

Automatic Assignment of SNOMED Categories: Preliminary and Qualitative Evaluations

P. Ruch, I. Tbahriti, J. Gobeill, R. Baud, C. Lovis and A. Geissbühler

Medical Informatics Service

1201 Geneva

Switzerland

patrick.ruch@sim.hcuge.ch

Abstract

In this paper, we describe the design and preliminary evaluation of a new type of browsing tools for the SNOMED CT terminology. The proposed system can be used either as a search tool to browse the terminology or as a categorization tool to support automatic annotation of textual contents with SNOMED concepts. The general strategy is similar for both tools and is based on the fusion of two complementary retrieval strategies with thesaural resources. The first classification module uses a traditional vector-space retrieval engine, which has been fine-tuned for the task, while the second classifier is based on regular variations of the term list. For quantitative evaluation purposes, the system uses a sample of MEDLINE and the subset of SNOMED concepts restricted to Medical Subject Headings (MeSH) using the SNOMED-MeSH mapping provided by the UMLS (version 2006). Preliminary results show that performances of the hybrid system are significantly improved as compared to each single module. For top returned concepts, a precision of more than 80% is observed. In addition, a manual and qualitative evaluation on a dozen of MEDLINE abstracts suggests that SNOMED CT could represent an improvement compared to existing broad medical terminologies such as the MeSH. Although the precision of the SNOMED categorizer is sufficient for semi-automatic coding tasks, it is concluded from our results that SNOMED-based benchmarks as well as usability studies are needed to assess the relevance of the text-to-SNOMED associations provided by the tool.

1 Introduction

SNOMED CT, the Systematized Nomenclature of Medicine Clinical Terms, represent an important advances in the field of biomedical terminological resources. Its broad coverage could be useful for several healthcare and management applications, in particular when deep semantic interoperability is of strategic importance: assisted generation of patient summary, automatic processing of patient record, billing, fine-grained epidemiology studies, data and text mining. However, this broad coverage can also be regarded as a major issue for SNOMED. Because it contains more than 380 000 concepts, total-

izing about 800 000 synonyms, the practical use of SNOMED will demand new types of tools to search and navigate intuitively in the term collection. The tool we describe in this paper is basically a terminology search tool. In a unique interface it combines: an interactive browser, which returns a set of possible matches in the terminology given a short input text; a categorizer, which attempts to assign a ranked set of category given a long input text (such as the abstract of an article, a full-article or a clinical narrative from a patient record); a passage retrieval tools, which associates each proposed SNOMED category with a short passage in the input document, so that the user can assess the quality of the automatic categorization by looking at the context, as explored for automatic curation (Ehrler et al., 2005) of the Swiss-Prot databases in molecular biology.

An example of the output of the tool is given in Figure 2 for the categorization mode. In that figure, the abstract in Figure 1 has been sent to the system: the association score, which expresses the similarity between the input text and the category, is represented by a progress bar, then the term of the category follows (the top returned concept in this document is *Burkholderia cepacia*), together with its SNOMED identifier, and finally a short passage (*The production of exopolysaccharides (EPSs) by a mucoid clinical isolate of Burkholderia cepacia involved in infections in cystic fibrosis patients, was studied. Depending on the growth conditions...*), which gives the context of the category in the input text is displayed. If only the title of the abstract is given to the system, i.e. when a short text is entered, the resulting output follows the browser mode, i.e. the ranked list contains more codes and the prediction no passage is given (Figure 3).

If we attempt to compare the tool with the well-known CLUE Browser, we should observe that hierarchical visualization is not available in our tool, while it is an important functionality in the CLUE system, which can be seen as complementary to our system. On the opposite, the CLUE interface, as a strict browsing system, cannot accept as input a full document. In addition, the ranking and matching power of our categorization system, which is based on an information retrieval engine, cannot be compared to the limited CLUE skills with respect to string approximation (morphological flexions and

The production of exopolysaccharides (EPSs) by a mucoid clinical isolate of *Burkholderia cepacia* involved in infections in cystic fibrosis patients, was studied. Depending on the growth conditions, this strain was able to produce two different EPS, namely PS-I and PS-II, either alone or together. PS-I is composed of equimolar amounts of glucose and galactose with pyruvate as substituent, and was produced on all media tested. PS-II is constituted of rhamnose, mannose, galactose, glucose and glucuronic acid in the ratio 1:1:3:1:1, with acetate as substituent, and was produced on either complex or minimal media with high-salt concentrations (0.3 or 0.5 M NaCl). Although this behavior is strain-specific, and not cepacia-specific, the stimulation of production of PS-II in conditions that mimic those encountered by *B. cepacia* in the respiratory track of cystic fibrosis patients, suggests a putative role of this EPS in a pathologic context.

