Jeferson L. Bitencourt, Pindaro S. Cancian, Edson J. Pacheco, Percy Nohama, Stefan Schulz:

Thesaurus Anomaly Detection by User Action Monitoring

Preprint Version

To appear in:

MEDINFO 2007 - Proceedings of the 12th World Congress on Health (Medical) Informatics – Building Sustainable Health Systems
Thesaurus Anomaly Detection by User Action Monitoring

Jeferson L. Bitencourt\textsuperscript{a}, Píndaro S. Canciana, Edson J. Pacheco\textsuperscript{a}, Percy Nohama\textsuperscript{a}, Stefan Schulzb

\textsuperscript{a}Pontifical Catholic University of Paraná (PUCPR), Curitiba, Brazil
\textsuperscript{b}University Hospital, Freiburg, Germany

Abstract

The construction and maintenance of a medical thesaurus is a non-trivial task, due to the inherent complexity of a proper medical terminology. We present a methodology for transaction-based anomaly detection in the process of thesaurus maintenance. Our experiences are based on lexicographic work with the MorphoSaurus lexicons, which are the basis for a mono- and cross-lingual biomedical information retrieval system. Any "edit" or "delete" actions within these lexicons that undo an action defined earlier were defined as anomalous. We identify four types of such anomalies. We also analyzed to which extent the anomalous lexicon entries had been detected by an alternative, corpus-based approach.

Keywords: Thesaurus, Quality Control

Introduction

A thesaurus is a collection of semantically related terms, used to represent document contents with the purpose of improving document classification and content retrieval \cite{1,2,3}. Especially in Medicine and Health, there is an ever-growing volume of texts, scientific publications, electronic textbooks, web-based health information for laypersons, and last but not least, a huge amount of text in electronic patient records. Medical terminology is complex, rapidly changing, and highly productive. New acronyms, abbreviations, single and multi word compounds are continuously generated, and English terminology increasingly permeates non-English medical documents. Multilingualism is an important issue here, due to the global tendency of using English as the primary language of research, whilst the local idioms are used for patient-related documentation and communication \cite{4}. All these factors hamper the use of simple text retrieval techniques such as popularized by Web search engines. So we claim that text retrieval in the biomedical domain should be supported by an underlying thesaurus that covers intra- and interlingual synonymy.

Thesaurus engineering is an iterative process, generally directed by guidelines, but also encompass individual arbitrariness. Controversies especially come up with relation to boundary decisions such as:

- Whether to consider or not a term pertinent to the domain and therefore relevant for the thesaurus;
- Whether to include composed or derived terms if their components or base forms are already in the thesaurus;
- Whether to consider two terms as synonymous;
- Which other semantic relations to use between terms;
- Which senses to be needed to be distinguished when dealing with ambiguous terms;
- Whether to consider additional senses which are of marginal importance for the domain.

The most common method to track modifications in a database is the procedure register or log \cite{5}. The log system has the property of recording all state changes. In this work maintenance problems in a medical thesaurus are analyzed, and a method is proposed which helps to detect anomalous change patterns in order to increase the efficiency of the thesaurus building process.

Material and Methods

The Subword Approach

The application context of this work is given by MorphoSaurus, a large multilingual thesaurus for clinical medicine \cite{6,7}. The main difference between MorphoSaurus and any other medical thesaurus such as MeSH \cite{8} is that the inclusion of lexical entries into the repository is strictly guided by criteria of semantic atomicity. This resulted in the introduction of a new kind of lexical item, the so-called subword entry, assuming that neither fully inflected nor automatically stemmed words constitute the appropriate granularity level for lexicalized content descriptions \cite{9,10}.
The building of a subword thesaurus requires a scrutiny of the following aspects of language:

- Orthographic variations such as in *oesophagus* vs. *esophagus*;
- Derivations that modify a word’s part of speech and/or its meaning, e.g.: *diabetes*, *diabetic*, *antidiabetic*;
- Proper string delimitation of what constitutes a valid subword entry, e.g. *nephr*, *nephr*, or both;
- Composition that builds complex word forms out of simple ones, such as *high blood pressure* (multi word composition) or *hyperprebetalipoproteinemia* (single word composition);
- Synonyms and translations: *nephro*, *renal*, *rim*, *kidney*, *niere*...;
- Acronyms: *AVC*, *ECG*, *DPOC*, *AIDS*,... and their possible expansions;
- Proper Names, such as substance names, trade names, or eponyms: *Diclofenac*, *Viagra*, *Parkinson*...;
- Ambiguous terms, i.e. terms which have different meanings, such as *head*, *ventricular*...;

The MorphoSaurus System

Particularly in the medical sublanguage we observe complex word forms such as in ‘pseudo|hypo|para|thyroid|ism’, ‘append|ectomy’, or in ‘tooth|ache’. The MorphoSaurus system does not register these complex terms, as long as their meaning can be derived from their components. Instead it focuses on so-called *subwords* as lexicon entries, which are registered together with attributes such as language and subword type (word stem, prefix, suffix, infix, invariant word). Instead of medical *terms*, MorphoSaurus includes the *building blocks* of medical terms.

