MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-949

Extracting Medication Information from French Clinical Texts

Louise Deléger, Cyril Grouin, Pierre Zweigenbaum

LIMSI-CNRS, Orsay, France

Abstract

Much more Natural Language Processing (NLP) work has been performed on the English language than on any other. This general observation is also true of medical NLP, although clinical language processing needs are as strong in other languages as they are in English. In specific subdomains, such as drug prescription, the expression of information can be closely related across different languages, which should help transfer systems from English to other languages. We report here the implementation of a medication extraction system which extracts drugs and related information from French clinical texts, on the basis of an approach initially designed for English within the framework of the i2b2 2009 challenge. The system relies on specialized lexicons and a set of extraction rules. A first evaluation on 50 annotated texts obtains 86.7% F-measure, a level higher than the original English system and close to related work. This shows that the same rule-based approach can be applied to English and French languages, with a similar level of performance. We further discuss directions for improving both systems.

Keywords:

Natural language processing, Information extraction, Drug prescriptions, Computerized medical records systems, Information storage and retrieval/methods

Introduction

The information contained in Electronic Health Records (EHRs) may take the form of coded data or may be written in free text. Free text is indeed the easiest and most natural way for physicians to convey information [1]. It cannot, however, be used as is by health information systems [1,2]. It is also time-consuming for clinicians to read narrative sections in order to find relevant information. Natural Language Processing (NLP) techniques — more specifically information extraction methods — have therefore been proposed to gain easier access to this information [3]. Among the useful types of information to extract from narrative reports is information related to treatment such as medications [4–9].

Within the framework of the i2b2 NLP2009 challenge[1], we developed a medication extraction system for English narrative patient records. This system extracts and links medications with related information such as dosage, mode of administration, frequency, duration and reason for treatment. According to Harris [10], "The structure of each science language is found to conform to the information in that science rather than to the grammar of the whole language". This was exemplified further by Borst *et al.* [11] when they designed a system for analyzing French discharge summaries starting from Sager's [12] original system designed for English. In contrast, Friedman *et al.* [13] highlight the differences found when developing systems for two distinct domains (clinical and biomolecular), although both in English. Closeness or distance of information structure is thus stronger than closeness or distance in languages. If the same principle applies to the language of medications, it should then be possible to transfer our initial English medication extraction system to one for French medications with limited modifications. In this paper, we test this hypothesis and report the implementation of a similar system to deal with medications in French medical records.

This work is part of the AKENATON project which addresses information extraction in the domain of telecardiology. In this context, extracting information from clinical texts would allow physicians to link more easily automatic alerts to patient data, including coded data obtained from electronic health records and that obtained from free text. Medications are one type of useful information to be extracted in this regard.

This paper is structured as follows. We begin by a review of existing work. We then describe our corpora of clinical texts and detail the implementation of our system. We finally present its evaluation and discuss the results.

Related work

A few approaches, all dedicated to English, have addressed medication extraction from free text.

Some approaches focus on extracting a specific type of medication information, such as drug names or dosage. Levin *et al.* [4] developed a system based on lexicons (drug names and medical abbreviations) and regular expressions to extract drug names (generic or trade names). In order to deal with misspelled drug names, the authors used a phonetically based matching module, thus allowing them to increase the extraction by 7%. Also centering on drug name recognition, Sirohi *et al.* [5] studied the importance of determining the best lexi-

[1] https://www.i2b2.org/NLP/Medication/

con list to use in order to improve the quality of drug name extraction. They used the commercial software FreePharma and experimented with filtering criteria to refine drug lexicons, more specifically to eliminate ambiguous entries or to take into account abbreviated forms. Shah and Martinez [6] focused on recognizing dosage from the free-text field of a database of patient records specifying dosage instructions. The extracted information is then classified according to an existing format, which includes daily dose, frequency, units, duration. This system does not detect drug names, as they are contained in a structured field of the database.

