

RESEARCH

Open Access

# A context-blocks model for identifying clinical relationships in patient records

Rezarta Islamaj Doğan\*, Aurélie Névéol, Zhiyong Lu

From Machine Learning for Biomedical Literature Analysis and Text Retrieval in the International Conference for Machine Learning and Applications 2010

Washington, DC, USA. 12-14 December 2010

## Abstract

**Background:** Patient records contain valuable information regarding explanation of diagnosis, progression of disease, prescription and/or effectiveness of treatment, and more. Automatic recognition of clinically important concepts and the identification of relationships between those concepts in patient records are preliminary steps for many important applications in medical informatics, ranging from quality of care to hypothesis generation.

**Methods:** In this work we describe an approach that facilitates the automatic recognition of eight relationships defined between medical problems, treatments and tests. Unlike the traditional bag-of-words representation, in this work, we represent a relationship with a scheme of five distinct context-blocks determined by the position of concepts in the text. As a preliminary step to relationship recognition, and in order to provide an end-to-end system, we also addressed the automatic extraction of medical problems, treatments and tests. Our approach combined the outcome of a statistical model for concept recognition and simple natural language processing features in a conditional random fields model. A set of 826 patient records from the 4th i2b2 challenge was used for training and evaluating the system.

**Results:** Results show that our concept recognition system achieved an F-measure of 0.870 for exact span concept detection. Moreover the context-block representation of relationships was more successful (F-Measure = 0.775) at identifying relationships than bag-of-words (F-Measure = 0.402). Most importantly, the performance of the end-to-end system of relationship extraction using automatically extracted concepts (F-Measure = 0.704) was comparable to that obtained using manually annotated concepts (F-Measure = 0.711), and their difference was not statistically significant.

**Conclusions:** We extracted important clinical relationships from text in an automated manner, starting with concept recognition, and ending with relationship identification. The advantage of the context-blocks representation scheme was the correct management of word position information, which may be critical in identifying certain relationships. Our results may serve as benchmark for comparison to other systems developed on i2b2 challenge data. Finally, our system may serve as a preliminary step for other discovery tasks in medical informatics.

\* Correspondence: Rezarta.Islamaj@nih.gov

National Center for Biotechnology Information, National Library of Medicine,  
National Institutes of Health, Bethesda, Maryland, USA

Full list of author information is available at the end of the article

## Background

The era of Electronic Health Records (EHRs) brings the necessity of automatic recognition for clinical concepts, and the relationships that tie them together. Patient records contain comprehensive accounts of the patients' visits to the hospital. Such information can be invaluable for pharmaco-vigilance, detection of adverse effects, comparative effectiveness studies, etc. Patient records contain a wealth of information regarding what is the patient's discomfort, what medical measures are taken and what procedures are performed with what results. These documents differ from the general biomedical text data in these aspects. First, clinical documents are of a sensitive nature. Automatically removing personal information from these documents is a research problem in itself [1,2]. For this reason, obtaining sufficient amounts of de-identified clinical data as required to build a robust machine learning model is difficult. Second, patient records and doctor notes often are not well-structured documents. The syntactic structure, presentation style and terminology used in EHRs significantly differ from those used in a published research paper. The language is often similar to daily discourse, and it contains a lot of (mostly non-standard) abbreviations. For this reason, obtaining real clinical data instead of possible synthetic variations is important in order to build reliable systems.

The i2b2 challenges [3] are one of the most important recent community efforts to develop scalable informatics frameworks that will enable scientists to use existing clinical data for discovery research. The first step in the automatic processing of a clinical document is the recognition of text phrases which refer to clinically relevant concepts such as: medical problems, treatments and tests. Medical problems are observations about the patient's clinical health. Treatments include procedures, or medications administered to patients. Tests include lab procedures or measurements prescribed to patients. The second step is the identification of relationships among the recognized concepts. A relationship between two clinical concepts identifies how problems relate to treatments, tests and other medical problems in text. In this study, we focus on the identification of relationships as defined in the 4<sup>th</sup> i2b2 challenge [3], and we build an end-to-end system that first performs concept recognition, next identifies relationships. Figure 1 shows a diagram of the possible relationships found in clinical documents. Table 1 also lists these relationships with mappings to i2b2 notation, and example sentences from annotated patient records.

