MEDTAG: Tag-like Semantics for Medical Document Indexing

Patrick Ruch, M.S.[1], Judith Wagner[1], Ph.D., Pierrette Bouillon, Ph.D.[2], Robert H. Baud, Ph.D., Anne-Marie Rassinoux, Ph.D.[1], Jean-Raoul Scherrer, M.D.[3]
[1]Medical Informatics Division, University Hospital of Geneva, Switzerland
[2]ISSCO, University of Geneva [3]Geneva Faculty of Medicine

Medical documentation is central in health care, as it constitutes the main means of communication between care providers. However, there is a gap to bridge between storing information and extracting the relevant underlying knowledge. We believe natural language processing (NLP) is the best solution to handle such a large amount of textual information. In this paper we describe the construction of a semantic tagset for medical document indexing purposes. Rather than attempting to produce a home-made tagset, we decided to use, as far as possible, standard medicine resources. This step has led us to choose UMLS hierarchical classes as a basis for our tagset. We also show that semantic tagging is not only providing bases for disambiguisation between senses, but is also useful in the query expansion process of the retrieval system. We finally focus on assessing the results of the semantic tagger.

INTRODUCTION

Already Hyppocrates recommended to his disciples to take written notes about their patients. At the age of the electronic patient record, such wise advice reaches its limit if accessing the relevant information in a reasonable time turns into a Holy Grail quest. We believe that NLP technologies are the best solution to provide such access. State-of-the-art in information retrieval (IR) systems using NLP can be summarized by two propositions:

- Part-of-speech (POS, i.e. syntactic categories) tagging information is mostly irrelevant in the query expansion process, while semantic information is of major importance for adding synonyms and semantically related terms.
- Vice versa: part-of-speech tagging is necessary for good disambiguisation while semantic disambiguisation was claimed to be hardly tractable.

Lately, new tools using tag-like representation and technologies, but dedicated to semantics have emerged. However, semantic tagging has mostly been considered as nothing more than disambiguisation to be performed along the same lines as syntactic tagging. It means, given x lexemes each with y senses, to apply linguistic rules or probabilities - represented as matrices in the case of taggers [1] using HMMs (Hidden Markov Model)- to keep the most likely meaning for each lexical item [2], also in medical informatics[3]. We decided to design a tagset, which will be used for both carrying out the disambiguisation and expanding the query.

We first present the methodology and results related to the semantic tagset, then we focus on the tagging process and performances. While our studies were made on French corpora, most of the examples are given in English for the sake of clarity.

METHODS

The probabilistic approach, although not always more efficient[4] than the rule-based one, was chosen for two reasons. First, for development time: HMM taggers are data-driven and known to be easy to train. Second, for ignorance of semantic rules: unlike syntax, semantic rules and heuristics have not been deeply explored yet[5]. From an epistemic point of view, four main hypotheses are guiding the MEDTAG project:

a. syntax can help to distinguish meanings of words having different syntactic categories;
b. syntactic analysis can be done by a probabilistic tagger and, more speculatively;
c. semantic ambiguity can be solved (mutatis mutandis) by a probabilistic tagger;
d. semantic expansion can also be tags-assisted.

These hypotheses have been tested in the following way: texts are first annotated with our two HMM taggers[6] to make explicit the syntactic and semantic analysis of the words. Then, this information serve to index the text, in order to improve the search results.

Syntactic tagging performances and retrieval improvements (and so hypotheses a, b and d) will not be deeply discussed. Instead the paper will show how to use existing medical terminological resources, and probabilistic tools, for syntactic and semantic annotation in order to improve semantic analysis.

Choice of a corpus

In order to introduce methods for tagging texts in the medical domain, a set of texts has to be carefully selected. In a first approach, these texts should be part of a specific domain and a large number of documents should be available. It was also important to select documents with a large part of free text. Finally, they had to be provided in a defined format, preserving the underlying text structure. We finally picked operation reports from the digestive surgery domain.

