Linguistic and semantic annotation for information extraction and characterization

Thierry Hamon[1], PhD, Amandine Périnet[1], MA, Jérôme Nobécourt[1], PhD, Natalia Grabar[2,3], PhD

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

[2]Centre de Recherche des Cordeliers; Université Paris Descartes, UMR_S 872, Paris, F-75006; INSERM, U872, Paris, F-75006 France

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

Abstract

The 2010 I2B2 NLP challenge concentrated on extraction of three types of information from clinical records: medical concepts, their certainty status and relations between them. For participation in this challenge, we designed an automatic NLP system exploiting terminological resources and a rule-based approach. An attemp was also made to apply knowledge engineering methods. Our system provides modest results for extraction of concepts and relations, while the extraction of the information status appears to be satisfying.

INTRODUCTION

Electronic health records (EHR) contain huge amount of medical information on patients: health condition, illness history, family and social history, medical observations, treatments and procedures, etc. The whole clinical process is a complex picture with a rich network of entities and relations. The objective of the 2010 I2B2 NLP challenge was to tackle this picture and to make it explicit, structured and thus more easily accessible to computer programmes and human users.

The 2010 challenge[a] addressed three tasks: (1) Extraction of medical concepts and their typology as medical problems, laboratory or imaging tests, and treatments. (2) Extraction of information on the assertion status of medical problems. According to the challenge guidelines, the assertion status may be one of the following: *present*, *absent*, *possible*, *conditional*, *hypothetical* or *associated with someone else* but not the patient. (3) Extraction of relations existing around medical problems: between medical problems and treatments; medical problems and tests; and medical problems and other medical problems. In each case, several types of relations may exist. For instance, a treatment may improve, worsen or cause a medical problem. It may be prescribed or not prescribed because of a health condition. From a methodological point of view, we took advantage of our previous work and participation in these challenges[1,2] and to exploit linguistic and semantic annotations and a rule-based system in order to extract the aimed information. The main competence upon which our participation was based comes from the Natural Language Processing (NLP) field. An attemp was also made to use knowledge representation and engineering skills. Besides, the only medical knowledge exploited was the one provided by challenge annotations from EHRs of the training set.

Previous research work has already addressed the problem of relation extraction between entities. First of the kind were the tasks proposed during the MUC conferences[3] which aimed mainly on extraction of structured information from financial and security corpora. Then followed other challenges on information extraction such as ACE[4], Protein-Protein Interaction task of BioCreative[b] or the shared task on biomedical event extraction[5]. These provide extensive platforms for the tuning and evaluating of systems designed for processing of biomedical scientific literature. Extraction of information and of complex events from clinical records have been less studied by researchers, but have been addressed thanks to tools like MedLEE[6] or Medsyndikate[7]. More recently, the UK project CLEF (Clinical E-Science Framework)[8] aims at extraction, integration and presentation of clinical information. In this project, a special task has been planned for extraction of complex medical events[9] from clinical discharges. Besides, one of the tasks of the 2010 challenge is concerned by extraction of certainty associated with medical concepts: this area presents an extensive previous research work[10,11,12] and has also been addressed by the shared task on event extraction[5].

The I2B2 challenge follows similar objectives and provides a framework for tuning and evaluation of systems for extraction of complex clinical events: it aims at extraction of clinical concepts, their certainty and the relations among them. It seems that the complexity and exhaustivity of the aimed information are particularly complete and difficult in this challenge.

MATERIAL

Discharge summaries

826 discharge summaries have been devided into train-

[a]https://www.i2b2.org/NLP/Relations

[b]biocreative.sourceforge.net/biocreative_2_ppi.html

ing (n=349) and test (n=477) sets. Training set was made available for a duration of three months, while the test set had to be processed during three days, *i.e.* one task per day had to be achieved and the results submitted.

Terminological and semantic resources

We prepared and used terminological and semantic resources specific for each task (concepts, assertions and relations). For concept tagging, we exploited three main sources. 316,368 terms from the UMLS[13] which belong to several semantic axes and were helpful in annotation of (1) medical problems (B2.2.1.2.1 *Disease or Syndrome*, A2.2.2 *Sign or Symptom*, B2.3 *Injury and Poisoning*, A1.2.2 *Abnormality* and A1.1.5 *Bacteries*), (2) tests (B1.3.1.1 *Diagnostic procedures* and B1.3.1.2 *Laboratory procedures*) and (3) treatments (B1.3.1.3 *Therapeutic or prevention procedures*). We also exploit 243,869 entries from RxNorm[14] used for detection of medication names (treatments). And finally 6,069, 2,608 and 3,697 I2B2 annotation entries are used for annotation of problems, tests and treatments respectively. Notice that the recovery between I2B2 concepts and UMLS/RxNorm entries is very low and does not exceed 3,000 entities.

