Revista Brasileira de ENGENHARIA BIOMÉDICA
BRAZILIAN JOURNAL OF BIOMEDICAL ENGINEERING
Volume 28, Número 1, p. 11-22, 2012

Artigo Original
DOI: 10.4322/rbeb.2012.002

Quality issues in thesaurus building: a case study from the medical domain

Percy Nohama*, Edson José Pacheco, Roosewelt Leite de Andrade, Jeferson Luiz Bitencourt, Kornél Markó, Stefan Schulz

Abstract To ensure the quality of a medical thesaurus is a non-trivial task, due to the inherent complexity of medical terminology. The peculiarities of the medical sublanguage and the subjectivism of lexicographers' choices complicate the thesaurus construction process. Our experience is based on the MorphoSaurus lexicon, the basis of a biomedical cross-language indexing and retrieval system. We describe two complementary maintenance approaches, *viz*. i) corpus-based error detection, and ii) thesaurus anomaly detection. These techniques were developed to detect so-called dynamic and static errors, which are committed by the lexicographers during the construction and maintenance process. Considering multilingual parallel corpora, the distribution of semantic identifiers should be similar whenever comparing related texts in different languages. In the first approach, those semantic identifiers are identified that exhibit greatest frequency variations when comparing text pairs. A manual review of these search results is supposed to spot content errors, which are subsequently classified and fixed by the lexicographers. The second approach analyses transaction-based anomalies, which are identified by interpreting the log of lexicographers' actions during thesaurus maintenance. This methodology highlights the four most common types of this kind of anomaly and evaluates the effectiveness of the corpus-based detection techniques. The overall quality improvement of the thesaurus was evaluated using the OHSUMED IR benchmark.

Keywords Multilingualism, Semantics, Natural language processing, Information storage and retrieval, Thesaurus engineering.

Questões de qualidade em construção de thesaurus: um estudo de caso do domínio médico

Resumo *Assegurar a qualidade de um dicionário médico não é uma tarefa trivial, devido à complexidade inerente à terminologia médica. As peculiaridades da sublinguagem médica e o subjetivismo das escolhas dos lexicógrafos complicam o processo de construção do dicionário de sinônimos. Nossa experiência baseia-se no léxico do sistema MorphoSaurus, uma plataforma básica de indexação e recuperação biomédica para vários idiomas. Neste artigo, descrevem-se duas abordagens complementares de manutenção: detecção de erros baseada em Corpus e detecção de anomalia de Thesaurus, que são usados para detectar os chamados erros dinâmicos e estáticos, introduzidos pelos lexicógrafos durante o processo de construção e manutenção. Considerando corpora paralelos multilinguais, a distribuição dos identificadores semânticos devem ser semelhantes, sempre quando textos relacionados são comparados em diferentes idiomas. Na abordagem proposta, a pesquisa é feita para identificadores semânticos que têm maiores variações entre pares de textos. A análise desses resultados de pesquisa expõe os identificadores de itens lexicais e que pode revelar erros, que são posteriormente classificados e fixados pelo lexicógrafos. Outro ponto é que as análises baseadas em anomalias baseadas em transações que são gerados pelo log de ações lexicógrafos durante a manutenção de sinônimos. Esta metodologia destaca os quatro tipos mais comuns de anomalia e avalia a eficácia das técnicas de detecção baseado em corpus. A melhoria da qualidade global do dicionário de sinônimos foi avaliada utilizando o benchmark OHSUMED IR e todo o processo apresenta uma melhoria considerável da qualidade de recuperação para os idiomas testados.*

Palavras-chave *Multilinguismo, Semântica, Processamento de linguagem natural, Armazenamento e recuperação da informação, Engenharia de thesaurus.*

*e-mail: percy.nohama@gmail.com
Recebido: 10/09/2010 / Aceito: 22/12/2011

Rev. Bras. Eng. Biom., v. 28, n. 1, p. 11-22, março 2012
Braz. J. Biom. Eng., 28(1), 11-22, March 2012

Introduction

Text retrieval in the medical domain presents several challenges, as medical terminology follows its own rules, and medical text collections are highly diverse, dependent on their genre, e.g. textbooks, websites, scientific articles, or routine documentation in medical record systems. Professionals and laypersons use highly different jargons, and the purposes for which the texts are produced highly impacts on their grammatical and orthographic correctness.

It is increasingly recognized (Markó *et al.*, 2005b) that the growth of data in the health care and life sciences domain demands a consensus on the terms and language used in documentation and communication. In spite of recent advances in biomedical terminologies, classifications and ontologies (Freitas *et al.*, 2009), most relevant information is still conveyed by free-text documents only. Coding and semantic annotation using controlled vocabularies is costly when done manually (hence limited to specific use cases like literature indexing and disease encoding), and error-prone when done automatically.

On a global scale, multilingualism of medical documents is an important issue: the global tendency of using English as the primary language in research is contrasted with the use of local languages for patient-related documentation and communication (Schulz and Hahn, 2000).

Finally, medical language is extremely dynamic. New terms, mostly single or multiword compounds and acronyms are constantly created, and English terminology increasingly permeates non-English medical documents.

All these factors hamper the use of simple text retrieval techniques such as popularized by Web search engines for efficient medical information retrieval.

After one decade of intensive research, search and recovery techniques for multilingual content have undergone a considerable evolution (Gey, 2001). These techniques are usually based on the application of either bi or multilingual dictionaries or of collections of texts, paragraphs, or sentences in two different languages, so-called parallel *corpora*. Due to the difficulty in getting sufficiently large parallel corpora for a specific domain, cross language information retrieval (CLIR) mechanisms are mostly based on dictionaries (Oard, 1997). Although the popularity of cross-language document retrieval platforms is still limited, CLIR constitutes, nevertheless, an intense research area (Peters, 2006).