Burkholderia cepacia*; Carbohydrate Conformation; Carbohydrate Sequence; Comparative Study; Culture Media*; Cystic Fibrosis*; Glucose; Glycerol; Human; Molecular Sequence Data; Onions; Phenotype; Polysaccharides, Bacterial*; Temperature

Figure 1: Citation with MeSH terms provided by professional indexers for PMID: 11506920.

Figure 2: Output of the tool (categorization mode): this six categories are associated to the abstract shown in Figure 1. Some strings are duplicated because they refer to different concepts. The two top ranked concepts (*Burkholderia cepacia*; *Cystic fibrosis*;) are precisely those expected by the manual MeSH annotation of the article. Categories proposed at lower ranks (*glucuronic acid*; *fibrosis*) are irrelevant regarding the manual annotation performed by NLM (National Library of Medicine) librarians.

derivations as in *expressions*, *expresses*, *expressed*, *expressive*...) and statistical ranking functions.

The remainder of this paper is organized as follows: the next section presents the data and metrics used in our experiments. Then, we present the methods used to perform the categorization task. Further, we propose a preliminary evaluation of the categorizer based on MEDLINE records, together with a qualitative evaluation, based on a few examples, which tries to exhibit SNOMED-specific features. Finally, we conclude on our experiments and suggest some future works to deliver an integrated and user-friendly system.

2 Data and metrics

Because MEDLINE is indexed with Medical Subject Headings (MeSH) rather than with SNOMED codes, we need to transform MeSH terms, which are

Figure 3: Output of the tool (browsing mode): nineteen categories are displayed. The score associated with every predicted category drops after one or two terms, showing that the categorization is not reliable.

used to index MEDLINE records, into SNOMED terms. This translation is done thanks to the MeSH-SNOMED mapping table provided by the Unified Medical Language System. However, the translation process is not bijective: it means that several MeSH terms cannot be mapped appropriately to SNOMED codes and vice versa. Thus, non medically-specific MeSH terms have usually no equivalent in SNOMED. Thus, technical categories methods, such as *Storage*, *data*, biological and pharmacological entities, such as *chaperonin* -which is mapped to *protein*- have often no appropriate equivalent in SNOMED. In all our experiments, we will assume that such an information loss is minor from a biomedical perspective and that the UMLS mapping between MeSH and SNOMED is sufficient to provide a basic evaluation of our SNOMED categorization tool. In order to assess our SNOMED categorizer, we apply the system on the Cystic Fibrosis¹ (CF) collection (Shaw et al., 1991), The CF collection is a collection of 1239 MEDLINE citations. From this collection, 239 records were used for tuning our system and 1000 were used to evaluate our system. In the collection, MeSH items listed in MeSH fields are replaced by SNOMED codes. We demand that SNOMED codes are unique so that

¹Available on Marti Hearst's pages at <http://www.sims.berkeley.edu/hearst/irbook/>

when two or more MeSH are mapped to the same SNOMED code, only one category is accepted. For each citation, we used the content of the abstract field as input for the categorizer. The average number of concepts per abstract in the collection is 12.3 and the average number of major terms is 2.8. From our MEDLINE records, only terms marked as major (with a star) are considered in our experiments.

Following (Larkey and Croft, 1996) and as it is usual with retrieval systems, the core measure for the evaluation is based on mean average precision. The top precision (interpolated $Precision_{at\ Recall=0}$), which is of major importance for a fully-automatic system, is also given.

3 Method

Two main modules constitute the skeleton of our system: the regular expression component, and the vector space component. The former component uses tokens as indexing units and can take advantage of the thesaurus, while the latter uses stems (i.e. strings such as *expression*, *expressed* are replaced by *express*). Each of the basic classifiers uses known approaches to document retrieval. The first tool is based on a regular expression pattern matcher. Although such approach is less used in modern information retrieval systems², it is expected to perform well when applied on relatively short documents such as SNOMED terms. It is to be observed that some SNOMED codes are particularly lengthy, at least when compared to MeSH terms: while MeSH terms are shorter than 6 tokens, there are several terms SNOMED with more than a dozen of tokens. The second classifier is based on a vector-space engine (Ruch, 2002). This second tool is expected to provide high recall in contrast with the regular expression-based tool, which should privilege precision.

