Deep Parsing at the CLEF2014 IE Task (DFKI-Medical)

Tigran Mkrtchyan and Daniel Sonntag

German Research Center for AI (DFKI)
Stuhlsatzenhausweg 3, 66123 Saarbruecken, Germany
firstname.lastname@dfki.de

Abstract. We present an information extraction system for patient records which has been submitted to the ShARe/CLEF eHealth Evaluation Lab 2014 Task 2. The task was information extraction from clinical text in terms of a disease/disorder template filling process. The system uses a lexicalized parser to annotate grammatical relations between diseases, disorders, and other constituents on a sentence level. Grammatical pattern matching rules are applied in order to annotate the specifics of individual disease/disorder cases. High accuracy is most important for clinical decision support; the comparative results suggest that a deep parsing approach is suitable for this task, as we achieved $acc = 0.822$ and $acc = 0.804$ for the two runs of the system.

1 Introduction

We engage in medical information extraction in the medical domain and are supported by several national and international ”smart data” initiatives[1]. We are particularly interested in medical records which can be data mined in combination with clinical sensor data and image data towards multimedia information extraction and knowledge capture in ontologies[2] and medical cyber-physical systems[3]. Technically, we use a deep parsing approach (dependency parsing) which we will tune towards high-precision real-time information extraction in the next 3 years. Deep linguistic processing approaches differ from ”shallower” methods in that they yield more expressive and structural representations which directly capture long-distance dependencies and underlying predicate-argument structures. The main objective is to provide a hybrid information extraction (IE) platform based on handwritten rules in combination with semi-supervised machine learning approaches. Constituency and dependency parsing is key to our approach, revealing a multitude of linguistic features for making both rule-writing and classifier induction more effective. The features we obtain from the

[1] Federal Ministry of Education and Research (BMBF), Federal Ministry for Economic Affairs and Energy (BMWi), and European Institute of Innovation & Technology (EIT)

[2] http://www.dfki.de/~sonntag/courses/SS14/IE.html

[3] http://www.dfki.de/MedicalCPS/

sentence-level parsing step include, most notably, bilexical affinities, distant dependencies, and verb head information to identify more complex relations (not yet fully evaluated at CLEF) according to valency information of heads and dependents.

We are interested in extracting information from unstructured text, particularly in the medical domain. Medical reports contain huge amounts of data about medications, recommendations, procedures, etc. which are expressed mostly as narrative text. Such form of information is difficult to access. One of the approaches is identifying medical semantic relations [11]. A technology which extracts important data from text is the cornerstone for many clinical applications, such as populating multimedia databases (PACS, picture archiving and communication system) and summarizing medical records, required medical insurance reporting, or clinical decision support [8]. Extracting modifiers for given disease/disorder is an important task. Structured information can be more effectively accessed and processed, which will result in construction of a more intelligent medical system. We have implemented a system which extracts targeted information from clinical reports.

Thereby we extend Task 1 from [9], focusing on Disease/Disorder template filling. The challenge of the ShARe/CLEF eHealth Task 2 consists of extracting 10 semantic attributes from unstructured medical texts (440 patient records); a list of expected attribute values has been provided (such as 'yes' and 'no' for negation indicator.) For the Body Location Indicator, for example, UMLS concept unique identifiers (CUIs) should be extracted if mentioned, see `http://clefehealth2014.dcu.ie/task-2`. In this task, a patient record consists of 2 documents: one unstructured text file of the patient record itself and another pipe delimited template with disease/disorder annotations (disorder text span indexes are provided as well as default values of the attributes that modify the disorder). Each patient record contained 60 disease/disorders on average. Because of a prevalent sentence structure where a disease/disorder is mentioned, we expect that the usage of NLP techniques like POS tagging, chunking and especially syntax parsing are appropriate and key for obtaining a high accuracy (which is most important for clinical decision support).

The task of filling the IE template consists of providing slot values for each given disease/disorder combination. This results in the task of checking whether the sentences in the record contain modifiers in terms of attribute types (in our deep NLP approach, the dependents). Table 1 describes the IE task in terms of example sentences, given attributes and their norm slot values.

2 Approach

We will adopt the following terminology from [7] to refer to special types of NLP components. Language Resources (LRs) refer to data-only resources such as lexica, corpora, thesauri or ontologies. Processing Resources (PRs) refer to resources whose character is principally programmatic or algorithmic, such as text classifiers, part-of-speech taggers (POS taggers), named entity recognizers

Table 1. Attribute types with example sentences and their norm slot values.

