Received 31 March 2016
Revised 13 June 2016
Accepted 13 June 2016

Feature Engineered Relation Extraction – Medical Documents Setting

Ioana Barbantan, Mihaela Porumb, Camelia Lemnaru and Rodica Potolea
Technical University of Cluj-Napoca, Cluj-Napoca, Romania

Abstract

Purpose – Improving healthcare services by developing assistive technologies includes both the health aid devices and the analysis of the data collected by them. The acquired data modeled as a knowledge base give more insight into each patient's health status and needs. Therefore, the ultimate goal of a health-care system is obtaining recommendations provided by an assistive decision support system using such knowledge base, benefiting the patients, the physicians and the healthcare industry. This paper aims to define the knowledge flow for a medical assistive decision support system by structuring raw medical data and leveraging the knowledge contained in the data proposing solutions for efficient data search, medical investigation or diagnosis and medication prediction and relationship identification.

Design/methodology/approach – The solution this paper proposes for implementing a medical assistive decision support system can analyze any type of unstructured medical documents which are processed by applying Natural Language Processing (NLP) tasks followed by semantic analysis, leading to the medical concept identification, thus imposing a structure on the input documents. The structured information is filtered and classified such that custom decisions regarding patients' health status can be made. The current research focuses on identifying the relationships between medical concepts as defined by the REMed (**R**elation **E**xtraction from **Med**ical documents) solution that aims at finding the patterns that lead to the classification of concept pairs into concept-to-concept relations.

Findings – This paper proposed the REMed solution expressed as a multi-class classification problem tackled using the support vector machine classifier. Experimentally, this paper determined the most appropriate setup for the multi-class classification problem which is a combination of lexical, context, syntactic and grammatical features, as each feature category is good at representing particular relations, but not all. The best results we obtained are expressed as F1-measure of 74.9 per cent which is 1.4 per cent better than the results reported by similar systems.

Research limitations/implications – The difficulty to discriminate between TrIP and TrAP relations revolves around the hierarchical relationship between the two classes as TrIP is a particular type (an instance) of TrAP. The intuition behind this behavior was that the classifier cannot discern the correct relations because of the bias toward the majority classes. The analysis was conducted by using only sentences from electronic health record that contain at least two medical concepts. This limitation was introduced by the availability of the annotated data with reported results, as relations were defined at sentence level.

Originality/value – The originality of the proposed solution lies in the methodology to extract valuable information from the medical records via semantic searches; concept-to-concept relation identification; and recommendations for diagnosis, treatment and further investigations. The REMed solution introduces a learning-based approach for the automatic discovery of relations between medical concepts. We propose an original list of features: lexical – 3, context – 6, grammatical – 4 and syntactic – 4. The similarity feature introduced in this paper has a significant

International Journal of Web Information Systems
Vol. 12 No. 3, 2016
pp. 336-358

1744-0084
DOI 10.1108/IJWIS-03-2016-0015

influence on the classification, and, to the best of the authors' knowledge, it has not been used as feature in similar solutions.

Keywords Text mining, Data mining, Concept relation, Data correlation, Dependency tree parser

Paper type Research paper

1. Introduction

The rapidly growing interest in the assistive medical technology which supports patients to cope with their suffering such as hearing loss or hand tremors (Lee, 2015) or those in need of physical therapy (Smith, 2014) led to notable findings for the healthcare industry. But providing improved healthcare services using assistive technology is not limited to the medical devices intended for patients. The patients can benefit from upgraded medical care when the medical records enclosing their medical history, illnesses, allergies, interventions and several other related health characteristics become accessible at any time by the physicians. On these grounds, the electronic health record (EHR) systems have been introduced to deliver advanced medical services.

The consequences of a healthcare system are quantified by three factors: patient suffering, medical costs and time. The collection of data about patients modeled as a knowledge base gives more insight into each patient's health status and medical needs. The existence of a knowledge base of former patients previously investigated and diagnosed, benefits along all dimensions: healthcare, costs, diagnosis and hospitalization time. Therefore, the ultimate goal of a medical system is obtaining recommendations provided by an assistive decision support system using such knowledge base. The benefits, to name a few, are the decrease of a patient's suffering and a decrease in the number of medical investigations, qualified both as costs and time interval between the patient's hospitalization and start of treatment, thus initiating the healing process.

A source of trustworthy data is included in the EHRs. They enable identifying the relation between medical concepts, predicting epidemics or detecting cases of rare diseases (Fung *et al.*, 2014). Domain-oriented ontologies such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) (SNOMED-CT, 2012) or Unified Medical Language System (UMLS) (Bodenreider, 2004) encode the information captured in the EHRs that enable making predictions driven by assistive decision support systems. A clinical discharge document in raw format informs about document structuring into chapters containing grouped information concerning: symptoms, diseases, diagnosis, patient's historical information, medical procedures (Long, 2005), medication (Halgrim *et al.*, 2011), investigations, demographic data or follow-up information (Rudd *et al.*, 2010).

Knowledge extraction from unstructured medical records is an important task in the development of medical decision support systems. In this attempt, structuring documents and identifying relevant items in free text is the first challenge which in turn faces, among others, the difficulty of detecting negated terms.

The analysis of the medical textual data offers information like predicting adverse reactions by analyzing the interaction of drugs (when combined) or identifying co-morbidity risks, forecasting possible conditions that may occur based on previous studies and cases. The outcome can be a solution for recommending investigations for a thorough diagnosis, suggesting diagnosis or follow-up appointments. The availability

of large amounts of data seems convenient, but it may divert from focusing on the significant data.

The rest of the paper is organized as follows. Section 2 sets the background for our research with references to similar solutions proposed for handling the information extracted from the EHRs and relation identification. In Section 3, the conceptual approach for making predictions from discharge notes is introduced along with the relation extraction methodology. Section 4 describes the instantiation of such a solution, while the results are discussed in Section 5. The conclusions of work are drawn in Section 6, while the future enhancements for our proposed approach are discussed in the Section 7.

2. Background

Along with the EHR and EHR systems' adoption, several studies were conducted to evaluate their impact and the users' satisfaction (Edsall and Adler, 2008). By 2011, in the USA, the EHR systems had been adopted by 54 per cent of the physicians and 85 per cent of the adopters reported being contented with the systems, while one-half of the physicians not using an EHR system said they were planning on purchasing one. The indicators are provided in Jamoom *et al.*'s (2012) study. The increasing trend on EHR system adoption is introduced by Hsiao and Hing (2014) who report an 18 per cent EHR system adoption by the office-based physicians in 2001 that reached up to 78 per cent in 2013 in the same medical cohort.

For exploiting the information captured in the EHR, numerous annotation tools have been made available. Jonquet *et al.* (2009) present an ontology-based Web service for annotating biomedical textual information. A collection of over 200 biomedical ontologies and terminology repositories was integrated coming from the UMLS ontology repository (Bodenreider, 2004) and the National Center for Biomedical Ontology (NCBO) bioportal ontologies (Musen *et al.*, 2012). The authors propose a two-step mapping approach. First, a syntactic concept recognition step is employed using a dictionary of terms generated from the UMLS ontology repository and the NCBO bioportal ontologies. Then, the annotations were augmented with the knowledge extracted from other ontologies. A semantic distance method was computed to create new annotations considering the sibling relations defined in the ontologies, while an ontology-mapping component propagated the annotations based on the mappings between the ontologies. One challenge faced when mapping text to ontology is the ontology selection, as a consequence of the increasing number of available ontologies, as reported by Jonquet *et al.* (2010).

