Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification

Anne-Lyse Minard, Anne-Laure Ligozat, Asma Ben Abacha, Delphine Bernhard, Bruno Cartoni, Louise Deléger, Brigitte Grau, Sophie Rosset, Pierre Zweigenbaum, Cyril Grouin

LIMSI—CNRS, Orsay Cedex, France

Correspondence to
Cyril Grouin, BP 133, 91403
Orsay Cedex, France;
cyril.grouin@limsi.fr

Received 1 February 2011
Accepted 8 April 2011
Published Online First
19 May 2011

ABSTRACT

Objective This paper describes the approaches the authors developed while participating in the i2b2/VA 2010 challenge to automatically extract medical concepts and annotate assertions on concepts and relations between concepts.

Design The authors'approaches rely on both rule-based and machine-learning methods. Natural language processing is used to extract features from the input texts; these features are then used in the authors' machine-learning approaches. The authors used Conditional Random Fields for concept extraction, and Support Vector Machines for assertion and relation annotation. Depending on the task, the authors tested various combinations of rule-based and machine-learning methods.

Results The authors'assertion annotation system obtained an F-measure of 0.931, ranking fifth out of 21 participants at the i2b2/VA 2010 challenge. The authors' relation annotation system ranked third out of 16 participants with a 0.709 F-measure. The 0.773 F-measure the authors obtained on concept extraction did not make it to the top 10.

Conclusion On the one hand, the authors confirm that the use of only machine-learning methods is highly dependent on the annotated training data, and thus obtained better results for well-represented classes. On the other hand, the use of only a rule-based method was not sufficient to deal with new types of data. Finally, the use of hybrid approaches combining machine-learning and rule-based approaches yielded higher scores.

INTRODUCTION

The i2b2/VA 2010 challenge addressed the extraction of medical concepts, the annotation of assertions on medical problems, and the detection of relationships between concepts. This kind of information summarizes the content of medical reports and is in line with the i2b2 2009 challenge, which aimed to allow easy access to medication information in medical reports.

In this paper, we present the LIMSI participation in this challenge. After a short reminder of the challenge requirements and corpora, we describe the pipelines we developed for each task, focusing our efforts on the use of hybrid approaches (machine-learning and rule-based systems). We then present their evaluation and discuss their results.

CHALLENGE REQUIREMENTS

The fourth i2b2/VA challenge consisted of three routes: first, the extraction of three types of medical concepts (problems, tests, and treatments); second, the annotation of assertions made on medical problems; and finally, the annotation of relations between concepts.

Three types of concepts are targeted, which mainly correspond to sets of UMLS semantic types: (1) problems concern observations made about the patient if thought to be abnormal or caused by a disease; (2) treatments describe all methods used to resolve a medical problem; and (3) tests refer to examinations and procedures carried out for a medical problem.

Assertion annotations must be provided only for medical problems and consist of six categories: the patient experiences the medical problem (present) or does not (absent); the patient may have a problem that is uncertain (possible) or that occurs only under certain conditions (conditional); the patient may develop the problem (hypothetical), or the problem is mentioned in relation to someone else (not associated with patient).

Relation annotations must describe relationships between: (1) a problem and a treatment where the treatment can improve (TrIP), worsen (TrWP), or cause (TrCP) the problem, where it can be administered (TrAP) or not (TrNAP) for the problem; (2) a problem and a test where the test reveals (TeRP) or allows a physician to investigate (TeCP) the problem; and (3) a problem that indicates another problem (PIP).

The corpus includes discharge summaries from three institutions and progress notes from another. The training corpus consists of 349 reports, while the ground truth corpus consists of 477 reports. We split the provided training corpus into three subcorpora: training (241 reports) to create linguistic resources, development (54) to tune our models, and test (54) to test these models. In each subcorpus, we preserved the original distribution of hospital sources. No cross-validation was performed while the final model for the overall evaluation has been built over the whole corpus.