For a short introduction on automatic text categorization in MEDLINE, the reader is referred to the NLM's indexing initiative (Aronson et al., 1999); for a detailed presentation of our vector space engine and a comparison with state-of-the-art systems, including NLM's tools, see (Aronson et al., 2006)³ (Ruch et al., 2003). For a complete overview and evaluation of our categorization system applied on Medical Subject Headings and on the Gene Ontology, see (Ruch, 2006).

3.1 SNOMED pre-processing

To be able to better associate SNOMED terms and textual entities, it is necessary to perform some pre-processing normalization steps. This includes removing meta-abbreviations, which are common in most terminological systems, such as NOS (Not otherwise specified) or NES (Not elsewhere classified),

²With a notable exception, the GLIMPSE system (Manber and Wu, 1994).

³In this joint evaluation between four retrieval systems, our engine showed competitive performances.

Table 1: Results for RegEx and (tf.idf) classifiers. weighting schemas. For the VS engine, tf.idf parameters are provided: the first triplet indicates the weighting applied to the "document collection", i.e. the concepts, while the second is for the "query collection", i.e. the abstracts.

System or parameters	Top precision	11pt Average precision
RegEx	0.641	0.400
tf.idf (VS)		
inc.atn	0.696	0.35525
anc.atn	0.691	0.3545
ltc.atn	0.75	0.33525
ltc.lnn	0.637	0.2775

NOC (Not otherwise classifiable), NFQ (Not further qualified)... But we also need to handle and expand more than fifty SNOMED specific abbreviations, often inherited from Read codes, such as *ACOF*; *ADVA*; *AR*; *CFIO*; *CFSO*; *FB*; *FH*; *FHM*; *HFQ*; *LOC*; *MVNTA*; *MVTA*...

3.2 Vector space system

The vector space (VS) module is based on a general information retrieval engine⁴ with *tf.idf* weighting schema. In this study, it uses stems (Porter-like, with minor modifications) as indexing terms, and an English stop word list. While stemming can be an important parameter, whose impact is sometimes a matter of discussion (Hull, 1996), we did not notice any significant differences between the use of tokens and the use of stems. However, we noticed that a significant set of semantically related stems should have been opportunely conflated in the same indexing unit: for example, the morpheme *immun* is found in 48 different stems, and morpheme-based word conflation (Baud et al., 2005) could be helpful to enhance the recall of the current method, especially in multilingual contexts (Ruch, 2004). Altogether, we counted 72 402 unique stems in the SNOMED vocabulary.

A large part of this study was dedicated to tuning the VS engine, and tf.idf weighting parameters were systematically evaluated. The conclusion is that cosine normalization was especially effective for our task. Thus, in table 1, the top-4 weighting function uses cosine as normalization factor. We also observed that the *idf* factor, which was calculated on the whole SNOMED collection performed well, it means that SNOMED is large enough to effectively underweight non-content words (such as *disease* or *syndrome*), which are very frequent in medical vocabularies but convey little meaning in the domain. Calculating the *idf* factor on a large collection of abstracts could have been investigated, but such solution may have resulted in making the system more collection-dependent (Gobeill et al., 2006).

⁴The engine, easyIR, is available on the first author's homepage.

3.3 Regular expressions and synonyms

The regular expression (RegEx) pattern matcher is applied on the SNOMED concepts (376 212) augmented with its synonyms (the total includes 787 091 terms). In this module, text normalization is mainly performed by removing punctuation (hyphen, parenthesis...). The manually crafted transition network of the pattern-matcher is very simple, as it allows one insertion or one deletion within a SNOMED term, and ranks the proposed candidate terms based on these basic edit operations following a completion principle: the more tokens are recognized, the more the term is relevant. The system hashes the abstract into 6 token-long phrases and moves the window through the abstract. The same type of operations is allowed at the token level, so that the system is able to handle minor string variations, as for instance between *diarrhea* and *diarrhoea*. Let us observe that the most frequent morphological variations are usually provided by SNOMED synonyms.