Other approaches aim at extracting a more complete set of information elements related to medications. The first system specifically designed to extract drug and dosage information was that of Evans *et al.* [7]. The authors defined a model of the drug-dosage information to be extracted that included drug name, dose level, route, frequency and necessity. They detected this information using a set of extraction rules relying on lexicons (both general and specialized) and NLP steps including stemming, part-of-speech tagging and semantic category assignment. In their system, a drug name was extracted only if associated with at least one related piece of information (*e.g.* dose, frequency, duration, etc). More recently, Gold *et al.* [8] built Merki, a parser for extracting a similar set of information. The process consists in identifying drug names using a lexicon as a first step, and in applying regular expressions to detect associated elements as a second step. Xu et al. [9] used a more detailed drug model and detected medication information performing semantic tagging and parsing.

Materials and Methods

Development and test corpora

The corpora used in this experiment consist of a total of 17,412 French EHRs from the cardiology unit of a French University Hospital, written between 2004 and 2006. They include discharge summaries, consultation reports and surgical reports. This set is divided into two corpora: a development corpus of 17,362 documents which we used to implement our system and a test corpus of 50 documents which we manually annotated to use as a gold standard. The test corpus contained 253 medications, plus the associated information elements.

A rule-based approach to medication extraction

Our approach to medication extraction is rule-based, as is often the case in information extraction. A set of lexicons define the relevant vocabulary, and a set of extraction rules encode the grammar of medication expressions.

Lexicons

Lexicons associate linguistic information to words. Here, each lexicon entry is associated to a semantic category which corresponds to one of the target information types: drug, dosage, mode of administration, frequency, duration, and sign or symptom which is the reason for prescription. We group these entries into three lexicons according to the sources of information used to compile them. The first two are acquired from external sources, while the third is mostly obtained from a development corpus: a *drug lexicon*, used to recognize drug names; a *list of signs and symptoms*, to detect reasons; a *list of abbreviations and expressions*, which is used by the extraction rules to identify information related to medications: dosage, mode of administration, frequency, and duration.

System outline

The system first segments the text into sentences based on typographical clues, *i.e.* certain types of punctuation (here we considered only full stops). We use these punctuations to determine sentence boundaries, while also taking care of exceptions, mainly periods which occur in abbreviations ("etc.") or within numbers ("1.5").

Then, the first stage of our extraction algorithm iterates over the sentences to recognize drug names. The process consists in a lexicon look-up. A drug is extracted when an exact match is found between the text and an entry of the drug lexicon. This stage terminates when the text has been completely scanned.

The second stage starts by dividing each sentence into subparts according to the detected drug names. That is, each resulting text span is composed of a medication name and the text which follows that name. We call this process drug span segmentation. The underlying assumption is that most information associated to a drug occurs in the text span which follows the drug name.

The second stage of the algorithm thus consists in looking for information related to medications within each of these text spans. If needed, we also extend the search to other parts of the sentence in which a drug name occurs, especially to the text closely preceding the drug name, in order to deal with cases where a piece of information does not follow a drug name. This second stage relies on lexicons and on a set of extraction rules implemented by regular expressions.

French implementation of the system

The system was originally designed for English in the context of the i2b2 challenge. The above-described principles and outline apply to both the English and French versions of the system. We do not detail the specificities of the English system any further and we focus on the French implementation from now on. This version was modified and extended to take into account the specificities of the language and the corpus.

Corpus study

A first study of the corpus highlighted structural similarity between French and English EHRs. At the document level, information is generally structured as follows: illness and social history, allergies, medications on admission, examinations, and discharge medications. At the local level, drug names are often followed by related information (dosage, frequency, mode of administration, etc.) in the same sentence. These shared patterns point to the same direction as previously observed by Harris [10] and Borst *et al.* [11]. They give positive signs that our hypothesis could be valid, and that it might thus be possible to obtain a localization from English to French with limited modifications.

Lexicons

Three French lexicons were constituted according to the three previously defined types. The *drug lexicon* was compiled from the Internet using drug lists provided by three different sources: Vidal[2], Eureka Santé[3] and Doctissimo[4]. We also completed these lists with drug names not found in these sources, namely drug name abbreviations (“*avk”* for “*anti-vitamine K”*, *i.e. oral anticoagulants* in English), common orthographic and grammatical variations of names (“*bétabloquant”*, “*bêtabloquants”*, *i.e. beta-blockers* in English), and drug names mentioned in the records but absent from the Internet sources since they are not used anymore nowadays. Finally, we added substance names from the Biam database[5]. The resulting lexicon is composed of 33,371 drug names.