1091—8280/99/$5.00

Comparing Public Resources

As the perfect medical ontology is not available and probably will not be available in the immediate future, we will report on our exploration of various terminological systems: SNOMED, GALEN, and UMLS. Other general semantic approaches [7] have also been considered, but not selected as they seemed too general. We also looked at ICD, but the items of ICD, and of classifications in general, do not necessarily correspond to the items found within texts, as they try to group them in classes. Moreover, such entities are too complex (multi-word phrases) for our purposes.

SNOMED would be an extremely interesting source for indexing. Tags could be selected at a higher level within each of the 11 axis, although the links between the items are formalized to a very limited extent. But the content of SNOMED is limited to the medical domain and, as such, does not provide tags for general vocabulary. Nevertheless, SNOMED remains an interesting source to be considered, especially when comparing the content coverage of major clinical classifications with the content of patient records. Although SNOMED (nomenclature, but designated as classification in this comparison) obtained excellent scores [8], some recent studies showed that UMLS had a better content coverage. [9]

The GALEN project aimed at developing a concept reference (CORE) model of medical concepts. It serves as interlingua, and so is language independent. It represents the concepts used in medical records or referred to by other coding systems or nomenclatures. Formally, the GALEN CORE model would provide the best basis for our purposes, as the concepts can furthermore be annotated with words or terms in several languages. These annotations allow concepts to be found via the lexicon entries. The hierarchy would allow tags to be defined at a more general level, and indexes to be attributed at the most detailed level. The hierarchies being multiple, it would be possible to find several aspects of a same concept. At the same time, we could find all the real ambiguities, if a same annotation in a language is available for several distinct concepts. However a fundamental problem with using GALEN for our purposes is its limited domain coverage: about 13000 concepts are contained in the model today, and few were relevant for our texts (digestive surgery reports) when we started our investigations. Furthermore, one major disadvantage of GALEN compared to UMLS or SNOMED is that we do not know about the future and the maintaining of the model.

Finally, we have decided to rely on the semantic types of UMLS, which may be considered as a basic ontology for the domain, as these are quite general and allow the tagset to be limited. The current version of the semantic network contains around 130 classes, and around 50 dyadic relationships. Every entry of the metathesaurus is attributed to one or several classes. Although this network is sometimes regarded as being too general for medical purposes, it seemed to be at the right level for our purposes. Unfortunately, the French part of the metathesaurus is too limited (and without any accent!) to be able to automatically find the semantic type.

Building the tagset

A list of 3641 words was extracted from 90 surgery reports. This list was enriched with the syntactic information provided by the lemmatizer (considering a word, the lemmatizer returns all its possible lexical entries with its lexical features). 20% of words were absent from the 20000 entries of our lexicon and were added manually.

For 600 items of the list, a search has been made for a corresponding entry in the UMLS metathesaurus, and the associated semantic types have been added to the lexical entries. We sometimes gathered different UMLS classes, which were semantically close in order to keep a minimal tagset (cf. tab.1), as one major constraint of HMMs is the size of matrices: a size of 50 is usually considered as the upper limit. We also sometimes had to slightly modify the strict definition of some UMLS classes: for example, in the metathesaurus hierarchy the tag *neop* "is a" *dis*, while the tagset classes are clearly exclusive. Finally we created some domain specific tags (*thers* for surgery procedures). Remaining entries of our lexicon were then tagged on this basis.

It was also necessary to introduce tags for relationships and modal expressions (tagged *mod*, as for example *maybe*). Only some very general ones (the general relationship *rel* for example) have been introduced at this level. As it was not possible to attribute some of the UMLS classes to some categories of words (auxiliary, determiner), we created some very general syntactic tags: *aux* for auxiliary verbs, *def* and *indef* for determiners.

RESULTS

In the lexicon, 86.8 % of items have been provided with exactly one tag (tab.1), while 13.2% have more than one tag (tab.2, 3 tags are very seldom). In both cases, we counted only the tag attributed to the whole form of a lexical item, as for example *hepatectomy* (lexically tagged *loc* + *thers* = *thers*) is counted as a *thers* (cf. Lexicon). The column "label" (tab.1) provides a short definition for each tag.

