In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 111–114, Lisbon, Portugal, 2000.

Minimal Commitment and Full Lexical Disambiguation: Balancing Rules and Hidden Markov Models

Patrick Ruch and **Robert Baud** and **Pierrette Bouillon** and **Gilbert Robert***
Medical Informatics Division, University Hospital of Geneva
ISSCO, University of Geneva
{ruch, baud}@dim.hcuge.ch, {bouillon, robert}@issco.unige.ch

Abstract

In this paper we describe the construction of a part-of-speech tagger both for medical document retrieval purposes and XP extraction. Therefore we have designed a double system: for retrieval purposes, we rely on a rule-based architecture, called minimal commitment, which is likely to be completed by a data-driven tool (HMM) when full disambiguation is necessary.

1 Introduction

Nowadays, most medical information is stored in textual documents[1], but 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 (Hersh and al., 1998), and on medical texts (Hersh, 1998), are quite critical towards linguistic engineering, we believe that natural language processing is the best solution to face two major problems of text retrieval engines: expansion of the query and lexical disambiguation. Disambiguation can be separated between MS (morpho-syntactic, i.e. the part-of-speech (POS) and some other features) and WS (word-sense) disambiguation. Although we aim at developing a common architecture for processing both the MS and the WS disambiguation (Ruch and al., 1999), this paper focuses on the MS tagging.

* We would like to thank Thierry Etchegoyhen and Erik Tjong Kim Sang for their helpful assistance while writing this paper. The Swiss National Fundation supported the present study.

[1]While our studies were made on French corpora, the examples are provided in English -when possible- for the sake of clarity.

2 Background

Before starting to develop our own MS tagger, some preliminary studies on general available systems were conducted; if these studies go far beyong the scope of this 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[2] the property of a system, which does not attempt to solve ambiguities when it is not likely to solve it well! Such property seems important for IR purposes, where we might prefer noise rather than silence in the recall process. However, it must remain optional, as some other tasks (such as the NP extraction, or the phrase chunking (Abney, 1991)) may need a full disambiguation.

2.1 Data-driven tools

We adapted the output of our morphological analyser for tagging purposes (Bouillon et al., 1999). We trained and wrote manual biases for an 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 the ambiguities were attributed a wrong tag! We attempted to set a confidence threshold, so that for similarly weighted transitions, the system would keep the ambiguity, as in (Weischedel and al., 1993), but results were not satisfying.

2.2 Constraint-based systems

We also looked at more powerful principle-based parsers, and tests were conducted on

[2]The first one using this expression was maybe M. Marcus, lately we can find a quite similar idea in Silberztein (1997).

Token	Lemma	Lexical tag(s)
fast	fast	a
section	section	nc[s]
of	of	sp
the	the	dad
internal	internal	a
faces	face/to face	nc[p]/v[s03]

Table 1: Tag-like representation of MS lexical features

Token	Lexical tags	Disambiguated tag
fast	a	a
section	nc[s]	nc[s]
of	sp	sp
the	dad	dad
internal	a	a
faces	nc[p]/v[s03]	nc[p]

Table 2: Example of tagging

FIPSTAG[3] (a Government and Binding chart-parser (Wehrli, 1992)). Although this system performed well on general texts, with about 0.7% of errors, its results on medical texts were about the same as stochastic taggers. As we could not adapt our medical morphological analyser on this very integrated system, it had to cope with several unknown words.

3 Methods

In order to assess the system, we selected a corpus (40000 tokens) based equally on three types of documents: reports of surgery, discharge summaries and follow-up notes. This ad hoc corpus 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.

3.1 Lexicon, morphological analysis and guesser

The lexicon, with around 20000 entries, covers exhaustively the whole ICD-10. The morphological analyser is morpheme-based (Baud et al., 1998), it maps each inflected surface form of a word to its canonical lexical form, followed by the relevant morphological features. 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. 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. Thus, the ending *able* characterises very probably an adjective, but does not provide any semantic information[4] on it.

Let us consider two examples of words absent from the lexicon. First, *allomorph*: the prefix part *allo*, and the suffix part, *morph*, are listed in the lexicon, with all the MS and the WS features, therefore it is recognized by the first-stage guesser. Second, *allocution*, it can not be split into any affix, as *cution* is not a morpheme, but the ending *tion* refers to some features (noun, singular) in the second-stage guesser. As the underlying objective of the project is to retrieve documents, the main and most complete information is provided by the first-stage guesser, and the second-stage is only interesting for MS tagging, as in (Chanod and Tapanainen, 1995). Finally (tab. 1), some of the morpho-syntactic features provided by the lemmatizer are expressed into the MS tagset[5], to be processed by the tagger (tab. 2).

[3] 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

[4] A minimal set of lexical semantic types, based on the UMLS, has been defined in (Ruch and al., 1999).

[5] The MS tagset tends to follow the MULTEXT lexical description for French, modified within the GRACE action (http://www.limsi.fr/TLP/grace/doc/GTR-3-2.1.tex). However, it is not always possible, as this description does not allow any morpheme annotation.