SYSTEM DESCRIPTION

Expert-knowledge-based and machine-learning methods

Medical natural language processing techniques are generally of two kinds. On the one hand, expert-knowledge-based techniques have been used for a long time for concept extraction and other tasks[1–4] including assertion classification.[5–9] They require much work while providing reliable results. On the other hand, machine-learning approaches are increasingly being used[6–8] because they provide a fast path to results, once corpora have been annotated. A combination of the two can also improve performance[10] and can take multiple forms.

We have taken advantage of this challenge to test different approaches to the challenge tasks: independent expert-knowledge-based and machine-learning approaches for concept extraction, using expert knowledge as a feature for assertion detection, and merging the results of both with priority to expert knowledge for relation detection.

Task 1: concept extraction

Concept extraction has been addressed by defining rules and gazetteers[1] or using linguistic resources obtained from the Unified Medical Language System (UMLS).[11–14] A few approaches also used the structure of the discharge summaries to extract test and treatment concepts.[2] We developed two pipelines for the concept extraction task: the first is mainly based upon MetaMap, and the second uses a machine-learning method.

Expert-based method based on MetaMap

In order to test rule-based methods for concept extraction, we used the MetaMap biomedical annotation tool.[15–17] MetaMap was designed at the National Library of Medicine to locate medical terms and their corresponding concepts and semantic types from the UMLS Metathesaurus and Semantic Network. MetaMap is widely used in medical language processing. However, it has some residual problems at the noun-phrase segmentation level and for the recognition of several treatments, diseases, and tests. In our experiments, direct application of MetaMap to the i2b2 evaluation corpus obtained an F-measure of 0.158 (0.161 precision and 0.155 recall).

We proposed an enhanced use of MetaMap which adds several preliminary steps (run C3): (1) sentence segmentation into noun phrases with treetagger-chunker; (2) noun phrase filtering using lists of stopwords and common MetaMap errors; and (3) a search of the located terms in lists of medical problems, tests, and treatments obtained from the training corpus, Wikipedia, Health on the Net, and Biomedical Entity Network. The noun phrases which were not located in the provided lists were then passed to MetaMap, which outputs concepts and semantic types. We map these semantic types to the target concept types (ie, treatment, problem, test) through tables (eg, 12 UMLS semantic types are mapped to 'problem'). For a given noun phrase, MetaMap can provide several candidate UMLS concepts and semantic types with the same score. In that case, we apply a voting procedure which selects the target type (treatment, problem, or test) that is most frequent among those output by MetaMap for this noun phrase. In case of a tie, the first returned type is selected.

Machine-learning method

We defined the following pipelines for machine-learning-based concept extraction. Each of them first performs a limited linguistic analysis, whose output is represented as features, which a machine-learning algorithm then uses to make decisions on concept boundaries and types. These features were defined for each token as follows:

1. N-grams of tokens, that is, sequences of n 'words' including the current word;
2. Typographic properties of a token: letter case and four binary character type features are defined according to the presence of alphabetic characters, digits, punctuation, or date.
3. Syntactic tags: we performed a morpho-syntactic analysis using Tree Tagger[18]; POS tags and lemmas are thus associated with each token. We then performed a syntactic tagging using a specific lexicon of 62 263 adjectives and 320 013 nouns based on the UMLS Specialist Lexicon. These lists specify the types of adjectives (*relational and qualitative*) and nouns (*proper name, countable and uncountable*), and the possible positions of adjectives in a sentence (*attributive, postnominal or predicative*).
4. Semantic tags: semantic tagging was performed with 11 major semantic types: anatomy, laboratory analysis (*creatinine, hematocrit*), examination (*angiography, biopsy, scan, x-ray*), pre- or postexamination mark (*follow-up…, physical…, repeat…, …culture, …evaluation, …levels*), general anatomical location (*lower, upper, right, left*), medication, mode of administration, medical artefact (*cannula, drain, pacemaker, stent*), procedure (*amputation, blood transfusion, dialysis*) and dosage. We created these categories thanks to lists drawn from the UMLS,[14] from Sager's work,[3 4] and from those we compiled for the i2b2 2009 challenge.

We built a model from the training corpus using CRF++,[19] a machine-learning tool based on Conditional Random Fields. We applied this model to the test corpus. This pipeline was used for our first submission (run C1).