Morphological, lexical, contextual and structural markers

We exploit morphological, lexical, contextual and structural markers. Morphological markers are substrings within the analyzed concepts, lexical markers occur as strings within them, while contextual markers occur around the concepts. Structural markers correspond to section names within the discharge summaries.

In order to define types of medical concepts, in addition to the resources, a total of 450 morphological markers are used for detection of problems (n=126), tests (n=118) and treatments (n=206). Morphological markers state for instance that a noun phrase is a treatment if it contains substrings like *-stomy*, *-plasty*, *-ectomy* or *-lysis*, or strings like *implant*, *repair*, *care* or *drug*. For detection of medical problems, adjectives are also used as morphological markers: thus, adjectives like *benign*, *persistent*, *metastatic* or *chronic* indicate that a noun phrase is a medical problem. Among morphological markers we distinguish prefixes and suffixes, and other substrings which may appear in any position. We exploit a total of 181 contextual markers (91 for problems, 19 for tests and 71 for treatments). According to these markers, if a noun phrase is preceded by contexts such as *absence of*, *complaint of*, *due to*, *show* or *worrisome for*, this noun phrase is categorized as a medical problem. Thus, in the sentence *it is noted there was nodularity in the common bile duct which is worrisome for cholangiocarcinoma*, the underlined context *worrisome for* indicates that the following noun phrase *cholangiocarcinoma* is a medical problem. As for structural markers, section names such as *Laboratory data on admission* or *Procedure* indicate that noun phrases occurring in these sections may belong to the test category.

A set of normality markers has also been defined. It includes entries like *regular*, *normal* or *good*. These markers appear for instance in this kind of medical problems: *good condition*, *regular rate and rhythm*, *good bowel sounds*. They mean that health condition is good or that an examination provides normal observations. In this case, health observations are not considered as medical problems.

For the detection of assertion, we exploited (1) NegEx[c] resource for negation; (2) and additional resources for detection of other types of assertion. These resources contain sets of lexical, morphological, contextual and structural markers. Among the lexical markers, here are some examples: *questionable* for possible, *on palpation* for conditional. Among the morphological markers, we have markers such as *un-* and *a-*, stating for the absent assertion, like in this example: *She is afebrile with stable vital signs*. As for the structural markers, section name *Family history* indicates that a problem occuring in this section is certainly not relevant to the patient but is associated with someone else, such as in this example: *FAMILY HISTORY: The patient's father died of a myocardial infarction at age 40, and a brother who died of a myocardial infarction at age 40.* Two lists of markers are used: complete (n=342) and reduced (n=227). The reduced list contains only those markers that are not ambiguous among the assertion categories. For instance, *may* and *should* markers are ambiguous between possible and hypothetical categories.

We defined also contextual rules (n=615) for the detection and typology of relations. They belong to one of the eight relations addressed in the challenge: (1) PIP (n=118), where a medical problem causes other medical problems; (2) TeCP (n=106), where a test is done in order to investigate a medical problem but the outcome is not known; (3) TeRP (n=74), where a test reveals a medical problem; (4) TrAP (n=131), where a treatment is administered for a medical problem; (5) TrCP (n=85), where a treatment causes a medical

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

problem; (6) TrIP (n=31), where a treatment improves a medical problem; (7) TrNAP (n=43), where a treatment is not administered due to a medical problem; and (8) TrWP (n=27), where a treatment worsens or not improves a medical problem. For detection of PIP relations, patterns such as *PB as a cause of PB* or *PB responsible for PB* are used. Detection of TeCP relations is performed with patterns such as *TE performed due to PB* or *PB shown after TE*. While within TeRP relations, the relation between tests and medical problems is explicit: *TE show PB*, *TE is notable for PB*, *evidence of PB on TE*.

METHODS

Pre-processing of discharge summaries

During the pre-processing step, documents are converted into the XML format, required by the annotation platform[15]. This format is used for encoding of information on document structure: document sections and lists are thus marked.