Extracting semantic relations from text is a crucial step toward natural language understanding and creating a structured representation of the content. Although the relation extraction task is a well-known problem, it is still not trivial. Applied to the healthcare domain, it gets even more difficult because of the lack of grammar rules and jargon-rich nature of the text. Some of the approaches dealing with relation identification between concepts in discharge summaries are reviewed below.

Two major lines of work in supervised approaches to relation extraction exist: feature-based methods, which propose a good set of features to use in the classification process, and kernel methods, which attempt to avoid the explicit computation of features by developing methods that are able to compare structured data (sequences, graphs and trees). Bunescu and Mooney's (2005) observations led to a kernel solution

based on the shortest path (SPK) between entities in a dependency graph. The kernel is based on the hypothesis that the words between the candidate entities or connecting them in a syntactic representation are expected to carry information regarding the relation. They proved their idea to be valid, and the subsequence kernel, which is an extension of the SPK, outputs very interesting results and even today it is still pointed out as a kernel with a very good performance in relation extraction tasks.

The task of relation identification is common in automated and semi-automated ontology development. Doing-Harris *et al.* (2015) exploit the synonymy and hierarchical relation existing between the concepts and use them to generate semantic vectors based on the tf/idf frequency.

Another use case for relation identification between concepts was proposed by Henriksson *et al.* (2014). They presented a solution for establishing relations between synonyms and abbreviations and their corresponding concepts from the medical domain. The generalization of the proposed approach derived from the use of semantic spaces extracted from two different corpuses of medical data, namely, a corpus of clinical documents and a corpus of medical journal articles. The performance measurements of the study are reported as recall: 0.39 for abbreviations to long forms, 0.33 for long forms to abbreviations and 0.47 for synonyms.

Albin *et al.* (2014) propose a method to identify the relations between medical concepts exploiting the UMLS ontology collection and implementing the onGrid Web platform that handles efficient transitive queries and conceptual relation. The relations were evaluated between any two sets of biomedical concept relations and the relations within one set of biomedical concepts. Their solution is exemplified on the disease–disease relation. The closeness between concepts is computed based on the semantic type of the concepts as defined in the UMLS. The relations are defined as weak when the concepts belong to the abstract types found closer to the root of the UMLS semantic network. For ordering the relations, they define a formula to identify the closeness between concepts consequently generating a relation matrix.

3. Knowledge extraction from medical documents

Knowledge extraction from unstructured medical records is an important task in the development of medical decision support systems. In this attempt, structuring documents and identifying relevant items in free text is the first challenge. Exploiting the medical documents (unstructured data) to identify the relations between the medical concepts, involves several methodologies from text mining to statistical methods, to supervised or unsupervised machine learning tasks. The decision of which methodology or ensemble is the most appropriate is made assessing both the benefits and the drawbacks. For example, a tradeoff is essential when selecting a rule- or learning-based approach which is in need of large amounts of labeled data, not that easy to acquire (text mining versus supervised machine learning). Medical diagnosis can be modeled as a combination of conditions and symptoms and their interaction. The relation between these medical findings helps discriminating the overlapping diagnoses, same as the presence of high fever leads to the diagnosis of pneumonia instead of the regular flu.

3.1 Methodology for implementing a medical assistive decision support system

The solution we proposed in our previous study (Bărbănţan and Potolea, 2015) for implementing an assistive decision support system follows the strategy represented in Figure 1. The approach takes as input any type of unstructured medical documents from EHRs to radiology reports or medical prescriptions. The input documents are first sent to a document analysis module where specific natural language processing (NLP) tasks are applied, followed by the semantic analysis. The output of the first module is represented by semantically enhanced data that can be used to extract the medical concepts. As soon as the concepts have been identified and assigned to a category, a structure for the input documents can be settled. In the attempt to provide a structure to the documents, the information must be grouped into sections, such as symptoms, diagnosis, mediation, follow-up appointments, investigations and medical history. The obtained structured information is filtered and classified such that custom decisions about the health status of the patients can be taken. The final objective triggers the type of solution.

The problems that can be solved using the proposed system are extracting the medical concepts and assessing the category they belong to, asserting the influence a medical concept has upon the current health status of the patient and predicting the diagnosis or treatment for a new patient. The extraction solution provides an outcome to the association between the medical concepts and their corresponding categories, asserting also whether the concepts are affirmative or negated in each context. Asserting the valence of the concepts was accomplished by a negation identification module, presented in more detail by Bărbănţan and Potolea (2014). The outcome of the extraction solution can be fed to other two modules dealing with knowledge extraction and prediction, one related to identifying patients with similar conditions and the other being focused on identifying the relations between the medical concepts.

The prediction task introduces the general setup for inferring knowledge based on previously analyzed and classified similar data. The approach is introduced by Bărbănţan and Potolea (2015). Prior to inferring data from the present EHR, two steps need to be considered. The information about previously investigated patients' needs to be loaded into a knowledge base to create a reference model. The knowledge base allows making informed decisions about the health status of the current patient, the treatment

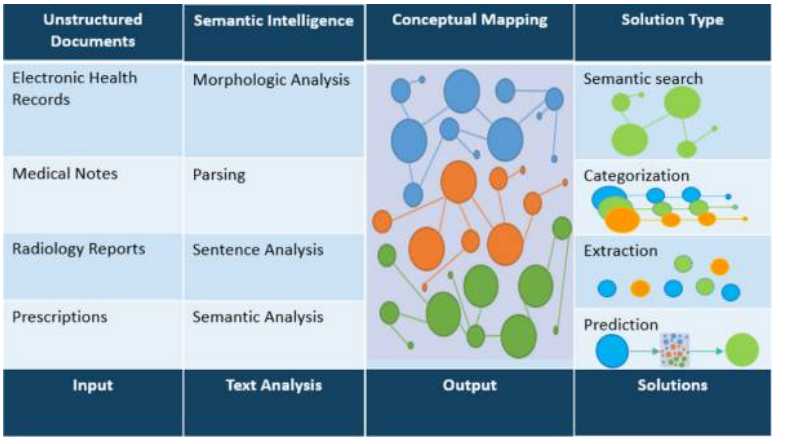

Figure 1.
An assistive decision support system

that should be administered or whether more investigations are required such that an accurate evaluation can is completed.

3.2 Approaches for relation extraction from electronic health records

The solution we are providing solves the task of identifying semantic relations between concepts in medical documents, more specifically medical discharge summaries. The starting point of our research was the 2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records, which involved three tasks:

(1) the extraction of medical concepts from patient reports;

(2) an assertion classification task focused on assigning assertion types for medical problem concepts; and

(3) a relation classification task focused on assigning relation types that hold between medical problems, tests and treatments.

The i2b2 and the VA provide an annotated reference standard corpus for the three tasks[1]. Uzuner *et al.* (2011) propose the relation extraction challenge aimed at recognizing three types of relations: treatment-problem, test-problem and problem-problem. The relations with examples are shown in Table I. Three general classes of relations between concepts are defined, each of them containing a different number of relation subtypes: for the treatment-problem relation – six relation subtytpes, for the test-problem – three relation subtytpes and for the problem-problem relation – two relation subtytpes.

