Indexing by statistical tagging

Pierrette Bouillon[1], Robert Baud[2], Gilbert Robert[1], Patrick Ruch[2]

[1] TIM/ISSCO - ETI - Université de Genève - 40 Bvd du Pont-d'Arve - CH-1205 Genève - Suisse

[2] DIM - University Hospital of Geneva, 24 Micheli-du-Crest, 1211 Geneva - Suisse

Abstract

Lexical ambiguity is a fundamental problem in Information Retrieval (IR), especially in the medical domain. Many systems use a subset of the words contained in the document to represent the content, but they are faced with the problem of ambiguity. In this paper, we propose a method for disambiguation based on existing medical terminological resources on the one hand, and statistical tools for linguistic annotation on the other, in order to develop more satisfactory indexing techniques for patient reports. The main hypothesises guiding this method are that: (i) Syntax can help to distinguate meanings of words that are polyfunctional. (ii) Syntactic analysis can be done by a probabilistic tagger (HMM, Hidden Markov Model) and, more daringly, (iii) remaining semantic ambiguity can also be solved (mutatis mutandis) by an HMM tagger.

Keywords: semantic disambiguation, statistical tagging, information retrieval, medical patient records

1. Introduction

Lexical ambiguity is a fundamental problem in Information Retrieval (IR), especially in the medical domain. Many systems use a subset of the words contained in the document to represent the content. Such systems are faced with two main problems ((Salton and McGill, 1983); (Krovetz, 1995); (Krovetz and Croft, 1992)). Firstly, words are ambiguous out of context and this ambiguity will cause documents to be retrieved that are not-pertinent; secondly, the user is not so much interested in retrieving documents with exactly the same words, as in retrieving those containing words with a similar meaning. Retrieval programs generally address these problems by expanding the query words by related terms from a thesaurus. But again this is only possible if the meaning of the word is unambiguously known (Towell and Voorhees, 1998).

If we accept the hypothesis that resolving ambiguity is essential and will lead to an improvement in the performance of these IR systems, the question is how to disambiguate the words. In this research, we propose a method based on existing medical terminological resources on the one hand, and statistical tools for linguistic annotation on the other, in order to develop more satisfactory indexing techniques for French patient reports. The main hypothesises guiding the project are that: (i) Syntax can help to distinguish meanings of words that are polyfunctional[1] (see also (Wilks and Stevenson, 1996); (Yarowsky, 1992); (Ceusters et al., 1996)). (ii) Syntactic analysis can be done by a probabilistic tagger (HMM, Hidden Markov model; (Rabiner, 1989); (Kupiec, 1992); etc.) and, more daringly, (iii) remaining semantic ambiguity can also be solved (mutatis mutandis) by an HMM tagger.

[1] i.e have different syntactic categories, as *bouche* in French which can be the indicative of the verb *to stop up* and the noun *mouth*.

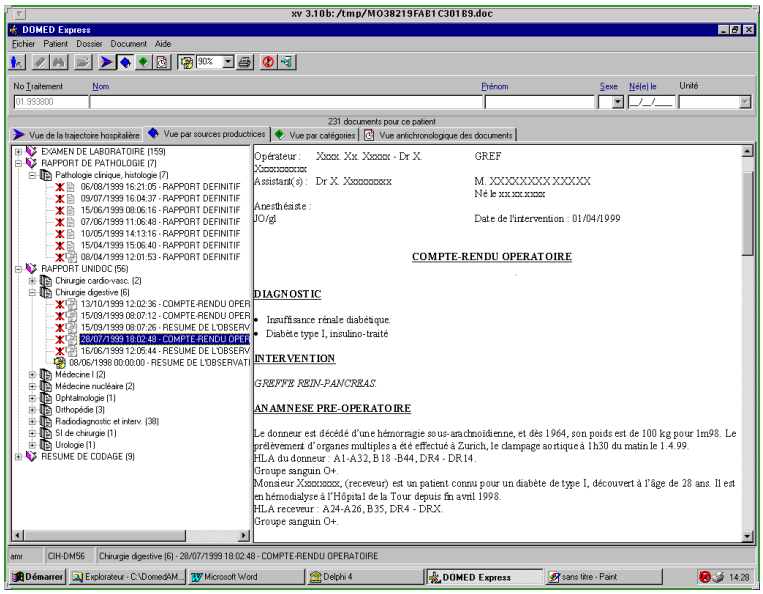

Figure 1: Anonymization tool

These hypothesises have been tested in the following way. The text is first annotated with ISSCO's corpus annotation tools ((Armstrong et al., 1995); (Armstrong, 1996)) that assign the syntactic and semantic analysis (tag) to the words. This information is then used to index the text and to improve the performance of the search engine. In this paper, we describe the first phase of the project, namely the method of linguistic annotation (section 2), its evaluation (section 3) and how the annotation is used for indexing (section 4). The evaluation of the search engine itself is foreseen for the second phase of the project and will not be described here.