Domain-specific collections of semantically related terms are generally named thesauri. Their main purpose is the semantic representation of a domain terminology in order to support classification and content retrieval (Frakes and Baeza-Yates, 1992; Hersh, 1996).

Thesaurus engineering is an iterative process, which usually involves experts who are familiar with the domain. Their activities are generally directed by guidelines, which, however, never avoid individual arbitrariness. Controversies especially arise in relation to boundary decisions such as:

- whether a term is pertinent for a given domain and therefore relevant for the thesaurus;
- whether to include composed or derived terms once their components or base forms are already in the thesaurus;
- whether to recognize two terms as synonymous;
- whether senses need to be distinguished when dealing with ambiguous terms;
- whether to adopt additional senses which are of marginal importance to the domain.

Usually, the thesaurus development and maintenance process depends on a team of domain experts (also named thesaurus curators). Such teams may consist of ten or more people that collaborate simultaneously. Therefore, care must be taken to assure that the decisions on borderline cases occur in consensus and in accordance with pre-established guidelines. It must be avoided that modifications done by one curator are undone by another one without communication and discussion.

The most common method to track modifications in a database is to analyze register or log files (Bernstein *et al.*, 1987). Log systems have the property of recording all state changes in a database, and they should facilitate the linking between changes and users as well as between state changes and data objects. In a thesaurus, such data objects are terms, semantic identifiers, comments, and relations.

Quality control of the thesaurus content is another issue. According to ISO 9000, quality is the degree to which a set of inherent characteristics fulfils the requirements of a product. As the rationale of a thesaurus is the support of information retrieval (IR), the overall thesaurus quality might be measured by assessing the results of a standardized IR benchmark. This kind of evaluation, called summative evaluation (Alkin *et al.*, 1990) is also suited to track the thesaurus quality across time. However, it is not very helpful for specifically identifying concrete errors or misspecifications that occur during construction and maintenance ("formation" of the thesaurus, hence "formative evaluation"). In order to do this, other approaches must be identified.

In this paper we present typical thesaurus maintenance and quality assurance problems in the context of the MorphoSaurus system, a large multilingual medical thesaurus. In particular, we propose two different fundamental error detecting techniques on the one hand, and a summative benchmark evaluation of the thesaurus' quality on the other hand.

Materials and Methods

MorphoSaurus – subwords as atomic meaning identifiers

The conventional view on human language builds on the hypothesis that words are the basic building blocks of phrases and sentences. In syntactic theories, words constitute the terminal symbols. However, looking at the sense of natural language expressions, evidence can be found that semantic atomicity frequently does not coincide with the word level, which bears methodical challenges even for pretended 'simple' tasks such as tokenization of natural language input. As an example, considering the English noun phrase "high blood pressure", the word limits reflect quite well the semantic composition, whereas this is not the case in its literal translations "verhoogde bloeddruk" (Dutch), "högt blodtryck" (Swedish) or "Bluthochdruck" (German). Especially in domain specific sublanguages such as the medical one, atomic senses are encountered at different levels of fragmentation or granularity. An atomic sense may correspond o word stems (e.g., "hepat" referring to "liver"), prefixes (e.g., "anti-", "hyper-"), suffixes (e.g., "-logy", "-itis"), larger word fragments ("hypophys"), words ("spleen", "liver") or even multi-word terms ("yellow fever"). The possible combinations of these word-forming elements are immense and ad-hoc term formation is common.

As a consequence, a high coverage of a domain-specific lexicon can only be expected if lexical units are restricted to units of atomic senses, which then can be used as building blocks for composed terms at any level of granularity. Therefore, the identification of atomic sense units from texts in order to achieve a basis for the (lean) semantic interpretation of natural language texts is an important requirement of document retrieval, information extraction, and text mining.

The definition of atomic sense units underlies the MorphoSaurus indexing and retrieval system (Markó *et al.*, 2004; 2005b; Schulz and Hahn, 2000), which is the framework of our further deliberations. It maps the content of domain-specific text onto a concept-like interlingua, which entails a semantic standardization which facilitates the retrieval of documents in multilingual collections (http://www.morphosaurus.de).

A sequence of characters is regarded as semantically atomic if the sense conveyed (in a given language and a given domain context) is not univocally derivable from the senses of its constituents (Markó, 2008). The constitution of words is governed by word-forming operations such as inflexion, derivation and composition. Lexical units may have multiple senses (homonymy), and one sense can be expressed by different surface forms (synonymy). For instance, "molar" has one sense in obstetrics ("molar pregnancy"), another one in lab medicine ("molar mass"), or in dentistry ("fractured molar"). "Operation" means "surgical procedure" in the medical domain, opposed to different senses in mathematics or business. In such cases, the local context of the word in focus generally helps to select the right sense.

Besides ambiguity, lexical units may have overlapping senses. Quasi-synonymy relations can hold between terms of different languages (Latin "caput" *vs*. English "head") or different language registers ("belly" *vs*. "abdomen"). Complete identity in sense (strict synonymy) which holds throughout all possible uses of a word is rare.

In order to establish classes of synonymous expressions, clear commitments to the environment in which the expressions can be regarded as synonyms have to be made, *viz*. defining the domain context. Moreover, an agreement has to be found on a sense deviation tolerance which is still compatible with the formal properties of an equivalence relation, *viz*. reflexivity, transitivity, symmetry: If "disease" is considered as synonymous to "illness" and "illness" as a synonym of "sickness", then "disease" and "sickness" are synonyms, as well. The tolerance depends also on the relevance of subtle sense distinctions in the chosen domain context. In medicine "neoplasm", "cancer" and "carcinoma" would hardly be considered synonyms but a different decision may be taken in another domain. A counterexample would be to equalize "excis-", "remov-" and "-ectom-" in a domain of general medicine, neglecting subtle distinctions of surgical techniques.

