MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Comparing General and Medical Texts for Information Retrieval Based on Natural Language Processing: An Inquiry into Lexical Disambiguation

Patrick Ruch, Robert Baud, Antoine Geissbühler, Anne-Marie Rassinoux

University Hospital of Geneva

Abstract

In this paper we compare two types of corpus, focusing on the lexical ambiguity of each of them. The first corpus consists mainly of general newspaper articles and literature excerpts, while the second belongs to the medical domain. To conduct the study, we have used two different disambiguation tools. First, each tool was validated in its respective application area. We then use these systems in order to assess and compare both the general ambiguity rate and the particularities of each domain. Quantitative results show that medical documents are lexically less ambiguous than unrestricted documents. Our conclusions emphasize the importance of the application area in the design of NLP tools.

Keywords:

Natural language processing; Lexical disambiguation; Electronic patient record

Introduction

Although some large-scale evaluations carried out on unrestricted texts ([1],[2]), and also on medical documents [3], conclude in a quite critical way about using NLP tools for information retrieval, we believe that such tools are likely to solve some lexical ambiguity issues. We also believe that some special settings -particular to the application area- must be taken into account while developing such natural language processing (NLP) tools.

Let us recall two major problems while retrieving documents with information retrieval (IR) engines [4]:

1-Expansion: the user is generally as interested in retrieving documents with exactly the same words, as in retrieving documents with semantically related words (synonyms, generics, specifics...). Thus, a query based on the word *liver*,[1] should be able to retrieve documents containing words such as *hepatic*. This step is usually thesaurus-based. The thesaurus can be built manually or automatically [5].

2-Disambiguation: a search based on tokens may retrieve irrelevant documents since tokens are often lexically ambiguous. Thus, *face* can refer to a body part, as a noun, or an action, as a verb. Therefore a query searching for the first concept must ignore the second.

Finally, this latter problem may be split into two sub problems. The disambiguation task can be based on part-of-speech (POS) or word-sense (WS) information, but the chronological relation is still a discussion within the community. Although, the target of our work ([6][7]) is a fine-grained semantic disambiguation of medical texts for IR in electronic patient records, we believe that the POS disambiguation is an important preliminary step. Therefore this paper focuses on POS tagging, and compares morpho-syntactic lexical ambiguities (MSLA) in medical texts to MSLA in unrestricted corpora.

Although the results of the study conform to preliminary naive expectations, the method is quite original[2]. Most of the comparative studies, dedicated to corpora, have addressed the problem by applying metrics on words entities or word pieces (as in studies working with n-gram strings), or on special sets of words (the indexing terms) as in the space-vector model (see [10], for a survey of these methods), whereas the present paper attempts to compare corpora at a morpho-syntactic (MS) level, as it is clearly a preliminary step for improving IR accuracy.

Method

Validating each tagger into its respective domain

In order to conduct the comparative study, we used two different morphological analyzers; each one has a specific lexicon tailored for its application field. The first system is specialized for tagging medical texts [11], while the second

[1] When possible, examples are provided in English for sake of clarity.

[2] We do not claim to be pioneer in the domain, as others authors ([8],[9]) are exploring similar metrics. However, it is interesting to notice that for these authors the adaptation of the NLP tools has rarely been questioned in a technical point-of view, and in order to feed back the design of NLP systems, and a fortiori to improve systems tailored for medical texts.

is a general parser (based on FIPS, cf. [12]).

For comparing lexical ambiguities on a minimal common base, we define a minimal common tagset (MCT). The output of each morphological analyser is first mapped into its respective tagset (more than 300 fine-grained tags for FIPSTAG, and about 80 for the morphem-based medical tagger). The tagsets are then converted into an intersection of each tagset. Finally, about 50 different items constitute this MCT, which will serve for comparing both corpora.

We collected two different sets of documents to be tagged at a lexical level via the predefined MCT: this step provides a set of tags to every token. This set of tags may come from the lexicon or from the POS guesser. As we are using guessers, the empty set (or the tag for unknown tokens) is forbidden. However, first of all, it is necessary to verify the lexical coverage of each system for each corpus, as we need to be sure that the lexical ambiguities provided by each system are necessary *and* sufficient.