Tag	Freq.(%)	UMLS	Label (definition)
1-qual	10.1	T080	qualifier
2-acto	9.5	add	general act
3-loc	9.3	T023/T029	organ, body location
4-spat	8.7	T082	spatial concept
5-temp	5.3	T079	temporal concept
6-mod	5.1	add	modal
7-quant	4.7	T081	quantitative concept
8-papr	4.5	T046	pathological process
9-find	4.2	T033	signs or symptoms
10-cpt	4.1	T078	other concept
11-ther	4.0	T061	therapeutic procedure
12-mdev	3.6	T074	medical device
13-thers	3.6	add	surgery procedure
14-hca	2.1	add	physician's act
15-diap	1.9	T060	diagnosis procedure
16-rel	1.9	add	relationships (other)
17-medi	1.7	T121/T103	drugs and chemicals
18-name	1.6	add	for medical techniques
19-dis	1.4	T047	disease or syndrome
20-bosp	1.4	T030	body space or junction
21-bopr	1.3	T039/T040	body process
22-rconj	1.2	add	conjunction relation
23-bosu	1.0	T031	body substance
24-obj	0.9	T073	general object
25-rspat	0.8	T135	spatial relation
26-actp	0.7	add	patient's act
27-mpr	0.7	T041	mental process
28-rtemp	0.7	T136	temporal relation
29-spec	0.7	T091/T097	medical speciality
30-occup	0.7	T090	occupation
31-neop	0.6	T191	neoplastic process
32-pers	0.6	T016	person
33-tiss	0.5	T024	tissue
34-labo results	0.5	T059/T034	laboratory or test
35-subst	0.4	T167	substance (other)
36, 37, 38 respectively aux, def, and indef for respectively auxiliary, definite or non-definite determiner, freq << 0.1%			

Tab.1: Distribution of the semantic tagset within the lexicon, with UMLS classes.

Ambiguity	Freq.(%)	Example
1-hca/acto	11.7	to break
2-actp/thers	9.6	to open
3-thers/bosp	8.3	section
4-rcaus/rtemp	7.9	for
5-qual/find	7.9	fatty
6-mdev/obj	6.7	blade
7-find/dis	5.0	fever
8-mdev/hca	4.0	protection

Tab. 2: Most frequent lexical semantic ambiguities (total: 61.1% of all the ambiguities).

Tuning the tagset

Before being split into 3 tags, there was a tag *act* (UMLS T052, labelled activity) with the highest occurrence (almost 13%). It was due to the fact that many verbs where attached to *act*. Instead of this tag, we created 3 new ones: *actp* for patient's act (like *to suffer*), *hca* for doctor's act (*to diagnose*), and *acto* for any other acts. The second most frequent tag, *qual*, includes many qualifiers, several of which are likely to belong to the signs or symptoms class (UMLS T033, tag *find*) in some special context, but whose meaning is too general out of this context. As for example, *yellow* has a very special meaning when followed by *fever* (cf. tab.2, example 7). The tag for organs and body locations (*loc*) is the most frequent tag that is rather particular to the medical domain. Location modifiers, such as *lateral*, are part of the *spat* tag, which is the next important tag. Tags for therapeutic procedures (*ther*) and surgical therapeutic procedures (*thers*) are more frequent than for diagnostic procedures (*diap*), which is not surprising, as we started our investigations on digestive surgery reports.

Semantic Lexicon

We have two types of lexicon, one for full form words and multi-word expressions (tab.3), and a second for morphemes (tab.4). All of these databases are implemented in a Berkeley-DB library. Here are some records where we removed the syntactic features, we see for example that *appendix* is ambiguous between the organ (*loc*), and an appendix at the end of a document (*obj*).

rec.	full form	morpheme	semantic tag
3120	postcardiotomy	post_cardio_tomy	temp+loc+thers=thers
3127	appendix	appendix	obj/loc

Tab.3. Words and multi-words lexicon

rec.	morpheme	semantic tag
98	hormono	bosu
101	hepat	loc

Tab.4. Morphemes lexicon

As more than 10% of items were compounds, we sort (tab.5) the list of the most frequent lexical compounds:

Compound	Freq.(%), example
1 spat+loc = loc	12.7, perianal
2 loc+loc = loc	11.5
pneumoperitoneum	
3 loc+thers = loc	9.0, gastrectomy
4 temp+thers = thers	8.3, postoperative
5 loc+papr = dis	8.3, nephritis

Tab. 5: Most frequent compounds (total: 50% of all the compounds)

In addition to the semantic tags, we also added links to associated lexical items (kinds of synonyms). As in English for the concept *eyelid*, we have three associated lexical items: *eyelid*, *palpebral* and *blepharo*.