3.2.1 Rule-based relation extraction. The initial approaches for relation identification between concepts were rule-based. One representative example, SemRep (Rindflesch *et al.*, 2000), was developed to identify branching of anatomical relations from reports and for detecting relations between medical problems and their treatments. The MedLEE approach presented by Friedman *et al.* (1994) is a combination of pattern matching rules and semantic grammars used to detect the nature of the relations. Rule-based approaches are not very robust, mainly because of the lack of generalization capacity; consequently, more recent approaches are focused on machine learning methodologies – both supervised and weakly supervised.

3.2.2 Feature-based relation extraction. Uzuner *et al.* (2011) used support vector machines (SVMs) for classifying semantic relations in medical discharge summaries. They presented a feature-based, fully supervised system, evaluated with macro F-score between 0.60 and 0.85 depending on the evaluated data. The features used for training included: surface features (distance, ordering of the concepts), lexical features (lexical trigrams, tokens in concepts) and syntactic features (verbs, syntactic bigrams).

The winning teams at the i2b2 workshop trained their solutions using SVMs (Grouin *et al.*, 2010; Patrick *et al.*, 2010; Roberts *et al.*, 2010; Solt *et al.*, 2010), thus SVMs become the first choice for a relation identification task. Anick *et al.*'s (2010) system used lists of n-grams; Demner-Fushman *et al.* (2010) used UMLS concept unique identifiers (CUIs) and exercised feature reduction through cross-validation; and Grouin *et al.* (2010) complemented their machine learning component with hand-built linguistic patterns and made use of simplified representations of text. Last but not least, de Bruijn *et al.* (2010) corrected for the label imbalance in the training data, calculated the "relatedness" of two concepts using pointwise mutual information in Medline and bootstrapped with

unlabeled examples. The most appropriate solutions to the relation extraction task belonged to Roberts *et al.* (2010), who used a supervised approach, and reported 0.737 F-measure, and the second best to de Bruijn *et al.* (2010), who developed a semi-supervised method and reported 0.731 F-measure. The authors trained three separate classifiers to classify treatment-problem, test-problem and problem-problem relations. They extracted context features similar to the ones by Uzuner *et al.* (2011), which were augmented with features extracted from MetaMap (Aronson and Lang, 2010) and cTakes (Savona *et al.*, 2010) taggers. Moreover, they approximated the relatedness of two concepts by calculating the pointwise mutual information between concepts as found in the Medline abstracts. They also submitted a semi-supervised system by applying bootstrapping on the unlabeled data, and they showed this added 0.4-point gain. Roberts *et al.* (2010) used a single SVM classifier to identify relations between concepts. They used several external resources such as Wikipedia, WordNet, General Inquirer and a relation similarity metric in the classification process. The lexical and contextual features proved to be very important in the relation extraction strategy as the F-score value decreases with 4 per cent when these features were not included in the training phase.

Relation type	Description and example
Type 1: treatment-problem relations	
TrIP	Treatment improves problem [Solu-Medrol]/tr was given for [tracheal edema]/pr
TrWP	Treatment worsens problem who presented with [acute coronary syndrome]/pr refractory to [medical treatment]/tr and [TNK]/tr
TrCP	Treatment causes problem [Allergies]/pr included [PENICILLIN]/tr and [IODINE]/tr
TrAP	Treatment administered for problem [antibiotic therapy]/tr for presumed [right forearm phlebitis]/pr
TrNAP	Treatment is not administered because of medical problem He was a poor candidate for [anticoagulation]/tr because of his history of [metastatic melanoma]/pr
NTrP	No relation between a treatment and a problem
Type 2: test-problem relations	
TeRP	Test reveals problem patient noted to have [acute or chronic hepatitis]/pr by [chemistries]/te
TeCP	Test conducted to investigate problem [chest xray]/te done to rule out [pneumonia]/pr
NTeP	No relation between a test and a problem
Type 3: problem-problem relations	
PIP	Medical problem indicates medical problem [Resting regional wall motion abnormalities]/pr include [mild inferior hypokinesis]/pr
NPP	No relation between two medical problems

Table I.
Relations between medical concepts referring diseases

4. Relation extraction from medical documents – the Relation Extraction from Medical documents solution

Concept relation identification represents a step toward establishing the structure of documents leading to a conceptual map for representing the documents. The part of speech tagger and the dependency parser give detailed information about the grammatical relation between the words and the grammatical units, while the relation between concepts offers valuable information that can be further exploited for modeling a domain or the documents used in the analysis. The concept relation identification process requires an initial preprocessing step where relevant concepts are identified and then fed to the relation identification module. Identifying the relations assists in predicting future behaviors or trends and recognizes the patterns in data. Nevertheless, identifying the relations between the concepts can be exploited as a learning tool. They are useful for identifying comorbidities and help understanding and learning medical conditions and inferring new relations between them.

This section shows in more depth the way in which we selected and computed the values for the proposed features. The input data were collected from two sources: the EHR content and the annotated medical concepts and their medical category. To create a data set for learning patterns, the data need to be paired among the two sources. The solution we proposed received as input clinical documents along with the annotated medical concepts. In the i2b2/VA-2010 challenge, manually annotated data were used, which allow the study of the relation identification problem, without worrying about the noise introduced by concept detection.

We proposed the REMed (**R**elation **E**xtraction from **Med**ical documents) (Porumb *et al.*, 2015) solution for extracting the semantic relations between medical concepts. The solution was modeled as a multi-class classification problem, as shown in Figure 2.

4.1 Data set description

The workshop organizers have provided two sets of discharge summaries, one obtained from the Beth Israel-Deaconess Medical Center (BIDMC), Boston, MA, and the other

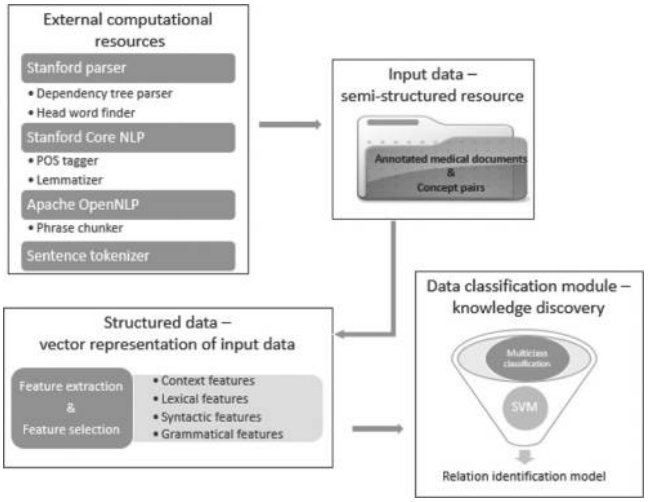

Figure 2.
The REMed flow

from Partners Healthcare, Boston, MA. The data sets consisted of the training set and the testing set. The volume of the data is 170 training documents with 3,118 relations and 256 documents for testing, with 6,292 relations. The instance distribution along the types of relations is presented in Figure 3. For the negative examples, we included 7,114 instances collectively referred to as None.

An example of how the relations are extracted from the sentences is presented in the following. Considering the input sentence S, which has been annotated with medical concepts and their types, we study all the possible concept pairs P1-P6 that will be further classified as relations, based on the assumptions:

S: If you experience [clear drainage] PROBLEM from [your wounds] PROBLEM, cover them with [a clean dressing] TREATMENT and stop showering until [the drainage] PROBLEM subsides for at least 2 days.

