Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

Tagging medical texts: a rule-based experiment

Patrick Ruch[1], Pierrette Bouillon[2], Gilbert Robert[2], Robert Baud[1], Anne-Marie Rassinoux[1],
[1]*Medical Informatics Division, University Hospital of Geneva, Geneva, Switzerland*
[2]*ISSCO, University of Geneva, Geneva, Switzerland*

Abstract: In this paper we describe the construction of a part-of-speech tagger for medical document retrieval purposes, therefore we have designed a specific architecture called *minimal commitment*. The system uses local grammatical rules for conducting the disambiguation task. Four evaluations are conducted, with and without taking unknown words into account. In between each evaluation the modules (lexicon, guesser, rules) of the system are incrementally improved.

1. Introduction

Nowadays, most medical information is stored in textual documents, but with the age of the electronic patient record, such large amount of data may remain useless if retrieving the relevant information in a reasonable time becomes impossible. Although some large-scale information retrieval (IR) evaluations, made on unrestricted corpora [1,2] and in the medical domain [3], are quite critical towards linguistic engineering, we believe that natural language processing applied to IR is the best solution to face the problem of lexical ambiguity. The two major problems, when searching information among text databases with automated agents, are first, the expansion of the query and second, the ambiguity:

- Expansion: the user is both interested in retrieving documents with exactly the same words, and in retrieving documents with the same meaning: for example *hepatic* is relevant when searching about *liver*.
- Disambiguation: retrieving strictly the words of the query may be insufficient, as words are often ambiguous out of their context. For example the word *face*, may be a body part (as a noun), or an action (as a verb), and a word like *section* (always a noun) may be a surgery procedure as well as a spatial concept. *Face* represents a good example of a word-sense ambiguity expressed via a morpho-syntactic (MS) ambiguity. *Section* provides an example of a pure word-sense (WS) ambiguity, without MS ambiguity.

Although the purpose of the project is to build up a common light architecture for processing both the MS and the WS disambiguation stages, the present paper reports exclusively on our investigations concerning the MS disambiguation task. We earlier [4] studied the relevance of using data-driven techniques (Hidden Markov Models or HMM) for both MS and WS tagging, and concluded [5] that rule-based approaches could be opportunely investigated. While our studies were made on French corpora, we decided to provide the examples in English for the sake of clarity[a].

Background

Before starting to develop our own tagger, some preliminary studies on general available systems were conducted; if such studies go far beyond the scope of this

[a] If the rules we wrote are tailored for the French language, the tagging toolkit and the rule-writing formalism is clearly language-independent, at least for most of European languages.

paper, we would like to report on the main conclusions. Both statistical taggers (HMM) and constraint-based systems were assessed. Two guidelines were framing the study: performances and *minimal commitment*. We call *minimal commitment*[b] the property of a system, which does not attempt to solve ambiguities when it is not likely to solve it well! This property seems particularly important for IR purposes, where we often prefer noise rather than silence.

Data-driven tools

We adapted the output of our morphological analyser for tagging purposes [4]. We trained and wrote manual biases for a bi-gram HMM tagger, but results were never far above 97% (i.e. about 3% of error); with an average ambiguity level of around 16%, it means that almost 20% of ambiguities were attributed a wrong tag! We could have attempted to set a threshold, so that for equiprobable or similarly weighted transitions, the system would keep the ambiguity (as in [6]), but it is particularly difficult to determine a priori such threshold. Tri-gram taggers or Brill type tools [7] may have performed better, but the minimal commitment axiom would have remained unsatisfied.

Constraint-based systems

We also looked at more powerful principle-based parsers, and some tests were conducted on FIPSTAG ([8], and on-line [9]). Although this system performed impressively on general texts (about 0.7% of errors!), its results on medical texts were about the same as simple taggers. The adaptation of such an integrated system is much heavier[c]. Thus, we could not adapt on it our own morphological analyser. Therefore, the system had to cope with several unknown words (drugs and medical morphological compounds...), and particular manners: capital letters, which usually indicate proper nouns in written French, are frequently used in clinical documents for medical devices, drugs, chemicals and diagnosis.