We tried to refine the output of this model by designing a few postediting rules to correct errors observed when testing on the development corpus. A token with 'medication' as feature is tagged as a treatment concept if not already detected. Assuming a 'one sense per corpus' principle, we also regularized the resolution of some ambiguities by selecting the most frequently assigned concept type in cases where different concept types had been assigned to the same string in different locations in the same text. This pipeline was used in our second submission (run C2).

Task 2: assertion annotation

Assertion classification has also used both expert-knowledge-based approaches, which involve listing and detecting indicative phrases or specific syntactic dependencies for a given type of assertion,[5–9] and machine-learning approaches, which rely on annotated data to train a supervised classification system.[6 9] It is also closely related to hedge classification,[20] which aims at detecting speculation in natural language texts. This task has been addressed with weakly supervised machine-learning.[20]

The corpus was also preprocessed to cope with coordination and to tag each concept with its type. Our study of the development data showed that many problem concepts are coordinated with commas or coordinating conjunctions—for example, 'pleural effusion or pneumothorax.' These sequences of coordinated problems might lead to obtaining reduced left and/or right contexts, mostly containing other coordinated problems. In this case, important cues for a specific assertion type may fall outside the scope of the contextual window. The important role of coordination has been highlighted before for event extraction.[8] We therefore preprocessed the data to identify coordinated problems and redefine the offsets for left and right token windows.

Expert-knowledge-based method: extension of NegEx

This method (run A2) was based on an extension of the NegEx[5] algorithm (recent releases of MetaMap now include a switch for

NegEx.) which locates trigger terms indicating a negation (eg, 'never had') or a probability (eg, 'possibly') and determines whether the concepts fall within the scope of these triggers. Then, we extended the General ConText Java implementation of NegEx (http://code.google.com/p/negex/) to deal with the categories *conditional, hypothetical, and not associated with the patient*, which NegEx does not handle.

Triggers for these categories were first manually defined based on a corpus study. The lists of triggers were also completed thanks to the results of the machine-learning system described in the next subsection: the attributes which were most useful for the classification were manually selected to be part of the trigger lists.

Machine-learning method

We also addressed assertion identification as a classification task, with the six assertion types as target classes (run A1). We trained a Support Vector Machine (SVM) classifier with the libsvm tool[21] based on binary feature vectors. We automatically selected the optimal parameter values using cross-validation. (This step was performed with the easy.py script provided with libsvm.) We focused on three types of features: contextual lexical features, trigger-based features and target concept internal features:

- Contextual lexical features consist of token and stemmed token unigrams in a five-word window to the left and to the right of the target concept. We also experimented with POS unigrams, as well as token bigrams and trigrams, but these did not lead to significant improvements.
- Triggers consist of manually defined phrases which are indicative of a given assertion class. We used the triggers collected for our extension of GenConText, with few additions. These triggers were identified before and after the problem concept, again in a five-word window. We also identified some concept-internal triggers (ie, words within the term that denotes the problem) such as 'on exertion' which is indicative of the conditional assertion class when it occurs within an annotated concept.
- Target internal features comprise problem tokens, stemmed problem tokens, and the presence of the 'non' negative prefix in one of the problem words.

Task 3: relation annotation

The extraction of semantic relations from medical texts has been the subject of an increasing stream of work in the last decade. Some approaches use linguistic methods based on patterns or extraction rules,[22 23] or machine-learning techniques.[24 25] Others proposed hybrid approaches which combine two or more techniques.[26]

Given two argument concepts, we considered relation identification as a nine-way classification task, with the eight relation types (TrIP, TrWP, TrCP, TrAP, TrNAP, TeRP, TeCP, and PIP) and the non-relation case. We used a hybrid approach which combines machine-learning techniques and linguistic pattern matching. We trained an SVM with the libsvm tool and constructed linguistic patterns manually. For the first run (run R1), before the prediction of relation types with libsvm, we used patterns to identify four relations: TrIP, TrWP, TrNAP, and TeCP for which there are few examples in the training set. The predictions of the patterns have priority over the predictions of machine learning. The second run (R2) uses supervised learning from simplified texts. Finally, the last run (R3) is a combination of the first two results with priority to run R1 if it detects a relation, else fall back on the results of run R2.

