Term variation and semantics for document classification and detection of obesity and its co-morbidities cases

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

[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]LIPN – UMR 7030, Université Paris 13 – CNRS, 99 av. J-B Clément, F-93430 Villetaneuse, France

Abstract

Within the translational medicine area, researchers aim at connecting more closely medical knowledge and individual patient data, in order to improve patient care quality. The 2008 I2B2 Obesity NLP challenge is to promote such research and to evaluate such systems within the same framework. We participated at this challenge. Our approach bets on use of semantics and term variants. The results obtained by our system are superior to the mean challenge results and seem to be promising, although there is room for improvements which we present and discuss in this paper.

Introduction

Recent developments in the medical area aim at connecting more directly scientific research and individual patient care. Thus, in the case of biomedical research or drug development, the translational medicine branch refers for instance to a better exploiting of the patient clinical documents and taking into account various parameters, specific to patient cases and possibly relevant for a given clinical case. Among these parameters let's cite risk factors, co-morbidities, medicine and chemicals, environment, etc. The very objective of the translational medicine is to improve the quality of individual health care.

The I2B2[a] (Informatics for Integrating Biology and the Bedside) Center is organizing yearly NLP challenges in order to propose common testbed for various tasks related to translational medicine. The goal of these challenges is to evaluate the participating systems on their ability to recognize whether a clinical documents (and the patient) is relevant to the questions asked. The 2008 NLP challenge is focused on the detection of obesity and its co-morbidities: systems have to detect who is obese and what co-morbidities they exhibit. An additional difficulty is related to the fact whether these co-morbidities are likely or definitely acquired. 28 systems, of which our system, participated at the Obesity challenge. Our results are superior to the mean challenge results and several improvements are still possible. In this paper, we describe our approach, compare its efficiency to other systems and discuss them, and point out some possible improvements.

Material

We used three types of material: discharge summaries to be processed, terminologies for annotation of these clinical documents and negation markers for definition of the status (negative, positive) of annotations.

Discharge summaries and Gold standard

The challenge data consists of discharge summaries from Partners Healthcare. They are written in English. All records have been fully de-identified.[1] Obesity and co-morbidities information have been marked at a document level as present, absent, questionable, or unmentioned in the documents. If a document is not annotated by a pathology, it is non relevant for this pathology. For each patient, both textual judgments, *i.e.*, what the text explicitly states about obesity and co-morbidities, and intuitive judgments, *i.e.*, what the text implies about obesity and co-morbidities, are provided. A total of 1233 documents have been used in this challenge, split into training and test sets. The training set is composed of 730 discharge summaries and the test set is composed of 507 discharge summaries. All documents have been manually annotated by two medical experts, and discussed until an agreement is reached. Annotation provided by professionals is the Gold standard used for evaluation of systems.

Terminologies

We used three types of resources for annotating discharge summaries (a total of 114'448 entries): (1) List of the 16 pathologies aimed in this challenge: *obesity* and 15 co-morbidities (*asthma*, *CAD*, *CHF*, *depression*, *diabetes*, *GERD*, *gallstones*, *gout*, *hypercholesterolemia*, *hypertension*, *hypertriglyceridemia*, *OA*, *OSA*, *PVD* and *venous insufficiency*). (2) 108'531 terms from six axes (A, D, F, M, P and T) of the Snomed International.[2] For these terms, each of them is associated to its semantic type as defined by the Snomed International. For instance, *decreased thickness* and *obstructive sleep apnea* are Snomed terms