Translation is a special case of synonymy in which words of different languages are linked. In this case, equivalence can be defined as well, e.g. consisting of English "disease" and "illness", German "Krankheit", Spanish "enfermedad", French "maladie", Swedish "sjukdom", as well as Portuguese "doença".

Not only the grouping of lexical units into synonymy classes, but also their proper delimitation depends on the domain context. "Leukemia", e.g., literally means "white blood", and "neurosis" literally

Rev. Bras. Eng. Biom., v. 28, n. 1, p. 11-22, março 2012
Braz. J. Biom. Eng., 28(1), 11-22, March 2012

means "nerve disease". This may be plausible in a historic view on medicine, but it provides an incomplete description when related to modern medicine. Thus, a composite sense may be ascribed in the historic context, and an atomic one in the present one.

The novel approach of MorphoSaurus is the introduction of so-called *subwords* as lexical units, based on the assumption that neither fully inflected nor automatically stemmed words constitute the appropriate granularity level for lexicalized content description. Especially in scientific sublanguages, we observe a high frequency of complex word forms such as in "pseudo" + "hypo" + "para" + "thyroid" + "ism" (Markó *et al.*, 2006). Assuming that subwords are semantically minimal, we can consider the term "*hepat + itis"* as a composition of two subwords, because their meaning results from the meaning of their constituents, in opposition to "*hypophysis*" whose meaning can not be derived from "*hypo" + "physis*". Subwords therefore tend to be less granular than linguistic morphemes but shorter than words.

In the MorphoSaurus system each subword entry is characterized by attributes such as language (English, German, French, Spanish, Portuguese, Swedish, Italian) and morphosyntactic type, distinguishing between:

- Stems (ST), like "gastr", "hepat", "diaphys", "head", the primary content carriers in a word, which can be prefixed, linked by infixes, and suffixed, some of them may also occur without affixes;
- Prefixes (PF), like "de-", "re-", "in-", "anti-" , "hyper-", which precede a stem once or more;
- Proper Prefixes (PP) such as "peri-", "hemi-", "down-", which are prefixes that themselves cannot be prefixed;
- Infixes (IF), like "-o-", in "gastr-o-intestinal", which are used as a (phonologically motivated) glue between stems;
- Suffixes (SF) such as "-a", "-io", "-ion", "-tomy", "-itis", which follow a stem or another suffix; and
- Proper Suffixes (PS), mostly verb endings like "-ing", "-ed", which are suffixes that cannot be further suffixed.

All these lexeme types are used for segmentation of inflected, derived and composed words, taking into account their compositional constraints. In contrast,

- Invariants (IV), like "ion", "gene", proper names as "aspirin" and acronyms such as "WHO" or "AIDS"

coincide with words and are not allowed as word parts. In most cases, these are short words which would cause artificial ambiguities if they were made available as possible constituents in the deconstruction of complex words.

The semantic layer of the MorphoSaurus system is represented by equivalence classes, identified by so-called MIDs (MorphoSaurus identifiers). Each lexical entry is associated with exactly one equivalence class. Equivalence classes group lexical variants, synonyms and translations which are considered to share the same meaning, in all languages considered. Additionally, MIDs can represent disjunctions of different senses. This is the case when ambiguous lexical units are addressed. To restate the example from above, the disjunction of the different senses of "molar" is represented by one MID, and the non-ambiguous senses by another MID each. Secondly, all lexical units which are assigned to one MID must be fully interchangeable. For example, {'head', 'caput', 'cabec', 'cabez', 'cefal', 'cephal'} would not be a proper reference for one MID, since "head" (in the example denoting a relative anatomical location) has additional senses, at least in a domain context which includes the meaning of "head" as a person.

Given a subword lexicon, a high performance extraction of subwords from large amounts of text is best achieved by the application of finite-state decomposition, derivation and deflection techniques such as described in (Schulz and Hahn, 2000). The MorphoSaurus segmenter is therefore a crucial component of the MorphoSaurus system.

It turned out that lexicon builders' decisions about proper subword delimitation must be driven not only by formal linguistic criteria but also by the proper functioning of the segmentation procedure. This is especially relevant with long and composed words where different valid segmentations are possible. For example, "nephrotomy" may be segmented into* nephr$^{[en,ST]}$ (#kidney) + o$^{[en;sp;pt]IN}$ + tomy$^{[en]SF}$ (#incision) but also in nephr$^{[en]ST}$ + oto$^{[en]ST}$ (#ear) + my$^{[en]ST}$ (#muscle). If the word segmentation routine prefers here a long match starting from the left, the second (erroneous) segmentation would be preferred. Only costly knowledge and deep language processing routines (which are not available in general) would be expected to detect this kind of errors. A pragmatic solution is to include additional synonymous lexeme variants. This means in our example that the sense #kidney is not only represented by nephr$^{[en]ST}$ but also by nephro$^{[en]ST}$ (as well as by nefr$^{[sp;pt]ST}$ and nefro$^{[sp;pt]ST}$).

The MorphoSaurus system uses two types of relations for linking equivalence classes, *viz.* "has_word_part" and "has_sense" (Figure 1):

* en – English, sp – Spanish, pt – Portuguese, ST – Stem, IN – Infix, SF –Suffix.