Attribute Types	Example Sentences	Norm Slot Values
Negation Indicator	*Denies* numbness	yes
Subject Class	*Son* has schizophrenia	family-member
Uncertainty Indicator	*Evaluation* of MI.	yes
Course Class	The cough *worsened* last two weeks	worsened
Severity Class	He noted a *slight* bleeding	slight
Conditional Class	Return *if* fever	true
Generic Class	*Pain* while standing.	true
Body Location	Patient has *facial* rush	C0015450: Face
DocTime Class	Patient had *tumor* removed	before
Temporal Expression	The rash was present *for* 3 *days*	duration

(NERs) or grammatical parsers. PRs typically include LRs such as a lexicon. For the information extraction task in medical domain, we employ specific LRs and PRs.

1. In our system, a first preprocessing step identifies the sentence in which a mention of disease/disorder exists. Given the narrative text of a patient record and the start and end indexes of the disease/disorder span, we capture the exact sentence of the mentioned disorder.
2. The second step is POS tagging: we use an implementation of the Stanford Log-linear Part-Of-Speech Tagger [10], which first tokenizes the text, then generates the word lemmata for all tokens in the corpus and finally labels tokens with their POS tag. These POS tags are then used by almost every PR in the pipeline.
3. Then we run a rule-based PR for recognizing temporal expressions in order to annotate temporal expression attributes based on SUTime. SUTime [2] is a library for recognizing and normalizing time expressions, which outputs temporal tagging features. SUTime is a rule-based system which can be easily extended and adapted to special temporal expression extraction needs of idiosyncratic datasets such as the provided patient records. We use SUTime to find out the time, date and duration occurrences within the records. We have added several patterns as medical LRs for recognizing medical time and date expressions, such as 14/10, 08-09 etc.; the rules do also capture mentionings of type pairs of DD/MM or DD-MM.
4. The fourth and most important step is syntactic sentence parsing. First we run a constituency parser, which outputs the noun and verb chunks of the sentence. We consider only those chunks where the disease/disorder is mentioned. After constituency parsing, a dependency parser is used to output the grammatical relations in the sentence. Here we check for the disease/disorder to be the governor or the dependent in the relation. The Stanford Parser [5] PR is used for both parsing steps. This parser is probabilistic, which means that it outputs the most likely analyses of the sentences, retaining a lot of ambiguity in the result set.

5. The fifth step consists of running constituency-tree based regular expressions on constituency trees and semantic graph based regular expressions on dependency trees. Tregex [6] is a utility for performing pattern matching on tree structures and tries to match regular expression on constituency parse tree nodes (the name is short for tree regular expressions); Semgrex is a utility for identifying patterns in Stanford Dependencies[3]. These pattern matching approaches work very similar to simple string based regular expression matching, i.e., regex or regexp, but run on acyclic graph structures instead. Benefits are that expressions may not only involve usual regular expressions, but the grammatical relations in the sentence; POS tags of words and their named entities can be used in the patterns, too, which allows for very detailed and highly accurate extraction patterns. Figures 1 and 2 show medical examples of IE-relevant constituency and dependency parses, respectively. As one can see in figure 1, both diseases (underlined red) belong to different noun and verb phrase constituents (chunks).

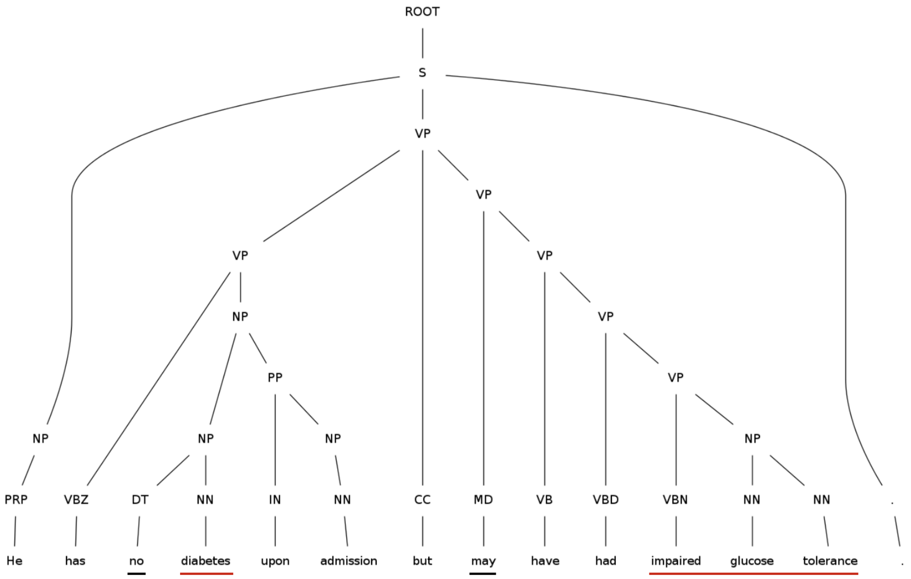


Fig. 1. Example of Parse approach: Constituency Parse Tree

6. The sixth step is running a look-up through our personal vocabularies (medical LRs), where synonyms of class labels, keywords of remaining classes are gathered, and if an occurrence of such a word is found in the output of Tregex and Semgrex the default attribute value is changed to the class label. In parallel to vocabulary look-up, we run MetaMap[1] via a web service API and check whether the lexical constituents within the disease/disorder

Fig. 2. Example of Parse approach: Dependency Parse Tree

noun chunks and their dependencies for any mentionings of "Body Location" can be identified and normalized (mapping of biomedical text to the UMLS metathesaurus). If this lookup is successful, we output the concept CUI as the body location feature. The IE pipeline is depicted in figure 3.