INDEXING BY TAGGING

Before the tagging, a preliminary task consists of training the tagger. A language model -i.e. the probabilities of HMM matrices- is automatically calculated from a hand-tagged sample. The model is refined manually using probabilistic rules (biases).

The first step of the indexing process is to tag the text, first syntactically to disambiguate the part-of-speech, then semantically. This is done sequentially.

POS tagging

Once the text is segmented into words (token) and sentences, the syntactic tag-like information [a] is added. It synthesizes some of the lexical features for each lexical item (these hypothetical tags are separated by |). On this tag-like basis [a], the syntactic tagger finally picks the best tag [b], which provides the part-of-speech [c] for each token. As in the example, *maladie du sigmoïde* (*disease of the sigmode*): *du* (equivalent to *of the* in English) will be finally tagged PREP as preposition, DET-SG (determiner singular) is discarded. For *sigmoïde*, NOUN-SG (SG for singular) is preferred to ADJ-SG (ADJ for adjective):

token	lexical tag(s)[a]	tag[b]	POS[c]
maladie	NOUN-SG	NOUN-SG	noun
du	DET-SG|PREP preposition	PREP	
sigmoïde	ADJ-SG|NOUN-SG	NOUN-SG	noun

Semantic tagging and medical compounds

The final step of our analysis is the semantic tagging, processed along the same lines as the syntactic one. Once the part-of-speech [a] of the token is given, a lexical access provides one or more possible semantic tags[b]. On this basis, the semantic tagger keeps the most likely tag [c]:

token	POS[a]	lexical tag(s)[b]	tag[c]
painful	adjective	find	find
appendix	noun	obj/loc	loc

However, for semantics, a major difference concerns the decomposition of compounds (10% of the lexicon). The basic form is split into its basic morphemes (for example hepatitis = *hepat* + *itis*) by the segmenter. Each morpheme is then linked to one or more tags. For example, the class for pathological function or process (UMLS class T046, tag *papr*) is assigned to the morpheme *itis,* and the class organ (UMLS classes T023/T029, tagged *loc*) is assigned to the morpheme *hepat*, finally the full form is also assigned a compound tag *dis* (UMLS class T047):

hepatitis hepat\loc | itis\papr hepatitis # noun\dis

The indexing

A Boolean search engine was developed for the project that makes use of the various types of indexes. Texts have been indexed by all the content words (i.e. nouns, adjectives and verbs) and their basic form, syntactic category and semantic tag. *Disease* for example is indexed by the following key words:

index1: key= "disease" index2: key= "diseases"

index3: key= "disease(*Noun*)" index4: key= "disease(*dis*)"

A last index (index 5) allows a word to be found from its semantic tag, for example *itis*, *infection* and *inflammation* from *papr*:

infection, inflammation, itis : index5: key= "papr"

As the word is segmented into its different morphemes, these segments can be indexed like other words. Moreover, as the words are disambiguated, it is possible to add into the query the synonyms present in the thesaurus, for example for *liver* the thesaurus will return the set { *hepar*, *hepat*, *liver*}. But, as to add synonyms is often insufficient for a good recall in IR, we rely on the index 5 to improve the query expansion, as this index is likely to provide more terms to enrich the query.

PERFORMANCES

Assessments of the word-sense tagging (tab.6) were made on a sample of 2113 words, also belonging to the digestive surgery domain; therefore we did not process any unknown words. The sample was first tagged manually, we then compared it with the output of the tagger.

	Correct	incorrect	success rate %
semantic tagging	**2036**	**77**	**96.4**
syntactic tagging	2060	53	97.5

Tab.6: Semantic tagging results. Syntactic tagging results are also given.