2. Linguistic annotation

In this project, linguistic annotation is done by statistical tagging. This is carried out sequentially in different stages by a start-up kit of corpus annotation tools for morphological analysis, syntactic tagging and semantic tagging:

a. **Anonymization:** The medical texts are first automatically anonymized to delete patient or doctor identification (cf. Figure 1). This is a necessary step to gain access to large corpora of confidential information, that was given as a severe constraint by the Hospital of Geneva, in order to make available the corpus of medical texts.

b. **Segmentation:** The text is segmented into sentences, words (TOK for token) and other units (PTERM, for punctuation), for example:

```
(TAG   <S>
TOK    Maladie
TOK    diverticulaire
TOK    du
TOK    sigmoïde
PTERM  .
)TAG   </S>
```

c. **Morphological lookup:** Each word is then annotated with its lexical description(s) with the morphological analyzer *Mmorph*, extended to cover the medical texts ((Petitpierre and Russell, 1994); (Bouillon et al., 1998))[2]. A short example of text after morphological analysis is given below:

```
TOK  Maladie        maladie\Noun[ gender=fem number=sing ]
TOK  diverticulaire =\Adj[ gender=fem number=sing degree=pos ]
                    |=\Adj[ gender=masc number=sing degree=pos ]
TOK  du             de\Det[ gender=masc number=sing type_s=art ]
                    |de\Prep[ ]
TOK  sigmoïde       =\Adj[ gender=fem number=sing degree=pos ]
                    |=\Adj[ gender=masc number=sing degree=pos ]
                    |=\Noun[ gender=masc number=sing ]
```

Associated to each word is the base form[3] and a morphosyntactic description of the word. Alternative analyses are separated by "|". For example *sigmoïde* can be a feminine or masculine adjective, or a noun.

d. **Conversion of lexical information to syntactic tags:** A fundamental distinction is made here between lexical information and syntactic tags. The former is the output of the morphological analyzer, as given in (c); the latter specifies the information that must be disambiguated by the tagger. These tags depend on the application (the information that needs to be made explicit) and on the tagger itself. A mapping table has therefore been defined for establishing the correspondence between the two kinds of information, namely each set of attribute-value pairs associated to a given word and the corresponding atomic tag(s), as in the following:

```
Noun[ gender=fem number=plural ]    NOUN-PL
Noun[ gender=fem number=sing ]      NOUN-SG
...
```

A sample result of the text segment *Maladie diverticulaire du sigmoïde* after conversion is shown below:

```
TOK  Maladie  NOUN-SG
         Noun[ gender=fem number=sing ]
TOK  diverticulaire ADJ-SG
         Adj[ gender=fem number=sing degree=pos ]
         |Adj[ gender=masc number=sing degree=pos ]
TOK  du DET-SG|PREP
         Det[ gender=masc number=sing type_s=art ]
         |Prep[ ]
TOK  sigmoïde ADJ-SG|NOUN-SG
         Adj[ gender=fem number=sing degree=pos ]
```

[2]The *Mmorph* tool is publicly available and can be dowloaded at http://www.issco.unige.ch/tools/.

[3]= is used when the word is identical to the base form.

```
|Adj[ gender=masc number=sing degree=pos ]
|Noun[ gender=masc number=sing ]
```

e. **Training and syntactic tagging:** The syntactic tagging is done by ISSCO HMM tagger *Tatoo* (Warwick et al., 1995)[4]. The language model is calculated in three stages. First the syntactic model is automatically built on the basis of the ambiguous text (following the Baum-Welch algorithm). The matrices are then refined with a set of linguistic rules (bigrams), very useful to bias the model when training data are not sufficient to make generalizations. Finally a small part of the text (5000 words), manually disambiguated, is used to reestimate the model. With this method, the error rate for syntactic tagging is less than 3 %.