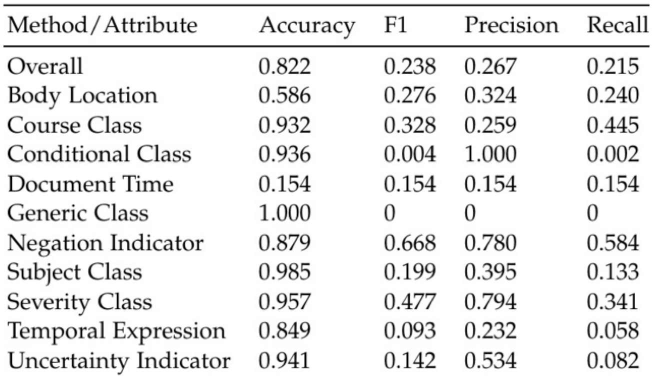
The derived relation extraction method based on syntax parsing has many advantages over purely string-based methods: instead of writing a huge amount of rules (for long dependencies) we can simply extract negation information with the correct scope, "no" is a modifier for the diabetes (Fig. 2). Dependency parsing captures the semantic predicate argument relationships between the entities in addition to the syntactic relationships (e.g., the scope of negation information). From the dependency parse tree we can imply that the modifier "may" (modal verb) has a grammatical relation to the head word "impaired", which is part of the disorder according to the governing head-driven dependency structure. This indicates a sentence structure driven appearance of an uncertainty indicator (UI) with a clearly defined scope and complements lexical uncertainty indicators like "evaluation of x". Our dependency rule experiments suggest that, unlike syntactic parsing, the semantic predicate argument relationships between the entities in addition to the syntactic relationships based on dependency parsing are useful in this domain (similar medical applications of dependency parses are reported in [4]). In addition, the rule set can be heavily reduced: Having written simple relation patterns between modal verb and the disorder can already annotate a comprehensive number of uncertain disorder indications. Overall, we have only written 25 generic patterns to annotate the entire set of attributes in the IE task (10 attributes) of the given disease/disorder, which results in a "low-cost" manual rule-writing ratio of about 2.5 patterns per attribute for the medical domain adaptation.

Fig. 3. IE pipeline

3 Results

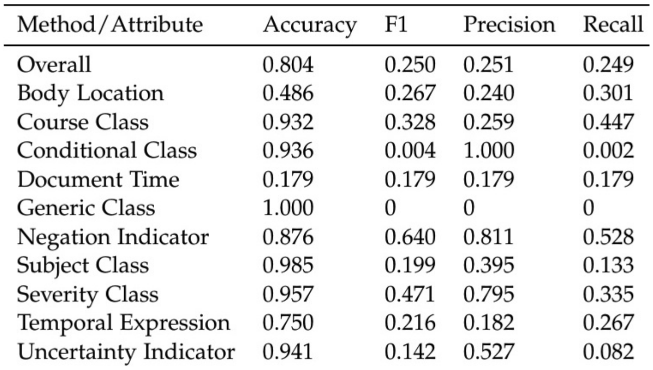
The main evaluation score was overall average accuracy for the task. Accuracy is the fraction of the number of correctly identified attribute:value slots (true positives and true negatives) and the total number of slot attributes. The IE task has been evaluated as per attribute type and on average.

The overall accuracy of our system is $acc = 0.822$ for the second run and $acc = 0.804$ for the first run. The attribute based recall, precision, and F-measure scores are shown in figures 4 and 5.



Method/Attribute	Accuracy	F1	Precision	Recall
Overall	0.822	0.238	0.267	0.215
Body Location	0.586	0.276	0.324	0.240
Course Class	0.932	0.328	0.259	0.445
Conditional Class	0.936	0.004	1.000	0.002
Document Time	0.154	0.154	0.154	0.154
Generic Class	1.000	0	0	0
Negation Indicator	0.879	0.668	0.780	0.584
Subject Class	0.985	0.199	0.395	0.133
Severity Class	0.957	0.477	0.794	0.341
Temporal Expression	0.849	0.093	0.232	0.058
Uncertainty Indicator	0.941	0.142	0.534	0.082