The *list of signs and symptoms* was obtained by querying the UMLS (version 2008AA) for French terms with the *Sign or Symptom* semantic type. It contains 3,988 entries.

The *list of abbreviations and expressions* was adapted from the English list, by translating existing entries and adding new ones when necessary. This list includes 68 abbreviations and expressions for 4 types of elements: dosage, mode of administration, frequency, and duration (see Table 1).

Table 1 – Excerpt of the list of abbreviations and expressions

Entry	Attribute
mg	DOSE
iv	MODE
h	FREQUENCY
heure (*hour*)	FREQUENCY
semaine (*week*)	DURATION

Since we can easily access such kind of information for French, there was no problem in creating these lexicons. Nevertheless, a small amount of post-processing of the lists was performed to remove entries that were ambiguous (for instance, *eau* (*water)* was listed as a pharmacologic substance) or too general (e.g. *"mal": pain*, *illness, disease,* in the signs and symptoms list); this should allow us to reduce over-extraction and thus increase precision.

Extraction rules

The extraction rules were designed using both the initial English rules and examples from the development corpus. For each item to be extracted (*i.e.* dosage, mode, frequency, duration and reason), regular expressions are applied in combination with a lexicon look-up. The list of abbreviations and expressions is used to detect dosage, mode, frequency and duration, while the purpose of the lexicon of signs and symptoms is to identify reasons. Examples of extraction rules are given in Table 2, where uppercase words represent semantic categories as obtained through lexicon lookup.

Table 2 – Example extraction rules (FREQ = frequency)

Rule	Sample matched phrase
[0-9]+[.,0-9]* DOSE	2,5mg
[0-9]+ DOSE [0-9]/[0-9]	2 cp 1/2 (*2 tabs 1/2*)
[0-9]+ FREQ / FREQ	3 fois / j (*3 times a day*)
[0-9]+ FREQ / [0-9]+	5 jours / 7 (*5 days out of 7*)
Pendant [0-9]+ DURATION	pendant 3 semaines (*for 3 weeks*)

These rules were either adapted from the English ones or rewritten based on the corpus study. The adaptation of the rules was often almost direct. For instance, the first rule of table 2 corresponds to the English rule *[0-9]+[.0-9]* DOSE*. In this case, a comma has simply been inserted in the expression since figures such as 2.5 are usually written as 2,5 in French. The last rule of Table 2 also shows a basic modification of the initial English rule *for [0-9]+ DURATION*: here, the English word *for* has been translated by the French word *pendant.*

The implementation of this French system was performed in a relatively short period of time: about 10 hours, of which approximately 1/3 was devoted to lexicon compilation and 2/3 to the adaptation and development of rules. Preparing the gold standard was comparatively a more time-consuming task (approximately 10 hours to annotate the 50 documents).

Evaluation

We evaluated our system against the test corpus, in terms of recall (the ratio between the number of correct extractions and the number of expected extractions), precision (the ratio between the number of correct extractions and the total number of extractions), and F-measure (the weighted harmonic mean of recall and precision, with a weight set to 1 to give recall and precision the same importance) computed at different levels: an horizontal level which assesses all information as a whole (drug names and their associated information) and specific levels (referred to as the vertical level) which evaluate each item separately (medication, dosage, frequency, mode, duration, reason), as per the i2b2 medication extraction guidelines.

Results

Table 3 shows that the F-measure is high on the horizontal level, as well as on the levels of medication, dosage, frequency and duration, but rather low for mode and reason.

[2] http://www.vidal.fr/fiches-medicaments

[3] http://www.eurekasante.fr/medicaments.html

[4] http://www.doctissimo.fr/html/medicaments/articles/medicaments_loupe.htm

[5] http://www.biam2.org/ (last update: May 2001).