[a] www.i2b2.org

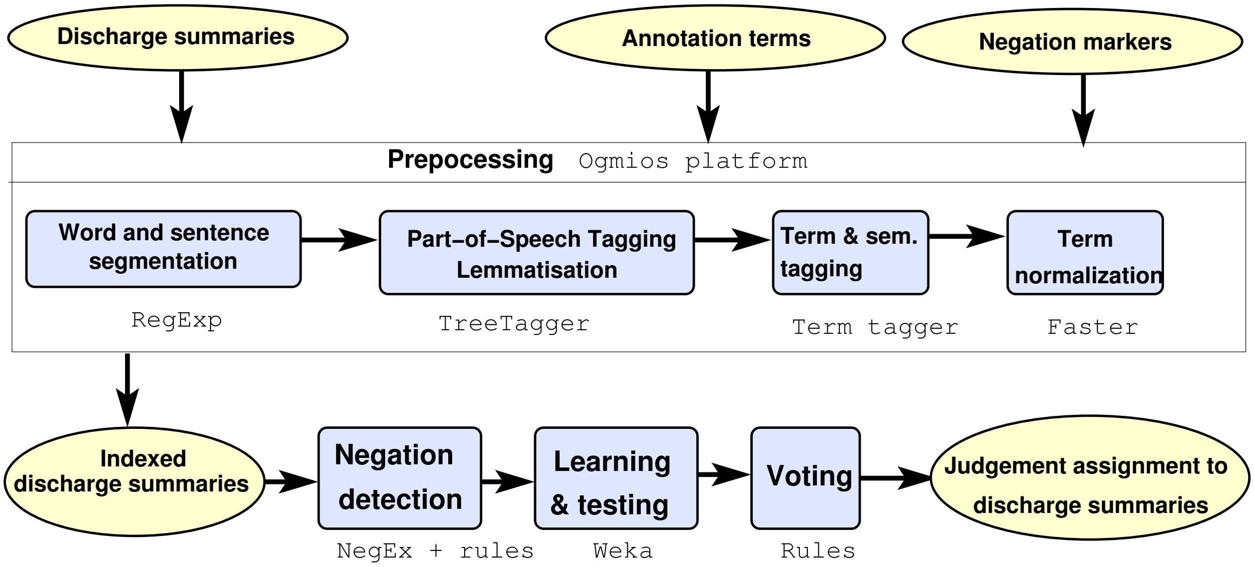

Figure 1: General scheme of the approach.

belonging respectively to Morphology (M) and Diagnostics (D) axes. (3) List of 5’716 International Nonproprietary Names (INNs) of medicine. They are extracted from the Thériaque resource.[b] Use of drug names is motivated by the fact that the prescriptions in discharge summaries are done with INNs. Each INN is associated with its therapeutic class as defined by Thériaque. For instance, *Allopurinol* is an antigout drug.

Negation

For the detection of negation, our main source is the NegEx resource available on-line.[c] Few additional markers have been extracted from the discharge summaries. Among these markers, pre and post-negation is distinguished as well as diminishing markers which decrease the scope of negation phrases. We use a total of 179 markers.

Methods

Figure 1 presents the general scheme of our approach. Five main steps are distinguished: preprocessing of discharge summaries, indexing of documents, detection of negation, categorization of documents with learning algorithms and voting of these algorithms.

Document preprocessing and indexing

The aim of preprocessing of clinical documents is to index them with terminological resources and to detect variants of their terms. We use the Ogmios platform,[d] suitable for the processing of large amounts of data and tunable to specialized areas. Through the platform, we perform: (1) segmentation of documents into words and sentences; (2) POS-tagging and lemmatization;[3] (3) term tagging;[e] (4) detection of term variants.[4] Steps 2 to 4 allow to detect noun phrases (possible terms), to normalize them through the lemmatization and variant detection and thus to index discharge summaries. In this way, to each processed document is associated a set of entities (challenge co-morbidities, Snomed International terms, INNs and negation markers). These entities are sorted in chronological order within sentences.

Detection of negation

Negation is detected within each sentence. The algorithm is close enough to the one previously proposed:[5] detected entities (terms and INNs) are characterized by pre or post-negation markers. The scope of these markers can be delimited by diminishing markers. In this way, each indexing entity can have positive or negative status.

Categorization of documents

We use algorithms proposed by the WEKA[6] platform. We describe various sets of features used, the learning algorithms exploited and models we decided to apply.

