Concurrent linguistic annotations for identifying medication names and the related information in discharge summaries

Thierry Hamon[3], PhD, Natalia Grabar[1,2], PhD

[1]**Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR_S 872, Paris, F-75006; Université Paris Descartes, UMR_S 872, Paris, F-75006; INSERM, U872, Paris, F-75006 France**

[2]**HEGP AP-HP, 20 rue Leblanc, Paris, F-75015 France**

[3]**LIM&BIO (EA3969), Université Paris 13, 74, rue Marcel Cachin, 93017 Bobigny Cedex France**

Abstract

The 2009 I2B2 NLP challenge concentrated on extraction of medication-related information (medication name, its dosage, frequency, mode of administration and reason for prescription). For participation in this challenge, we designed an automatic NLP system exploiting terminological resources and a rule-based approach. We considered the challenge task as annotation and annotation selection problem. Thus, several annotations can be generated concurrently and then modules of the postprocessing step perform cleaning, disambiguation and establishing of dependency relations between medication names and the related information. Our system provides good results for the annotation and extraction of medication names, their frequency, dosage and mode of administration, while information on duration and reasons is poorly annotated and extracted.

Introduction

In this paper, we describe the architecture of the system we designed and developed in order to participate in the I2B2 challenge 2009. This year's challenge aim consisted into extraction of information related to medication from discharge summaries: medication names prescribed to a patient and additional information on their dosage, frequency, mode of administration and reason for the prescription. Besides, automatic systems had also to indicate whether the extracted information was extracted from narrative or list sections, and to provide offset information (line and token numbers). We used an NLP system which exploits terminological resources and a rule-based approach.

Building of material and of annotation resources

Discharge summaries. Challenge data consists of discharge summaries from Partners Healthcare. All records, written in English, have been de-identified.[1] A total of 1,249 documents have been used in this challenge, split into training (n=696) and test (n=553) sets. Within the training set, there was only 17 manually annotated documents illustrating the annotation guidelines.

Resources. We use two types of resources for the annotation of discharges summaries (a total of 290,243 entries):

1. 243,869 entries from RxNorm[2] for detection of medication names as the main source. The medication list has been cleaned up and enriched. Moreover, we added therapeutic classes and groups of medications as they have been provided by the FDA website. Besides, as we observed that at least 108 RxNorm entries are ambiguous (*i.e.*, *red blood cells*, *magnesium*, *iron*), we consider them as medication in specific contexts only: when they appear in list sections.
2. 45,898 terms from the Diagnosis and Morphology axes of the Snomed International[3] for detection of reasons. We use the Snomed International terminology because it proved to be an efficient source for the NLP processing[4]. As for the reason list, it has been enriched with 476 terms issued from the training set documents.

Negation Markers. Detection of negation is performed with the NegEx resource available on-line.[a] Among these negation markers, pre and post-negation are distinguished as well as diminishing markers which decrease the scope of negation phrases. Some additional markers have been added, for a total of 284 markers.

Contextual rules. Annotation of frequency, dosage, duration and mode of administration is based on recognition of named entities and on contextual rules. They have been built manually and are encoded as automata. In order to increase the coverage of reason identification, 52 additional rules allow to characterize extra-Snomed International noun phrases as reasons.

Method for concurrent linguistic annotations

Our system (see figure 1 foe the global architecture) is built upon the Ogmios platform[b], suitable for the processing and annotation of large amounts of data and tunable to specialized areas. Through this platform,