Table 1 shows that the single RegEx (mean average precision = 0.4 and top-precision = 0.64) system performs better than the different settings tested for the vector space module, with $tf.idf^5$ (term frequency-inverse document frequency) and length normalization (cosine...) factors, so that the thesaurus-powered pattern-matcher provides better results than the basic VS engine for SNOMED mapping.

4 Results

The hybrid system combines the regular expression classifier with the vector-space classifier. Unlike (Larkey and Croft, 1996) we do not merge our classifiers by linear combination, because the RegEx module does not return a scoring consistent with the vector space system. Therefore the combination does not use the RegEx’s score, and instead it uses the list returned by the vector space module as a *reference* list (RL), while the list returned by the regular expression module is used as *boosting* list (BL), which serves in order to improve the ranking of terms listed in RL . A third factor takes into account the length of terms: both the character’s length (L_1) and the byte size (L_2 , with $L_2 > 3$) of terms are computed, so that long and/or multi-word terms appearing in both lists are favored over short and/or single word terms. We assume that the reference list has exhaustive coverage, and we do not set any threshold on it. For each term t listed in the RL , the combined Retrieval Status Value (RSV) is⁶:

$$RSV_{Hybrid} = \begin{cases} RSV_{VS}(t) \cdot Ln(L_1(t) \cdot L_2(t) \cdot k) & \text{if } t \in BL, \\ RSV_{VS}(t) & \text{otherwise.} \end{cases}$$

⁵We use the (de facto) SMART standard representation in order to express these different parameters, cf. (Singhal et al., 1996) for a detailed presentation. For each triplet provided in table 1, the first letter refers to the *term frequency*, the second refers to the *inverse document frequency* and the third letter refers to a *normalization factor*.

⁶The k parameter is set empirically using the tuning data

Table 2 shows that the optimal $tf.idf$ parameters *lnc.atn* for the basic VS classifier does not provide the optimal combination with RegEx. The optimal combination is obtained with *ltc.lnn* settings⁷. We also observe that the *atn.ntn* weighting schema maximizes the top candidate (i.e. *Precision at Recall=0*) measure, but for a general purpose system, we prefer to maximize average precision, since this is the only measure that summarizes the performance of the full ordering of concepts. However, in the context of a fully automatic system, the top-ranked concepts are clearly of major importance, therefore we also provide this measure.

Table 2: Results of the system when combining the vector space and the regular expression modules.

Weighting function concepts.abstracts	Top Precision	Average Precision
Hybrids: $tf.idf$ (VS) + RegEx		
ltc.lnn	0.800	0.4545
lnc.lnn	0.791	0.453
anc.ntn	0.787	0.4515
atn.ntn	0.823	0.4485

The top precision (82.3%) is in the range of what has been reported elsewhere (Friedman et al., 2004) (Lussier et al., 2001), while the search space of our tool (800 000 terms and 1000 documents) is much larger than in these experiments, which work with some hundreds categories and use sentences rather than abstracts for the categorization. Such a precision means that the top-returned category is one of the expected categories in 8 cases out of 10. The measured mean average precision of almost 50% (0.45%) means that half of the expected categories are proposed by the system.

4.1 Qualitative evaluation and discussion

To conduct the qualitative evaluation, we looked at a sample of twelve abstracts. A unique judge manually controlled the top three categories provided by the SNOMED categorizer to find if there were non-MeSH categories which could have been relevant for indexing the abstract. For eight abstracts, a relevant category was found in the SNOMED ranking. A typical example is given in Figures 2 and 1: thus, the SNOMED terminology contains the concept *glucuronic acid*, which could have been chosen to index the content of the article based on its abstract, but the concept does not exist in the MeSH. We also can observe that the current setting of the system, which favors exact match concepts (via the regular expression module) and content-bearing features (via the document frequency factor) seems somehow able to discard some very lengthy SNOMED concepts such as *nontraffic accident involving collision*

⁷For the augmented term frequency factor (noted a , which is defined by the function $\alpha + \beta \times (tf/\max(tf))$), the value of the parameters is $\alpha = \beta = 0.5$.

of motor-driven snow vehicle, not on public highway, driver of motor vehicle injured. This qualitative observation suggests that the conceptual coverage of SNOMED can be larger than the MeSH and that automatic indexing could be improved regarding recall by using at least some SNOMED codes. This observation must be balanced by our preliminary remarks: several MeSH categories cannot be appropriately mapped into SNOMED CT.