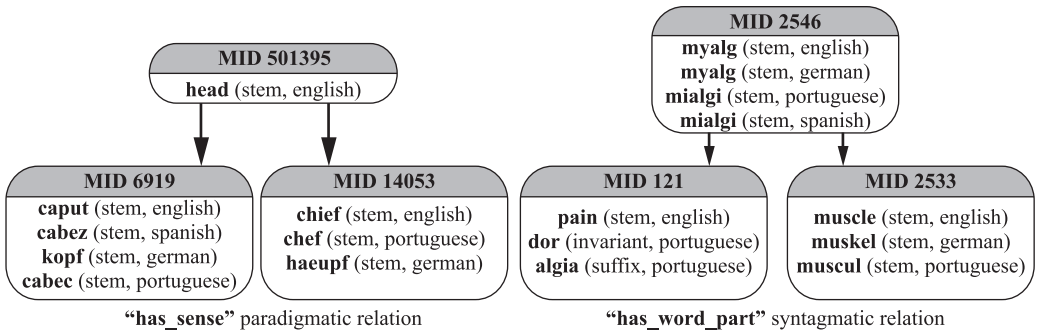

Figure 1. Semantic relations supported by the MorphoSaurus thesaurus. Left: treatment of lexical ambiguity (two alternative senses of English "head"). Right: pre-coded subword combination ("myalg") to preclude erroneous segmentation results.

- **has_word_part** links one MID to an ordered list of MIDs (at least two elements) in order to make a hidden semantic composition explicit. It is generally applied for component terms that cannot be properly split by the segmentation routine, e.g., due to missing characters (e.g., in the word "urinanalys") or due to very short subword components (such as "my" in "myalgia");
- **has_sense** relates an ambiguous MID to at least two other MIDs. This type of relationship is used to correlate it to its possible meanings.

The delimitation of semantic classes is a task that requires considerable knowledge of the domain terminology and therefore cannot be fully automated (Schulz and Hahn, 2000).

For the construction of the thesaurus we have invested more than seven years of work, initially focusing on English and German and then adding Portuguese, Spanish, Swedish, Italian, and French. In this process we capitalized on the fact that there were substantial similarities between medical terms in different languages, so-called cognates. We developed a semi-automated approach to acquire lexical entries of a new language thus optimizing the lexicon acquisition process. Using the Portuguese lexicon, identical and similarly spelled Spanish subword candidates (cognates) are generated. As an example, the Portuguese word stem 'estomag' ('stomach') is identical with its Spanish cognate. An example for a pair of similar stems is 'mulher' ('woman') (Portuguese) *vs*. 'mujer' (Spanish). Similar subword candidates were generated by applying a set of 45 string substitution rules (e.g. lh-j, in Portuguese "*mulher*" to Spanish "*mujer*") as a result of identifying common-language Portuguese-Spanish cognates in a commercial dictionary (Schulz *et al.*, 2004).

The current Morphosaurus version has 3,515 has_sense and 1,506 has_word_part relations, 27,488 English terms, 24,489 German terms, 16,490 Portuguese terms, 14,330 Spanish terms, 10,145 French terms, 15,783 Swedish terms and 8,187 Italian terms.

Figure 2 depicts the embedding of the MorphoSaurus system into a document retrieval framework. First, the orthographic normalization step removes insignificant words and character substitutions are applied (e.g. elimination of capitalization or accents). A morphosyntactic parser then splits each remaining word into subwords. In the semantic normalization step, these units are finally linked to a language-independent MID representation. Queries are processed in an analogous way, thus allowing multilingual search.

The main stages of this process, with English, German and Portuguese examples are shown in Figure 3.

Lexical ambiguity is treated after the normalization process. For example, the Portuguese "*lobo*" may denote either an animal (*wolf*) or a brain structure (*lobe*). The MorphoSaurus system contains a lexical sense disambiguation routine, trained by multilingual corpora. The disambiguation routine chooses the most appropriate meanings of an ambiguous word. A well-known probabilistic model, the *maximum likelihood estimator* was used. For each ambiguous subword at position k with n readings, we examined a window of ± 2 and ± 6 surrounding items. It has been proven that disambiguation substantially increases the overall performance of the system with no manual sense tagging (Markó *et al.*, 2005a).

Underlying methods

Since a lexical resource and the number of its curators constantly grow, maintenance problems also increase

Rev. Bras. Eng. Biom., v. 28, n. 1, p. 11-22, março 2012
Braz. J. Biom. Eng., 28(1), 11-22, March 2012

and therefore require a guided solution. Although the curation of the Morphosaurus lexical database has been based on written guidelines, many borderline decisions cannot be unambiguously subsumed by these guidelines and often produce arbitrary and even irreproducible results.

Two kinds of errors

We here distinguish two main types of errors, *static* and *dynamic* ones. *Static errors* are, predominantly, thesaurus design errors which affect the overall quality of any system that uses MorphoSaurus technology as a component for semantic indexing. In contrast, *dynamic errors* are deficiencies introduced in the curation process due to insufficient coordination and communication between lexicographers. Although the latter kind of error may also have an impact on the overall performance of the Morphosaurus system, it is characterized, above all, by the inefficient use of human resources.

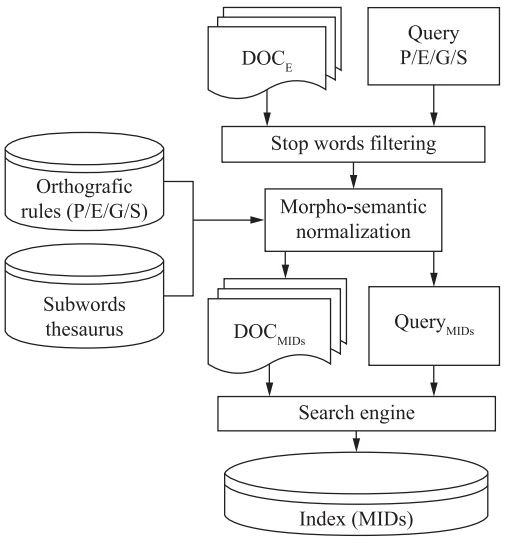

Figure 2. MorphoSaurus' embedding into a document retrieval framework.