[a] www.dbmi.pitt.edu/chapman/NegEx.html

[b] http://search.cpan.org/∼thhamon/Alvis-NLPPlatform/

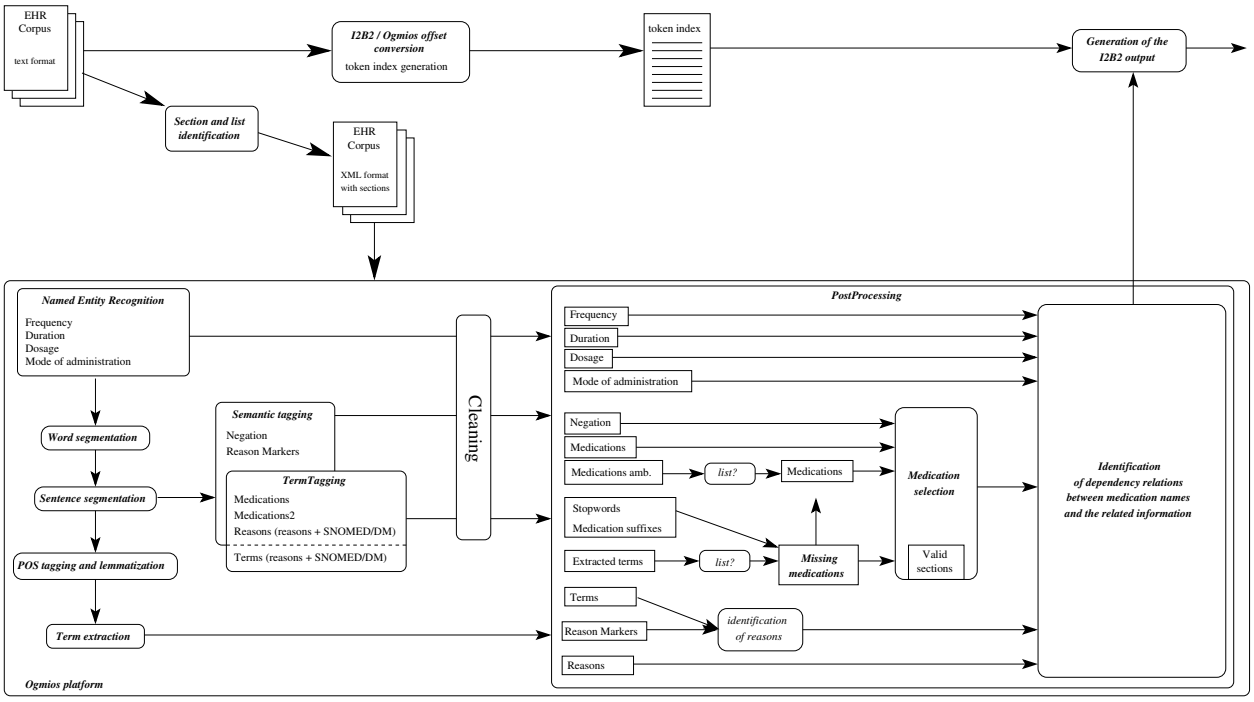

Figure 1: System architecture for extraction medication-related information and establishing dependencies among the annotations

we perform a three step processing of documents: pre-processing, processing and postprocessing.

Pre-processing step. During the pre-processing step, documents are converted into the XML Alvis format. This format is used for encoding information on document structure: document sections and lists are marked.

Processing step. We considered the challenge task as annotation and annotation selection problem. Thus, several annotations can be performed concurrently by existing modules (semantic and term tagging) or by newly created modules (such as Named Entity Recognition (NER)). The generated annotations are further disambiguated by specific modules before the postprocessing step. NER module is dedicated to the identification of information on frequency, dosage, duration and mode of administration. The related information can be found thanks to automata implemented as a set of regular expressions. Preliminary cleaning step is performed in order (1) to avoid multiple annotations within nested strings (*e.g.*, a frequency entity can be nested within dosage entity), and (2) to merge adjacent named entities of the same semantic type (medication name, dosage, ...). Term and semantic tagging modules are dedicated to the annotation of medication names and of reason terms, but also to the annotation of negation and reason markers. This step is based on linguistic information (word and sentence segmentation) and on previously performed named entity annotations and term tagging. Dedicated cleaning step manages some specificities of the I2B2 guidelines: nested terms, parenthesed medication names (e.g. *nitroglycerin 1/150 (0.4 mg)*, *singulair (montelukast)*), etc.