After empirical observations on the training corpus, we only kept the patterns of four relation types, since the others did not offer satisfying results. The advantage of such a hybrid approach lies in the fact that some relation types do not have enough annotations to feed the automatic classifiers.

Expert-knowledge-based method: relation patterns

This approach uses a manually constructed set of lexical patterns for each semantic relation. The constructed patterns are regular expressions describing a set of matching sentences containing medical entities at specified positions with a more or less specific lexical context. More precisely, each pattern consists of a sequence of words, tags corresponding to the three concept types and generic markers representing a length-limited character sequence (eg, _chars_). The patterns were constructed from the training corpus, and external electronic dictionaries were used to enrich them with synonyms of important words. Table 1 shows the number of constructed patterns and some simplified pattern examples.

Machine-learning method

Our three variant methods for relation extraction are based on supervised learning using SVMs. The features of our SVM are as follows:

- Surface features: order of the argument concepts, distance (ie, number of tokens—a token is a word or a punctuation mark) between them, and presence of other concepts.
- Lexical features: tokens and stemmed tokens in argument concepts, left and right trigrams (of stemmed tokens) of the two concepts, stemmed tokens between them, verbs in a three-word window before and after each concept and between them, prepositions between concepts, headword of concepts (the headword of a noun phrase is approximated as the token before a preposition, else as the last token).
- Syntactic features: part-of-speech in a three-word window to the left and to the right of the argument concepts, presence of a preposition, presence of a coordinating conjunction between concepts and punctuation signs.

Table 1 Sentence/pattern examples and number of constructed patterns for four relations

Relation	Studied example sentences	Constructed pattern examples	Patterns
TrIP	Her pain *resolved after* surgery, and she has been doing well since	PROBLEM _chars_ resolved ((after|with))? TEST	43
TrWP	Prolong hospitalization can *exacerbate* some of her Axis I and II conditions	TEST _chars_ exacerbate _chars_ PROBLEM	27
TrNAP	The plan was to treat heart failure with intravenous diuretics, as medical therapy coronary syndrome *is limited due to* thrombocytopenia and guaia-positive stools	TEST _chars_ (limited|discontinued|stopped) (secondary to|because of|due to)? _chars_ PROBLEM	25
TeCP	The patient subsequently underwent a CT scan angiogram with Pancreas protocol *to assess* the pseudocyst versus enlarged pancreas head	TEST _chars_ to (assess|evaluate) _chars_ PROBLEM	56

TeCP, test conducted to investigate medical problem; TrIP, treatment improves medical problem; TrNAP, treatment is not administered because of medical problem; TrWP, treatment worsens medical problem.

Table 2 Concepts: Recall (R), Precision (P), and F-measure (F) (class exact span)

	Run C1			Run C2			Run C3		
	R	P	F	R	P	F	R	P	F
All concepts	0.725	0.825	0.772	0.726	0.826	0.773	0.487	0.565	0.523
Problem	0.741	0.795	0.767	0.742	0.799	0.769	0.530	0.608	0.567
Treatment	0.724	0.844	0.779	0.723	0.843	0.778	0.619	0.520	0.565
Test	0.703	0.853	0.770	0.705	0.851	0.771	0.285	0.567	0.379

- Semantic features: UMLS semantic type of tokens in a three-word window on either side of each argument concept, i2b2 types of the concepts (problem, test, or treatment) and Levin's class of the verbs (from VerbNet; http://verbs.colorado.edu/~mpalmer/projects/verbnet.html).[27]

Files are preprocessed and normalized. First, we replaced abbreviations with their expansions—for example *h.o.* is converted to *history of* and *p.r.n.* into *as needed*. Then, we substituted the person's name (or eg, **NAME[VVV]), the date (or eg, **DATE[Jan 06 2008]), the person's age and other numbers respectively with <NAME>, <DATE>, <AGE> and <NUM>. Finally, files were POS-tagged by the TreeTagger.