2. Methods

From an epistemic point of view, two main hypotheses are guiding the project:

a. syntax can help to distinguish meanings of words having different syntactic categories;
b. syntactic ambiguities can be solved within a light rule-based framework, using very local rules (cf. [10], for a quite similar approach).

These hypotheses have been tested in the following way: we first pick a corpus of 40000 words. Then, this sample is split into 5 equivalent sets. The first one (set A, 8520 words) will serve to write the basic rules of the tagger, while the other sets (set B, 8480 tokens, C, 7447 tokens, D, 7311 tokens, and E, 8242 tokens), will be used for assessment purposes and incremental improvements of the system.

Lexicons and corpus

Our lexicon, with around 20000 entries, covers exhaustively the whole ICD-10 (cf [11], for a detailed assessment of the lexical coverage). The source lexicon is stored in relational databases. The operative version of the lexicon is transformed into a letter-

[b] The first one using this expression was maybe M. Marcus [17], lately we can find a quite similar idea in [18].
[c] A future release of FIPSTAG should solve some of these issues.

tree [12], before being minimised into a direct acyclic graph. In order to assess methods for tagging texts in the medical domain, a set of texts has to be carefully selected. On the one side, in order to implement the foreseen WS disambiguation, it was useful to rely on texts belonging to a narrow domain. On the other side, it was also important to select documents with a large part of free text, in order to build up a scalable system. Finally, a large number of documents should be available. We finally picked reports from the digestive surgery domain.

Morphological analysis and guessing

The morphological analyser is based on morphosemantemes [13]. Using finite-state automata techniques, it allows a fast lexical access. This lemmatizer maps each inflected surface form of a word to its canonical lexical form followed by the relevant morphological features (tab1).

Tab1. Example of lemmatisation providing the MS features

Faces	lemma: face	
		Category: noun
		MS features:
		number: plural
	lemma: to face	
		Category: verb
		MS features:
		number: singular
		person: 3
		mood: indicative
		tense: present

Words absent from the lexicon follow a two-step guessing process. First, the unknown token is analysed regarding its respective morphemes, if this first stage fails then a last attempt is made to guess the hypothetical MS tags of the token. The first stage is based on the assumption that unknown words in medical documents are very likely to belong to the medical jargon, the second one supposed that neologisms follow regular inflectional patterns. Both guessing stages are likely to point to various categories.

If regarding the morpho-syntax, both stages are functionally equivalent, as each one provides a set of morpho-syntactic information, they radically behave differently regarding the WS information: For guessing WS categories only the first stage guesser is relevant, as inflectional patterns are not sufficient for guessing the semantic of a given token. As for the ending *ly*, which characterises very probably an adverb, but which may refer to almost any kind of adjective without indicating whether it is a finding (*feverishly*, tagged *find*), or a temporal qualifier (*quickly*, tagged *temp*).

Let us consider three examples of words absent from the lexicon. First, *allomorph*: the prefix part *allo*, and the suffix part, *morph*, are present in the lexicon, with all the MS and the WS features. Second, *allomorphly*: *morphly* does not occur into the morpheme database, while the ending *ly* occurs. These words are recognized by the first-stage guesser. Third *allocution*, recognized by the second-stage guesser: it can not be split into any affix, as *cution* is not a morpheme, but whose ending (*tion*) refers to the following features in the second-stage guesser: noun, singular. Let us notice that the ending *tion* (like the ending *ly*) will not provide any WS information. As the

underlying objective of the project is to retrieve documents, the main and most complete information is provided by the first-stage guesser, while the second-stage is only interesting for MS tagging (the second-stage guesser is purely a MS guesser).

Categories and morpho-syntactic features provided by the lemmatizer are then expressed into the MS tagset (annexe A provides some items of the tagset). The MS tagset expressed only some of the features provided by the lexicon. Therefore, the lexical information provided by the lemmatizer is over-specified for the MS tagset. Thus, the tense feature does not appear into the MS tagset. Here is a short example, where lexical ambiguities and lemma are separated by a '/':

Tab2. Tag-like representation (lexical tags) of the MS lexical features

Token	Lemma	Lexical tag(s)
Section	Section	nc[s]
of	Of	sp
Internal	Internal	s
Faces	face/to face	nc[p]/v[s03]

Note: nc, v, s, p, and 03 respectively stand for common noun, verb, singular, plural and third person, cf. annexe A, for a short description of the MS tagset.