In the MorphoSaurus database, each entry is assigned to one synonymy class. In the beginning of the thesaurus building process, there is one class for each entry, but in the following classes are increasingly fused, containing not only intralingual synonyms but also interlingual translations. Each synonym class is identified by one language-independent descriptor, called MorphoSaurus identifier (MID). Every MID represents a unique meaning. MIDs are further divided between those which are relevant for indexing (here marked by #) and those which have grammatical functions only (e.g. auxiliary verbs, inflection suffixes, conjunctions, etc., marked by %).

For instance, the MID #liver={liver, leber, hepar, hepat, figad, …} represents the sense of “liver”, whereas the MID %ness={-ness,-idade ,-keit} groups semantically irrelevant suffixes, just as %be={be, is, are, sao, e, ist, …} groups forms
of the verb “to be”, which should be ignored for document indexing.

MorphoSaurus provides two semantic relations between equivalence classes:

- **has_sense**: This relation links ambiguous MIDs to their senses, e.g., the MID #head is linked to the MIDs #caput and #boss by has_sense;
- **has_word_part**: This relation links an MID having a composed meaning to each of its parts, e.g., the MID #myalg={“myalg-”, “mialg-”, “mialg-”} to both #muscle={myo-, muscle, mio, muscul, …} and #pain={pain, -alg, -algi, -algia, dor, schmerz,...}. The reason for this is to deal with composed meanings even in cases where a compound cannot be properly dissected.

The MorphoSaurus indexing engine converts text into an interlingual representation by a three-step procedure, as depicted in Fig. 1 for a parallel text. In the first step, the first step deals with general and language specific orthographic normalization. In a second step, the text is split into sequences of semantically plausible subwords, which are then checked for morphological plausibility using a finite-state automaton (Morphological Segmentation). Thus invalid segmentations such as segmentations without stems or ones beginning with a suffix are rejected. Finally, each meaning-bearing subword is replaced by its MID. This constitutes the interlingual output representation of the system. The bold MIDs in Fig. 1 are evidence of the high interlingual correlation on the level of semantic atomicity.

**Pragmatics of Thesaurus Building in MorphoSaurus**

Modeling decisions for a subword thesaurus are more complex than decisions for word-based thesauri. Thus, the proper delimitation of what constitutes a lexical entry is often subject to argument. These boundary decisions are even influenced by the performance of the word segmentation algorithm. The following example may provide a flavor of the daily problems in the maintenance of MorphoSaurus: The word *nephrotomy* used to be erroneously divided into *nephro*tomy instead of the correct *nephro*to*my*. This was solved by the inclusion of *nephro* as a synonym of *nephro* into the thesaurus, which may be considered linguistically unsound, but which produced the desired result. Another controversial issue is how to deal with utterly short stems. For instance, the MID *muscle* can be erroneously extracted from many words due to the frequency of the character combination “my”. The decision made in this case was not to include *my*, but to add the common compounds (myosis, myalg, ...) instead and to use, in a second step, the has_word_part relations as introduced above in order to represent their meanings.

**Anomaly Detection in Logged Data**

With the purpose to collect data about the changes in the MorphoSaurus database during a given time interval we collected 86 backups of the MorphoSaurus database for the period between July 14th, 2005 and March 30th, 2006, representing regular intervals of approximately three days. A script extracted all alterations for each single data object. This data volume then served as a basis for the detection of anomalies.

We here introduce the notion of thesaurus management anomalies as sequences of such actions done by the thesaurus curators that consume effort without any positive impact on the quality of the thesaurus.

We distinguish the following four anomaly types:

- A relationship anomaly is defined as a sequence of editing steps in which a semantic relation (has_sense or has_word_part) between two equivalence classes is first eliminated and later restored,
- A type anomaly is defined as a sequence of editing steps in which a lexicon entry is first moved from one equivalence class to another one and later happens to be moved back into the original class,
- A delimitation anomaly is defined as a sequence of editing steps in which the string delimitation of a lexicon entry is modified in a first step but is later restored to its original form,
- A permanence anomaly is defined as a sequence of editing steps in which an existing lexicon entry is first deleted, but recreated in a later phase.

According to these anomaly types, the whole body of logged user data from the backup data was analyzed.