Static error detection

Targeted detection of problematic thesaurus entries

According to the current thesaurus curation workflow, the thesaurus curators are using a moderated mailing list in order to facilitate the communication of supposed errors and to support consensus decisions in difficult modeling issues. However, as we have observed, this process tended to be guided rather by expertise than by systematic contemplations. Therefore, one of our objectives was to improve this process by using a more principled thesaurus error detection approach.

This approach is based on the hypothesis that in *closely related corpora* (Fung, 2000; Rapp, 1995) – that is, texts that deal with the same subject-matter in different languages – the statistical distribution of semantic identifiers exhibits a high degree of correspondence. In consequence, any exception to this expected conformity should indicate errors either in the text segmentation and indexing routines or in the representation of semantics, such as weaknesses in the delimitation of equivalence classes or questionable semantic relations between classes. These equivalence classes capture intralingual as well as interlingual synonymy. E.g. Class 512 contains German 'kardiak', 'herz', English 'heart', 'card', Portuguese 'corac', 'cardiac', Spanish 'corazon', 'card', French 'card', 'coeur', Swedish 'cord', 'hjärt', Italian 'card', 'cuor' and all others subwords related to the meaning "heart".

General proposal

Our proposal is to render the lexicographic activities more efficient through guiding lexicographers with a ranked list of supposedly problematic equivalence

Orthographic normalization *Ortographic rules*	**Morphosyntactic parser** *Subword lexicon*	**Semantic normalization** *Subword thesaurus*
High TSH values suggest the diagnosis of primary hypothyroidism...	high tsh values suggest the diagnosis of primary hypothyroidism...	high tsh value s suggest the diagnos is of primar y hypo thyroid ism	#up tsh #value #suggest #diagnost #primar #small #thyre
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose...	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose...	er hoeh te tsh wert e erlaub en die diagnos e einer primaer en hypo thyre ose	#up tsh #value #permit #diagnost #primar #small #thyre
A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário...	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario...	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primary o...	current# value# top# tsh# suggest# diagnos# hypo# thyroid# first#...

Figure 3. Morpho-Semantic Indexing Pipeline. The orthographic normalizer removes capitalized letters and diacritics, the parser identifies lexicalized subwords and the normalizer maps the subwords to language-independent semantic identifiers (MIDs).

(MID) classes. This list is generated out of the comparison between the "semantic extracts" of comparable corpora in different languages, i.e. the pipeline output such as depicted in Figure 3.

In parallel, we want to supervise the progress of the lexicon curation activities through the repeated execution of a summative quality metric. The metric we chose is based on an information retrieval benchmark that had already been applied in previous studies (Honeck *et al*., 2002). This benchmark mirrors the general appropriateness of the thesaurus for medical document retrieval.

Related bilingual Medical Corpora

In order to create frequency distributions between Morphosaurus semantic identifiers (MIDs), generated from a related multilingual corpus, the Merck, Sharp & Dohme (MSD) manual of clinical medicine, a reference handbook of clinical medicine, available for English (EN), Spanish (SP), Portuguese (PT), and German (DE) (http://www.merck.com), was used. This corpus was submitted to the Morphosaurus indexer and a MID frequency table was generated for each language.

Scoring of descriptors

For each language pair, a ranked list was generated which uses both relevance and imbalance measures for ranking:

$$S = \frac{2S_d + S_a}{3} \tag{1}$$

$$S_d = \frac{|f_1 - f_2|}{|f_1 + f_2|} \tag{2}$$

$$S_a = \frac{f_1 + f_2}{(f_{x1} + f_{x2})_{\max}} \tag{3}$$

In the above f_1 and f_2 are the MID frequencies in each corpus, S_d the degree of imbalance and S_a the relation between the frequency of the MID under scrutiny and the frequency of the MID with the highest frequency in both corpora. The overall score S is therefore predominantly influenced by the degree of imbalance but also gives an additional boost to highly frequent MIDs (one third).

Guided by the sequence of problematic MIDs presented by the frequency lists, the curators start to review the thesaurus. The modifications are put down in a spreadsheet that contains the following information: MID, problem description, problem class, solution, and motivation for modification.

Progress assessment

The progress in the thesaurus refinement can be checked in two ways: on the one hand, MID frequency lists can be periodically generated, expecting a decrease of indexes. On the other hand, the performance of a multilingual document retrieval system can be measured, using the MorphoSaurus system for indexing documents and queries.

In order to draw better conclusions for the proposed error detection methodology, we choose the second approach.

Precision and recall were chosen as performance parameters in an IR system. Precision is the proportion of relevant documents among all retrieved ones; recall is the rate of relevant documents retrieved. In IR systems which return all documents in a ranked output, it is possible to measure precision at different recall points, thus obtaining precision / recall diagrams. By interpolation, precision values are computed at defined recall points. As an overall assessment parameter we used the eleven point average value (AvgP11), defined as the arithmetic mean of the precision values at the eleven recall points 0.0, 0.1, …, 0.9, 1.0.

As a benchmark, the OHSUMED collection (Markó *et al.*, 2005b) was used, a subset of Medline abstracts that had been manually annotated with regard to their relevance to a set of authentic user queries. In order to use this resource for benchmarking a cross-language retrieval system, all queries had been previously translated to Portuguese, Spanish, and German.

During the correction period (three months), ten thesaurus backups were produced. Each one of these backups was used for a complete IR experiment with the OHSUMED corpus and produced an AvgP11 benchmark value for each of the four languages.

Dynamic error detection

We introduced the concept of dynamic errors as deficiencies in the thesaurus curation process due to insufficient transparency and communication between the curators. As the most serious dynamic error we identified the so-called "do-undo" actions. They represent a very general kind of interaction problem wherever complex resources are maintained by a group of people. "Do-undo" actions consist in the fact that one curator reverses an action done by another curator.