The candidate pairs are the following:

- *Pair1*: (clear drainage, your wounds);
- *Pair2*: (clear drainage, a clean dressing);
- *Pair3*: (clear drainage, the drainage);
- *Pair4*: (your wounds, a clean dressing);
- *Pair5*: (your wounds, the drainage); and
- *Pair6:* (a clean dressing, the drainage).

The task is to correctly extract the relations between each candidate pairs. In this case, the solution is: P1- PIP, P2-TrAP, P3-None, P4-TrAP, P5-None, P6-None, where None stands for no relation.

4.2 **R***elation* **E***xtraction from* **Med***ical documents feature definitions*

The goal of the REMed solution is to identify patterns that lead to the identification of concept pairs and the classification of the concept-to-concept relations. Nevertheless, identifying the patterns leads to finding other remarkable relations in the data.

The solution of extracting knowledge from EHRs follows the general mining process as stated also by Alag (2009). The first step in developing the learning strategy is understanding the purpose of the solution and defining a strategy of achieving it in the setup imposed by the content. Identifying the relations between concepts becomes a

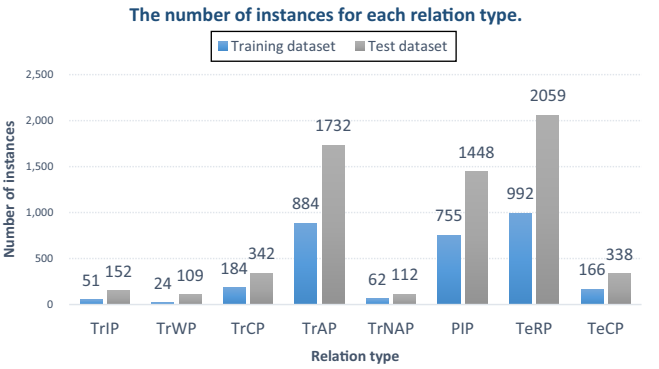

Figure 3.
The number of instances for each relation type in the training and test data sets

problem of finding patterns in data. The generated data set is highly dimensional and sparsely populated, and is made up of document vectors. The created data set was evaluated in several feature setups, and to define the REMed model, a best feature setup has been identified by considering the tradeoff between precision and recall.

As the objective is to identify relations between concepts, we conducted our analysis using only the sentences from the provided data that contain at least two medical concepts, thus limiting the scope within the sentence (i.e. no relations assumed between concepts in different sentences). This limitation is consistent with the approaches in literature (Uzuner *et al.*, 2011). The types of candidate concepts are: *problem*, *test*, and *treatment* and a relation is computed by any combination of problem and any of the three concepts. The relation identification approach relies on feature engineering. For each pair of concepts, several features were extracted. The feature vector used in our approach was built starting from the bag of words representation of the input data and progressively enhanced with features grouped into the following categories: context, lexical, syntactic and grammatical. The features were typically Boolean, but for a few, the type was integer or real. When extracting an n-gram feature, the name of the feature became the n-gram and the value assigned, Boolean. We will refer to this feature type as Boolean n-gram. To exemplify the content of the feature categories, in the statements below, the following notations are used:

concept(x) – *x* – any medical concept: problem, test, treatment
token(x) – *x* – any word or punctuation mark
trigram(x) – *x* – any list of 3 consecutive words in the input sentence
t(x) – *x* – either a test or treatment concept

4.2.1 Context features. The context features capture the word position in the sentence and the distance between the concepts. The number of concepts in a sentence influences the types of existing relations, and our analysis showed that a sentence can contain more than two concepts. The context features category includes the following features.

Number of concepts (integer). Counts the number of concepts in a sentence:

$$\forall x, concept(x) \rightarrow count(x)$$

Exactly two concepts (Boolean). Although redundant, indicates the existence of other concepts (besides the two used to define the pair) in the evaluated concept pair:

$$\forall x, y,\ concept(x),\ concept(y) \rightarrow \nexists z\ such\ that\ concept(z)$$

Inner concepts (Boolean n-gram). The feature marks the existence of concepts in the list of terms between the selected concepts in the pair:

$$\forall x, y, z, concept(x), concept(y), concept(z), pair(x, y) \rightarrow precedes(z, x) \wedge precedes(y, z)$$

Concepts distance (integer). Counts the number of tokens between the concepts:

$$\forall x, y, z_i, concept(x), concept(y), token(z_i) \rightarrow count(precedes(x, z_i) \wedge precedes(z_i, y))$$

>Concepts order (Boolean). The feature evaluates whether the order in which the concepts occur in the sentence is a problem followed by test or treatment:

$$\forall x, y, problem(x), t(y) \rightarrow sequence(x, y)$$

Relation type (integer). Once the concepts have been identified, an initial assumption about their category is made, while the actual existence of the relation between the concepts is established by the classification process:

$$\forall x, y, conceptPair(x, y), problem(x), test(y) \rightarrow 1$$
$$\forall x, y, conceptPair(x, y), problem(x), treatment(y) \rightarrow 2$$
$$\forall x, y, conceptPair(x, y), problem(x), problem(y) \rightarrow 3$$

4.2.2 Lexical features. The lexical features heavily increase the size of the feature vector. Each extracted value is a lemmatized n-gram that becomes the name of the feature and it has assigned a Boolean value.

Concept lemmas (Boolean n-gram). The feature extracts the lemmas of the concepts:

$$\forall x_i, concept(x_i) \rightarrow lemma(x_i)$$

Lexical trigrams (Boolean n-gram). The feature marks the existence of a particular sequence of lemmas in the surrounding area of a concepts:

$$\forall x, a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2, c_3, concept(x), concept(y), sequence(a_i, x, b_i, y, c_i) \rightarrow lemma(a_i) \cup lemma(b_i) \cup lemma(c_i)$$

Inner tokens (Boolean n-gram). The feature extracts all the consecutive tokens (words or punctuation marks) between the selected concepts:

$$\forall x, y, z_i, concept(x), concept(y), token(list(z_i)), sequence(x, list(z_i), y) \rightarrow list(z_i)$$

4.2.3 Syntactic features. In addition to the lexical features, the syntactic features capture more details about the text surrounding the concepts. For all the syntactic features, we used additional shallow syntactic information in the sentence.

Verb lemmas (Boolean n-gram). The feature extracts the lemma of the verbs identified in between the concepts:

$$\forall x, y, z_i concept(x), concept(y), verb(z_i), sequence(x, z_i, y) \rightarrow lemma(verb(z_i))$$

Inner prepositions (Boolean). The prepositions indicate a relation with the nouns or pronouns and while the concepts defining a pair are typically nouns, the existence of a preposition in the vicinity of concepts increases the likelihood of a relation:

$$\forall x, y, z_i, concept(x), concept(z_i), token(z) \rightarrow \exists preposition(z_i)$$

Inner conjunctions (Boolean). Conjunctions connect words or phrases; thus, their presence influences the existence of a relation between the concepts:

$$\forall x, y, z_i, concept(x), concept(z_i), token(z) \rightarrow \exists conjunction(z_i)$$

Phrase chunk (Boolean). The feature representation is a concatenation of phrase chunks encountered on the path between the relation arguments expressed as syntactically correlated groups.

$$\forall x, y, z_i, concept(x), concept(y), sequence(x, z_i, y) \rightarrow$$
$$phraseChunk(x) \wedge phraseChunk(z_i) \wedge phraseChunk(y)$$

4.2.4 Grammatical features. The grammatical features are constructed considering the grammatical relations at sentence level.