[b] www.theriaque.org

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

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

[e] http://search.cpan.org/∼thhamon/Alvis-TermTagger/

Feature definition. Three sets of features are used. They all are provided by the indexing step of the method (no feature selection is performed):

- T All the indexing terms (Weka algorithms use 4'165 of them) are taken into account for the generation of models. Values of attributes are the states of terms depending on whether they occur in positive (*y*) or negative (*n*) contexts, or whether they don't occur (*0*) at all within documents.
- TS Instead of terms, their semantic types are taken into account (440 attributes). In this case, value of these attributes is numeric, it corresponds to the frequency of semantic types (cumulated frequencies of the corresponding terms).
- TST In this set, we use medical terms (i2b2 categories and Snomed International terms) and semantic types (therapeutic classes) of INNs. This gives us a total of 3'048 attributes. Values of term attributes are their status (*y, n, 0*), values of semantic types are their frequencies.

Learning algorithms. We take advantage of the algorithms implemented within the Weka platform and decide to use several of them. They belong to four different families (names of algorithms are those mentioned by Weka): trees (`J48`, `REPTree` and `RandomForest`), bayes (`NaiveBayes`), functions (`SMO`) and rules (`OneR` and `DecisionTable`). `J48` algorithm, which is an implementation of the `C4.5` classification algorithm,[7] provides the decision trees, which allow to perform their qualitative analysis.

Tested models. We distinguished five models for classification of documents and their assignment to a given category (in our approach, all the processed summaries are analyzed by all these models):

- y (*yes* model) tries to detect whether documents should be annotated by a given co-morbidity. For creation of this model, summaries with Y statement are contrasted with the remaining documents.
- n (*no* model) tries to detect whether documents should be not annotated by a given co-morbidity. Summaries with N statement are contrasted with other documents
- q (*questionable* model) tries to detect whether a patient state is questionable according to a given comorbidity. Summaries with Q statement are contrasted with other documents
- i (*irrelevant* model) tries to detect whether documents should not be taken into account for a given co-morbidity. Not annotated summaries are contrasted with all the annotated documents.
- yn (*yes/no* model) tries to decide between Y and N annotations.

Voting

Voting is performed at document level through two steps: decision per algorithm and final decision for all the algorithms used. During the first step, for each document, we consider results provided by different models of each algorithm. If document is classified by only one model, this class, whatever it is, is recorded for this document. If document is classified as both *y* and *n*, the decision is made according to the *yn* model. Otherwise, the document class selection in based on the priority of models (*y*, *n*, *q*, *i*). During the voting step, in the simplest case, the most frequent class provided by various algorithms, is recorded as the final decision. If vote counts of *y* and *n* classes are equal and if at least one algorithm classifies it as questionable *q*, the analyzed document is classified *q*. If no *q* class is found, the final decision is based on the *yn* model and the ratio between *y* and *n* classes attributed. If none of the algorithms provide classification, document is classified as irrelevant *i*.

Evaluation

Evaluation of the results is performed by organizers of the challenge according to three measures: precision, recall and F-measure, in their micro and macro versions. Macro F-measure is the main measure for rating the results of all the participating systems.

Results and Discussion

All the participants had three months for tuning their systems for the Obesity challenge: time between the release of training and test data. During the test step, 507 discharge summaries from the test set have been processed through the `Ogmios` platform. We needed for this 53 min 19 sec, which gives 6.3 secs per summary. Three runs have been submitted to the challenge: voting$_{TST}$ (medical terms and semantic types for INNs with the voting of algorithms), voting$_{T}$ (all the annotation entities (medical terms and INNs and the voting of algorithms) and RepTree$_{T}$ (all the annotation entities and classification by the RepTree algorithm). These runs have been submitted to the intuitive judgment evaluation: they all are provided by machine learning algorithms and are likely to be based on non explicit data. The second run voting$_{T}$ shows our best results submitted. Surprisingly, the voting$_{TST}$ run, which uses more extensively the semantic information, showed lower results.

Table 1 presents performances of our system as compared to results obtained by other systems during this

	MiP	MaP	MiR	MaR	MiF	MaF
Our system	0.92	0.62	0.91	0.61	0.91	0.61
Mean	0.91	0.78	0.90	0.60	0.90	0.60
Std Dev	0.09	0.17	0.09	0.06	0.09	0.06

Table 1: All diseases, intuitive.