The tagger was able to disambiguate medical ambiguities as much as the general ones. For example *protection* (1) (*mdev* or *hca*) and *section* (2) (*thers* or *spat*) were tagged successfully. In the example, we give first the token (TOK), then the part-of-speech and finally, the semantic tag:

```
TOK Horizontal          adjective\spat
TOK section             noun\thers
        (1)
TOK of                  prep\rel
TOK the                 det\def
TOK aponeurosis         noun\bosp
[...]
TOK and                 conj\rconj
TOK installation        noun\hca
TOK of                  prep\rel
TOK a                   det\indef
TOK plastic             adjective\qual
TOK protection          noun\mdev       (2)
```

A 3-4% error rate is not considered as a bad result for word-sense taggers, but these results must be handled carefully. Considering that only 12.5% of lexemes in the sample were semantically ambiguous, the score must be put in perspective. Thus, 87.5 would have

been the score of a system choosing always a wrong one among the possible ambiguous tags, as 12.5% of tokens were ambiguous in the assessment extract. While an apparently honest score of 93.7% would represent in fact a random tagger, choosing any possible tags, and mistaking it half of the time! So before getting a 100% tagger, only half of the distance has been covered.

Besides the numeral results, we noticed numerous very interesting patterns, such as: [{rspat*,spat*,bosp*}loc], [{papr*,dis*}loc], or [hca{*diap,*thers,*ther}] (the Kleene star means that zero, one or more tags is optionally occurring).

Finally, we can also notice that semantic ambiguities were more difficult to solve than syntactic ones (with success rates of respectively 96.4% vs. 97.5%), what seems to confirm that semantics relies on a larger context. But this last point would need deeper investigations, as the syntactic assessment was not the main focus of the project.

CONCLUSION

In this paper we reported the construction of a semantic tagset and tagger for medical document indexing. Results were of two kinds: a lexicon with semantic tag-like features on the one hand, and a probabilistic tagger to process the tag-like information on the other. Whereas semantic tagging results open new perspectives in Medical Language Processing, mastering further semantic disambiguisation may require more adapted tools, likely to cope with long and very long (out of the sentence) distance dependencies, similar to what is done in semantic clustering.

Another problem arises from the maintaining of the probabilistic tagger. As biases may have important negative side effects, a simple rule-based assistant could improve performances significantly. Therefore some patterns extracted by the tagger once expressed in a symbolic formalism could serve as a basis for a future semantic rule-based tagger. At this level, the next step will be to analyze more precisely these patterns. Last but not least, a larger lexical coverage, less domain-specific, will be necessary to provide an evaluation of the tag-like approach for indexing and retrieval of large medical corpora.

Acknowledgments

The Swiss government (FNRS – Swiss National Foundation for Research) funds the MEDTAG project.

References

1. Kupiec J, 1992, Robust Part-of-Speech Tagging using a Hidden Markov Model. in: *Computer, Speech and Language*, vol6, pp. 225-242.
2. Yarowsky D, 1992, Word sense disambiguisation using statistical models of roget's categories trained on large corpora. *In Proceedings of COLING92*, pages 454-460.
3. Ceusters W, Spyns P, De Moor G, Martin W, 1998, *Syntactic-Semantic Tagging of Medical Texts: the Multi-TALE Project*, IOS Press.
4. Results of the GRACE action for syntactic taggers on http://m17.limsi.fr/TLP/grace/
5. ACL workshop «Tagging Text with Lexical Semantics: Why, what and how», April 4-5, 1997, Washington, D.C., USA. Available from: URL:http://www.ims.uni-stuttgart.de/~light/tueb_html/semtag_ws_papers.html
6. Armstrong S, Petitpierre D, Robert G, Russell Gr, 1995, An Open Architecture for Multili ngual Text Processing. in *Proceedings of Sigdat Workshop*, Dublin, pp. 30-34.
7. Buitelaar P, 1998, Corelex Systematic Polysemy and Underspecification, PhD. Thesis, Computer Science, Brandeis University.
8. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR, 1996, for the Computer-Based Patient Record Institute's Work Group on Codes & Structures. The Content Coverage of Clinical Classifications. *J Am Med Informatics Assoc* 1996, 3(3): 224-233.
9. Hersh WR, 1996, Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical Records: a lexical analysis, *J Am Med Informatics Assoc* 1996; 580-584.