With the purpose of detecting changes in the MorphoSaurus database during a time interval, 86 MorphoSaurus backups were collected. These backups covered a time interval of nine months and represented regular intervals of approximately

Rev. Bras. Eng. Biom., v. 28, n. 1, p. 11-22, março 2012
Braz. J. Biom. Eng., 28(1), 11-22, March 2012

three days between each backup. A script detected all alterations between related data objects in all consecutive backup pairs. Its results were then used as a basis for the automatic anomaly detection. The concept of thesaurus management anomalies is introduced as sequences of actions taken by the thesaurus curators that consume effort without any positive impact on its quality. We distinguish four anomaly types:

- Relationship anomaly: defined as a sequence of editing steps in which a thesaurus relation (*has_sense* or *has_word_part*) between two equivalence classes is first eliminated and later restored;
- Type anomaly: understood 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;
- Delimitation anomaly: considered as a sequence of editing steps in which the string delimitation of a lexicon entry is modified in a first step and later restored to its original form;
- Permanence anomaly: defined as a sequence of editing steps in which an existing lexicon entry first deleted is later recreated.

The whole user data collected from the backups were analyzed and checked whether the MIDs were involved in some editing anomalies cases matched with those detected by the other method.

Results and Discussion

Correction process based upon MID distribution

During the problem analysis process and correction, it became clear that most of the highly scored MIDs in the ranked list spotted real problems which could be solved. Three extreme examples of imbalance between Portuguese and English due to missing MIDs in one of the languages are listed in Table 1. For example, the preposition "from" belongs to a MID marked for indexing, but its Portuguese analogue "de" is marked as a stop word (it is a word that is bypassed in the text analysis) and it is, therefore, ignored for indexing. The most frequent problems are depicted in Table 2:

- The ambiguity is mainly due to ambiguous lexemes (and, accordingly, MIDs) in one language but not in another. In some cases, the ambiguous MIDs found were not mapped to the unambiguous ones and therefore used for indexing. The normal procedure, which consists of substituting an ambiguous MID with the MID which represents its non-ambiguous senses, did not take place. This is a problem that can easily be corrected by including the missing "has_sense" links;
- Missing or dispensable MIDs were common in borderline cases where a lexical entry had a very specific sense which resulted in the fact that an entry was given a semantic identifier in one language but not in another one. An example is the preposition *from* (see discussion above and Table 3). The solution was the creation of a consensus about what is to be considered stop (sub)word, i.e. a lexicon entry which is excluded from indexing;
- The same sense was expressed by different MIDs (which generally did not contain lexemes of all languages). This problem could be solved by merging MIDs;
- Different senses were found in the same MID, and at least one of the senses was also present in another MID (usually with a focus on a different language). The solution consisted of splitting the non-uniform MID and redistributing its entries.

Other problems occurred with a quite low frequency, e.g. problems of string delimitation.

Table 1. MID frequencies (f_1: English text, f_2: Portuguese text), related parameters and ranking score (S).

MID	EqClass	f_1	f_2	S
peopleriixypa	500783	6352	0	0.7155
fromiwiixxa	060077	4676	0	0.7026
icasikprrr	023555	0	3022	0.6899

Table 2. Problems identified during the MID corrections.

Reason for MID high score	pt[2] / en (%)	ge / en (%)	sp / en (%)
Ambiguities	23	38	14
Missing or dispensable MID	49	18	53
Same sense in different MIDs	6	12	19
One MID with different senses	4	5	6
No problem	11	10	4
Unclassified	7	17	4

[2]pt – Portuguese, en – English, ge – German, sp – Spanish.

Table 3. Number of anomaly occurrences found by log analysis. Numbers in parentheses indicate anomalies picked up also in the discussion forum.

Anomaly type	Count
AR – Relationship anomaly	76 (28)
AT – Type anomaly	18 (18)
AD – Delimitation anomaly	0 (0)
AP – Permanence anomaly	5 (4)

These problems had rarely any impact on the MID distribution.

The results of the summative evaluation are depicted in Figure 4 that illustrates the evolution of the thesaurus, using the IR benchmark previously described. The AvgP11 values were calculated at 10 points within an evaluation period of nine weeks and during this time about one hundred working hours were invested by experienced thesaurus curators.

For none of the languages under scrutiny there was a monotonous performance increase. Comparing the first and the last AvgP11 value, there is a relatively insignificant growth for Portuguese values (1.8%), and for German (2.6%). This improvement occurred mainly due to the addition of relations between MIDs and to the rearrangement of existing MIDs. We also found an IR performance decrease of 1.9% for English. This value, as the increase of German and Portuguese, lies within the range of expected variation, especially considering that the benchmark does not measure the whole information space, but precisely the IR performance of a sample with 106 queries. Simple modifications should not make a considerable difference in a consolidated resource. In contrast, the Spanish benchmark increment amounted to impressing 53%. This increase supports the hypothesis that prioritization of curation tasks – as done by our error detection approach – can result in a good performance boost.

The different degrees of maturity between the language-specific thesauri became obvious when comparing the values from Table 1. The main difference is the rather low rate of missing or unnecessary MIDs for German/English. This fact stems from the maturity of the German data in the thesaurus, and also to a more concordant treatment of stop words in this language pair.

Another interesting fact is that 10% of the MID disparities could not be attached to any thesaurus error, but are a consequence of a lexical ambiguity in one language that is not paralleled to the other language. The resolution of an ambiguous MID can give rise to one high-frequency MID in one language but not in the other one. As we disambiguate using the expected frequency (as described above) one or more readings happen to be ignored. For instance, if for the English noun "head" the reading *caput* is preferred over the reading *boss* (due to the frequency distribution in the other languages), the latter is supposed to occur with a lower frequency (*viz.* only where there is a occurrence of the word "boss") compared to other languages in which there is no ambiguous term analogous to "head".