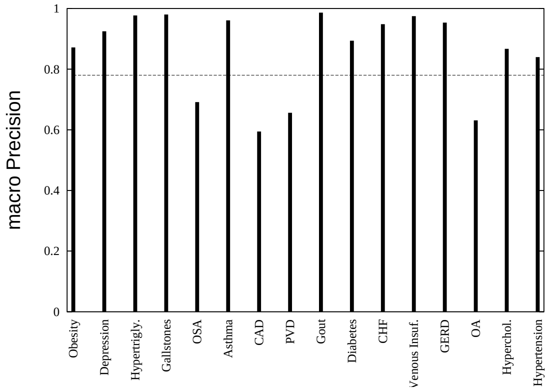

Figure 2: Macro precision values by disease, intuitive.

challenge. As we noticed, three measures (precision, recall and F-measure) are used in their micro and macro versions. According to micro values, our system shows performances slightly superior to the challenge mean performances. Situation is different when analyzing macro values, especially for macro precision, although recall and F-measure are still slightly superior to the mean performances. Thus, macro precision, which values have the most important standard deviation, shows to be one of lowest with our system. This means that some categories (diseases), with a small number of discharge summaries, are difficult to recognize: they have small effect on micro values, but their effect is more important on macro values. Since our macro F-measure still remains superior to the mean challenge values, we assume this task was difficult for other systems as well. As the macro precision is the weak point of our system, we will have a closer look at our performances by disease.

Analysis of results by disease

Figure 2 presents performances obtained by our system for each of the 16 co-morbidities. The horizontal line corresponds to the mean macro precision of the challenge, for all diseases and intuitive judgment. For this illustration, we use the *yn* models. Four comorbidities (OSA, CAD, PVD and OA) show performances inferior to the mean values. Distribution of documents among the classes (in both training and test sets) is homogeneous, therefore we assume the failures are due to the content of documents and our ability to process it. Let's analyze some decision trees built by the `J48` algorithm in order to get closer to what we missed to detect. We cannot present and analyze all the decision trees, thus we concentrate on two diseases: *gout* as a good example of using the semantic information, and *OSA* which has low macro precision.

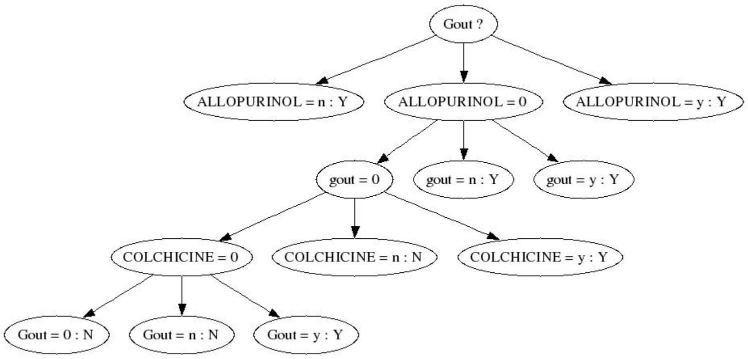

Figure 3: `J48` decision tree for *Gout*.

The *gout* decision tree (figure 3) is composed of only nine leaves, which appears to be sufficient for a successful classification: macro precision (and other measures) are high for this disease. It is based on both medical and drug information: diagnosis *gout* and two drugs (*Allopurinol* and *Colchicine*) are involved. Both drugs are antigout chemicals which explains their suitability and successful use for this disease. The `J48` trees present the decision information: according to the attribute value (*y, n, 0*) each leave either allows to make a decision (Y: annotate by the disease, N: do not annotate by the disease), or introduces the further decision level. For instance, on figure 3, the first level says that if *Allopurinol* drug appears in positive or negative contexts, the patient is concerned with *gout*, otherwise `J48` must observe the attribute *gout*, etc.