For the second variant (R2), preprocessing of the text consists of a syntactic simplification, which involves deletion of some syntactic phrases between the argument concepts. The aim of the simplification is to delete useless information for the relation identification process, rather than to obtain grammatically correct sentences. Before the simplification process, concepts are substituted with their types (problem, test or treatment), and each sentence is duplicated for each candidate relation (if there are three concepts in a sentence, it is written three times, once for each pair of concepts). Then, texts are analyzed by the Charniak/McClosky self-training parser.[28] Simplification proceeds in two steps. First, if an argument concept is at the beginning of its noun phrase, all words after the concept in the noun phrase are deleted. Second, if there is a prepositional phrase, an adjectival phrase, a phrase with a conjunction, a relative pronoun, or a coordination conjunction (followed by a noun phrase) between the concepts, it is replaced with its syntactic category (<PP>, <ADJP>, etc).

RESULTS

Task 1: concept extraction

The ground-truth corpus contains 45 009 concepts to be extracted (18 550 problems, 13 560 treatments, and 12 899 tests).

Run C1 is machine-learning based, while run C2 applies correcting rules to the previous output. Run C3 is a rule-based method using MetaMap. Run C2 was only marginally better than C1 (see table 2), which means that the correcting rules had limited impact.

We examined the origin of errors in our best run C2, distributing them into the following categories. Concept *insertions* are concepts not present in the reference (9.1% of results), and concept *deletions* are concepts present in the reference but not in system output (21.3%). Concepts present in the reference may have been found with erroneous *boundaries* (3.7%) or with erroneous *types* (2.5%).

Task 2: assertion annotation

The ground-truth corpus is composed of 18 550 assertions on medical problems (13 025 present, 3609 absent, 883 possible, 717 hypothetical, 171 conditional, and 145 associated with someone else). Our system (run A1) ranked fifth out of 21 participants at the i2b2/VA 2010 challenge with a 0.931 F-measure (see table 3).

The supervised machine-learning system yields very good results. If we compare the two systems, we notice that the machine-learning system tends to have better precision than recall. It also obtains better precision overall than the rule-based system and is characterized by a better F-measure, except for the 'Associated with someone else' category. In this case, the triggers used by the rule-based system lead to a very high recall of 0.95, showing that it has a very good coverage of this phenomenon. This comparison highlights the complementary nature of the two systems.

We examined the number of annotations per category assigned by each system, and noticed that the machine-learning system has a strong tendency to overannotate the categories 'present' (13 405 annotations) and 'absent' (3673 annotations). This result is not surprising, since both categories also have the largest amount of occurrences in the training corpus. The rule-based system tends to annotate many assertions from the 'present' class as being 'conditional,' leading to very low results for the 'conditional' category. It is also the category for which both systems have the lowest overlap: they share only 31 annotations in this category.

Task 3: relation annotation

The ground truth corpus is composed of 9069 relationships (198 TrIP, 143 TrWP, 444 TrCP, 2486 TrAP, 191 TrNAP, 1986 PIP, 3033 TeRP, and 588 TeCP). Our system (run R3) ranked third out of 16 participants at the i2b2/VA 2010 challenge with a 0.709 F-measure (see table 4).

The three runs are machine-learning based. Run R1 uses rules for four relations, run R2 applies simplification to input sentences,

Table 3 Assertions: Recall (R), Precision (P), and F-measure (F) (exact span with matching assertion)

	Run A1			Run A2		
	R	P	F	R	P	F
All assertions	0.931	0.931	0.931	0.898	0.898	0.898
Present	0.970	0.942	0.956	0.948	0.917	0.932
Absent	0.947	0.931	0.939	0.853	0.934	0.891
Possible	0.538	0.738	0.622	0.572	0.614	0.592
Hypothetical	0.830	0.928	0.876	0.741	0.863	0.797
Conditional	0.240	0.745	0.363	0.275	0.287	0.281
Associated with someone else	0.779	0.856	0.816	0.952	0.758	0.844

Table 4 Relations: Recall (R), Precision (P), and F-measure (F) (exact span)