Path length (integer). The number of elements contained in the grammatical path between the concepts, generalizes the previously extracted context feature, "*Concepts distance*" changing the actual terms to the corresponding POS sequence:

$$\forall x, y, concept(x), concept(y) \rightarrow$$
$$countGrammaticalTerms(shortestPath(x, y))$$

Path sequence (Boolean n-gram). The feature value is a Boolean that marks the existence of the grammatical path between the concepts:

$$\forall x, y, concept(x), concept(y) \rightarrow shortestPath(x, y)$$

Shortest path similarity (real). The shortest path similarity measures a similarity based on the shortest path between the relation arguments on the dependency graph:

$$\forall x, y, concept(x), concept(y) \rightarrow computeShortestPathSimilarity(x, y)$$

Head word (Boolean n-gram). When a concept consists of several terms, one of them is the more relevant in that particular concept phrase:

$$\forall x, conceptPhrase(x) \rightarrow headWord(x)$$

4.3 **R***elation* **E***xtraction from* **Med***ical documents feature extraction methodology*

The context and lexical features are easily computed by analyzing the relative position of the concepts in the sentence and the lexical structure of the sentences. For the grammatical and syntactic features, however, additional resources were required such as the lemmatization module from the Stanford Core NLP toolkit (Manning *et al.*, 2014) used to identify the part of speech tags.

We used the Stanford Dependency Parser (de Marneffe *et al.*, 2006) for extracting the grammatical structure of the sentence. We exploited the Stanford Dependency Parser for extracting three features, all having as input the labeled dependency graph representation of the sentences. The advantage of using the dependency parser is that a

representation of a sentence as a labeled dependency graph contains rich semantic information that can indicate possible relations between the entities in the sentence. The Stanford dependencies map straightforwardly onto a directed graph representation in which the words in the sentence become nodes and the grammatical relations become edge labels (de Marneffe and Manning, 2008). For the feature representation, we used the Stanford Parser's collapsed representation for typed dependencies.

The parser provides the Universal Dependencies (UDs) presented in detail by de Marneffe *et al.*, 2014, the Stanford Dependencies output and the phrase structure trees. In our experiments, we used the 3.5.2 parser version which, by default, outputs the UDs. We chose the UD representation from the Stanford Parser. This representation of grammatical relations might be considered a standard for describing the grammatical structure of a sentence. The UD topological representation of relations is relevant to us because we used it in the edge similarity measure for the shortest dependency path similarity feature.

Although the feature-based methods lead to good performance, we wanted to exploit the kernel methods, as well. That is why, we extracted a similarity feature based on the shortest path between the relation arguments on the dependency graph. We chose the UD representation from the Stanford Parser to evaluate the feature. Similar to the presented solution by Uzuner *et al.* (2011), we proposed a similarity feature that extracts information by comparing two dependency paths corresponding to different instances. First, we extracted the path between the head words of the relation argument using the CollinsHeadFinder solution. The Stanford algorithm implements a "semantic head" variant of the English HeadFinder introduced in Michael Collins' 1999 thesis (Collins and Mitchell, 1999). Using a similarity metric, we searched in the training data set for the most similar instances to the current instance (in our example the TREATMENT-PROBLEM). Experimentally, we considered that the first 20 instances with the highest similarity score are relevant in the next computation step. Based on these 20 training instances, we computed the frequency for each relation category. In the end, the similarity feature indicates the percentage of similar relations for each relation type. We used two different similarity metrics for the token and relation sets from the path representation (1), and the final score was obtained using formula in equation (1):

$$\begin{aligned} SimilarityScore &= \begin{Bmatrix} 0, m \neq n, \\ \prod_{i=0}^{n} similarity(x_i, y_i), m = n. \end{Bmatrix} \\ similarity(x_i, y_i) &= \begin{Bmatrix} tokenSim(x_i, y_i), i\%2 \neq 0, \forall 1 \leq i \leq n \\ edgeSim(x_i, y_i), i\%2 = 0. \end{Bmatrix}. \\ tokenSim\,(x_i, y_i) &= |x_i \cap y_i| \end{aligned} \tag{1}$$

$$edgeSim(x_i, y_i) = \begin{Bmatrix} 6, \ if\, x = y, \\ 3, \ if\, x, y \in same\ category\ in\ UD\,(eg.\ nsubj, dobj), \\ 2, \ if\, x, y \in same\ category\ in\ UD\,(eg.\ nsubj,\ csubj), \\ 1, \ otherwise \end{Bmatrix} \tag{2}$$

For tokens' set, the similarity is computed as a simple intersection, and for the relation set, we used an edge similarity formula, defined in equation (2).

5. Evaluation of the Relation Extraction from Medical documents solution

To prove the strength of the proposed feature categories, we evaluated our solution in several experimental setups, assessing the impact of each feature category, the significance of each feature using a ranking approach and finally reporting our results in contrast to the similar relation extraction solutions.

5.1 Evaluation of the **R***elation* **E***xtraction from* **Med***ical documents solution on each feature category*

We evaluated our proposed relation extraction solution REMed in two setups as defined by Roberts *et al.* (2010), namely, M1 and M2 evaluation types. M1 evaluates the performance of the REMed solution when identifying each individual relation type (2010 i2b2 Challenge evaluation method), while M2 evaluates the discriminative power of the solution when having to discriminate whether a relation exists between two concepts or not. The micro-averaged F1-measure is computed as a harmonic average of the micro-averaged precision and micro-averaged recall. Because in our task precision and recall were considered equally important, we conducted the analysis using the F1-measure. To evaluate the discriminatory power of the features, we performed several experiments and reported the results as precision (Figure 4), recall (Figure 5), F1-measure (Figure 6) and an overall evaluation is reported as M1 measure in Figure 7.

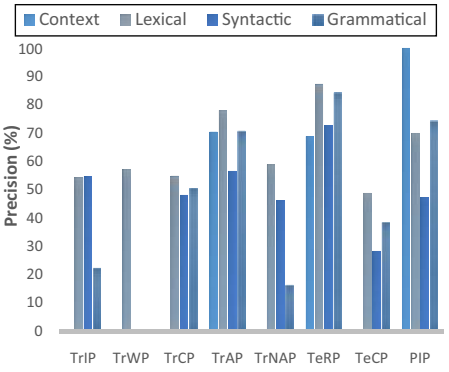

Figure 4.
Evaluation of feature effectiveness – Precision

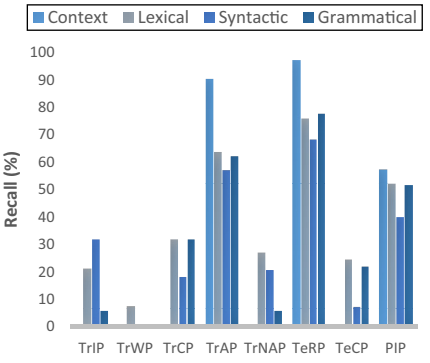

Figure 5.
Evaluation of feature effectiveness – Recall

In Figures 4-6 on the abscissa are represented the relation categories with evaluations for each feature category: context, lexical, syntactic and grammatical, while the ordinate represents the performance measurement as per cent.

5.1.1 Lexical features. The best results are achieved by the lexical features, with an overall F1-measure of 66.52 per cent. The lexical features are good for identifying the TrNAP features, and are the only ones that can identify the TrWP relation.