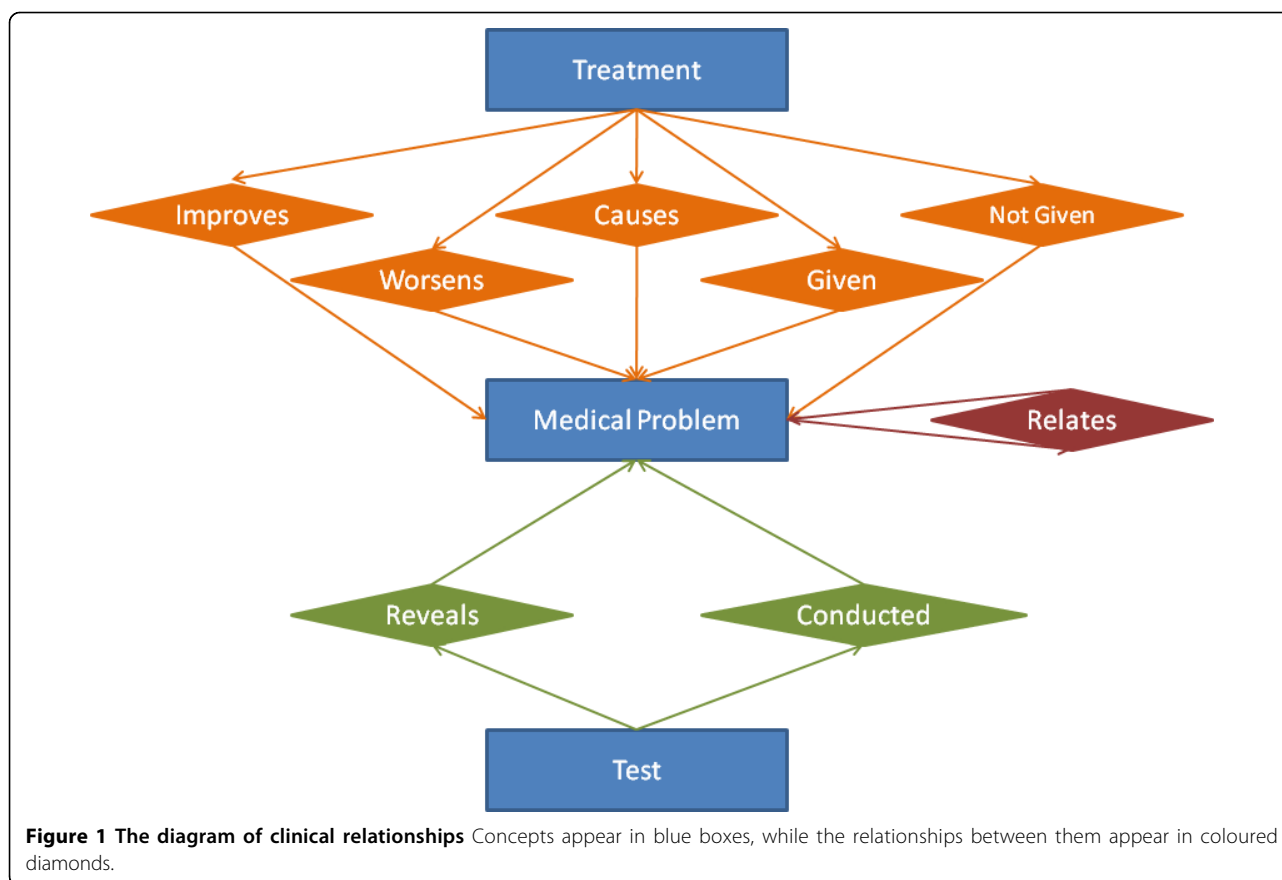
Extraction of relevant clinical concepts is an ongoing problem in the clinical domain, and several tools have been developed for this purpose. For example, MedLee

[4] maps clinical text to Unified Medical Language System<sup>®</sup> (UMLS<sup>®</sup>) concepts, whereas MedEX [5] focuses on extracting medication information from discharge summaries. Machine learning methods can also be used successfully for concept extraction from biomedical text. For instance, Hidden Markov Models (HMM) and Conditional Random Fields (CRF) were compared for the extraction of Proteins and Cell types [6]. Machine learning techniques have also been shown effective in mining patient smoking and medication status [7,8] from unstructured patient records.

Extraction of relationships between biomedical concepts has also produced a significant body of literature in the biomedical domain [9-14]. However, this research mostly addresses the extraction of relationships between biological entities (e.g. protein-protein interaction). In contrast, fewer studies are found on the extraction of relationships between diseases, symptoms, and medication in patient records. The proposed methods are typically based on co-occurrence statistics, semantic interpretation, and machine learning. For example, Chen *et al.* [15] proposed an automatic method for extracting disease-drug pairs which applied MedLEE for identifying associations. More recently, similar methods were developed for the identification of association between symptoms and diseases [16] and the detection of adverse drug effects [17,18].

Relationships between clinical entities can be identified with co-occurrence based methods. However, such methods are unable to further characterize specific relations. A significant research effort addressing the extraction of relationships between biomedical entities has resulted in the development of the semantic representation program SemRep [19]. SemRep exploits linguistic analysis of biomedical text and domain knowledge in the UMLS. This tool achieved competitive performance in [20,21] for extracting drug-disease treatment relationships from biomedical text. However, the set of relationships that can be extracted by SemRep does not match those of our dataset. Similarly, the set of relationships defined in [22] and [23] do not match those in our dataset, so that a direct comparison is not possible.