Studying ambiguities

Our first investigations aimed at assessing the overall ambiguity of medical texts. We found that 1227 tokens (14.4% of the whole sample) were ambiguous in set A, and 511 tokens (6.0%) were unknown. We first decided not to worry about unknown words, therefore they were not taking into account in the first assessment (cf. Performances). Then, we realised that some frequent words (even some functional ones, such as *on*, equivalent to *one* in English) were missing, so that together with the MS guesser, we would improve the guessing score by adding such very frequent words. Thus, adding 232 words into the lexicon and linking the lexicon with the Swiss compendium [14] allows an average unknown word rate of less than 3%. This result includes also the pre-processing of patients and physicians proper nouns [15].

Concerning the most frequent ambiguities: we found that 5 tokens were responsible for half of the ambiguities, while in unrestricted corpora this number seem close to 16 [16]. Another interesting observation is that the most frequent ambiguous words are usually words, which are in general domain-independent, i.e. words that be can be exhaustively listed (determiners, pronouns, preposition), auxiliaries or common verbs. However, together with classical and expected ambiguities of the French language - most determiners (*l', le, la, les* equivalent to *the*) are likely to be clitic pronouns (equivalent to *it, him, her...*)- we found a very medical ambiguity within this very special set of 12 items: *Patient*, which is ambiguous between a noun and an adjective (like in English, but in French the feminine form is also a verb: *to wait*) has been found 40 times (3.3% of all the ambiguities). *Patient* is ranked at the 6th position between the set of the 12 more ambiguous tokens.

Local rules

We separated set A in smaller sets (8 subsets of around 1000 tokens) in order to write our rules. We wrote around 50 rules (which generated around 150 operative rules) for the first subset, while on the 8th, only 12 rules were necessary to reach a score close to 100% on set A. Rules may be classified into two categories: multi-level rules (as

for example: rule 1) and level-independent rules (as for example: rule 2). Finally, these rules are using intermediate symbols (such as *, the Kleene star[d]) in order to ease and improve the rule-writing process. These intermediate symbols are replaced when the operative rules are generated. Let's give one example of each category:

Rule 1: prop[**], v[**]/nc[**] --> prop[**];v[**]

This rule says 'when a token is ambiguous between (/) a verb (v), whatever (**) features it has (3rd or 1st/2nd person, singular or plural), and a common noun, whatever (**) features it has. And such token is preceded by a personal pronoun (prop), whatever (**) features has this pronoun (3rd or 1st/2nd person). So, the ambiguous token can be rewritten as a verb, keeping its original features (**)'.

Rule 2: prop[s03], v[s03]/nc[ms]{TOK:fait} --> prop[03];v[s03]

This rule is exactly equivalent to rule 1. But, is applied only when the ambiguous token is *fait* so that MS values (expressed by ** in rule 1) are now instantiated: *fait* may be a verb (3rd person of singular, in English: *does*, *makes*), as well as a noun singular masculine (in English: *fact*).

Finally let's give a short example of the kind of ambiguity, the tagger is likely to solve: the first column of the table (token) show the word as it appears in the text, the second (lexical tags) provides the tag attached to each token after a lexical access. Finally, column 3 provides the output of the tagger after removal of irrelevant tags. The meaning of each tag, together with some examples can be found in annexe A.

Tab3: example of tagging

Token	Lexical tags	Disambiguated tag
Fast	a	a
Section	nc[s]	nc[s]
Of	sp	sp
Internal	a	a
Faces	Nc[p]/v[s03]	nc[p]

Performances

Four successive evaluations were conducted; after each session, the necessary rules were added in order to get a tagging score close to 100%. In parallel, words were entered into the lexicon, and productive endings were added into the MS guesser. The second, third, and fourth evaluations were performed with activating the MS guesser. Moreover, translations phenomena [20, d'après Tesnière], which turn the lexical category of a word into another one, seem rare in medical text: thus, only 3 translations (2 from past participle to adjective, 1 from adjective to noun) were not foreseen in the lexicon[e]. Here are the results of each evaluation (GC stands for good candidates):

[d] -the ** will be replaced by any MS features attached to the category of the token, thus, nc[**] will generated 4 tags: nc[ms], nc[fs], nc[mp], nc[fp].
-The syntax of rules has some relation with definite clause grammars (the infix operator --> is used as rewriting operator), but in this formal language only terminal symbols and are allowed, i.e. the tags. The other main difference concerns the presence of the ambiguity operator '/'.

