpretation of the β parameter is that a negative value shows a tendency toward term extinction. However, such an interpretation is only possible in the context of other synonyms of the term. This is the case because we analyze a domain specific corpus and we want to make sure not to include situations such as a temporary disappearance of a term or phenomenon inside the domain literature (e.g. no publications representing a specific disease have been registered during a certain period of time). Only when other synonyms of the same term continue to occur can we talk about extinction of that specific term. The situation of one term showing a tendency to disappear (negative β value) when its synonyms continue to be used (positive β value) is called term replacement (Grigonyte et al., 2012A, 2012B).

Second, the positive value of the β parameter shows an increase in term occurrences over time. The larger parameter means that the term is used proportionally more than its synonyms and its use is therefore increasing with time.

5. Results

We analyzed the terminology of diseases in the UMLS 2012 release. All terminological entries come under the semantic group of disorders.[4] The set of disease terminology concepts that contain at least two synonymous terms comprises 17,410 concept entries. Each concept entry in the UMLS database has several synonymous terms. One or more of them is marked as the ‘preferred term’.

For evaluation purposes we chose the annotation of MeSH which has only one ‘preferred term’ for each concept.[5]The test set was therefore left with 2,966 concepts

[4] Semantic tags of disorders: T020, T190, T049, T019, T047, T050, T033, T037, T048, T191, T046, T184. For more information see: http://semanticnetwork.nlm.nih.gov/SemGroups/

[5] We chose UMLS term entries that match the MeSH Descriptor record.

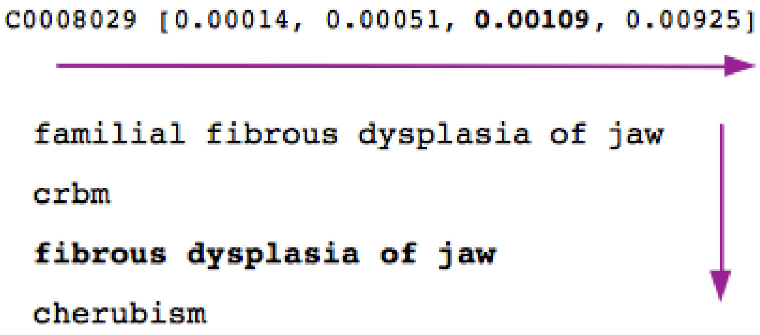
that have synonymous terms and one 'preferred term' tag.

The evaluation was performed by comparing the highest ranking synonym against the manually assigned 'preferred term' tag in the UMLS. We used two methods: a) the highest ranked synonymous term modelled by our approach *linreg*; and b) the most frequently occurring synonym *maxocc* (see Table 2).

# of concepts that have synonym synsets	17410	
# of synsets with MESH 'preferred term' tag	2966	
# of cases of 'preferred term' match by **linreg**	1805	60.86%
# of cases when a different 'preferred term' is suggested by **linreg**	1161	39.24%
# of cases of 'preferred term' match by **maxocc**	1852	62.55%
# of cases when a different 'preferred term' is suggested by **maxocc**	1114	37.45%

Table 2: Results of term preference evaluation.

Both approaches yielded very similar results. The agreement between *linreg* and *maxocc* is 88%. Around 60% of the preferred UMLS terms match with the most preferred terms used in domain corpora. However, for a substantial number of term entries both methods would also suggest other preferred terms. For instance, the concept *C0008029* has four synonyms, of which '*fibrous displasia of jaw*' is the manually assigned preferred term. The highest ranking synonym according to *linreg* and to *maxocc* methods is '*cherubism*'.



C0008029 [0.00014, 0.00051, **0.00109**, 0.00925]

familial fibrous dysplasia of jaw

crbm

fibrous dysplasia of jaw

cherubism

Figure 2: Synonym preference by *linreg* method.

Examples of different suggestions between *linreg* and *maxocc* are:

seasonal allergic rhinitis	hay fever
rheumatic disease	rheumatism

The large proportion of preferred terms not matching the manually assigned 'preferred terms' can be explained by at least two contributing factors. First, we performed the 'hard match' between the highest ranking term and the UMLS term, which included only exact matching strings, no orthographical deviations were allowed. Second, we only compared one preferred term from the UMLS entry instead of analyzing all preferred terms against the top preferred term suggested by the *linreg* method.

6. Conclusions

We present an approach for term preference detection that relies on term usage in the chronological reference corpus.

The *linreg* method was tested against manually assigned preferred terms. For the task of synonym preference detection the *linreg* method showed similar results to the *maxocc* method which can be partially explained by *linreg* modeling the tendency of a synonym as having increasing usage in the future. However a term preferred by the *linreg* method also indicates that it might not necessarily reflect the most frequently used term.

Lexicographers and terminologists could use the preference ranking of terms for a validation of the contents of existing term bases. As an outlook for employing the *linreg* method, a terminology expert should look at cases where the predictions and the actual preferred term are different. The method described in this paper can be used as a diagnostic tool in terminography, i.e. increases, decreases and temporary absence of term occurrences can assist an interpretation of domain terminology change.

The proposed approach could be implemented in different domains, provided that domain terminologies and large reference corpora spread over many years are available, e.g. legislative and political domains.

7. Acknowledgements

This research was supported by the Conference of Swiss University Rectors organization, Sciex research funding grant 11.002. The authors thank O. Bodenreider and anonymous reviewers for valuable comments.

8. References

Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating Biomedical terminology. Nucleic Acids Research, vol. 32(1), p. 267-270.

Grigonyte, G., Rinaldi, F., Volk, M. (2012A). Term evolution: use of biomedical terminologies. In Proceedings of AAAI-2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text, p. 79-80.

Grigonyte, G., Rinaldi, F., Volk, M. (2012B). Change of biomedical domain terminology over time. In: Human Language Technologies – The Baltic Perspective, p. 74-81.

Pratt, W., Yetisgen-Yildiz, M. (2003) A Study of Biomedical Concept Identification: MetaMap vs. People. AMIA Annual Symposium Proceedings, p. 529–533.