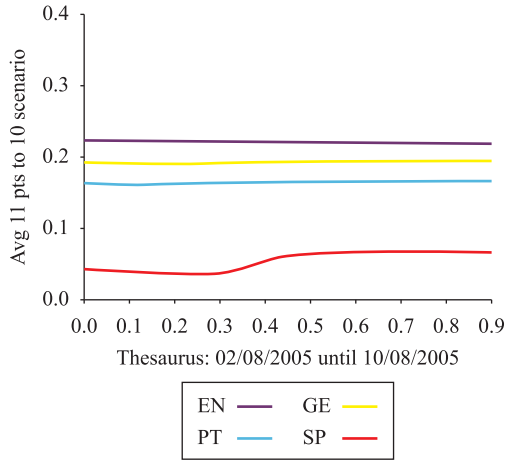

Figure 4. Average of eleven point average value (AvgP11) evolution for English, Portuguese, German, Spanish.

Analysis of anomalies

The log analysis yielded a total of 146 anomalies. There were many occurrences of relationship anomalies, and 23 MIDs were identified which exhibited a relationship anomaly more than once during the observation period. Counting the multiple occurrences only once, we got 99 different anomalies, as shown in Table 3.

We juxtaposed this data to the discussion forum focusing on English, Portuguese and German where 325 problems had been addressed as shown in Table 4. The acquisition of the anomaly data had been completely independent of the problem analysis discussion. By the comparison of the two sources we can analyze the anomalous MIDs which were chosen from the discussion forum.

The collected data show that more than a third (36 of the 99) anomalous MIDs were not addressed in the discussion forum, and also that these discussions covered much more cases than identifiable by the log analysis (cf. values in parentheses in Table 3).

In Table 4, the values in parentheses indicate the frequency of those MIDs which had also been spotted by the log analysis. Eventually, Table 5 gives a closer view on the multiple occurring anomalies.

The correct handling of lexical ambiguities is the most error-prone step in the thesaurus management process. This was evidenced not only by the relationship anomaly frequency (AR, cf. Table 2) but also by its occurrence in the discussion forum. This anomaly was also the only one where up to seven editing repetitions occurred, a fact that highlights a considerable waste of resources and lack of communication in the thesaurus building process.

Rev. Bras. Eng. Biom., v. 28, n. 1, p. 11-22, março 2012
Braz. J. Biom. Eng., 28(1), 11-22, March 2012

Table 4. Number of problem occurrences in the discussion forum.

Problem type	Count
PR – An expected relation between an ambiguous MID and MIDs (*has_sense* type) or an expected expansion relation (*has_word_part* type) was missing.	86 (24)
PC – Entries assigned to one MID did not cover all languages.	80 (6)
PU – The same sense is represented by two unrelated MIDs.	70 (8)
PM – Lexicon entries assigned to one MID diverge in meaning.	11 (1)
PL – Language-specific entries do not translate to other languages.	11
PO – Orthographic errors.	16 (1)
PI – Similar senses are represented by two unrelated MIDs, one of them of the type "excluded from indexing".	31
PD – Errors caused by incorrect subword delimitation.	16
PS – Errors caused by incorrect functioning of the segmentation engine.	4

Numbers in parentheses indicate the cases where the same problem was posted to the forum and had been identified by the log analysis independently.

Table 5. Anomalies AR: multiple changes related to one MID found by log analysis (left column). Number of MIDs which exhibit multiple changes (right column).

#Changes	Count
2	4 (1)
4	16 (6)
5	2 (2)
6	2 (0)
7	23 (10)

Numbers in parentheses indicate the anomalies picked up in the discussion forum.

The assignment of lexemes to equivalence classes (MIDs) was also subject to changes (AT anomaly), but all these cases were addressed in the discussion forum. This shows the effectiveness of the corpus-based detection of discrepancies in the MID distribution and its good take-up by the thesaurus curators.

It was quite surprising that no string delimitation anomaly (AD) could be observed. This is probably due to the fact that observed segmentation problems were always solved by adding new string variations (e.g. "*-otomy*" in addition to "*-tomy*") instead of modifying the existing ones, in accordance to 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 largest discrepancy in the corpus analysis, due to the fact that the lexemes with more semantic importance occur with a higher frequency. For instance, the case where the preposition "from" was assigned to a "*for indexing*" MID, and the Portuguese translation "de" was assigned to a "*not for indexing*" MID, was the first on the ranked list of lexicon discrepancies, and was therefore preferentially addressed in the discussions. These types of "stop words" are frequent and equally distributed across the whole document space. Hence, they are irrelevant for distinguishing documents in text retrieval scenarios (Frakes and Baeza-Yates, 1992), but 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 the case the preposition "*from*" was neglected.

The anomaly detection process would be more valuable if the anomalous MIDs could be detected in execution time, in such a way that the thesaurus curators could get an immediate feedback whenever an action executed earlier was undone. A new version of the MorphoSaurus editing tool, currently under development, is implementing this functionality.

Conclusions

Thesaurus management is a highly dynamic process and the kind of decisions which have to be continuously taken imposes challenges on the community of thesaurus curators.

The waste of resources due to the phenomenon that one person undoes an action which another person has previously performed is considerable and, unfortunately, no guideline for thesaurus management can ever foresee all borderline cases. Such cases can only be solved by consensus. The proposed technique of edition-based anomaly detection is useful to discover these problems. However, such *process quality oriented* auditing techniques should be complemented by *thesaurus content oriented* methods, such as the analysis of frequency distribution patterns in comparable corpora.

In the context of the MorphoSaurus system, we could provide empirical evidence that the analysis and correction of the most relevant unevenly distributed MIDs had an important impact for the least developed language in our thesaurus – Spanish – on the performance of a text retrieval system supported by MorphoSaurus.