Linguistic processing of discharge summaries

The challenge tasks are tackled though the linguistic annotation and application of contextual rules. Annotation with several resources are performed concurrently by various modules (semantic and term tagging, Named Entity Recognition...). We also apply POS-tagging[16] and term/noun phrase extractor[17].

Rule-based approach for extraction and typology of entities and relations

After the annotation of discharge summaries with resources, markers and syntactic analysis (detection of noun phrases), a rule-based system is applied. The designed rules are responsible for typology and disambiguation of concepts, assertions and relations. We describe here some of these rules.

Together with the term extraction, we apply also statistical methods for their filtering, such as *iTer*[18] or *NC-Value*[19]. Such measures take into account frequency of noun phrases, their length, frequency and number of the larger including noun phrases. We combined these measures with other indicators, among which (1) noun phrase is not included in larger noun phrases, (2) it is not a stop word (grammatical or lexical), (3) it does not contain a normality marker, or (3) it does not occur in titles (except for tests). A threshold, specific for each measure, is then set. If a noun phrase's weight exceeds this threashold, this noun phrase is not considered as medical concept and is not proposed as relevant.

For automatically extracted noun phrases, definition of their category (test, treatment, problem) is based upon morphological, lexical, contextual and structural markers. If a noun phrase is ambiguous, its final category corresponds to the category which shows the maximal weight according to: $weight(cat_i) = \sum_{j \in \{lex,adj,mor,m-p,m-s\}} \alpha_j \times freq_j$, where the weights assigned to markers are: $\alpha_{lex} = 1$ (lexical markers), $\alpha_{adj} = 0.8$ (adjectival markers), $\alpha_{m-p} = 0.5$ (prefixes), $\alpha_{m-s} = 0.5$ (suffixes), and $\alpha_{mor} = 0.25$ (other morphological markers). $freq_j$ corresponds to the frequency of a noun phrase. Structural markers are also used for the categorization of noun phrases. For instance section named *Preoperative laboratory studies* may contain laboratory tests and results; and *Allergies/sensitivities* section may contain tratments and possibly medical problems.

Decision on assertions is based upon the exploitation of the lists of markers complete or reduced, upon the section names, contextual patterns and weighting of the markers. If ambuguous, weights of assertion subcategories are computed according to a formula similar to concept categorization (see the paragraph above). Patterns are used for a more contextual management of assertion markers. Thus, in the sentence *two ct scans had shown a question of an enlargement of the head of his pancreas*, assertion is processed through a pattern where *question* marker is related to the possible category through the verb *show*. Besides, an assertion order is established according to the challenge specifications: present is the default value, it is followed by absent, then by possible, conditional, hypothetical and non associated with the patient. Finally, within the same list, assertions may be propagated to other elements for which no assertion has been detected yet.

Detection of relations is mainly based upon contextual markers, but also on a knowledge base and knowledge engineering-inspired rules. The knowledge base contains 1,040 known associations between tests, problems and treatments. These associations have been extracted from the annotated training set data. For instance, *bradycardia* may be treated with *atropine*, *dopamine*, *iv fluids*, *pacemaker* or *pacer*, which means that a TrAP relation may be established between such entities. Another example is that medication *fentanyl* may cause medical problems such as *acidosis*, *jaw clenching*, *rigidity* or *tongue biting*, and the TrCP relation can thus be established when they occur in the same sentence. As for the TrNAP relation, medications like *bactrim* and *diovan* should not be prescribed in presence of *acute renal failure*. Concerning the knowledge engineering approach, it takes into account data such as the status of the medical problem (namely,

	Training set			Test set		
Runs	P	R	F	P	R	F
Con R1	0.761	0.675	0.715	0.654	0.553	0.599
Cla R1	0.741	0.657	0.696	0.629	0.533	0.577
Con R2	0.512	0.304	0.382	0.523	0.280	0.364
Cla R2	0.403	0.240	0.301	0.414	0.221	0.288
Con R3	0.672	0.699	0.685	0.590	0.604	0.597
Cla R3	0.646	0.672	0.659	0.569	0.555	0.562
Ast R1	0.824	0.824	0.824	0.800	0.800	0.800
Ast R2	0.763	0.763	0.763	0.777	0.777	0.777
Asr R3	0.799	0.799	0.799	0.800	0.800	0.800
Rel R1	0.524	0.486	0.504	0.520	0.444	0.479
Rel R2	0.508	0.492	0.500	0.501	0.445	0.471
Rel R3	0.508	0.492	0.500	0.501	0.445	0.471

Table 1: Precision, recall and F-measure figures obtained with training and test sets, exact match.

its assertion), punctuation (such as coordination), relative positions of the two possibly related entities.