The corpus of the unrestricted texts consists of 16003 tokens: about one third of newspaper articles (*Le Monde*), one third of literature excerpts (provided by the InaLF, *http://www.inalf.fr*), and a smaller third being mainly texts for children. Approximately a quarter (3987 tokens) of this corpus is used for evaluating FIPSTAG tagging results (the tool together with some explanations can be found at http://latl.unige.ch). In parallel, we chose three types of medical texts to make up the medical corpus: it represents 16024 tokens, with 3 equal thirds: discharge summaries, surgical reports, and laboratory or test results (in this case, tables were removed). Again, a regularly distributed quarter (4016) of this corpus is used for assessing the medical tagger.

The test samples used for assessing the results of each tagger are annotated manually before measuring the performances, but in both cases we sometimes had to modify the word segmentation of the test samples. Indeed, FIPSTAG handles several acceptable but unusual collocations (which gather more than one 'word'), as for example *en avion* (in Eng. *by plane*), which is considered as one lexical item, tagged as an adverb. For the lexical tagger we had to modify the 'word' segmentation in the other direction (for tagging items smaller than 'words'), as morphems [13] were also tagged. Table 1 gives the results for FIPSTAG, and table 2 gives the results for the medical tagger. In the case of the medical tagger, together with the error rate and the success rate, we provide results of the residual ambiguity rate: the basic idea is that the system does not attempt to solve what it is not likely to solve well [14].

Table 1: Evaluation of FIPSTAG

1 Correct tag	3959 (**99.3%**)
1 Incorrect tag	28 (**0.7%**)

A comparison of the tagging scores (99.3 vs. 98.5) confirms that both systems behave in an equivalent way in their respective application area[3].

Table 2: Evaluation of the medical tagger

1 Correct tag	3962 (**98.5%**)
1 Incorrect tag	12 (**0.4%**)
2 or more tags, at least 1 is correct	39 (1.0%)
2 or more tags, 0 correct	3 (0.1%)

Morphological analyzers, lexicons and guessers

Lexical ambiguities have two origins: the lexicon, and the guessing stages for unknown tokens. However, all the ambiguities considered in this study are strictly lexical, and so translation phenomena [15], which turn the syntactic category of a lexical item into another category, are not considered here.

Medical lexicon

The medical lexicon is tailored to parse biomedical texts, thus, with about 20000 lexemes, it covers exhaustively ICD-10. The biomedical language is not only a 'big' sub language, as its morphology is also more complex. This high level of composition (at least compared to regular French or English languages) concerns about 10% of tokens within clinical patient records; therefore the lexicon contains also about 1200 affixes. For example, the token *postileojejunostomy* is absent from the lexicon, however, this type of token may be recognized via its compounds (see [13] for more details): *post*, *ileo*, *jejuno*, and *stomy*.

Morphological analysis and medical morphology

The morphological analysis associates every surface form (word) with a list of morpho-syntactic features. When the surface form is not found in the lexicon, it follows a two-step guessing process: the first level (oracle1) is a more complex morphological analyzer, based on morphems, while the second level guesser (oracle2) attempts to provides a set of MS features looking at the longest ending (as reported in [16]).

The importance of these two levels is not clear for POS tagging, but becomes manifest when dealing with sense tagging. Let us consider three examples of tokens absent from the lexicon: *allomorphique*, *allomorphiquement* (equivalent to *allomorphic* and *allomorphically* in Eng.

[3] Out of curiosity, we ran each tagger on a small sample of the other domain. The tests were made without any adaptation. FIPSTAG made 27 errors in a medical sample of 849 tokens, i.e. an error rate of 3.2%. The medical tagger made 18 errors in a general sample of 747 tokens, which means an error rate of 2.4%. However, in the case of the medical tagger, 41 tokens remained ambiguous after disambiguation; the residual ambiguity is therefore about 5.5%. In this sample, and before disambiguation, the number of ambiguous tokens was 150, which means an ambiguity rate of 20%. Thus, even using a given lexicon, the ambiguity rate seems higher for general corpora than for domain-specific ones (about 16%).

language) and *allocution*. In the first case, the prefix *allo* and the suffix *morphiques* are listed in the morphem database (MDB). In the second case, *morphiquement* is not listed within the MDB, but *ment* can be found in it. In these two cases, oracle1 is able to provide both the MS and the WS information associated. The latter example cannot be split into any morphems, as *cution* is absent from the MDB. Thus, oracles1 is unable to recognize it, and finally oracle2 will be applied and will provide some MS features regarding exclusively the endings. The major role given to oracle1 and the semantic features it provides is obvious for IR purposes.