[e] Translation phenomena in French have been explored within [20], where the relevance of translations between past participles and adjectives has been questioned.

Tab. 4: Results for each evaluation

Evaluation	1-Set B	2-Set C	3-Set D	4-Set E
Tokens with lexical ambiguities	1178 (13.9)	1273 (17.1)	1132 (15.5)	1221 (14.8)
Tokens correctly tagged	8243 (97.2)	7177 (96.4)	7137 (97.6)	8082 (98.1)
Tokens still ambiguous, with GC	161 (1.9%)	183 (2.5)	136 (1.9)	101 (1.2)
Tokens ambiguous, without GC	-	9 (0.1)	2 (0)	9 (0.1)
Tokens incorrectly tagged	76 (0.9%)	78 (1.0)	36 (0.5)	51 (0.6)

A success rate of 98.1% (tab. 4, evaluation 4) is not a bad result for a tagger, but the main result concerns the error rate, with less than 1% of error, the system seems particularly *minimally committed*! Finally, we also observed that the system performed better when considering unknown words than when not! This can be explained by the fact that ambiguities -caused by the unknown words- have important side effects on the residual ambiguity rate even for words present into the lexicon. Therefore, the improvement of the correctly tagged tokens between the assessments 1-4 (0.9%) can be distributed: about 1/3 from the incorrectly tagged tokens, and about 2/3 from the ambiguous ones. We must say that for assessment 1, the system had about 1000 operative rules, while the assessment 4 was conducted with more than 2000 operative rules. Another interesting result concerns the residual ambiguity (tokens still ambiguous, with GC): in the set E, at least half of these ambiguities could be handled by writing more rules. However some of these ambiguities seem clearly untractable in the close context and would demand more lexical information on verbal selections, as in *le patient présente une douleur abdominale brutale et diffuse [...]* (the patient shows an acute and diffuse abdominal pain/the patient shows an acute abdominal pain and distributes*[f]), where *diffuse* could be both an adjective or a verb.

3. Conclusion and future works

We have presented a rule-based tagger for electronic medical records. The target of this tool is the disambiguation for IR purposes, therefore we decided to design a system without any heuristics. Rules were written very simply, when we considered that the rules were good enough (almost 100% of tokens were tagged correctly), we ran the tagger on a first test sample, results were satisfying, mainly with respect to the minimal commitment axiom, but the unknown words were not taken into account. Then, we decided to handle the unknown words and we conducted two more evaluation procedures. The results decreased sensitively, both regarding the error rate and the residual ambiguity. Each time, the system was improved regarding the rules modules, the lexicon, and the guesser, and a last evaluation showed that the system performed even better when taking unknown words into account than when not. A last question concerns the scalibility of the approach out of such a narrow domain (digestive surgery reports). Therefore we would like to run the tagger on other medical reports, and on totally unrestricted test, as both tests may be very interesting for validating the added-value of the minimal commitment paradygm.

[f] The lexical information on the valence + OBJECT is necessary for disambiguating the verb form of *diffuse*.

Annexe A

Tag	Freq.(%)	Label	Example
1-v[***]	2.4	verb	eats, does, has
3-v[pp]	0.5	verb present participle	leaving
4-v[**]	1.8	verb past participle	hidden
5-nc[**]	18.2	common noun	finger, eyes
6-np	0.6	proper noun	George, USA
6-a[**]	10.4	adjective	big, fat
7-dad[**]	7.0	definite determiner	the
8-dai[**]	3.4	indef. determiner	a
9-prop[***]	1.7	personal pronoun	we, she, it
11-sp	12.3	preposition	of,
12-r	2.1	adverb	happily
13-cc	3.0	coordination	and", or
14-f	6.7	punctuations	., ?, :
24-x	6.4	unknown words	mlHg

** represents MS features such as ms, mp, fs, fp, where m, f, s, and p mean respectively masculine, feminine, singular and plural. *** represents MS features such as 12, s03, p03. 12 refers to first and second person of both plural and singular. s03 and p03 refer respectively to 3rd person singular and 3rd person plural.