RESULTS AND DISCUSSION

The results we obtained for the three challenge tasks are indicated in table 1. The designed rule-based system seems to perform not as well as machine learning systems did in this challenge. We observe that the variation between training and test results is small, although test results are usually lower that training results, except for assertion runs. We detail here our best runs: (1) ConR1 (detection of concepts) and ClaR1 (categorization of concepts) consisted in exploitation of terms from existing terminologies. Its combination with term extraction (R3) decreases slightly the results, while the use of term extraction alone (R2) appears to be catastrophic. (2) AstR1 consisted in exploitation of the complete list of markers. Use of reduced list, as it is done in other runs, decreases the results. and (3) RelR1 relies upon the exploitation of relations markers and patterns. Its combination with the knowledge base (R2) improves the recall but drops the precision figures, while application of knowledge engineering methods (R3) has no impact on results. These results seem to indicate especially that our knowledge engineering methods or term extraction tools show currently feeble results, and that an additional research is necessary to adjust them to the aimed tasks.

Figure 1 presents more detailed results for the best runs for each task: we can observe performances for each sub-category. Within the ClaR1, the three concept types show close performances. While there is more variation between assertions or relations subcategories. Thus assertion sub-categories present, absent and associated with someone else are well recognized, while the conditional sub-category remains difficult to detect. The picture is even more difficult with relation sub-categories: results are acceptable with TeRP, TrAP, PIP and TrCP relationships, but are feeble with TrIP, TrWP, TeCP and especially TrNAP relationships. An additional work and possibly exploitation of other methods are necessary for a better processing of clinical dischange summaries.

Application of such methods to clinical documents and extraction of structured information from them can be used further for a graphical representation of medical history of patients, as it was done within the CLEF project[20]. Thus, figure 2 proposes such a representation for record-17 from the training set. Graphical conventions are indicated in the caption of the figure. For this record, we obtain one complex graph and several simple graphs. Complex graph describes detection and treatment of *COPD* and of the related diagnoses. Simple graphs are related to other medical problems and also to *COPD*. Distribution of management of the same diseases within several graphs is due to the fact that, at this point, there is no co-reference among information extracted from different sentences: problems like *copd exacerbation* and *his copd exacerbation* provided by different sentences remain isolated and not linked between them. Notice that in the majority of discharge summaries mainly simple and disconnected graphs are generated. For a more complete picture and for the detection of inter-sentence relations, an additional work is necessary. Besides, addition of temporal relations between events would contribute to a better representation of data. Additionally, positionning of the medical events according to human anatomy may also be interesting.

REFERENCES

1. Grabar N and Hamon T. Impact des informations smantiques sur la catgorisation automatique des documents cliniques. In: TIA 2009, 2009.
2. Hamon T and Grabar N. Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc* 2010;17(5):549–.
3. Grishman R and Sundheim B. Message understanding conference. A brief history. In: COLING, 1996:466–71.
4. Doddington G, Mitchell A, Przybocki M, et al. The automatic content extraction (ace) program - tasks, data, evaluation. In: Proc LREC, 2004:837–40.
5. Tsujii J, ed. *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*. ACL, Boulder, Colorado, June 2009.
6. Friedman C, Alderson PO, Austin JH, Cimino JJ, and Johnson SB. A general natural-language text processor for clinical radiology. *JAMIA* 1994;1(2):161–74.

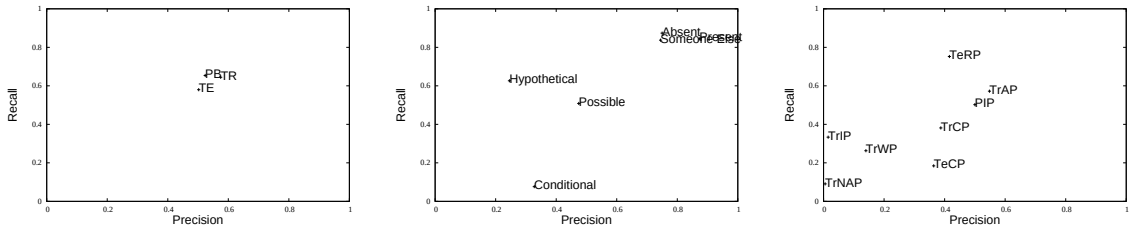

Figure 1: Results for each sub-category of the best runs (ClaR1, AstR1 and RelR1), exact match.

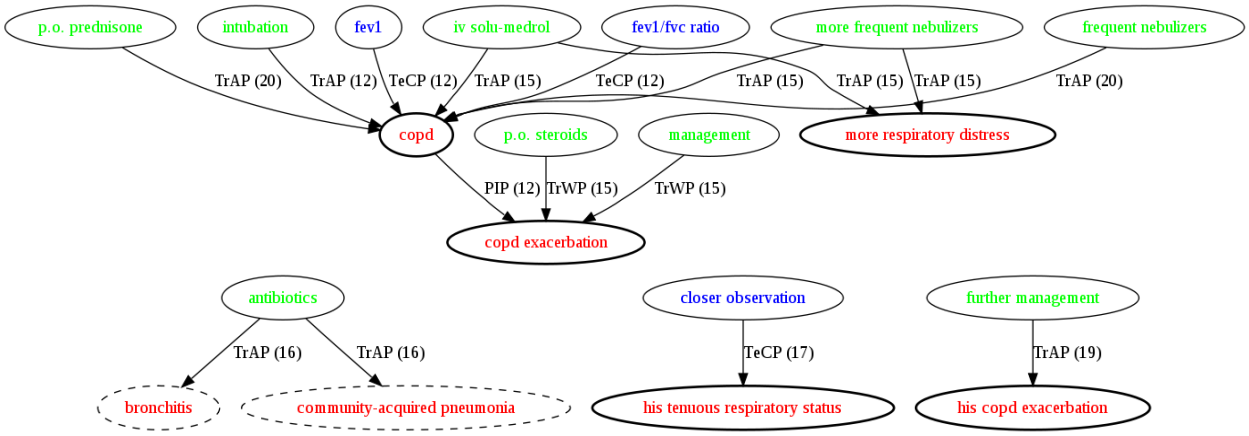

Figure 2: Network of semantic relations between medical problems (red), tests (blue) and treatments (green) as extracted from the record-17 of the training set. Concepts are encircled, arrows are labelled with the relations (between parentheses we indicate number of sentence where relations occur). Bold circles indicate present concepts, and dashed circles indicate possible, hypothetical or conditional concepts.

7. Hahn U, Romacker M, and Schulz S. Medsyndikate–a natural language system for the extraction of medical information from findings reports. *Int J Med Inform* 2002;67(1-3):63–74.

8. Rector A, Rogers J, Taweel A, et al. Clef: joining up healthcare with clinical and post-genomic research. In: Proc of UK e-Science All Hands Meeting, 2003:264–7.

9. Roberts A, Gaizauskas R, Hepple M, and Guo Y. Mining clinical relationships from patient narratives. *BMC Bioinformatics* 2008;9(11):3–.

10. Hyland K. The author in the text : Hedging scientific writing. *Hong Kong papers in linguistics and language teaching* 1995;18:33–42.

11. Light M, Qiu XY, and Srinivasan P. The language of bioscience: facts, speculations and statements in between. In: ACL WS on Linking biological literature, ontologies and databases, 2004:17–24.

12. Mercer RE, Marco CD, and Kroon FW. The frequency of hedging cues in citation contexts in scientific writing. In: in Computer Science LN, ed, CSCSI. Springer Berlin, 2004:75–88.

13. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 2004;32:267–70.

14. 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*.

15. Hamon T, Nazarenko A, Poibeau T, Aubin S, and Derivière J. A robust linguistic platform for efficient and domain specific web content analysis. In: RIAO 2007, Pittsburgh, USA. 2007.

16. Tsuruoka Y, Tateishi Y, Kim JD, et al. Developing a robust part-of-speech tagger for biomedical text. *LNCS* 2005;3746:382–92.

17. Aubin S and Hamon T. Improving term extraction with terminological resources. In: Salakoski T, Ginter F, Pyysalo S, and Pahikkala T, eds, FinTAL 2006, number 4139 in LNAI. Springer, August 2006:380–7.

18. Drouin P. *Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés*. Phd thesis, Université de Montréal, 2002.

19. Maynard D and Ananiadou S. Identifying terms by their family and friends. In: Proceedings of COLING 2000, Saarbrucken, Germany. 2000:530–6.

20. Hallett C, Power R, and Scott D. Summarization and vizualisation of e-health data repositories. In: Proceedings of AHM, 2006:69–76.