f. **Decomposition of words into morphemes** and **linking word-semantic tags:** The base form is decomposed into morphemes (*otite* = *ot* + *ite*) by the Geneva University Hospital's segmenter. Each morpheme/word is then linked manually to one or more UMLS codes (Unified Medical language System; (UMLS, 1998); (Ruch et al., 1999)) that define the base word semantic types (EAGLES-group, 1998).

g. **Training and semantic tagging:** The HMM tagger, using a model trained on the semantic tags defined in UMLS, resolves the remaining semantic ambiguities. As we are in a restricted domain after syntactic analysis, homonyms are very rare; what need to be disambiguated here are polysemes whose senses are related in a systematic way (Pustejovsky, 1995), as *préparation* that denotes both the process of preparing something and the result of this process. These are particulary suitable for this kind of method as by definition the correct sense can be identified by the context around the word and their disambiguation does not require pragmatic disambiguation. At this stage, the training is done as in (e), but the model is simplified as much as possible. Best results (around 7% for interesting classes; see next section) are obtained when adverbs, determiners and the preposition *de* are not taken into consideration. These are either too ambiguous or not relevant for the disambiguation of content words.

A short example of text after semantic tagging is given below:

```
Maladie                                    maladie#Noun!pafu
diverticulaire                             diverticulaire#Adjective!abn
du                                         de#Prep!rel
sigmoïde                                   sigmoïde#Noun!loc
.                                          .#SPUNCT)!PUNCT
</S>
<S>
Sigmoïdectomie   sigmoïd\loc|ectomie\ther  sigmoïdectomie#Noun\ther
et                                         et#Conj\rconj
anastomose                                 anastomose#Noun\ther
...              ...                       ...
```

[4]The *Tatoo* tool is publicly available and can be dowloaded at http://www.issco.unige.ch/tools/.

Here, the second column indicates the decomposition of the word into its morphemes (when it is possible) and the third one the result of the syntactic/semantic analysis, i.e. the lemma, the part of speech of the word and its UMLS semantic tag. For example, *maladie* is related to the base form MALADIE, the syntactic tag NOUN and the semantic tag PAFU (for pathologic function). *Diverticulaire* is an anatomical abnormality (ABN), *sigmoïde* a part of body (LOC) and *sigmoïdectomie* a therapeutic or preventive procedure (THER). This last word has been decomposed into two morphemes, the prefix *sigmoïd* and the suffix *ectomie*. In the following section, we first give the evaluation of the method for semantic tagging; then we describe how this complex information is used for indexing the texts.

3. Results of semantic tagging

Up to now, 50 patient records from the domain of digestive surgery (approximately 20,206 words) have been processed in this way. Two different evaluations have been carried out.

Evaluation 1: A subset of 4000 words were manually tagged and compared with the output of the tagger. Of these, 1366 are determiners, adverbs or the preposition *de* and are not taken into consideration for training and semantic tagging. The results are given below:

number of words : 2634
number of ambiguous words : 443
errors of semantic tagging : 80
percentage of errors of semantic tagging : 80/2634 (3.04%)
percentage of errors of tagging for ambiguous words : 80/443 (18.06%)

These results are promising in two ways. First, as predicted, semantic ambiguities are very small after syntactic tagging (443/2634, 17%). Secondly a lot of the errors are caused by prepositions (33/80, 41%) that are not very useful for the IR point of view, but have been proven useful for disambiguating content words (for example, *au-dessus de* cannot be followed by an action).

Evaluation 2: The second evaluation only takes into consideration the most interesting ambiguity classes from the point of view of retrieval, but in the whole text:

Ambiguity classes	Examples	Errors of semantic tagging
health care activity/spatial concept	section, abord, etc.	2/59 (3,39%)
health care activity/body space or junction	ouverture, séparation, etc.	5/52 (9,52%)
health care activity/organization	administration, etc.	0/1 (0%)
health care activity/finding	décollement, déviation, etc.	3/24 (12,5%)
health care activity/substance	préparation, etc.	0/9 (0%)
Total of errors		10/145 (6,90%)

It is encouraging to notice that these results are better than those obtained in evaluation 1. Deeper analysis shows that some errors could easily be avoided by taking into consideration more data or by improving the biases; for example, the system should learn that in "*sans section*

de la veine", the word *section* more probably denotes an action (i.e without sectioning the vein) than a spatial concept (i.e without a piece of the vein). Other errors are more related to the limitation of using the HMM for tagging. *Ouverture* for example is ambiguous between a body space ('an opening') and an activity ('to open'). As the preposition *à* can be both spatial and temporal, two solutions are *a priori* equiprobable for a sequence as *"A l'ouverture de la cavité abdominale"*, namely "spatial preposition+body space" (i.e. at the top of the cavity) or "temporal preposition+activity" (i.e. when we open the cavity). However these cases are relatively rare in our corpus, as shown by the results.