		Run R1			Run R2			Run R3		
	No of relations in the training corpus	R	P	F	R	P	F	R	P	F
All relations	5264	0.634	0.797	0.706	0.626	0.718	0.669	0.708	0.711	0.709
TrIP	107	0.364	0.468	0.409	0.258	0.638	0.367	0.414	0.458	0.435
TrWP	56	0.161	0.742	0.264	0.021	0.600	0.041	0.168	0.774	0.276
TrCP	296	0.354	0.781	0.489	0.347	0.490	0.406	0.435	0.550	0.486
TrAP	1423	0.678	0.747	0.710	0.661	0.676	0.668	0.760	0.676	0.715
TrNAP	106	0.199	0.528	0.289	0.162	0.484	0.243	0.251	0.495	0.333
PIP	1239	0.538	0.791	0.641	0.565	0.667	0.612	0.645	0.670	0.657
TeRP	1734	0.836	0.870	0.853	0.818	0.822	0.820	0.881	0.813	0.846
TeCP	303	0.293	0.726	0.417	0.330	0.616	0.430	0.391	0.612	0.477

PIP, medical problem indicates medical problem; TeCP, test conducted to investigate medical problem; TeRP, test relations with medical problem; TrAP, treatment is administered for medical problem; TrCP, treatment causes medical problem; TrIP, treatment improves medical problem; TrNAP, treatment is not administered because of medical problem; TrWP, treatment worsens medical problem.

and run R3 is a combination of the results of run R1 and run R2. Our combination method, with priority to the most precise method and fall-back to the less precise method, logically improves recall at the cost of precision, balancing them more evenly.

Discussion

The concept extraction results obtained by run C3 are highly dependent on its chunker. We used the Treetagger chunker, whose output noun phrases matched 60.8% of the ground-truth concepts. This imposed a ceiling of 0.608 on the recall of the improved MetaMap method we designed. We obtained 0.48 recall with this method, which leads to the conclusion that the major part of missed concepts is due to chunking rather than concept type categorization. It is important to note that the problem is not specific to the Treetagger chunker: we tested other chunking tools (eg, OpenNLP, GeniaTagger), and they yielded worse results than Treetagger. Therefore, the main problem is the correct detection of medical entity boundaries. This problem may be improved by machine-learning techniques. Such techniques could also have benefits on the categorization into semantic types and provide more scalable solutions.

The CRF-based approach for concept extraction embodied in runs C1 and C2 determines morpho-syntactic and semantic information for each token and lets a state-of-the-art sequence classifier make concept type and boundary decisions. This allowed us to obtain a good basis. However, we missed 9569 concepts, including generic terms (*pain, fever, blood*) and abbreviations (*jvd, nt, dm2*) because of incomplete lists and lack of normalization, and long segments (*two units of packed red blood cells*), which are more difficult for the machine-learning system to capture with its limited-dependency model. Some type errors are due to a lack of context modeling: for instance, *blood coagulation* is a treatment in the reference while it was detected as a test. Taking the context into account could help resolve this ambiguity. The use of more accurate semantic information could help improve the set of semantic tags that we defined. Indeed, we consider that the use of a semantic grammar to capture term structure better could improve precision. Furthermore, we computed semantic annotations for each token without taking its context into account. There, too, context should allow us to disambiguate tokens and obtain semantic annotations that are more accurate.

For assertion annotation, with the machine-learning system, we achieved better results with well-represented classes (such as 'present' and 'absent' that total 89.7% of all assertions) than with smaller classes such as 'conditional' (only 0.9% of the assertions). Nevertheless, we achieved good results with the class 'associated with someone else' (0.8% of the assertions) thanks to the use of trigger words. For this specific class, the expert-knowledge-based method achieves better results than the machine-learning system: this shows that it can be beneficial even to combine both systems.

In relationship annotation, we also obtained better results for the well-represented classes in the training corpus, such as TeRP or TrAP. In contrast, for the TrWP class, the training corpus contains only 56 relations, so the system was unable to learn. Figure 1 orders the relations according to their number of training examples: it shows that except for the TeCP/TrCP pair, the F-measure of relation annotation system generally varies in the same direction as the number of training examples.