Fig. 4. System results for the second run



Method/Attribute	Accuracy	F1	Precision	Recall
Overall	0.804	0.250	0.251	0.249
Body Location	0.486	0.267	0.240	0.301
Course Class	0.932	0.328	0.259	0.447
Conditional Class	0.936	0.004	1.000	0.002
Document Time	0.179	0.179	0.179	0.179
Generic Class	1.000	0	0	0
Negation Indicator	0.876	0.640	0.811	0.528
Subject Class	0.985	0.199	0.395	0.133
Severity Class	0.957	0.471	0.795	0.335
Temporal Expression	0.750	0.216	0.182	0.267
Uncertainty Indicator	0.941	0.142	0.527	0.082

Fig. 5. System results for the first run

Teams were allowed to submit up to two runs of systems. In our case, the difference between the two runs is that in the first one for the first seven attributes we used predefined grammatical relations (domain-independent) in order to assess the performance of the IE engine on the attributes directly. For example, to annotate the severity class, the attribute was searched only in JJ (Adjective) relation within the disease/disorder governor node. For the Time Expression attribute we relied purely on SUTime; when looking for body location we sent all linguistic types of a sentence of the disease/disorder to the MetaMap API; and to output the document type attribute, we used document paragraph pattern output (i.e., if there is a mention of the word "history", then the document time is "before"). The second run works with more sophisticated and domain-specific patterns, e.g., looking for the attributes not only in grammatical relations among constituents, but also in noun phrase premodifiers. Here SUTime and MetaMap are also involved, but first we check if the word has a relation to the disease/disorder and consider them only in the positive case. Additionally, for the Document Time attribute, we rely on the verb's tense, which comes from the POS tagger, in the second run. The benefits of the parsing approach comes through when considering the accuracy for body location indicator. For the first run the accuracy was $acc = 0.486$ and for the second one $acc = 0.586$. Parsing the sentence and sending only the noun chunk to the MetaMap API increases the accuracy by 20%. Another important result is that the parsed attribute output results in higher precision, whereas the whole sentence approach results in higher recall.

Concerning temporal expression attribute results, we see that parser approach behaves better and has roughly 10% better overall accuracy when comparing to pure lexical search results. In the first run the temporal expression was being searched in the overall sentence, whereas in the second run only considers chunks where the disorder was mentioned. Just as for body location, the phrase chunked method results in higher precision in comparison to the whole sentence approach which results in higher recall. For the seven remaining attributes we can see that with deep parsing rules (second run) we have better accuracy in the majority of cases.

4 Summary

The system uses a lexicalized parser to annotate grammatical relations between diseases, disorders, and other constituents on a sentence level. The features we obtained from the sentence-level parsing step include, most notably, bilexical affinities, distant dependencies, and verb head information to identify more complex relations (major part of which is not yet evaluated at CLEF) according to valency information of heads and dependents. For writing domain-specific extraction patterns, we used Semgrex, an utility for identifying patterns in Stanford Dependencies [3]. Benefits are that expressions may not only involve usual regular expressions, but the grammatical relations in the sentence; POS tags of words and their named entities can be used in the patterns, too, which allows for very detailed and highly accurate extraction patterns: the benefits of the parsing approach are most evident when considering the accuracy for body location indicator. For the first run the accuracy was $acc = 0.486$ and for the second one $acc = 0.586$. Overall, we have (only) written 25 generic (Semgrex) patterns to annotate the entire set of attributes in the IE task (10 attributes) of the given disease/disorder, which results in a "low-cost" manual rule-writing ratio of about 2.5 patterns per attribute for the medical domain adaptation.

References

1. A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, pages 17–21, 2001.
2. A. X. Chang and C. D. Manning. SUTIME: A Library for Recognizing and Normalizing Time Expressions. In *Procedings of LREC*, 2012.
3. M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, pages 449–454, 2006.
4. A. O. Gunes Erkan and D. R. Radev. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of EMNLP-CoNLL*, pages 228–237, 2007.
5. D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430, 2003.
6. R. Levy and G. Andrew. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of LREC*, 2006.
7. D. Sonntag. Distributed NLP and Machine Learning for Question Answering Grid. In *Proceedings of the workshop on Semantic Intelligent Middleware for the Web and the Grid at ECAI*, 2004.
8. D. Sonntag, P. Wennerberg, P. Buitelaar, and S. Zillner. Pillars of ontology treatment in the medical domain. *Journal of Cases on Information Technology (JCIT)*, 11(4):47–73, 2009.
9. H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, and G. Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab. In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation, Multilinguality, Multimodality, and Visualization*, Lecture Notes in Computer Science, pages 212–231. Springer Berlin Heidelberg, 2013. 4th International Conference of the CLEF Initiative.
10. K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259, 2003.
11. S. Vintar, L. Todorovski, D. Sonntag, and P. Buitelaar. Evaluating context features for medical relation mining. In *Proceedings of the ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics*, 2003.