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)

Table 3: Results for each evaluation (GC stands for good candidates)

Statistical evaluation on the residual ambiguity	MFT	HMM
Tokens correctly tagged	8136 (98.7)	8165 (99.1)
Tokens incorrectly tagged	107 (1.3)	78 (0.9)

Table 4: Processing the residual ambiguity

3.2 Studying the ambiguities

Our first investigations aimed at assessing the overall ambiguity of medical texts. We found that 1227 tokens (14.4% of the whole sample[6]) were ambiguous in set A, and 511 tokens (6.0%) were unknown. We first decided not to care about unknown words, therefore they were not taking into account in the first assessment (cf. Performances). However, some frequent words were missing, so that together with the MS guesser, we would improve the guessing score by adding some lexemes. Thus, adding 232 entries in the lexicon and linking it with the Swiss compendium (for drugs and chemicals) provides an unknown word rate of less than 3%. This result includes also the pre-processing of patients and physicians names (Ruch and al., 2000). Concerning the ambiguities, we found that 5 tokens were responsible for half of the ambiguities, while in unrestricted corpora this number seems around 16 (Chanod and Tapanainen, 1995).

3.2.1 Local rules

We separated the set A in 8 subsets of about 1000 tokens, in order to write the rules. We wrote around 50 rules (which generated more than 150 operative rules) for the first subset, while for the 8th, only 12 rules were necessary to reach a score close to 100% on set A. These rules are using intermediate symbols (such as the Kleene star) in order to ease and improve the rule-writing process, these symbols are replaced when the operative rules are generated. Here is an example of a rule:

prop[**];v[**]/nc[**] $\rightarrow$ prop[**];v[**]

This rule says 'if 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 this pronoun has (3rd or 1st/2nd person), then the ambiguous token can be rewritten as a verb, keeping its original features (**)'.

4 Performances

4.1 Maximizing the minimal commitment

Four successive evaluations were conducted (tab. 3); 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. Let us note that translation phenomena (Paroubek and al., 1998), which turn the lexical category of a word into another one, seem rare in medical texts (only 3 cases were not foreseen in the lexicon).

A success rate of 98% (tab. 3, 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[7]. Another interesting result concerns the residual ambiguity (tokens still

[6]For comparison, the average ambiguity rate is about 25-30% in unrestricted corpora.

[7]Let us note that in the assessment 1, the system had

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 are clearly untractable with such contextual rules, and would demand more lexical information, 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*[8]), where *diffuse* could be adjective or verb.

4.2 Maximizing the success rate

A last experiment is made: on the set E, which has been disambiguated by the rule-based tagger, we decided to apply two more disambiguations, in order to handle the residual ambiguity. First, we apply the most frequent tag (MFT) model, as baseline, then the HMM. Both the MFT and the HMM transitions are calculated on the set B+C+D, tagged manually, but without any manual improvement (bias) of the model.

Table 4 shows that for the residual ambiguity, i.e. the ambiguity, which remained untractable by the rule-based tagger, the HMM provides an interesting disambiguation accuracy[9].

5 Conclusion

We have presented a rule-based tagger for electronic medical records. The first target of this tool is the disambiguation for IR purposes, therefore we decided to design a system without any heuristics. As second target, the system will be used for conducting NP extraction tasks and shallow parsing: the system must be able to provide a fully disambiguated output; therefore we used the HMM tool for completing the disambiguation task.

References

Abney, Steven. 1991. Parsing by chunks. In R. Berwick and S. Abney and C. Tenny, editors, *Principle-based parsing*, pages 257–278. Kluwer.

Robert Baud, C. Lovis, and AM. Rassinoux. 1998. Morpho-semantic parsing of medical expressions. In *Proceedings of AMIA'1998*, pages 760–764, Orlando.

Pierrette Bouillon, R Baud, G Robert, and P Ruch. 1999. Indexing by statistical tagging. In *Proceedings of the JADT'2000*, pages 35–42, Lausanne.

Jean-Pierre Chanod and Pasi Tapanainen. 1995. Tagging french: comparing a statistical and a constraint-based method. In *Proceedings of EACL'95*, pages 149–156, Dublin.

William, Hersh and S. Price and D. Kraemer and B. Chan and L. Sacherek and D. Olson. 1998. A large-scale comparison of boolean vs. natural language searching for the trec-7 interactive track. In *TREC 1998*, pages 429–438.

William Hersh. 1998. Information retrieval at the millenium. In *Proceedings of AMIA'1998*, pages 38–45, Lake Buena Vista, FL.

Paroubek, Patrick and G. Adda and J. Mariani and J. Lecomte and M. Rajman. 1998. The GRACE french part-of-speech tagging evaluation task. In *Proceedings of LREC'1998*, Granada.

Ruch, Patrick and J. Wagner and P. Bouillon and R. Baud and A.-M. Rassinoux and G. Robert. 1999. Medtag: Tag-like semantics for medical document indexing. In *Proceedings of AMIA'99*, pages 35–42, Washington.

Ruch, Patrick and R. Baud and A.-M. Rassinoux and P. Bouillon and G. Robert 2000. Medical document anonymization with a semantic lexicon. In *Proceedings of AMIA'2000*, Los Angeles.

Max Silberztein. 1997. The lexical analysis of natural languages. In Emmanuel Roche and Yves Shabes, editors, *Finite-State Language Processing*, pages 176–205. MIT Press.

Eric Wehrli. 1992. The IPS system. In *Proceedings COLING-92*, pages 870–874.

Weischedel, Ralph and M. Meeler and R. Shwartz and L. Ramshaw and J. Palmucci, 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19(2):359–382.

about 1000 operative rules, while the assessment 4 was conducted with more than 2000 rules.

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

[9]The accuracy of the HMM tagger, on the fully ambiguous version of set E, was 96.3%, while the MFT performed about 93.5%.