Later, we evaluated the machine-learning method alone, and we obtained a total 0.702 F-measure. With the hybrid method (ie, combination of pattern matching and machine-learning), we obtained an overall 0.706 F-measure. The use of patterns with machine-learning improves relation classification. For the TrWP relation, without patterns, the F-measure is null, while we have a 0.264 F-measure with the use of patterns. The F-measure increases from 0.237 to 0.409 for the TrIP relation, from 0.118 to 0.289 for the TrNAP relation, and from 0.375 to 0.417 for the TeCP relation.

Run R2 allowed us to find new relations compared to run R1. Indeed, run R3 obtains a better recall than run R1. We think that using syntactic simplification can really improve relation classification, but this requires the development of more precise simplification rules. Moreover, after error analysis, we believe that another possible improvement of our system is to add information about syntactic structure.

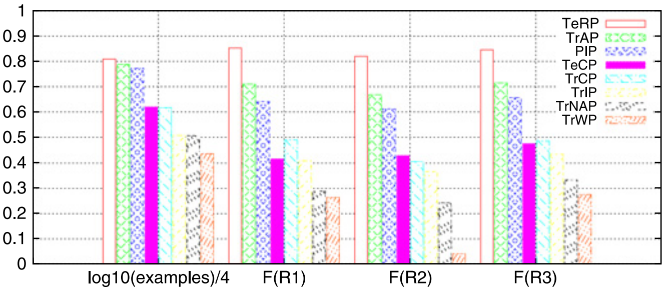

Figure 1 F-measure of relation extraction (runs R1, R2, and R3) and number of examples (normalized to fit within the same scale). PIP, medical problem indicates medical problem; TeCP, test conducted to investigate medical problem; TeRP, test relations with medical problem; TrAP, treatment is administered for medical problem; TrCP, treatment causes medical problem; TrIP, treatment improves medical problem; TrNAP, treatment is not administered because of medical problem; TrWP, treatment worsens medical problem.

CONCLUSION

In this paper, we investigated knowledge extraction from clinical texts. We particularly worked on three tasks: (1) extraction of medical concepts, (2) assertion annotation, and (3) extraction of semantic relations between medical entities. We tested several approaches for each task. Our experiments showed that the approaches combining rule-based and machine-learning methods led to higher scores than classical rule-based or machine-learning techniques if applied alone. We achieved F-measures of 0.773 in concept extraction, 0.931 in assertion annotation, and 0.709 in relation annotation.

For each task, all the tests we made with the combination of both rule-based and machine-learning methods led to higher scores than using only one approach. Nevertheless, there is still room for improvement, especially to reinforce their complementary nature.

Funding The 2010 i2b2/VA challenge and the workshop are funded in part by grant number U54-LM008748 on Informatics for Integrating Biology to the Bedside from National Library of Medicine. This challenge and workshop are also supported by resources and facilities of the VA Salt Lake City Health Care System with funding support from the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204. MedQuist, the largest transcription technology and services vendor, generously co-sponsored the 2010 i2b2/VA challenge meeting at AMIA. This work was partially funded by project Akenaton under grant number ANR-07-TECSAN-001, and partly realized as part of the Quaero Programme, funded by OSEO, FRENCH State agency for innovation.