Table 3 – Evaluation of French medication extraction (n = number of instances to be extracted, R = recall, P = precision, F = F-measure)

	n	R	P	F
Horizontal	257	0.839	0.896	0.867
Medication	257	0.887	0.934	0.910
Dosage	110	0.891	0.907	0.899
Mode	6	0.500	0.750	0.600
Frequency	122	0.795	0.858	0.825
Duration	4	0.750	0.750	0.750
Reason	23	0.391	0.563	0.462

Table 4 – Example extracted medications (an "nm" label is used when the information item is not mentioned in the text)

Original text	Extracted medication

nous conseillons donc la mise en place d'un traitement par Durogésic et débutons ce jour les patch de 25 µg/72 h
we thus advise treatment with Durogesic and start today patches of 25 µg/72h
	medication="durogésic"
	dose="25 µg"
	mode="patch"
	frequency="/72 h"
	duration="nm"
	reason="nm"

Paracétamol 1 g = x 4/j si douleur
Paracetamol 1g = x 4 a day if pain
	medication="paracétamol"
	dose="1 g"
	mode="nm"
	frequency="x 4/j"
	duration="nm"
	reason="douleur"

The results obtained by our localized French medication information extraction system are higher than those of the original English system[6]: at the horizontal level, we obtained an F-measure of 0.867 in French and of 0.773 in English. At the vertical level, medication name extraction yielded the best results for French (0.910) while it was only average for English (0.798). Dosage extraction also yielded much better results in French: 0.899 against 0.804 in English. Frequency detection was equally good for both systems (0.825 and 0.827). Reason extraction produced the worst results also for both systems: 0.462 in French and 0.299 in English (this was a common feature to all systems participating to the i2b2 challenge as well). There is a high difference in favour of the English system, however, regarding the extraction of the mode of administration: while it is the best type of information we extracted in English (F-measure of 0.836), we obtained a low F-measure in French (0.600). This difference is essentially due to the fact that there are very few occurrences of modes in the French corpus (only 6 occurrences over a total of 257 medication sets of information to extract).

Table 4 gives examples of medication information extracted by our system.

Discussion

Our system achieved good results for French, even higher than those of our initial English version. It should be noted, however, that this cannot constitute a fully accurate comparison between the two systems, mainly because we evaluate the French system on a much smaller reference corpus (annotating records is indeed time-consuming) than that used in the i2b2 challenge. It gives a fair indication, though, as to the quality of the system. Results are also close to existing systems working on English. Gold *et al.* [8], for instance, obtained a precision of 94.1% and a recall of 82.5%, and Xu *et al.* [9] reported F-measures over 93% for drug names, strengths, routes and frequencies. Our system is most comparable to these two works because we identify similar types of information. They do not extract, however, reasons for administration.

There is room for improvement, especially for the extraction of reasons. We mainly relied on proximity to associate signs and symptoms to prescriptions. However, more sophisticated Natural Language Processing modules, such as a part-of-speech tagger or a syntactic parser, could also be applied to the texts. Syntactically parsing the text is an interesting direction to investigate, as it would allow us, for instance, to identify prepositional and noun phrases and grammatical relations, which would be useful to link reasons to prescriptions more accurately. Another way to improve reason identification would be to rely on a knowledge base associating drug names with the symptoms they treat (*e.g. simvastatine* and *Zocor* for *hypercholestérolémie*). Based on this known association, if the reason *hypercholestérolémie* (or other signs or symptoms related to this one) is found in the neighborhood of *Zocor* or *simvastatine*, we could give it a more important weight: this might help improve the precision of reason detection.

Originally, the program was designed within the framework of the 2009 i2b2 challenge. Therefore, the pieces of information to extract are those defined in the challenge. We transposed our program from English to French using the same definition of the items to be extracted, namely the following six types of information: drug name, dosage, mode, frequency, duration and reason. However, when processing the corpus, we were confronted with the ambiguity of some types of information. *Dosage* refers both to the drug dose the patient has to take ("Previscan **1 cp** par jour": *Previscan **1 tab** a day*) and to the drug concentration ("Plavix **75 mg**"). Sometimes dose and concentration are both mentioned (e.g. "Levothyrox **150 µg 1 cp**": *Levothyrox **150 µg 1 tab***), in which case both pieces of information were extracted as dosage. It might be interesting, however, to separate them. Another ambiguity due to the chosen representation of information concerned *Frequency*: it can refer to the frequency with which the patient has to take the medicine ("Coversyl 8 mg/**jour**": *Coversyl 8*

[6] The English system was evaluated against a set of 256 annotated records. Those are the official results of the i2b2 challenge.