The final stage transforms some of the lexical features returned by the morphological analysis in a tag-like representation to be processed later by the tagger.

FIPSTAG tagger and lexicon

The FIPSTAG lexicon is a general French lexicon; therefore it contains most well formed French words. The overall structure of the lexicon is more or less stable, but the content is regularly updated in order to improve the coverage. Currently, the coverage is about 200000 words with around 30000 lexical items. The lexicon is designed for deep parsing, so that, together with classical morpho-syntactic features, we can also find sub categorization of verbs, semantic features, and some very specific grammatical classes.

As the system is claimed to be general, it is supposed to master efficiently any unknown words: the lexical modules supply, in an equiprobable way, all the possible lexical categories (i.e. nouns, verbs, adjectives, and adverbs), as other categories are supposed to be exhaustively listed in the lexicon. Consequently, the guesser does not rely on any morphological information, and only syntactic principles are applied to choose the relevant features.

Results and comparison of ambiguities

Previously, while attempting to assess the performance of our tools, only a sample of the ad hoc corpus we built up was used, whereas the following studies on the ambiguities will be carried out on the whole corpus. Like in the validation task, the lexical ambiguities are based on the morphological analysis of each tagger, expressed in the MCT. First of all, table 3 gives the general ambiguity rate in each corpora: it clearly states that the total ambiguity rate in general corpora is about twice as big as in medical texts.

Table 3: ambiguity rates according to the corpus

	medical corpus	general corpus
ambiguities	2532 (15.8%)	4657 (29.1%)

Table 4: Similarity measure for the most frequent classes of ambiguity.

Amb. class	**Si.**	**Fm.**	**Fg.**	**Ex. or BR**
proc/v[ms]	0	0	1	lui
nc[ms]/v[n]	0	0	1.3	être
d[fs]/nc[fs]	0	0	2.3	une
v[12]/v[s03]	0.2	1.3	7	semble,
sp/v[12]/v[s03]	0.2	0.2	1	entre, contre
prop[03]/cccs	0.2	0.3	1.7	s'
nc[ms]/v[12]/v[s03]	0.3	0.4	1.3	contrôle
r/v[12]/v[mp]	0.8	1	1.3	plus
d[ms]/nc[ms]	0.8	1.6	2	son
d[bp]/proc	0.8	5.5	7	les
d[ms]/proc	0.9	7.1	8.3	le
cccs/nc[ms]/r	1	1	1	bien
nc[ms]/v[s03]	1	1	1	fait
proc/prop[12]	1	1.7	1.6	nous
cccs/r	1	2.1	2.2	que
nc[ms]/r	1.1	4.9	4.6	pas
nc[ms]/v[s03]	1.2	5.3	4.5	est
nc[fs]/v[12] /v[s03]	1.3	2.6	2	sorte, mesure
proc/sp/cccs	1.6	7.5	4.6	en
d[bs]/proc	1.9	13.8	7.3	l'
d[fs]/proc	2.1	14.1	6.8	la
a/nc	4.2	1.7	0.4	patient
a/nc/v[4]	5.0	1.5	0.3	patiente

Note (tab. 4):

Column 1 gives the ambiguity class. Column 2 provides the ratio of similarity between the frequency of the considered ambiguity in medical (Fm.) and general texts (Fg.). For example, a similarity of 1 means that the ambiguity class is equally frequent in both corpora. A similarity of 5 means that the class is five times more frequent in medical texts, while a similarity of 0.2 means that the class is five times less frequent in medical texts. Columns 3 and 4 (resp. Fm. et Fg.) indicate the frequency of the ambiguity respectively in the medical texts and in the general texts. Column 5 provides some examples or the best representative (BR) of the ambiguity class, i.e. when one lexeme represents at least 80% of the class.

List of abbreviations for the syntactic categories: proc, clitic pronoun; v, verb; nc, common noun; d, determiner; sp, preposition; prop, personal pronoun; cccs, conjunction; q, numeral. List of abbreviations for the morpho-syntactic features and sub categorizations: ms, masculine singular; n, verbal infinitive form; fs, feminine singular; bs, masculine or feminine singular; 12, first and second person singular or plural; s03, third person singular; p03, plural third person.