Post-processing step. Finally, a postprocessing step aims at correcting annotations and at selecting relevant information among the concurrent annotations according to the guidelines. It establishes also dependency relations between medication names and the related information. The use of the Ogmios platform is helpful for the token offset computing as it handles natively such information: once the annotations are performed and selected, the generation of the challenge output only requires an offset conversion.

Results and Discussion

The designed system has been applied to the training and test sets. All the documents (n=1,249) have been fully annotated, and their annotations submitted through three runs.

With the postprocessing step, our system performs an important work on disambiguation, cleaning and

especially on establishing dependency relations between medication names and the related information. We describe with more details and discuss some of these modules. In the following examples, medication names are underlined.

1. Ambiguous medication names. Some medications, blood or other products (*i.e.*, *iron* in the following examples) are ambiguous. In certain contexts they may correspond to blood or organism laboratory data:

> *Heme. Anemia workup. Iron 49, TIBC 256, B12 555, folate normal, ferritin 102, reticulocyte 7.9, and Epogen level 19.*

while in other contexts they mean the medication:

> *HOME MEDS: methadone 20 bid, imdur 120 bid, hydral taking 25 bid, lasix 20 bid, coumadin, colace, iron, nexium 40 bid, doxazosin 2 qd, allopurinol 100 qod*

In order to manage this kind of ambiguity, the system first assigns a specific tag and then retag it in medication name if it appears in lists, like in the last example.

2. Allergies, etc. According to the challenge guidelines, our system reject medication names occurring within `ALLERGY` sections. Because in this case, medications do not correspond to the prescription but to known allergies of a patient, like in this example:

> *ALLERGY: prednisone, penicillins, tamsulosin, simvastatin, spironolactone*

3. Negative contexts. The system systematically reject medication names appearing within negative contexts, like in this example: *did not require medications for abdominal pain*. We exploited this functionality also for other contexts where a medication name can appear but does not mean a medication prescribed:

> *INR's will be followed by coumadin clinic*
> *insulin-dependent diabetic*

4. Missing medication names. Our system has a module for identification of missing medication names, because we consider that our the medication list is not exhaustive. In order to identify new medications, discharge summaries have been additionally morpho-syntactically analyzed with Genia POS tagger [5] and noun phrases have been recognized by the term extractor YaTeA[6]. Then, new medication names are identified through specific semantic patterns, such as:

```
m do mo?  f
```

where dosage (`do`), frequency (`f`) and optional mode of administration (`mo`) are known, the system can infer that entity at the first position may be a medication name (`m`). We check further the nature of this entity: we verify whether it is a stopword and whether its ending is typical of the medication names endings. Thus, Pavachol, missing in our list of medications, could be recognized thanks to this module:

> *Diovan 160mg PO BID, HCTZ 25mg PO QD, Imdur ER 60mg PO QD, NTG .4mg PRN CP, Norvasc 10mg PO QD, Pavachol 80mg PO QD.*

5. Window's size. Two types of windows are considered by the system: (1) large window, defined according to the guidelines, contains two lines before and after the line containing a medication name; (2) restricted window takes into account occurrence of other medications as well as sentence and section ends. According to contexts, information associated to medications is searched within large or restricted windows.

6. Collection of frequency, dosage, duration and mode of administration. The system then collects medication-related information (frequency, dosage, duration, mode of administration) within context of each medication name. This information can appear before or after medication names. Besides, several elements of the same semantic type can be collected for each medication name.

7. Collection of reasons. Reasons can be collected with two kinds of strategies: (1) tagged terms which are part of the reason list, and (2) noun phrases which appear within a reason context. Notice that several reasons can be collected for each medication name.

8. Segmentation of medication names containing other types of information. Segmentation of medication names which contain also dosage or mode of administration information. In the cases of such nested information, it is split and each semantic type of information informs the corresponding categories. For instance, our medication name list contains the entry *Lisinopril 5 mg*, which is correctly recognized within the following prescription *Lisinopril 5 mg p.o. q. day*. But, according to the guidelines, it should be split into medication name *Lisinopril* and its dosage *5 mg*.