4. The indexation and search engine

In this first phase of the project, the texts are first indexed on all words with the base form, the syntactic category and the semantic category. *Maladies* for example is indexed by the following key-words, namely the word itself (Word), its lemma (Lemma), its syntactic category (POS) and the semantic type (SEM):

```
Word       Lemma     POS      SEM    REF
maladies   maladie   Noun     pafu   {ref1, ref2, ...}
```

Moreover, as words are segmented into their different morphemes (*otite = ot + ite*; *laparoscopie = laparo + scopie*, etc.), they are also indexed on the morphemes and *vice versa*, for example:

```
Word            morpheme
otite           ot
otite           ite

Morpheme         word
ot              otite
ite             otite
```

Finally, a last index links the words/morphemes to their synonyms. It is created automatically by matching the base forms of the *Mmorph* lexicon to the existing resource GALEN (Cf. Ruch et al., 1999), for example:

```
Word         SEM    SYNONYMS
ot           loc    oreille
ouverture    ther   ouvrir
```

All of these databases are implemented in SQL-DB (see http://www.mySQL.com/). A search engine has been developed in Java that can make use of the various types of indexes, as shown in Figure 2. In this specific example the user searches for all texts that contain two words, namely a word tagged as a pathologic function **and** a noun with the lemma *péritoine*. The distance between the two words must be inferior to 3 words (<**3**). The results of the request include texts with *péritonite*, but also those with both a synonym or a compound of *péritoine* (*péritonéal, pneumo-péritoine*, etc.) and a pathologic function (like *enflammé, inflammation, épaississement*), for example: *le péritoine pariétal est enflammé, pas d'épanchement péritonéal*, etc.