Now, let's analyze the *OSA* (*obstructive sleep apnea*) decision tree (figure 4). It is composed of 11 leaves: functions (*sleep apnea*, *pain*, *apnea*) and diagnoses (*OSA*, *obstructive sleep apnea*). No relevant drug information is detected. *Pain* is a general medical term and may be not the most relevant for this disease. The remaining terms are directly related to the OSA comorbidity. We notice that the context (*y, n*) has not always effect on the decision: positive and negative status of *OSA*, *obstructive sleep apnea* and *pain* provide identical annotations. This may point out that either negation is not correctly assigned to these terms or these terms occurrence is important by itself. As for the node *apnea*, its appearance in negative context or its missing have the same effect on annotation (not as-

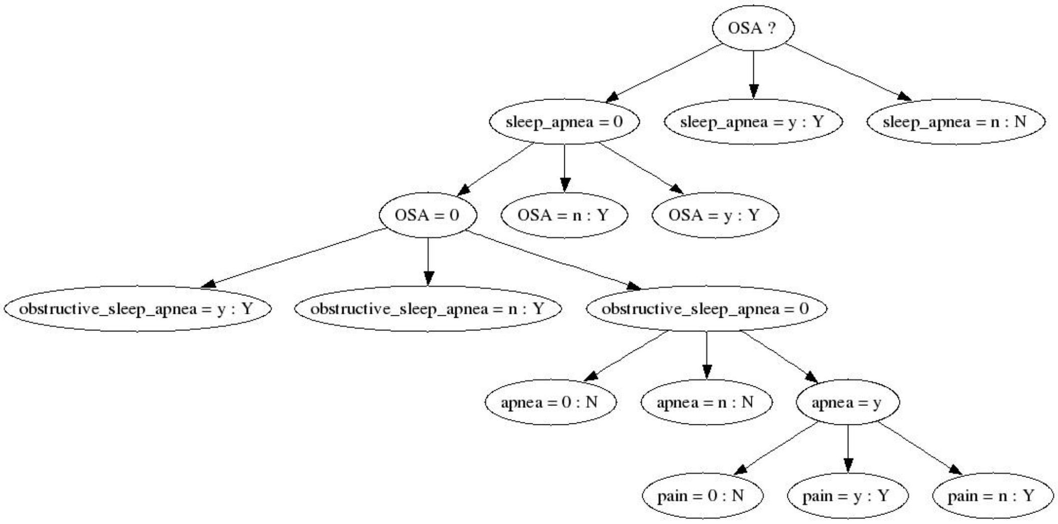

Figure 4: `J48` decision tree for *OSA*.

signment of *OSA*), and this seems to be medically relevant. The main limitation of this decision tree remains in the fact that lexical (and conceptual) extend of the decision space is limited to *apnea* and its lexically subsumed terms. We assume we failed to recognize other relevant concepts to this disease.

Conclusion and Perspectives

In this paper, we described the system we used in the 2008 I2B2 Obesity challenge and analyzed some of the results obtained. Through our system, designed and created for this challenge, we aim at extracting and exploiting the linguistic and semantic information (term variation and semantics, negation, ...). The global results obtained show to be slightly superior to the mean challenge results. This leaves us room for a more deep evaluation and improvement of our system. Thus we will analyze why the semantic information does not seem to be efficient for classification. We plan also to perform a detailed evaluation of the negation detection and possibly to extend the used set of negation markers. As noticed, we did not perform the feature selection, while it can be interesting (through the use of semantic information from terminologies or the statistic selection algorithms). Otherwise, use of the Weka platform may be relevant for designing the classification system (our current situation); as for its further development we plan to use other implementations of the learning algorithms. This may be suitable for their tuning and evaluation: within Weka only the micro precision values are available. Finally, we will implement more tools[8] and resourcesfor the detection of term variation and exploit the structure of the discharge summaries. Another interest is concerned by the certainty statement within clinical literature.

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. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, and Brochu L. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. College of American Pathologists, Northfield, 1993.
3. Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK. 1994:44–9.
4. Jacquemin C and Tzoukermann E. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In: Strzalkowski T, ed, *Natural Language Processing and Information Retrieval*. Kluwer, Boston, MA, 1999:25–74.
5. Chapman W, Bridewell W, Hanbury P, Cooper G, and Buchanan B. Evaluation of negation phrases in narrative clinical reports. In: Annual Symposium of the American Medical Informatics Association (AMIA), Washington. 2001.
6. Witten I and Frank E. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.
7. Quinlan J. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1992.
8. Hamon T, Nazarenko A, and Gros C. A step towards the detection of semantic variants of terms in technical documents. In: International Conference on Computational Linguistics (COLING-ACL'98), Université de Montréal, Montréal, Quebec, Canada. 1998:498–504.