5.1.2 Context features. Used alone, the context features are not able to identify the minority classes (TrIP, TrWP, etc.), but they achieve good results for the general relation categories (TrAP, TeRP, PIP). The context features show the best classification performance for the PIP relation, but completely fail to identify the TrIP, TrWP, TrCP, TrNAP and TeCP relations.

5.1.3 Syntactic features. Although the results obtained classifying the relations using the syntactic features individually are fairly worse compared to the lexical features, they help improve the identification of the TrIP relation. The syntactic features are the best at identifying the TrIP features.

5.1.4 Grammatical features. While the grammatical features do not show the best performance in identifying any of the relations, when used in conjunction with the other features, they lead to important improvements.

The lexical, context and syntactic feature categories are relatively computationally inexpensive (compared to the dependency features in the grammatical category), they bear great importance in real system construction, and we expect that by adding more advanced features, the overall performance improves. The experimental results proved

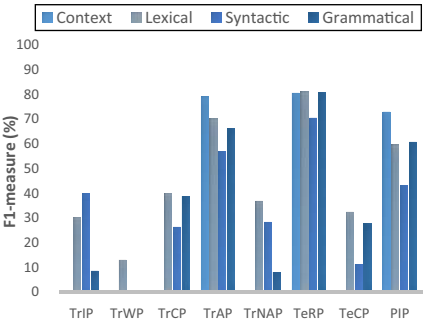

Figure 6.
Evaluation of feature effectiveness – F1-measure

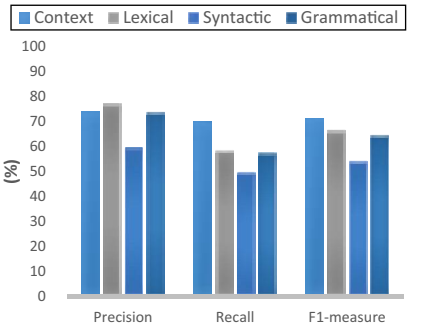

Figure 7.
M1 methodology for feature effectiveness

that the information extracted using the lexical and grammatical feature subsets only, led to achieving similar performances. The motivation of using a combination of feature categories is given by the classification results obtained by each single category in isolation. We consider the classes TrAP, TeRP, PIP as easily recognizable, as they are identified by each of the feature categories. Being the majority classes, thus well-represented, this is rather an expected behavior.

5.2 Feature ranking

The following experiment we conducted is oriented toward the analysis of the importance of the features using a ranking algorithm, and for this task, we used the GainRatioAttributeEval as attribute evaluator combined with the Ranker as search method in their default configuration form Weka Data Mining tool (Hall *et al.*, 2009). The outcome of the feature ranking showed that the most important features are the lexical trigrams, the verb lemmas and the tokens that appear between concepts. Our developed similarity features are ranked in the top 500 features from a total of almost 16,500 features that make up the feature vector. Because of the poor relation classification achieved in this setup (less than 60 per cent), we attempted fine-tuning the feature vector by examining the influence of each proposed feature on the overall relation classification task. Table II shows how subsets of features from different categories influence the overall classification results. The table does not contain all the stated features in Section 4.2, but only the ones whose presence showed significant differences in evaluating the performance.

Because in our task precision and recall were considered equally important and we do not want our solution to be biased toward precision or recall, we used the F-measure as the evaluation metric. The best feature vector according to these experiments includes a mix of features from each category (Table II – last line), as follows. All the features from the context feature category were included, as they proved to have a good discriminative behavior, improving the F-measure value with over 60 per cent. From the lexical features category, the lexical trigrams and the features related to the number of concepts between the concepts proved to have a positive influence in classification, while the words found between the concepts did not bring relevant information. The *Tokens in Concepts* feature does not improve the performance, but degrades it, in comparison to *Head Words*. This shows that the head word is important when analyzing a medical concept, but not all the tokens that are used to express it. For example, in the labeled concept is "her pain", what we are interested in "pain" whereas "her" could introduce noise that might drop the classifier's performance. We can also point out that the most important features are included in the context and lexical categories. Several tests were performed to identify the most appropriate combination of verb lemmas, and in the end the successful association was to extract the verbs to the left of the first concept (the table symbol <), between the concepts (< >) and the ones after the second concept (>) from the concept pair. The same observation applies to the lexical trigrams, as the most effective setup was to use a window of three words before and after the concept pair.

While all the grammatical features are used in the best feature subset combination, the *Syntactic trigrams* from the syntactic features category proved to have a rather disappointing behavior for the relation classification task. We performed multiple tests with small variations, such as: take into account up to three POS to the right and to the left of the concepts, but none of them was successful. Moreover, we tested with

Table II.
The impact of subsets of features on the classifier performance.

Features																				Results		
Context/surface features			Lexical features										Syntactic features						Grammatical structure features				
Words between concepts WBC	Concepts order sensitive COS	Concept between candidate concept CBCC	Relation type RT	Inter concept tokens ICT	Tokens in concepts TIC	Lexical Trigram uni LT	bi LT	tri LT	Syntactic Trigram uni ST	bi ST	tri ST	< VL	<> VL	> VL	Phrase chunks between concepts PCBC	Prepositions between concepts PBC	Conjunctions between concepts CONJ	Head words HW	Path length + Path itself LP	P	R	F
X	X	X	X	X	X	X	X	X				X	X	X	X					87.3	55.9	68.2
X	X	X	X	X	X	X	X	X				X	X	X	X	X				87.8	56.8	69
X	X	X	X	X	X	X	X	X				X	X	X	X	X		X		87.6	56.2	68.5
X	X	X	X	X		X	X	X				X	X	X	X	X		X		88	57.4	69.5
X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X		X		87	55.8	68
X	X	X	X	X		X	X	X				X	X	X	X	X	X	X		88	57.3	69.4
X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X		87.5	56	68.2
X	X	X	X	X		X	X	X				X	X	X	X	X		X	X	**88**	**57.5**	**69.6**

generalized POS (noun, verb, adjective, etc.) without any improvement in the classification performance. The results of the classification are shown in Table III.

5.3 Comparison of **R***elation* **E***xtraction from* **Med***ical documents with similar solutions*

In this section, we present the experimental results obtained by incorporating the knowledge sources mentioned in the features section. The overall results along with the influence of each feature on the classification of the individual relations are presented. The classification was performed using the LibLINEAR implementation for the SVM. LibLINEAR (Fan *et al.*, 2008) is an extension of the libsvm (Chang and Lin, 2011) library that implements a linear kernel for large data classification, to achieve significant speed gain. In our experiments, we set the epsilon termination parameter to 0.5. The class weight associated with the NPP class was set to 0.025 to decrease the significance of those concept pairs. The weight for the other classes was 1.0.

Our best setup achieves an overall 74.9 per cent micro-averaged F1-measure, which outperforms the first system (micro-averaged F1-measure of 73.7) submitted to the i2b2/VA-2010 challenge. This is mainly because of the similarity feature based on dependency path, as we will further show. Comparing our system to the one proposed in (Roberts *et al.*, 2010), in which it was additionally measured how well the system identified whether there is any relation between entities, it can be noticed, in Figure 8, that our solution performs better because of an improved F1-measure.