Distribution and description of some of the most frequent morpho-syntactic tags within the sample. The MS tagset tends to follow the MULTEXT lexical description [17]

References

1 Hersh WR, Price S, Kraemer D, Chan B, Sacherek L, Olson D, 1998, *A Large-Scale Comparison of Boolean vs. Natural Language Searching for the TREC-7 Interactive Track*. TREC 1998, pp. 429-438.

2 Sparck-Jones K, *What is the role for NLP in Text Retrieval*. In Strzalkowski (ed.) Natural Language Information Retrieval (Kluwer), pp. 1-25.

3 Hersh WR, *Information Retrieval at the MILLENIUM*. In C. Chute (ed.), American Medical Informatics Association Annual Symposium (AMIA'1998, ex-SCAMC). Orlando. Pp. 38-45

4 Bouillon P, Baud, R, Robert G, Ruch P, 2000, *Indexing by statistical tagging*. In Proceedings of the JADT'2000, Lausanne.

5 Ruch P, Bouillon P, Baud RH, Rasinoux A-M, Scherrer J-R. *MEDTAG: Tag-like Semantics for Medical Document Indexing*. In Proceedings fo the AMIA'99 Annual Symposium, Washington, DC, November 6-10, 1999.

6 Weishedel R, Meteer M, Schwartz R, Ramshaw L, Palmucci J, 1993. *Coping with ambiguity and unknown words through probabilistic models*. Computational Linguistics, Volume 19.

7 Brill E, 1992, *A simple rule-based part-of-speech tagger*. In Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy.

8 Wehrli E, 1992. The Interactive Parsing System, In ACL, Ed., *Actes de COLING-92*. 870-4. Nantes.

9 http://latl.unige.ch/latl/french/projets/presentation_fipstag_f.html.
For a MULTEXT-like description of the FIPSTAG tagset see Ruch P, 1997: *Table de correspondance GRACE/FIPSTAG*, available at http://latl.unige.ch/doc/etiquettes.ps

10 Laporte E, 1994, *Experiences in Lexical Disambiguation Using Local Grammars*. In Third International Conference on Computationnal Lexicography. Pages 163-172. Budapest.

11 Baud RH, Lovis C., Ruch P., Rassinoux A.-M., 2000, *A Toolset for Medical Text Processing*, accepted to MIE'2000

12 Roche E, Shabes Y., 1997, Deterministic Part-of-Speech Tagging with Finite-State Transducers, *in Finite-State Language Processing* (Roche E. and Shabes Y. Eds). p.206-239.

13 Baud RH; Lovis C; Rassinoux AM; Scherrer JR, 1998, *Morpho-semantic parsing of medical expressions*. Proc AMIA Symp 1998;:760-4.

14 http://www.kompendium.ch

15 Ruch P, Baud R, Rassinoux AM, Bouillon P, Robert G, 2000, *Medical Document Anonymization with a Semantic Lexicon*, Submitted to AMIA'2000 Annual Symposium. Los Angeles.

16 Chanod, JP and Tapanainen P, 1995, Tagging French: comparing a statistical and a constraint-based method. In Proc. *EACL'95*. pp. 149-156. Dublin.

17 http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX1.html

18 Marcus MP; Hindle D; and Fleck MM; 1983; *D-Theory: Talking About Talking About Trees*. Proc. 21st Annual Meeting of the Association for Computational Linguistics. 129-136.

19 Silberztein M, 1997, The Lexical Analysis of Natural Languages *in Finite-State Language Processing* (Roche E. and Shabes Y. Eds). p.206-239.

20 Rajman M, Paroubek P, Lecomte J, 1996, *Format de description lexicale pour le français – partie 2: Description morpho-syntaxique*, rapport GRACE GTR-3-2-1. (http://www.limsi.fr/TLP/grace/www/gracdoc.html).