*mg **a day***) as well as to the time of day when the drug should be taken ("Symbicort 1 bouffée **matin et soir**": *Symbicort 1 puff **in the morning and in the evening***). In this last example, we can deduce the frequency from the time (2 times a day), but we cannot deduce the time from the frequency. Grouping information into one type makes identification easier, but it would also make sense to represent each type of information separately.

It could also be interesting to extract additional information related to medications. Useful information would be, for instance, *temporal markers (i.e.* is the time of medication administration in the past, in the present, or in the future?*),* *events* (*i.e.* is the medication being started, stopped, or continued?), and *certainty (i.e.* is the medication suggested or compulsory?). These were considered by the i2b2 challenge at first, but later dropped to simplify the task. An interesting direction for future work would be to process such information.

Finally, the current output representation is the exact strings of words found in the input texts. Further work will address normalizing these strings into canonical forms: *e.g.* unique identifier for each drug, unique preferred form for *tab* and *tablet*, etc.; a task similar to that described in [14]. This will enable us to merge them with coded data obtained from EHRs.

Conclusion

In this paper, we presented our experiments to localize an existing medication information extraction system from English to the French language. This localization kept the same target information items and semantic categories. It was based upon the compilation of French lexicons and the adaptation to French of the regular expressions used to extract the different items. This last part represents most of the work, since it implies re-writing a certain number of rules. Nevertheless, the work done for English gave us pointers to the types of rules to define, so that we believe our approach saved time compared to creating a system from scratch. Also, the French medical texts exhibited some similarity to the English texts, which made the transposition of some of the rules almost direct.

An evaluation of this localization over a corpus of 50 French EHRs provided better results than those obtained by the English systems at the i2b2 challenge.

This work shows that in the case of a specific sublanguage, that of prescriptions, the same approach can be successfully applied to two different languages, English and French.

Acknowledgments

This work was partially funded by the Akenaton project under grant number ANR-07-TECSAN-001and by the Quæro project. English deidentified clinical records used first were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Özlem Uzuner, i2b2 and Suny.

References

[1] Spyns P. Natural language processing in medicine: an overview. Meth Inform Med 1996;35(4-5):285–301.

[2] Friedman C, and Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999 Aug;74(8):890–5.

[3] Meystre S, Savova G, Kipper-Schuler K, and Hurdle J. Extracting information from textual documents in the electronic health record: a review of recent research. In: Yearb Med Inform. 2008:128–44.

[4] Levin MA, Krol M, Doshi AM, and Reich DL. Extraction and mapping of drug names from free text to a standardized nomenclature. In: AMIA Annu Symp Proc 2007 Oct 11:438–42

[5] Sirohi E, and Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. In: Pac Symp Biocomput 2005:308–18.

[6] Shah AD, and Martinez C. An algorithm to derive a numerical daily dose from unstructured text dosage instructions. Pharmacoepidemiol Drug Saf 2006 March;15(3):161–6.

[7] Evans DA, Brownlow ND, Hersh WR and Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. In: Proc AMIA Annu Fall Symp 1996; 388–92.

[8] Gold S, Elhadad N, Zhu X, Cimino JJ, and Hripcsak G. Extracting structured medication event information from discharge summaries. In: AMIA Annu Symp Proc 2008: 237–41.

[9] Xu H, Stenner S, Doan S, et al. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc Jan-Feb 2010;17(1):19–24.

[10] Harris ZS. Language and information. New York: Columbia University Press, 1988.

[11] Borst F, Sager N, Nhàn NT, Su Y, Lyman M, Tick LJ, Revillard C, Chi E, and Scherrer JR. Analyse automatique de comptes rendus d'hospitalisation. Informatique et Gestion des Unités de Soins. Paris: Springer-Verlag, 1989;1:246–56.

[12] Sager N, Friedman C, and Lyman MS (eds.). Medical Language Processing: Computer Management of Narrative Data. Reading, Mass.: Addison Wesley, 1987.

[13] Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002 Aug;35(4):222–35.

[14] Morgan AA, Hirschman L. Overview of BioCreative II gene normalisation. In: Proc BioCreAtIvE II Workshop, Madrid, 2007; 17–27

Address for correspondence

Corresponding Author: Louise Deléger, LIMSI-CNRS, BP 133, F-91403 Orsay Cedex, France; E-mail: louise.deleger@limsi.fr