**Corpus-driven Error Detection**

For further validation of anomalies, we used the following additional resources:

- Related corpora analysis: in a parallel study, the MorphoSaurus database was checked against real data in order to detect imbalances of MID distribution. In particular, this study used closely related corpora in different languages. All these texts were indexed by the MorphoSaurus system (yielding an output as shown in Fig.1), and the distribution of MIDs was pairwise compared. The hypothesis was that anomalous MIDs exhibited the most unequal distribution patterns.
- Discussion Forum: a list of ill-distributed MIDs (ranked by degree of disproportion and overall frequency) was used as a basis for manual cleansing of the thesaurus by our team of lexicographers. Each of these MIDs was submitted to a thorough analysis, and the result was discussed in an online forum. Whenever the lexicographers reached a consensus, the MID entry was modified accordingly.

Joining the anomaly data from the log analysis with the data from the forum then provides a measure of relevance of the anomaly data.

**Results**

**Analysis of Anomalies**

The log analysis yielded a total of 146 anomalies. Interestingly, there were many multiple occurrences in the
Analysis of Problems Addressed in the Discussion

which exhibited a relationship anomaly more than once during the observation period. Counting the multiple occurrences only once, we got 99 anomalies, as shown in Table 1. The discussion forum focused on English, Portuguese, and German and addressed 325 problems as shown in Table 2.

Comparison between Anomalies and Discussed Problems

So far, the acquisition of anomaly data was completely independent from the analysis of problem discussions. Bringing together both sources we can now analyze the anomalous MIDs which were picked up in the discussion. The data show that 36 out of 99 anomalous MIDs were not addressed in the forum discussions, and that forum discussions covered far more cases than could be identified by the log analysis, cf. values in parentheses in Table 1.

An inverse view is given by Table 2. Here the values in parentheses show the frequency of those MIDs which had also been spotted by the log analysis. Finally, Table 3 gives a closer view of multiple occurring anomalies.

Discussion

The correct handling of lexical ambiguities is the most error-prone step in the thesaurus management process [11]. This was evidenced not only by the frequency of the relationship anomaly (AR, cf. Table 1) but also by its occurrence in the discussions. Moreover, this anomaly was the only one where up to seven repetitions of editing occurred. This sheds light at a considerable waste of resources and lack of communication.

The assignment of subwords to an equivalence class (MID) was also subject to changes (AT anomaly), but here all cases were addressed in the discussion forum. This shows the effectiveness of the corpus-based detection of disproportions in the MID distribution and its good take-up by the lexicographers.

It was quite surprising that no anomaly of string delimitation (AD) could be observed. This is probably due to the fact that observed segmentation problems were solved by adding new string variations (e.g. “-atomy” in addition to “-tomy”) and not by modifying the existing ones, in accordance with the guidelines used in the thesaurus building and maintenance process.

Finally, the awareness of the permanence anomaly was shown by its high coverage in the discussion forum. This kind of anomaly elicited the major disproportions in the corpus analysis, due to the fact that lexemes with a borderline semantic importance occur with high frequency. For instance, the phenomenon that the preposition “from” was assigned to the MID type “marked for indexing”, and the Portuguese translation “de” was assigned to another MID marked as “excluded from indexing”, put the first MID on top of the ranked list of lexicon disproportions, and was therefore preferentially addressed in the discussions. These types of “stopwords” are frequent and equally distributed across the whole document space. Hence, they are irrelevant for distinguishing documents in text retrieval scenarios [3,12]. On
the other hand they can acquire importance as context modifiers, such as in complex terms like *Removal of foreign body from stomach*, which could be matched to *Removal of stomach* in case that the preposition *from* is neglected.

The anomaly detection process would have an additional value if the anomalous MIDs would be detected in runtime so that the lexicographers would get an immediate feedback whenever an action executed earlier is undone. A new version of our lexicon editing tool MorphoEditWeb, currently under development, implements this functionality.

**Conclusion**

Thesaurus management is a highly dynamic process, and the kind of decisions which have to be done continuously imposes challenges on the lexicographers upon which the creation and continuing maintenance is incumbent. The waste of resources due to the phenomenon that one person undoes an action which another person has done before is considerable. Unfortunately no guideline for thesaurus management can ever foresee all borderline cases which can only be solved by consensus. The proposed technique of user action based anomaly detection is useful to discover them and is complemented by other techniques such as the corpus-based check for disproportion of semantic identifiers. A seamless integration of such quality assessment routines in the thesaurus management tools is of utmost importance for achieving higher process effectiveness.

**Acknowledgements**

This work was supported by the CNPq, Brazil, (project number 550830/05-7), and by German-Brazilian cooperation project DLR/PUCPR (BRA 03/013).

**References**


**Correspondence Address**

Jeferson Luiz Bitencourt - LER, Pontifical Catholic University of Paraná. Rua Imaculada Conceição, 1155 – CEP 80215-901, Curitiba, Brazil, jefersonbitencourt@yahoo.com.br.