A more precise table (tab. 4) provides at least two

[4] This class has only one representative within the medical corpus, the word *patient* (feminine *patiente*): an equivalent in the general corpus is *politique* (in Eng., both *political* and *politics*), but the former (0.5% of tokens) is more frequent than the latter (0.05%). The frequency of the word *politique* is consistent with the frequency list distributed by Jean Vèronis (*http://www.up.univ-mrs-fr/~veronis*).

remarkable results. First, it shows that in the general corpus, less than a dozen words are responsible for half of the global ambiguity rate. These results must be compared to [16], who situate this number around 15, while about six words generate the same level of ambiguity in the medical corpus! It means that medical texts are not only less ambiguous, but that the ambiguity distribution is more concentrated.

This table also shows that the distribution of the most frequent ambiguities is roughly domain-independent, however it is not totally true. Thus, the ambiguity d[fs]-[bs]/proc is twice more frequent in medical texts, and the ambiguity represented by the tokens *patient/patiente* (masculine and feminine form of *patient*; which may be a noun, an adjective, or some form equivalent to the verb *to wait*) is five times more frequent. On the contrary, some classes of ambiguity are simply absent or very rare in the medical domain (as for example v[12]/v[s03], or nc[ms]/v[n]). Such absence is important, as it is likely to cause both noise (if both alternatives are kept), and silent (if the disambiguation picks up the wrong category) in the retrieval process.

Let us notice that the overall lexical ambiguity rate is directly dependent on the lexicon size. Therefore, one of the main reasons why medical texts are less ambiguous is that we need smaller lexicons in order to parse them with the same accuracy (98-99%). However, even if we consider the same given lexicon, it remains that medical texts are 'less' ambiguous than general texts, as when we attempted to tag general texts with the medical lexicon, the ambiguity rate went up to 20% vs. 16% for medical texts (cf. footnote 3).

Table 5: Distribution of the most frequent morpho-syntactic categories : occurrences in general texts vs.occurrences in medical texts.

general texts		medical texts	
r	505	276	v[n]
v[n]	721	301	v[12]; v[s03]; v[p03]
cccs	765	550	q
v[12]; v[s03];	837	587	cccs
sp	1356	1283	a
d	1659	1529	f
nc	1707	1784	d
f	2179	2138	sp
-	-	3472	nc

Note (tab. 5) : f refers to the punctuations.

Finally, in table 5, we give the distribution of the most frequent syntactic categories according to the corpus. In this table, a particularly interesting result concerns the imbalance between categories of noun phrases (determiner, noun, adjective...) and categories of verb phrases (verb, adverb...); the former being much more frequent in medical texts, whereas the latter are more frequent in general texts. Here we verify a well-known stylistic manner: in clinical reports, verbs are frequently implicit, and nominalization is usually preferred[5]. As a corollary, noun categories (noun, adjective, determiners) are very frequent. Simple or complex numeral tokens (date, time, expressions with digits and measure symbols) are also much more frequent.

Discussion and conclusion

We have showed that the lexical ambiguity in medical texts is different to the one in general texts, both at a purely quantitative level, and at a deeper qualitative level. However, we must be aware of the artifacts introduced by the normalization method based on the MCT, which tends to make disappear some major differences between both corpora. All the part-of-speech (verb, noun, adjective... regardless on sub specifications: tense, mood...) of the FIPSTAG tagset have an equivalent in the medical tagset, while the contrary is not true[6]. Therefore some classes of ambiguity, quasi-specific to the medical domain, are not taken into account, as illustrated in the phrase *post and preoperative*, with the ambiguity *pref/nc/v* (prefix, common noun, verb):

post	pref/nc/v
and	cccs
pre	pref
operative	a

Another result concerns the difference in the distribution of the POS categories: noun phrases are more frequent in medical texts, while verb phrases are less frequent. All these particularities must be added to others: lexical, morphological, spelling and grammar errors. This last point has been rarely studied, but errors in documents, which are not intended for publication, may be quite impressive (the spelling error rate in our medical corpus was about 3%, i.e. up to one error every three sentences). Our conclusion is of two types: First, concerning the study, we showed that the use and comparison of taggers tailored for different corpora, supports a measure of the difference between these corpora; second, at a more methodological level, if it seems that the syntax may be -ceteris paribus- regarded as a domain-independent field (at least at a computational level, cf. [19]), we argued that real natural language processing applications require domain-adaptable tools. Therefore, the use of NLP tools by other research fields must be very carefully related to the design of these tools. We suggest that adaptability should be explored in at least three directions:

1. Systems must allow lexical items to be added (custom lexicon) *and* removed from the lexicon; therefore

[5] However, this is not always true and discharge summaries are often syntactically richer.