Provenance and peer review Not commissioned; externally peer reviewed

REFERENCES

1. **Mykowiecka A,** Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. *J Biomed Inform* 2009;**42**:923–36.
2. **Long W.** Lessons extracting diseases from discharge summaries. *AMIA Annu Symp Proc* 2007:478–82.
3. **Sager N,** Lyman M, Nhàn NT, *et al*. Medical language processing: applications to patient data representation and automatic encoding. *Meth Inform Med* 1995;**34**:140–6.
4. **Sager N,** Nhàn NT. The computability of strings, transformations, and sublanguage. In: Nevin BE, Johnson SM, eds. *The Legacy of Zellig Harris—Language and Information into the 21st century: Computability of Language and Computer Applications*. **Vol. 2**. Amsterdam: John Benjamins Publishing Company, 2002:79–120.
5. **Chapman WW,** Bridewell W, Hanbury P, *et al*. A Simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;**2001**:301–10.
6. **Light M,** Qiu XY, Srinivasan P. The language of bioscience: facts, speculations, and statements in between. In: Hirschman L, Pustejovsky J, eds. *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*. Boston, MA: Association for Computational Linguistics, 2004:17–24.
7. **Chapman W,** Dowling J, Chu D. ConText: an algorithm for identifying contextual features from clinical text. In: *Biological, Translational, and Clinical Language Processing*. Prague, Czech Republic. Association for Computational Linguistics, 2007:81–8.
8. **Kilicoglu H,** Bergler S. Syntactic dependency based heuristics for biological event extraction. In: *BioNLP'09: Proceedings of the Workshop on BioNLP*. Morristown, NJ: Association for Computational Linguistics, 2009:119–27.
9. **Uzuner O,** Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. *J Am Med Inform Assoc* 2009;**16**:109–15.
10. **Patrick J,** Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010;**17**:524–7.
11. **Friedman C,** Shagina L, Lussier Y, *et al*. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;**11**:392–402.
12. **Li Q,** Wu YFB. Identifying important concepts from medical documents. *J Biomed Inform* 2006;**39**:668–79.
13. **Denecke K.** Semantic structuring of and information extraction from medical documents using the UMLS. *Meth Inform Med* 2008;**47**:425–34.
14. **Lindberg DA,** Humphreys BL, McRay AT. The unified medical language system. *Meth Inform Med* 1993;**32**:281–91.
15. **Aronson AR,** Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229–36.
16. **Meystre SM,** Haug PJ. Comparing natural language processing tools to extract medical problems from narrative text. *AMIA Annu Symp Proc* 2005:525–9.
17. **Meystre SM,** Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008:128–44.
18. **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.
19. **Kudo T.** *CRF++: Yet Another CRF Toolkit*. http://crfpp.sourceforge.net/.
20. **Szarvas G.** Hedge classification in biomedical texts with a weakly supervised selection of keywords. In: *Proc ACL 2008: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008:281–9.
21. **Chang CC,** Lin CJ. *LIBSVM: A Library for Support Vector Machines*, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
22. **Lee C,** Khho C, Na J. Automatic identification of treatment relations for medical ontology learning: an exploratory study. In: McIlwaine I, ed. *Knowledge Organization and the Global Information Society: Proceedings of the Eight International ISKO Conference*. Würzburg, Germany: Ergon-Verlag, 2004.
23. **Ben Abacha A,** Zweigenbaum P. Automatic extraction of semantic relations between medical entities: application to the treatment relation. In: Collier N, Hahn U, Rebholz-Schuhmann D, *et al*, eds. *Proceedings of the Fourth International Symposium Mining in Biomedicine (SMBM2010)*. Cambridge, UK: CEUR Workshop Proceedings. 2010:**714**:1–8.
24. **Roberts A,** Gaizauskas R, Hepple M. Extracting clinical relationships from patient narratives. In: *BioNLP2008: Current Trends in Biomedical Natural Language Processing*. Columbus, OH: Association for Computational Linguistics 2008:10–18.
25. **Zhou G,** Su J, Zhang J, *et al*. Exploring various knowledge in relation extraction. In: *Proceedings of the 43rd Annual Meeting of the ACL*. Ann Arbor, MI: Association for Computational Linguistics 2005:427–34.
26. **Ben Abacha A,** Zweigenbaum P. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In: *12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING2011)*. Berlin and Heidelberg, Germany: Lecture Notes in Computer Science, Springer-Verlag 2011:**6608**:139–50.
27. **Kipper K,** Korhonen A, Ryant N, *et al*. A large-scale classification of English verbs. *Lang Res Evaluat J* 2008;**42**:21–40.
28. **McClosky D,** Charniak E. Self-training for biomedical parsing. In: *Proceedings of ACL 2008: HLT*. Columbus, OH. Association for Computational Linguistics, 2008:101–4.