5 Conclusion and future work

We have reported on the development and preliminary evaluation of a new type of categorization and browsing tools for SNOMED encoding. The system combines a pattern matcher, based on regular expressions of terms, and a vector space retrieval engine that uses stems as indexing terms, a traditional *tf.idf* weighting schema, and cosine as normalization factor. For top returned concepts, a precision of more than 80% is observed, which seems sufficient to help coding textual contents with SNOMED categories. A manual and qualitative evaluation on a dozen of MEDLINE abstracts suggests that SNOMED CT could represent an improvement compared to existing broad medical terminologies such as the MeSH. Clearly, further studies will be needed using documents directly annotated with SNOMED codes, as described in (Friedman et al., 2004) and (de Bruijn et al., 1999). Usability studies (Despont-Gros et al., 2004) are also needed to assess the relevance of the text-to-SNOMED associations provided by the tool from a coder perspective. Finally, the system will need integration with a hierarchical viewer in order to provide a semantic interoperability between textual entities, as found in medical reports, and terms, as listed in SNOMED CT.

Acknowledgements

The study has been partially sponsored by the European Union (SemanticMining Network of Excellence, IST FP6 Grant 507505) and the Swiss State Secretariat for Education and Research (Grant 03.0399).

References

- A Aronson, O Bodenreider, H Chang, S Humphrey, J Mork, S Nelson, T Rindfleisch, and W Wilbur. 1999. The indexing initiative. A report to the board of scientific counselors of the lister hill national center for biomedical communications. Technical report, NLM.
- A Aronson, D Demner-Fushman, S Humphrey, J Lin, H Liu, P Ruch, M Ruiz, L Smith, L Tanabe, and J Wilbur. 2006. Fusion of Knowledge-intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents. In *TREC 2005*.
- R Baud, M Nystrom, L Borin, R Evans, S Schulz, and P Zweigenbaum. 2005. Interchanging lexical information for a multilingual dictionary. *AMIA Symposium Proceedings*.
- LM de Bruijn, A Hasman, and JW Arends. 1999. Automatic SNOMED classification - a corpus based method. *Yearbook of Medical Informatics*.
- C Despont-Gros, H Mueller, and C Lovis. 2004. Evaluating user interactions with clinical information systems: a model based on human-computer interaction models. *J Biomed Inform*, 38(3):244–55.
- F Ehrler, A Jimeno, A Geissbühler, and P Ruch. 2005. Data-poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot. *BMC Bioinformatics, Special Issue on BioCreative: A Critical Assessment of Text Mining Methods in Molecular Biology*, 6 (suppl. 1).
- C Friedman, L Shagina, Y Lussier, and G Hripcsak. 2004. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*, 11(5):392402.
- J Gobeill, I Tbahriti, AL Veuthey, and P Ruch. 2006. Document Frequency Mixture for Effective Functional Annotation in Swiss-Prot. *SemanticMining Summer School*.
- D Hull. 1996. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society of Information Science*, 47(1):70–84.
- L Larkey and W Croft. 1996. Combining classifiers in text categorization. In *SIGIR*, pages 289–297. ACM Press, New York, US.
- Y Lussier, L Shagina, and C Friedman. 2001. Automating SNOMED coding using medical language understanding: a feasibility study. *J Am Med Inform Assoc (Symposium Suppl)*, pages 418–22.
- U Manber and S Wu. 1994. GLIMPSE: A tool to search through entire file systems. In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 23–32, San Francisco CA USA, 17-21.
- P Ruch, R Baud, and A Geissbühler. 2003. Learning-Free Text Categorization. *LNAI 2780*, pages 199–208.
- P Ruch. 2002. Using contextual spelling correction to improve retrieval effectiveness in degraded text collections. *COLING 2002*.
- P Ruch. 2004. Query translation by text categorization. *COLING 2004*.
- P Ruch. 2006. Automatic Assignment of Biomedical Categories: Toward a Generic Approach. *Bioinformatics*, 6.
- W Shaw, J Wood, R Wood, and H Tibbo. 1991. The cystic fibrosis database: Content and research opportunities. *LSIR*, 13:347–366.
- A Singhal, C Buckley, and M Mitra. 1996. Pivoted document length normalization. *ACM-SIGIR*, pages 21–29.