A seamless integration, available in many kinds of available tools, of such quality assessment routines in the thesaurus management tools is necessary for achieving higher process effectiveness.

Acknowledgements

This work was supported by the CNPq, Brazil (550830/05-7) and by the German-Brazilian cooperation project DLR-IB (BRA 03/013). Medical text retrieval technology based on the MorphoSaurus approach is available by Averbis GmbH (http://www.averbis.de).

References

Alkin MC, Weiss CH, Patton MQ. Debates on Evaluation. Newbury Park: Sage Publications; 1990.

Bernstein P, Hadzilacos V, Goodman N. Concurrency control and recovery in database systems. Massachusetts: Addison-Wesley; 1987.

Frakes WB, Baeza-Yates R. Information Retrieval: Data Structures & Algorithms. New Jersey: Prentice Hall; 1992.

Freitas F, Schulz S, Moraes E. Survey of current terminologies and ontologies in biology and medicine. RECIIS - Electronic Journal in Communication, Information and Innovation in Health. 2009; 3(1):7-18. http://dx.doi.org/10.3395/reciis.v3i1.239en

Fung P. A statistical view of bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In: Véronis J, editor. Parallel Text Processing: Alignment and Use of Translation Corpora. Dordrecht: Kluwer Academic Publishers; 2000. p. 219-36.

Gey FC. Research to improve cross-language retrieval - position paper for CLEF. In: Peters C, editor. Cross-Language Information Retrieval and Evaluation Workshop of Cross-Language Evaluation Forum: CLEF 2000. Lecture Notes in Computer Science 2001; 2069:83-8. http://dx.doi.org/10.1007/3-540-44645-1_8

Hersh WR. Information Retrieval: A Health Care Perspective. New York: Springer, 1996.

Honeck M, Hahn U, Klar R, Schulz S. Text retrieval based on medical subwords. Studies in Health Technology and Informatics. 2002; 90:241-5.

Markó K, Hahn U, Schulz S, Daumke P, Nohama P. Interlingual indexing across different languages.In: RIAO'04: Proceedings of the 7th International Conference "Recherche d'Information Assistée par Ordinateur" – RIAO'04; 2004 Apr 26-28; Avignon. Avignon; 2004. p. 82-99.

Markó K, Schulz S, Hahn U. Unsupervised multilingual word sense disambiguation via an interlingua. In: AAAI '05: Proceedings of the 20th National Conference on Artificial Intelligence – AAAI '05; 2005a Jul 9-13; Pittsburgh. Pittsburgh; 2005. p. 1075-80.

Markó K, Schulz S, Hahn U. MorphoSaurus – design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. Methods of Information in Medicine. 2005b; 44(4):537-45.

Markó K, Baud R, Zweigenbaum P, Borin L, Merkel M, Schulz S. Towards a multilingual medical lexicon. Proceedings of the AMIA Annual Symposium. 2006; 8(Pt 1):534-8. PMcid:1839525.

Markó K. Foundation, Implementation and Evaluation of the MorphoSaurus System: Subword Indexing, Lexical Learning and Word Sense Disambiguation for Medical Cross-Language Information Retrieval [thesis]. Jena (Germany): Friedrich-Schiller University; 2008. Available from: http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-16871/Marko/Dissertation.pdf

Oard DW. Alternative approaches for cross-language text retrieval. In: Hull D; Oard DW, editors. Working Notes of AAAI´97 Spring Symposiums on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence; 1997. p. 154-62.

Peters C. What happened in CLEF 2006. In: Peters C, Clough P, Gey FC, Karlgren J, Magnini B, Oard DW, Rijke M, Stempfhuber M, editors. Evaluation of Multilingual and Multi-modal Information Retrieval. Lecture Notes in Computer Science 2007; 4730:1-10.

Rapp R. Identifying word translations in nonparallel texts. In: Annual Meeting of the Association for Computational Linguistics: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics; 1995 Jun 26-30; Cambridge. Cambridge; 1995.

Schulz S, Hahn U. Morpheme-based, cross-lingual indexing for medical document retrieval. International Journal of Medical Informatics. 2000; 58-59:87-99. http://dx.doi.org/10.1016/S1386-5056(00)00078-2

Schulz S, Markó K, Sbrissia E, Nohama P. Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In: International Conference on Computational Linguistics: Proceedings of the International Conference on Computational Linguistics; 2004 Aug 23-27; Geneva, Switzerland. p. 813-9. http://dx.doi.org/10.3115/1220355.1220472

Rev. Bras. Eng. Biom., v. 28, n. 1, p. 11-22, março 2012
Braz. J. Biom. Eng., 28(1), 11-22, March 2012

Authors

Percy Nohama, Edson José Pacheco
Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná – PUCPR,
Rua Imaculada Conceição, 1155, Bloco CCBS, 2º andar, Prado Velho, CEP 80215-901, Curitiba, PR, Brazil
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial,
Universidade Tecnológica Federal do Paraná – UTFPR, Av. Sete de Setembro, 3165, Rebouças,
CEP 80230-901, Curitiba, PR, Brazil

Roosewelt Leite de Andrade
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial,
Universidade Tecnológica Federal do Paraná – UTFPR, Av. Sete de Setembro, 3165, Rebouças,
CEP 80230-901, Curitiba, PR, Brazil

Jeferson Luiz Bitencourt
Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná – PUCPR,
Rua Imaculada Conceição, 1155, Bloco CCBS, 2º andar, Prado Velho, CEP 80215-901, Curitiba, PR, Brazil

Kornél Markó
Institut für Medizinische Biometrie und Medizinische Informatik, Albert-Ludwigs-Universität, Freiburg, Germany

Stefan Schulz
Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria