Multi-Class SVM for Relation Extraction from Clinical Reports

Anne-Lyse Minard[1,2] **Anne-Laure Ligozat**[1,3] **Brigitte Grau**[1,3]
(1) LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
(2) Université Paris-Sud, 91400 Orsay, France
(3) ENSIIE, 1 square de la résistance, 91000 Évry, France
firstname.lastname@limsi.fr

Abstract

Information extraction in specialized texts raises different problems related to the kind of searched information. In this paper, we are interested in relation identification between some concepts in medical reports, task that was evaluated in the i2b2 2010 challenge. As relations are expressed in natural language with a great variety of forms, we proceeded to sentence analysis by extracting features that enable all together to identify a relation and we modeled this task as a multi-class classification based on an SVM, each type of relation representing a class. We will present the selection of the features used by our system and an error analysis. This approach allowed us to obtain an F-measure of 0.70, classifying the system among the best systems.

1 Introduction

Medical information systems have developed past years, and are used for the storage of the information to facilitate the access to data, to help to search medical information about the patient or to provide decision support to improve the quality of care. The information processed mainly concern medical literature and medical records of patients, such as the clinical reports and the consultation reports which contain a lot of information about the medical follow-up. A large part of this information is in texts. So an important issue is to convert automatically all this information into some structured knowledge as it is the starting point for the development of some semantic interrogation tools and high level processing of this information.

Extraction of medical information raises different problems related to the kind of information sought in texts: i) the recognition of medical terms, ii) related concepts and iii) relations between them. A terminologic analysis of documents lead to build semantic indexes used to search information (Jonquet et al., 2010). Identifying relations between concepts provides a more structured representation. That is useful for precise information retrieval, for example for Question-Answering systems (Tjongkimsang et al., 2005), (Embarek and Ferret, 2010).

In this paper, we present our work[1] on the identification of relations in clinical reports, task of the i2b2 2010 challenge[2]. One of the goals of the challenge was to identify several kinds of relations between concepts (treatment, test and problem). These relations are expressed in the reports by a wide range of wordings. The incompleteness of semantic knowledge bases combined with the difficulty of relating wordings with conceptual representations is an obstacle to the realization of a deep analysis of sentences which would highlight the relations between concepts.

Thus, we considered that a lot of sentence characteristics such as the words used, their syntactic category help to detect the presence of a relation. We realized a shallow analysis of sentences to extract the useful features for the detection of a relation, and we considered relation identification as a multi-class classification task, with each category of relation considered as a class. We will focus on the selection of features, that allowed us to rank 3rd with an F-measure of 0.70.

2 Related work

The first approaches for relation extraction were based on handmade patterns. In the medical domain, the SemRep system (Rindflesch et al., 2000) was developed to identify branching of anatomical relations from reports. It was also applied to detect relations between medical problems and their

[1]This work has been partially supported by OSEO under the Quaero program.

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

Proceedings of Recent Advances in Natural Language Processing, pages 604–609,
Hissar, Bulgaria, 12-14 September 2011.

treatments (Srinivasan and Rindflesch, 2002). The MedLEE system extracts relations from radiographic reports, biomolecular interactions (Friedman et al., 2001) and gene-phenotype relations (Chen and Friedman, 2004).

These approaches are not very robust and are mainly effective for precision without broad generalization capacity. So, other approaches are based on supervised machine learning. (Uzuner et al., 2010) use SVM (Support Vector Machines) to class relations between medical problems, tests and treatments in clinical reports. They defined surface features (ordering of the concepts, distance, etc.), lexical features (lexical trigrams, tokens-in-concepts, etc.), and shallow syntactic features (verbs, syntactic bigrams, syntactic link path, etc.). Results show an F-measure from 0.60 to 0.85, but for under-represented relations the classification did not work. (Roberts et al., 2008) also use a SVM to extract relations in the corpus of the Clinical E-Science Framework (CLEF) project that hold between entities (e.g. condition, drug, result) and modifiers (e.g. negation) in clinical records of cancer patients. There are seven classes of relations and each entity pair can be linked by one relation only (except between an investigation and a condition). So the classification task is considered as a binary classification (i.e. the detection of relation) between a type of relation and the non-relation class. The classification is also based on lexical, morpho-syntactic and semantic features.

In the general domain, (Zhou et al., 2005) use SVM to identify relations between people, organizations and places, etc. on the ACE corpus.

Our system also uses SVM to classify fine-grained relations. We make use of classical features as well as features specific to the domain, as the semantic types of the UMLS[3] and medical abbreviation lists, and features specific to the writing style of texts, for handling concept coordination.

3 Corpus

The corpus is made of reports from several medical centers in the USA. It was provided by i2b2 organizers. The texts were manually anonymized and annotated to build the reference. A first corpus was given before the evaluation phase, it consists of 350 documents. We divided this corpus in two parts: training corpus (4515 instances of relations) and test (749 instances of relations). For the final evaluation, i2b2 organizers gave participants a corpus of 477 documents (9070 instances of relations).

[3]Unified Medical Language System (http ://www.nlm.nih.gov/research/umls/)

TrIP	Treatment improves medical problem
	<pb> hypertension </pb> was controlled on <treat> hydrochlorothiazide </treat>
TrNAP	Treatment is not administered because of medical problem
	<treat>Relafen</treat> which is contra-indicated because of <pb>ulcers</pb>
TrWP	Treatment worsens medical problem
TrCP	Treatment causes medical problem
TrAP	Treatment is administered for medical problem
TeCP	Test conducted to investigate medical problem
	<test>an VQ scan</test> was performed to investigate <pb>pulmonary embolus</pb>
TeRP	Test reveals medical problem
PIP	Medical problem indicates medical problem
	<pb>Azotemia</pb> presumed secondary to <pb>sepsis</pb>

Table 1: The eight relations to identify

Three types of concepts were manually annotated in the corpora:

- Medical problems defined as the observations made by patients or clinicians about what are thought to be abnormal or caused by a disease.
- Treatments defined as the procedures, interventions, substances and drugs given to the patient to treat a medical problem.
- Tests defined as the procedures and examinations that are done to a patient or body fluid to control or rule out a medical problem.

Between these three kinds of concepts, eight relations can exist. The relations are described in Table 1.

The number of instances of each relation in the corpus is presented Table 2. We also report the inter-annotator agreement (IAA) calculated by the i2b2 organizers. The adjusted IAA was obtained after discussion on problematic cases. We can observe that the IAA is low for TrWP and TrIP relations.

The corpus is made of short sentences (on average 17 words per sentence in the training corpus). Clinical reports are often written using fragments of sentence (1) and enumerations (2).

(1) <pb> C5-6 disc herniation</pb> with <pb>cord compression</pb> and <pb>myelopathy</pb>.

(2) Revealed <pb>icteric sclerae</pb>, <pb>the oropharynx with extensive thrush</pb>, and <pb>an ulcer under his tongue</pb>.

Relation	training	evaluation	IAA	IAA adjusted
TrIP	107	198	0.44	**0.62**
TrWP	56	143	**0.30**	**0.58**
TrCP	296	444	0.50	0.82
TrAP	1423	2487	0.68	0.95
TrNAP	106	191	0.44	0.76
PIP	1239	1986	**0.35**	0.79
TeRP	1734	3033	0.70	0.96
TeCP	303	588	0.43	0.74
All	5264	9070	0.56	0.94

Table 2: Number of each instances of relations and inter-annotator agreement (IAA)

4 Method

4.1 Preprocessing of the corpus

Texts were preprocessed and normalized before the classification process. First, abbreviations were replaced with their meanings, thanks to a list. This list was built for the i2b2 2009[4] challenge by (Deléger et al., 2010) from the biomedical abbreviation list of Berman[5] and examples found in the i2b2 2009 corpus. For example, *h.o.* is converted in *history of* and *p.r.n.* into *as needed*. Then we substituted the anonymized data with the markups *NAME*, *DATE* and *AGE*, and numerical values (mainly proportions) are replaced with the markup *NUM*. Finally texts are part-of-speech (POS) tagged by the TreeTagger (Schmid, 1994) in order to have lemmas and POS categories.

4.2 Classification

The classification makes use of SVM implementation of LIBSVM tool (Chang and Lin, 2001) parametrized for a multi-class classification (*one-versus-one* voting). We chose a RBF kernel, which gave better results than a linear kernel. The parameters are chosen by the script *grid.py* provided with LIBSVM. The *c* parameter was set to 16 and the *gamma* parameter to 0.03125. We also tested a classification by pair of concepts by training a classifier for relations between a test and a medical problem, then between a treatment and a medical problem, and between two medical problems. But results were lower than when we learned with all the relations. The features used for the classification capture surface information, such as the position of the two candidate concepts, lexical information, for example the words which refer to the concepts and the relation, syntactic information as POS tags, and semantic information. The features are automatically computed, if necessary by using tools and external resources. Each feature has an unique identifier, which is set to one if it appears else zero.

4.2.1 Surface features

Ordering of the candidate concepts: the expression of the relation depends on the position of the test or treatment compared with the problem. In example (3) the test is uttered before the revealed problem, and conversely in example (4) the problem is uttered before the test.

(3) She had <test>a workup</test> by her neurologist and <test>**an MRI**</test> revealed <pb>a C5-6 disc herniation</pb> [...]

(4) The patient was <pb>thrombocytopenic</pb> with <test>**a platelet count**</test> of <NUM> on the <NUM>.

Distance (i.e. number of words[6]) between the candidate concepts: in the training corpus there is never more than 65 words between two related concepts. However two concepts which are not in relation can be separated by a maximum of 205 words. The value of this feature is a number.

Presence of other concepts between the candidate concepts: for 80% of the concept pairs in relation in the training corpus there are no other concepts between them.

4.2.2 Lexical features

In order to provide some structure to the information given in texts, we decompose sentences in three zones: left and right contexts of the two candidate concepts and the between part.

The words and stems[7] which constitute the concepts and the headword[8] of each concept. The stems are used to group inflectional and derivational variations together. The words of concepts can trigger relations. For example in (5) the adjective *recurrent* is the trigger of a TrWP relation (a treatment worsens a problem).

(5) He has had <NUM> week courses of <treat>antibiotics</treat> with <pb>**recurrent** bacteremia</pb>.

The stems of the three words in the left and right context of candidate concepts. After several experiments we chose a window of three words;

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

[5]http://www.julesberman.info/abbtwo.htm

[6]The words include also the punctuation signs.

[7]We use the PERL module lingua::stem to obtain the stem of the word.

[8]The headword is the word which precedes a preposition or the last word of the concept (see (Zhou et al., 2005)).

Relation	base	+dist	+conc	+dir	+verb	+prep	+intra	+types
TrIP	0.333	0.333	0.333	0.333	0.235	0.235	0.235	0.235
TrWP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TrCP	0.366	0.370	0.405	0.411	0.424	0.441	**0.526**	0.517
TrAP	0.620	0.638	**0.708**	0.721	0.708	0.706	0.737	0.726
TrNAP	0.620	0.620	0.620	0.620	**0.666**	0.666	0.666	0.620
PIP	0.611	0.613	0.664	0.664	0.654	0.671	0.618	**0.659**
TeRP	0.790	0.792	0.833	0.843	0.850	0.848	0.866	0.866
TeCP	0.253	0.253	**0.373**	0.351	0.373	0.351	0.333	0.285
All	0.647	0.652	**0.704**	0.712	0.711	0.713	0.724	**0.727**

Table 3: Variation of the F-measure according to the features (test corpus)

with bigger or smaller windows, precision lightly increases but recall decreases.

The stems of the words between candidate concepts; the most important information for the classification is located here.

The stems of the verbs in the three words at the left and right of candidate concepts and between them. The verb is often the trigger of the relation: for example in (6) the TeRP relation (a test reveals a problem) is expressed by *reveal*.

(6) <test>CT scan</test> was obtained and this **revealed** <pb>free air</pb> and <pb>massive ascites</pb>.

The prepositions between candidate concepts. In (7) the preposition *for* indicates a TrAP relation (a treatment is administered for a problem).

(7) She was treated with <treat>IVF</treat> **for** <pb>her ARF</pb>.

4.2.3 Morpho-syntactic features

The morpho-syntactic tags of the three words at the left and right of candidate concepts.

The presence of a preposition between candidate concepts, regardless of the preposition.

The presence of a punctuation sign between candidate concepts, if it is the only "word". This feature is useful for considering lists.

4.2.4 Semantic features

The semantic type (from the UMLS) of the three words at left and right of candidate concepts. In the example (3) *neurologist* has the semantic type *professional or occupational group*.

The types of candidate concepts (problem, test or treatment): it is the most important feature, because the relations are expressed differently between a test and a problem, a treatment and a problem, and between two problems.

The VerbNet's classes[9] (an expansion of Levin's classes) of the verbs in the three words at the left and right of candidate concepts and between them. For example *reveal* is member of the class *indicate-78-1-1* which contains also the verbs *show*, *prove*, *demonstrate*, etc. In examples (6) and (8) *reveal* and *show* are triggers of the same relation.

(8) <test>Recent chest x-ray</test> **shows** <pb>resolving right lower lobe pneumonia</pb>.

4.2.5 Coordination

Two concepts in relation can be separated by other concepts which do not carry information about the relation. So, we processed sentences before the feature extraction. We deleted other annotated concepts in coordination with candidate concepts, and we added three features: the number of deleted concepts, the coordination words that are the triggers of the deletion (*or*, *and*, a comma), and a feature which indicates that the sentence was reduced. Coordinations are often a sign of the non existence of relation, while they add information that are not useful to type it and even create some noise. In the training corpus the sentences have been reduced for 23% of the pairs of concepts (3819 pairs on 16437). In the example (6) for the pair *CT scan* and *massive ascites*, after reduction the sentence segment is: *CT scan was obtained and this revealed massive ascites*.

4.2.6 Feature relevance

We evaluated the usefulness of each feature with the same method as (Roberts et al., 2008). We observed the performances of the system on the test corpus by adding features class by class. Results are shown in Table 3.

[9] http://verbs.colorado.edu/∼mpalmer/projects/verbnet.html

The features are grouped in categories according to the information they describe. The category *base* contains the stems of the words, the morpho-syntactic tags of the three words at the left and right of the concepts, and the stems of the words between the concepts. Then we added the category *dist* (distance between the concepts), *conc* (the other concepts), *dir* (the ordering of the concepts), *verb* (the stems of the verbs and the VerbNet classes), *prep* (the prepositions between the concepts), *intra* (the constituent words and the headword of the concepts) and *types* (semantic types). The results of the last column in the table are the results of the system with all the features. This system corresponds to the system used for the evaluation. In this system we did not use the features about the coordination of concepts. We separately evaluated these features, which increase the F-measure of the final system of 0.002.

5 Evaluation

Table 4 shows the results obtained[10]. We achieved better results over well-represented relations (such as TeRP with an F-measure of 0.852) than over smaller classes of relations (such as TrCP relation with an F-measure of 0.489).

For the i2b2 challenge, we used this system (without the control of the coordination) and we combined it with some patterns to identify the four under-represented relations (patterns have priorities on the classifier). Our system obtains an F-measure of 0.709, and ranked 3rd out of 16 teams. In Table 4 we show the results of the 1st, 2nd and 4th systems and the median. For classification of non-relations, our system obtained a recall of 0.93, a precision of 0.84 and an F-measure of 0.89.

6 Error analysis

For relations occurring between a treatment and a medical problem, we studied the confusion matrice and observed that the misclassified relations are mainly classified in the TrAP category (treatment is administered for medical problem) or as a non-relation. For example 54% of TrIP relations (treatment improves medical problem) are classified as a non-relation and 31% as TrAP relation. It is sometimes difficult to differentiate a TrIP or TrAP relation, because the TrIP relation is a specific TrAP relation. Indeed if a treatment improves a medical problem so the treatment is administered because of a medical problem. It is the same for TrWP relation which includes cases where the treatment is administered for a medical problem but worsens it.

[10]The F-measure for "all relations" is the micro-averaged F-measure that weights each relation by its frequency.

Relation	Recall	Precision	F-measure
TrIP	0.156	0.861	0.264
TrWP	0.000	0.000	0.000
TrCP	0.369	0.725	0.489
TrAP	0.693	0.739	**0.715**
TrNAP	0.057	0.423	0.101
PIP	0.552	0.787	0.649
TeRP	0.835	0.870	**0.852**
TeCP	0.238	0.833	0.370
All relations	**0.628**	**0.803**	**0.705**
Median			0.664
1st system	0.753	0.720	0.736
2nd system	0.693	0.773	0.731
4th system	0.675	0.730	0.701

Table 4: Recall, precision and F-measure obtained on the evaluation corpus

For relations between two medical problems, we observed that 50% of PIP relations (medical problem indicates medical problem) were not detected. In the training corpus there are enough examples, but the description of the relation might not be precise enough (see IAA in Table 2). In example (9) a PIP relation was annotated between *symptoms* and *anxiety*, but not in the example (10) between *symptoms* and *dry cough*.

(9) She was hooked up with support services in Collot Ln, Dugo, Indiana <NUM> for <treat>further counselling</treat> and given <treat>Xanax</treat> for <pb>**symptoms**</pb> of <pb>**anxiety**</pb>.

(10) Pt was o/w in his USOH until <NUM> weeks ago when he developed <pb>a URI</pb> with <pb>**symptoms**</pb> of <pb>**dry cough;**</pb> no <pb>fever</pb> [...]

By studying sentences of misclassified relations we have found three types of errors:

- The relation is expressed by a verb or an expression, but this construction is not represented in the training corpus. In (11), the system classified the relation between *pulmonary nodules in his RML* and *fu imaging* as TeRP. Indeed *reveal* is a trigger of a TeRP relation, and the trigger of the TeCP relation is *which need*, but this last verb occurs only once in the training corpus.

(11) <test>CTS chest</test> was negative for <pb>PE </pb>, however it did **reveal** <pb>pulmonary nodules in his RML</pb> **which need** <test>fu imaging</test> in <NUM> months.

- The relation cannot be classified without using external resources or more training examples. In (12) the system wrongfully detected a relation, as it would need to know that there is no relation possible between incisions and obesity to correctly classify the relation.

(12) <pb>obese</pb> with <pb>multiple well healed surgical incisions</pb>, positive bowel sounds.

- The annotation of the relation is debatable. In (9) a relation between *symptoms* and *anxiety* has been annotated, but this two terms make reference to the same concept.

To improve the extraction of under-represented relations such as TrWP or TrIP, a bigger corpus is necessary, as these relations are represented by a few number of occurrences in the corpus. However there is no such annotated available corpus.

7 Conclusion

Relation extraction between concepts in clinical reports is a task that helps improve access to information in medical documentation. This task is based on the recognition of the several wordings that the relation can take in the sentences. This variability is very important as for the vocabulary variability as syntactic structures. So, we have taken into account these variabilities by defining different features, which can describe such kinds of sentences. We used features specific to the domain, the type of concepts for instance, features specific to the kind of texts and general domain features. We obtained very good results thanks to the selection of the features and the combination we made. The selected features are general enough that they can be used on corpora in other fields, with an adaptation of the domain dependent features (such as semantic types).

The results are low for not well-represented relations in the corpus. To have more representative instances of these relations, we could operate a reduction of the syntactic variability and a simplification of sentences before the learning stage.

References

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Lifeng Chen and Carol Friedman. 2004. Extracting phenotypic information from the literature via natural language processing. In *Medinfo 2004: Proceedings Of THe 11th World Congress On Medical Informatics*.

Louise Deléger, Cyril Grouin, and Pierre Zweigenbaum. 2010. Extracting medical information from narrative patient records: the case of medication-related information. *JAMIA*, 17(5):555–558.

Mehdi Embarek and Olivier Ferret. 2010. Can esculape cure the complex of œdipe in the medical domain? In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10.

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17.

Clement Jonquet, Paea LePendu, Sean M. Falconer, Adrien Coulet, Natalya F. Noy, Mark A. Musen, and Nigam H. Shah. 2010. Ncbo resource index: Ontology-based search and mining of biomedical resources. In *Semantic Web Challenge, 9th International Semantic Web Conference, ISWC'10*.

Thomas C. Rindflesch, Carol A. Bean, and Charles A. Sneiderman. 2000. Argument identification for arterial branching predications asserted in cardiac catheterization reports. In *AMIA Annu Symp Proc*.

Angus Roberts, Robert Gaizauskas, and Mark Hepple. 2008. Extracting clinical relationships from patient narratives. In *BioNLP2008: Current Trends in Biomedical Natural Language Processing*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Padmini Srinivasan and Thomas Rindflesch. 2002. Exploring text mining from medline. In *Proc AMIA Symp*, pages 722–726.

Erik Tjongkimsang, Gosse Bouma, and Maarten de Rijke. 2005. Developing Offline Strategies for Answering Medical Questions. In *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains, Pittsburgh, PA, USA*.

Ozlem Uzuner, Jonathan Mailoa, Russell Ryan, and Tawanda Sibanda. 2010. Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*, 50:63–73.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 427–434.