Our method obtained the best results on the following relations: TrWP, TrAP, TrNAP and PIP. TrNAP along with TrWP are the two relations with the smallest data available in both the training and test data sets (accounting only 1.7 per cent TrWP and 1.8 per cent TrNAP in the test set). The fact that we improved the F-measure for the

Evaluation	P (%)	R (%)	F1 (%)
TrIP	59.15	27.63	37.66
TrWP	50.0	4.58	8.4
TrCP	64.80	44.15	52.52
TrAP	85.08	74.76	79.59
TrNAP	56.0	25.0	34.56
TeRP	90.54	79.06	84.41
TeCP	63.05	29.28	40.0
PIP	95.38	62.77	75.71
Relation existence (M2)	81.6	81.8	81.6

Table III. Results of the relation identification and classification

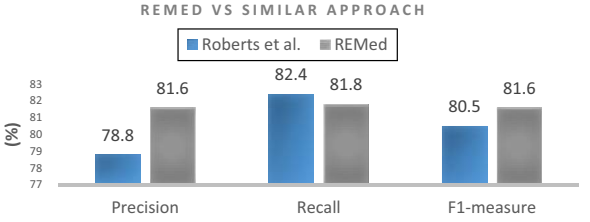

Figure 8. Comparative results on the M2 evaluation type between the system proposed by Roberts *et al.* (2010) and REMed

relation discovery of these two relation types shows that our method provides a small enhancement on discovering relations with very low frequency in the training set. But still, the results for these two relation types are very poor, and that can be explained by the very small number of training instances in these two cases.

5.4 Misclassification analysis

In the attempt to improve the performance, a close exploration of the misclassification sources was performed. The approach started with the analysis of the confusion matrices, and the following misclassification sources have been established. Most errors origin in the difficulty of discriminating among the minority classes (TrIP, TrNAP, TrWP, TrCP, TeRP) and classifying them in one of the majority ones (TrAP, TeCP, PIP, None). For example, for the TrIP relation (treatment improves problem), 27.63 per cent of the instances are correctly classified as TrIP, but 28.94 per cent of them are classified as TrAP, and 26.31 per cent as None. The difficulty to differentiate between a TrIP and TrAP relation resides in the hierarchical relation between the two classes as TrIP is a specific relation of TrAP. We state that the most likely reason the classifier cannot discern the correct relations is because of the aforementioned bias the classifier has toward its majority classes. Similar bias has been stated by other participants to the i2b2/VA 2010. A possible approach to reduce the negative outcomes because of the class imbalance could be applying informed sampling strategies such as synthetic minority over-sampling technique (SMOTE).

6. Conclusions

The research that we carried out demonstrated that the current solutions for relation identification between medical concepts are a combination of machine learning and pattern matching techniques. Similar to acknowledged approaches, we proposed the REMed solution which is expressed as a multi-class classification problem. To distinguish between the instances belonging to different classes and identify the similar instances, we developed an extended set of features starting from the bag of words representation of the training corpus. We enhanced this set with several features for which we propose an original classification into four main categories: context features, lexical features, syntactic features and grammatical features. Based on the experiments we performed with different subsets of features, we identified the most suitable setup for the multi-class classification problem, as shown in Table II, last line. In this particular selection of features, the similarity feature has a significant influence in the classification, and, to the best of our knowledge, it has not been used as a feature in similar solutions. The best results we obtained are expressed as F-measure of 74.9 per cent which is 1.4 per cent better than the results reported for similar systems. A more in-depth analysis was performed on the individual feature categories that proved that the categories that have the greatest discriminatory value are the lexical and context categories, while the features *Tokens in Concepts* (lexical feature category) and *Syntactic trigrams* (syntactic features category) did not improve the classification, but at the same time, it did not decrease the performance. Because the distribution of the instances in our training data is not homogeneous, the classes that are not well represented proved to have small performance.

The solution of identifying relations between medical documents can be further explored on a different dataset to identify the relation between treatment, genes and conditions. We propose using as source solution the REMed strategy for identifying problem-treatment-test relations and apply it on the target solution represented by the new medical data set.

7. Further work

Our strategy aims to be a step toward a medical assistive decision support system: starting from raw medical data, it infers the appropriate suggestion to each specific task (further investigations, diagnosis or medication). The medical documents which are usually stored in unstructured format can be structured using a terminology mapping technique. The required preprocessing steps proved to have a significant role in normalizing both the input text (unstructured data) and the terminology sources (structured data). The filtering step which discriminates between medical and non-medical concepts via the WordNet dictionary proves to be an efficient method for filtering the non-medical concepts. In the selection of the terminology sources (WordNet and SNOMED-CT), their ability to cover the biomedical domain and also to obtain accurate information was considered.

The current status covers complete solutions for automatically structuring medical documents and extracting relevant medical concepts via the PreNex (Bărbănţan and Potolea, 2014) and MedCIM (Bărbănţan *et al.*, 2015) strategies while the knowledge extraction and prediction tools are under development. Our efforts are focused on identifying a methodology for extending the solution to be able to address other relations between medical concepts. Also, including additional medical information like medical history or demographics could lead to the identification of further relations.

Switching the perspective to an unsupervised approach lead us to start investigating the benefits brought by the Word2Vec algorithm proposed by Mikolov *et al.* (2013). Our strategy consists in building a model from the EHRs and automatically generating patterns for building relations between medical concepts. The goal is represented by the identification of new relations between the medical concepts, thus extending the relation knowledge base. The expected output includes relations between the medical history and the health condition of the patient or the relation between demographics and the development of the patient's health status.

Note

1. De-identified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr Ozlem Uzuner, i2b2 and SUNY.

References

Alag, S. (2009), *Collective Intelligence in Action*, Manning Publications, Greenwich, CT.

Albin, A., Ji, X., Borlawsky, T.B., Ye, Z., Lin, S., Payne, P.R.O., Huang, K. and Xiang, Y. (2014), "Enabling online studies of conceptual relationships between medical terms: developing an efficient web platform", *JMIR Medical Informatics*, Vol. 2 No. 2, p. e23.

Anick, P., Hong, P. and Xue, N. (2010), "I2B2 2010 challenge: machine learning for information extraction from patient records", *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, *Boston, MA*.

Aronson, A. and Lang, F. (2010), "An overview of MetaMap: historical perspective and recent advances", *Journal of the American Medical Informatics Association*, Vol. 17 No. 3, pp. 229-236.

Bărbănţan, I. and Potolea, R. (2014), "Exploiting Word Meaning for Negation Identification in Electronic Health Records", *IEEE International Conference on Automation, Quality and Testing, Robotics*, *Cluj-Napoca*, pp. 283-289, doi: 10.1109/AQTR.2014.,6857880.

Bărbănţan, I., Lemnaru, C. and Potolea, R. (2015), "Concepts identification in medical documents", *17th International Symposium on Health Information Management Research – ISHIMR*, *New York, NY*.

Bărbănţan, I. and Potolea, R. (2015), "Towards cross language morphologic negation identification in electronic health records", in Ramón Agüero, T.Z.G.G. (Ed.), *Mobile Networks and Management, Chapter: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Vol. 141, Springer International Publishing, Würzburg, pp. 417-430, doi: 10.1007/978-3-319-16292-8_30.

Bodenreider, O. (2004), "The Unified Medical Language System (UMLS): integrating biomedical terminology", *Nucleic Acids Research*, Vol. 32 (Database issue), pp. D267-D270.

Bunescu, R. and Mooney, R. (2005), "A shortest path dependency Kernel for relation extraction", *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, *Vancouver, British Columbia*, pp. 724-731.

Chang, C. and Lin, C. (2011), "LIBSVM: a library for support vector machines", *Journal ACM Transactions on Intelligent Systems and Technology (TIST)*, New York, NY, Vol. 2 No. 3.

Collins, M. and Mitchell, P. (1999), *Head-driven Statistical Models for Natural Language Parsing*, University of Pennsylvania, Philadelphia, PA.

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J. and Zhu, X. (2010), *NRC at i2b2: One Challenge, Three Practical Tasks, Nine Statistical Systems, Hundreds of Clinical Records, Millions of Useful Features*, Boston, MA.

de Marneffe, M. and Manning, C.D. (2008), *Stanford Typed Dependencies Manual*, LREC, Manchester.

de Marneffe, M., MacCartney, B. and Manning, C.D. (2006), *Generating Typed Dependency Parses from Phrase Structure Parses*, LREC, Genoa.

de Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. and Manning, C. (2014), "Universal Stanford dependencies: a cross-linguistic typology", *Language Resources and Evaluation Conference (LREC)*, *Reykyavik*.

Demner-Fushman, D., Apostolova, E., Islamaj, D.R., Lang, F.M., Mork, J., Neveol, A., Shooshan, S., Simpson, M. and Aronson, A. (2010), "NLM's system description for the fourth i2b2/VA challenge", *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, *Boston, MA*.

Doing-Harris, K., Livnat, Y. and Meystre, S. (2015), "Automated concept and relationship extraction for the semi-automated ontology management (SEAM) system", *Journal of Biomedical Semantics*, Vol. 6, No. 15.

Edsall, R.L. and Adler, K.G. (2008), "User satisfaction with EHRs: report of a survey of 422 family physicians", *Family Practice Management*, Vol. 15 No. 2, pp. 25-32.

Fan, R., Chang, K.W., Hsieh, C.J., Wang, X.R. and Lin, C.J. (2008), "LIBLINEAR: a library for large linear classification", *Journal of Machine Learning Research*, Vol. 9, pp. 1871-1874.

Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J. and Johnson, S.B. (1994), "A general natural-language text processor for clinical radiology", *Journal of the American Medical Informatics Association*, Vol. 1 No. 2, pp. 161-174.

Fung, K.W., Richesson, R. and Bodenreider, O. (2014), "Coverage of rare disease names in standard terminologies and implications for patients, providers, and Research", *AMIA Annu Symposium Proceedimsgs eCollection 2014, Washington, DC*, pp. 564-572.

Grouin, C., Abacha, A., Bernhard, D. and Zweigenbaum, P. (2010), "CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches", *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data, Boston, MA*.

Halgrim, S., Xia, F., Cadag, E. and Uzuner, Ö. (2011), "A cascade of classifiers for extracting medication information from discharge summaries", *Journal of Biomedical Semantics*, Vol. 2 No. S3, p. S2.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009), "The WEKA data mining software: an update", ACM SIGKDD Explorations Newsletter, ACM, New York, NY.

Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V. and Duneld, M. (2014), "Synonym extraction and abbreviation expansion with ensembles of semantic spaces", *Journal of Biomedical Semantics*, Vol. 5, No. 6.

Hsiao, C. and Hing, E. (2014), "Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001-2013", *NCHS Data Brief*, Vol. 143 (January), pp. 1-8.

Jamoom, E., Beatty, P., Bercovitz, A., Woodwell, D., Palso, K. and Rechtsteiner, E. (2012), "Physician adoption of electronic health record systems: United States, 2011", *NCHS Data Brief*, No. 98 (July), pp. 1-8.

Jonquet, C., Nigam, H., Shah, H. and Musen, A.M. (2009), "The Open Biomedical Annotator", *AMIA Summit on Translational Bioinformatics*, pp. 56-60.

Jonquet, C., Musen, M.A. and Shah, N.H. (2010), "Building a biomedical ontology recommender web service", *Journal of Biomedical Semantics*, S1.

Lee, M. (2015), "New stroke therapy uses motion sensor video game to help rehabilitation", available at: www.metro.us/lifestyle/new-stroke-therapy-uses-motion-sensor-video-game-to-help-rehabilitation/zsJodo – NkqvJr2z246/ (accessed 10 January 2016).

Long, W. (2005), *Extracting Diagnoses from Discharge Summaries*, AMIA, pp. 470-474.

Manning, C.D., Surdeanu, M., Bauer, J. and McClosky, D. (2014), "The Stanford CoreNLP natural language processing toolkit", *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, At Baltimore, MD*, pp. 55-60.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), "Efficient estimation of word representations in vector space", *International Conference on Learning Representations 2013, Scottsdale, Arizona*.

Musen, M., Noy, N.F., Shah, N.H., Whetzel, P.L., Chute, C.G., Story, M.A., Smith, B. and NCBO Team (2012), "The National Center for Biomedical Ontology", *Journal of the American Medical Informatics Association*, Vol. 19 No. 2, pp. 190-195.

Patrick, J.D., Nguyen, D.H.M., Wang, Y. and Min, L. (2010), "I2b2 challenges in Clinical Natural Language Processing 2010", *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data, Boston, MA*.

Porumb, M., Bărbănţan, I., Lemnaru, C. and Potolea, R. (2015), "REMed – automatic relation extraction from medical documents", *7th International Conference on Information Integration and Web-Based Applications & Services - IIWAS, Brussels*.

Rindflesch, T., Rayan, J. and Hunter, L. (2000), *Extracting Molecular Binding Relationships from Biomedical Text*, Association for Computational Linguistics, Morristown, NJ, pp. 188-195.

Roberts, K., Rink, B. and Harabagiu, S. (2010), *Extraction of Medical Concepts, Assertions, and Relations From Discharge Summaries for the Fourth i2b2/VA Shared Task*, Boston, MA, JAMIA 2011, Vol. 18 No. 5, pp. 594-600, doi: 10.1136/amiajnl-2011-000153.

Rudd, K., Johnson, M. and Liesinger,. J.T. (2010), "Automated detection of follow-up appointments using text mining of discharge records", *International Journal for Quality in Health Care*, Vol. 22 No. 3, pp. 229-235.

Savona, G.K., Masanz, J.J., Ogren, P.V. and Zheng, J. (2010), "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications", *Journal of the American Medical Informatics Association*, Vol. 17 No. 5, pp. 507-513.

SNOMED-CT (2012), *SNOMED-CT: International Health Terminology Standards Development Organisation*, SNOMED-CT, available at: www.ihtsdo.org/snomed-ct/.

Smith, C. (2014), *Tracking Hand Tremors with Leap Motion*, Digital Hand Tremor Assessment, available at: http://blog.leapmotion.com/tracking-hand-tremors-leap-motion/ (accessed 10 January 2016).

Solt, I., Szidarovszky, F. and Tikk, D. (2010), "Concept, assertion and relation extraction at the 2010 i2b2 relation extraction challenge using parsing information and dictionaries", *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data, Boston, MA*.

Uzuner, O., South, B., Shen, S. and DuVall, S. (2011), "i2b2/VA challenge on concepts, assertions, and relations in clinical text", *Journal of the American Medical Informatics Association*, Vol. 18 No. 5, pp. 552-556.

Further reading

Lucey MD, C.R. (2015), *Clinical Problem Solving: Coursera*, University of California, San Francisco, CA.

Corresponding author

Ioana Barbantan can be contacted at: ioana.barbantan@cs.utcluj.ro