9. Coordination of medication names. When medication names appear within enumeration or coordination structures, such as in this example:

> *CV: 3VD s/p CABG x 3 in 2002; continued ASA, plavix, lisinopril, lopressor, statin*

during the processing step, they may be merged into the same unit because they share the same semantic type. In order to generate a more correct output,

such medication sequences are split on the punctuation marks by the current module.

10. Dependency relations between medication names and the related information. Identification of dependency relations between medication names and the related information, collected previously, is a core step for production of the challenge output. According to contexts, punctuation and ambiguities, the related information is associated with preceding or following medication names. Notice also that, according to the guidelines, medication name is the only required category together with its structural information (narrative or list section), which have to be informed. All other categories may remain empty. On the contrary, a given medication name can be associated with more than one series of the related information, which occurs when the dosage or frequency of administration change:

> *By the end of this hospitalization, the patient's INR was 1.7 on a dose of Coumadin 7 mg p.o. q.h.s. That dose was increased to 7.5 mg p.o. q.h.s.*

11. Generation of the I2B2 offsets. This module generates the I2B2 output by converting internal Ogmios offsets into the i2b2 offsets.

All these postprocessing modules constitute the core of the designed system for information extraction. Their creation has been conditioned by analysis of a large amount of documents from the training set. We decided to implement semantic resources and a rule-based approach, which was the most adequate solution: the training set contained very small number of the annotated documents: only 17 out of 696. The machine learning approach, which have to be trained on a large learning corpus, could not be applied by our team.

A positive result of our experience is that the obtained output performance proved to be stable within training and test set. Our system provide good results for four categories of the extracted information: medication names, their frequency, dosage and mode of administration. Detection of the structural information, list or narrative sections, is also efficient but seems not to be evaluated currently. As for the remaining two categories, duration and reasons, they are poorly annotated and extracted.

Conclusion and Perspectives

We presented in this paper the design of the system developed for the annotation and extraction of medication-related information from discharge summaries. We considered this task as annotation and annotation selection problem. We apply specific resources and a rule-based approach. Our system allows to perform concurrent annotations and then apply specific modules for the resolution of the ambiguities and specific cases. We developed also a module for detection of new medication names which are missing from the existing medication list. Our system provide good results for extraction of medication names, their frequency, dosage and mode of administration, while it performs poorly with durations and reasons.

One of the perspectives is the improvement of extraction of duration related information. For instance, we can search prepositional phrases (starting with words like *during*, *while*, etc.) in order to detect duration information. As for the improvement of reason related information, we can manage the extraction of extended noun phrases as reasons (including determinants) which would improve strict score of the evaluation. Otherwise, a specific reasoning module should be implemented for proposing knowledge-based associations between medical problems and medications, and further for a better extraction of reasons. Finally, this system has been developed for processing of discharge summaries in English. If we were to adapt this system to French language, it would require nearly as much time as we spent it for building the English system.

REFERENCES

1. Sibanda T and Uzuner O. Role of local context in de-identification of ungrammatical, fragmented test. In: Proceedings of the North American Chapter of Association for Computational Linguistics/Human Language Technology (NAACL-HLT 2006), New York, USA. 2006.
2. RxNorm, a standardized nomenclature for clinical drugs. Technical report, National Library of Medicine, Bethesda, Maryland, 2009. Available at *www.nlm.nih.gov/research/umls/rxnorm/docs/index.html*.
3. Côté RA, Brochu L, and Cabana L. SNOMED Internationale – Répertoire d'anatomie pathologique. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec, 1997.
4. Lussier YA, Rothwell DJ, and Cté RA. The SNOMED model : a knowledge source for the controlled terminology of the computerized patient record. *Methods in Informatics and Medicin (MIM)* 1998;37:161–4.
5. Tsuruoka Y, Tateishi Y, Kim JD, et al. Developing a robust part-of-speech tagger for biomedical text. *LNCS* 2005;3746:382–92.
6. Aubin S and Hamon T. Improving term extraction with terminological resources. In: Salakoski T, Ginter F, Pyysalo S, and Pahikkala T, eds, Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006), number 4139 in LNAI. Springer, August 2006:380–7.