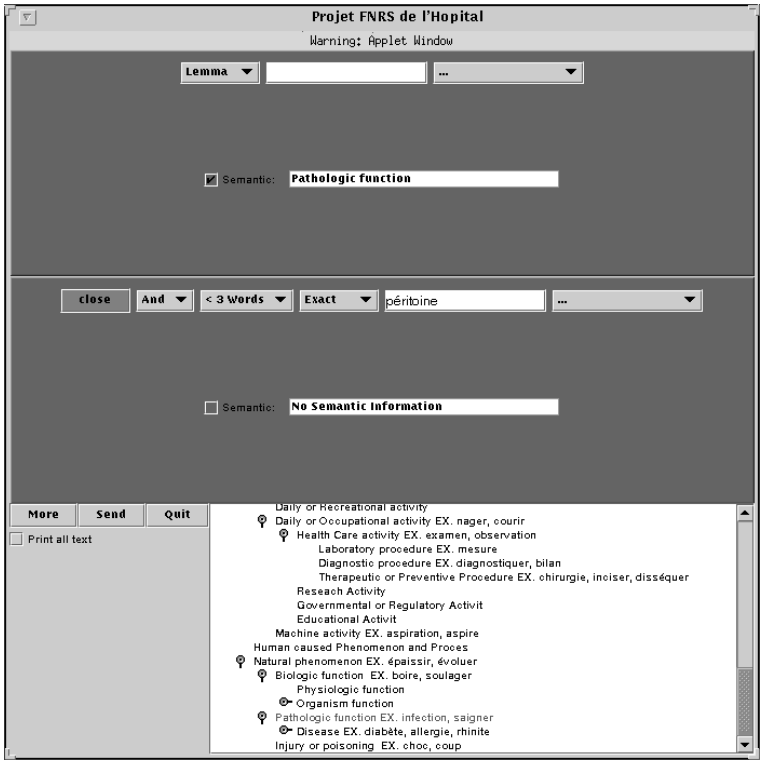

Figure 2: The search engine: query

On the theoretical side, three functionalities are therefore particulary interesting here: firstly, as words are disambiguated, it is possible to add into the request the synonyms present in the thesaurus, without increasing the noise, for example, *ouverture(action)* = {ouverture, ouvrir}; but *ouverture(spatial)* = {ouverture}. Secondly, as words are segmented into their morphemes and as most of the prefixes/suffixes are also linked in the thesaurus to their synonyms, it is therefore possible to retrieve texts with *otite* from a query like [inflammation **and** oreille]. Finally, it is possible to use the UMLS and POS tags in the request. For example: [pathologic function **and** foie] (all texts that deal with liver disease), [pince(**noun**)] (all texts with the noun *pince*), [pince(**verb**)] (all texts with the verb *pincer*), [section(**action**) **and** loc] (all texts that deal with the action of sectioning a body part), [section(**spatial**) **and** intestin] (all texts that deal with a section - piece - of intestin), etc. Their practical impact for IR will be examined in the next phase of the project.

5. Conclusion

In this paper, we proposed a method based on existing medical terminological resources on the one hand, and statistical tools for linguistic annotation on the other, in order to develop more satisfactory indexing techniques for patient reports. We showed that indexing can be done by the well-known technique of HMM tagging that can be used both for syntactic and semantic disambiguation. In this approach, semantic disambiguation by HMM tagging is made possible because the following conditions are met:

- most of the semantic ambiguities are solved at the syntactic level by syntactic tagging.

- The model for semantic analysis is improved by not taking in consideration certain categories or words that do not seem useful for the disambiguation.
- Finally, we do not disambiguate homonyms (very rare in a restricted domain after syntactic analysis), but polysemes.

The linguistic annotation seems very promising for indexing as it allows:

- the use of UMLS tags in the query and
- the expansion of the query with a thesaurus without increasing the noise (for exemple, a *préparation* (**liquid**) won't expand to *préparer*).

The main advantage of the method is to be adaptative: tools and matrices can be reused for new texts. The only resource which must be built manually is the lexicon that links the words to their corresponding UMLS tags. The next phase of the project will test the matrices on new texts and evaluate the impact from the IR point of view.

References

Armstrong S. (1996). Multext: Multilingual Text Tools and Corpora. In Feldweg H. and Hinrichs W. editors, *Lexikon und Text*. Tübingen: Niemeyer.

Armstrong S., Petitpierre D., Robert G., and Russell G. (1995). An open architecture for multilingual text processing. In *Actes de Sigdat Workshop*, pages 30–34, Dublin.

Bouillon P., Lehmann S., Manzi S., and Petitpierre D. (1998). Développement de lexiques à grande échelle. In *Actes du colloque de Tunis 1997 "La mémoire des mots"*, Tunis.

Ceusters W., Spyns P., DeMoor G., and Martin W. (1996). *Tagging of Medical Texts: The Multi-TALE Project*. Amsterdam:IOS Press.

EAGLES-group (1998). Eagles preliminary recommendations on semantic encoding. Technical report, Pisa.

Krovetz R. (1995). *Word Sense disambiguation for Large Text databases*. PhD thesis, Université of Mass.

Krovetz R. and Croft W. (1992). Lexical ambiguity and information retrieval. *ACM transactions on Information Systems*.

Kupiec J. (1992). Robust part-of-speech tagging using a hidden markov model. *Computer Speech and Language*.

Petitpierre D. and Russell G. (1994). Mmorph - the multext morphology program. Technical report, ISSCO, University of Geneva.

Pustejovsky J. (1995). *The Generative Lexicon*. Cambridge:MIT Press.

Rabiner L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In Waibel A. and Lee K. editors, *Readings in Speech Recognition*. San Mateo:Morgan kaufmann.

Ruch P., Bouillon P., Baud R.-H., Rassinoux A.-M., and Scherrer J.-R. (1999). MEDTAG: Tag-like Semantics for Medical Document Indexing. In *Actes de AMIA99 Annual Symposium*, Washington, DC.

Salton G. and McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Towell G. and Voorhees E. M. (1998). Disambiguating hightly ambiguous words. *Computational Linguistics 24:1*.

UMLS (1998). Umls knowledge sources. Technical report, National Library of Medecine.

Warwick S., Robert P., and Robert G. (1995). Tools for part-of-speech tagging. Technical report, ISSCO, University of Geneva.

Wilks Y. and Stevenson M. (1996). The grammar of sense: is word-sense tagging much more than part-of-speech tagging? Technical report, University of Sheffield, UK.

Yarowsky Y. (1992). Statistical models of Roget's categories trained on large corpora. In *Proceedings of Coling'92*, Nantes.