[6] When possible the medical tagset follows the MULTEXT [17] morpho-syntactic description, modified within the GRACE action. But we must notice that the original MULTEXT description and the GRACE version [18] for the French language (and English) have not been foreseen for annotating morphems.

access to the main lexicon must be available – at least negatively.

2. Systems must be optionally applied with a specialized morphological analyzer module (as for example for coping with medical morphem-based compounds).
3. MS description (tagset) should be parametrable, and this should include the ability to provide a mapping table.

Acknowledgments

This study was supported by the FNRS (Swiss National Science Foundation).

References

[1] Hersh WR. *Information Retrieval at the MILLENIUM.* In R MASY, Ed. *American Medical Informatics Association Annual Symposium (AMIA'1998).* Orlando. 1998.

[2] Spark-Jones K. What Is The Role for NLP in Text Retrieval. In: Strzalkowski, ed., *Natural Language Information Retrieval*, Kluwer Publishers, 1999. pp.1-25.

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

[4] Salton G. *Term-weighting approaches in automatic text retrieval.* McGraw.Hill. Vol. 24. New-York. 1988.

[5] Nazarenko A, Zweigenbaum P, Bouaud J. Corpus-based identification and Refinement of Semantic Classes. In R MASY, ed., *American Medical Informatics Association Annual Symposium (AMIA'1997, ex-SCAMC)*, 1997. pp. 585-589.

[6] Bouillon P, Ruch P, Baud R, Robert G. *Indexing by statistical tagging*. In: Rajman and al. ed. *Proceedings of the 7th JADT'2000. Lausanne*. Switzerland. 2000. Vol. 1, pp. 35-42.

[7] Ruch P, Wagner J, Bouillon P, Baud R. Tag-like semantics for medical document indexing. In N. M. LORENZI, ed. *American Medical Informatics Association Annual Symposium (AMIA'1999).* Washington. 1999. pp. 137-141.

[8] Biber D, Conrad S, Reppen R. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press. 1998.

[9] Folch H, Heiden S, Habert B, Fleury S, Illouz G, Lafon P, Nioche J, Prévost P. (2000) TypTex: Inductive Typological Text Classification by Multivariate Statistical Analysis for NLP Systems Tuning/Evaluation. In *Proceedings of the 3rd International Conference on Language Ressources and Evaluation (LREC'2000)*, Athenes. 2000.

[10]Kilgariff A. *Which words are particularly characteristic of a text? A survey of statistical approaches.* ITRI Technical report 96-08. (http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/publications.html). 1996.

[11]Ruch P, Bouillon P, Baud R, Rassinoux AM, Robert G. Tagging medical texts: a rule-based experiment. In: A. Hasman and al., ed. *Medical Infobahn for Europe (Proceedings of MIE'2000)*, IOS Press. 2000. pp.448-455.

[12]Wehrli E. The Interactive Parsing System, *In ACL, ed., Proceedings of COLING-92.* Nantes. France. 1992. pp. 870-4.

[13]Lovis C, Baud R, Michel PA, Scherrer JR. Morphosemantems decomposition and semantic representation to allow fast and efficient natural language recognition of medical expressions. In R MASY, ed., *American Medical Informatics Association Annual Symposium (AMIA'1997, ex-SCAMC). Washington.* 1997.

[14]Ruch P, Bouillon P, Robert G, Baud R. Minimal Commitment and Full Lexical Disambiguation: Balancing Rules and Hidden Markov Models. *In Proceedings of the 5th CoNLL Conference (ACL-SIGNLL).* Lisbon. Portugal. 2000. pp. 111-114.

[15]Tesnière L. *Elements de syntaxe structurale.* Klincksieck. Paris. 1959.

[16]Chanod JP, Tapanainen P. Tagging French: comparing a statistical and a constraint-based method. In ACL, Ed., *7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95).* Dublin. 1995. pp. 149-156.

[17]Nancy Ide N, Véronis J. MULTEXT: Multilingual Text Tools and Corpora. *In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan. 1994.

[18]Rajman M, Paroubek P, Lecomte J. *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). 1996.

[19]Wehrli E, Clark R. Natural Language Processing: Lexicon and Semantics, *Meth. of Inf. in Medicine*,. 1995. Vol. 34, pp. 68-74.

Address for correspondence

Patrick Ruch
University Hospital of Geneva - Medical Informatics Division
CH-1211 Geneva
Switzerland
ruch@dim.hcuge.ch