In this study we address the relationship identification task, expanding from [24], in an end-to-end system, starting with the recognition of concept phrases and then predicting possible relationships between two concepts found in the same sentence. This problem is very similar to the focus of the 4<sup>th</sup> i2b2 challenge, and our work is developed on the same dataset and annotations. The system we describe in this work was inspired by our own participation in this challenge and should be directly comparable to other work built on the same data.



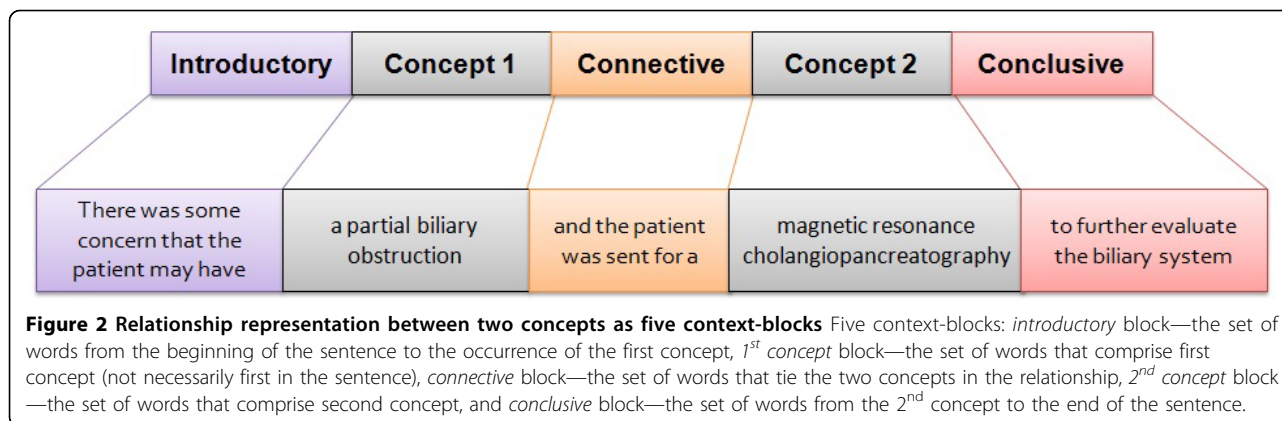
We developed a context-based scheme of representing relationships between two different kinds of entities. In this study, the relationship between two concepts is defined as a structure of five context-blocks (see Fig. 2). We characterized each of these five different blocks, and built a machine learning model that makes an informed

decision based on the present characteristics. In addition, in order to make this scheme easily applicable for real document processing, we implemented the automatic extraction of concepts using a CRF based model. We built a highly accurate concept recognizer which we used to predict concepts intended for relationship

**Table 1 Examples of relationships between medical concepts in patient records**

Relationship	Abbrev.	I2b2 Notation	Example Sentence
Treatment Improves Medical Problem	Improves	TrIP	<i>Her chest pain*</i> was controlled with <u>morphine</u> .
Treatment Worsens Medical Problem	Worsens	TrWP	He was started on <u>p.o. steroids</u> and to <u>CMED</u> for management of <i>COPD exacerbation</i> but he appeared in more respiratory distress overnight.
Treatment Causes Medical Problem	Causes	TrCP	During ER evaluation pt was noted to have some degree of <i>loss of short term memory</i> and <i>brief unresponsive period</i> after one dose of <u>IV Benadryl</u> .
Treatment is Administered for Medical Problem	Given	TrAP	<u>Lasix</u> 40 mg Tablet Sig : One ( 1 ) Tablet PO once a day as needed for <i>shortness of breath or wheezing</i>
Treatment is Not Administered because of Medical Problem	Not Given	TrNAP	There was some question regarding restarting of the patient 's <u>Coumadin</u> given her <i>positive lupus anti-coagulant status</i> .
Test Reveals Medical Problem	Reveals	TeRP	<b>Pathology</b> was reviewed revealing <i>invasive squamous cell carcinoma</i>
Test is Conducted to Investigate Medical Problem	Conducted	TeCP	There was some concern that the patient may have a <i>partial biliary obstruction</i> and the patient was sent for a <b>magnetic resonance cholangiopancreatography</b> .
Medical Problem Indicates Medical Problem	Relates	PIP	<i>Bilateral crackles at bases and midlungs , trace bilateral ankle edema</i> and CXR with <i>diffuse opacities</i> suggest possible <i>pulmonary edema</i> .

\* Medical problems are shown in *italics*, tests are shown in **bold** and treatments are underlined.



extraction in the test set of 477 clinical documents. The performance of the relationship extraction system using automatically extracted concepts was comparable to that obtained using the manually annotated concepts, and the difference was not statistically significant.

## Methods

In order for a relationship to be identified between two co-occurring concepts, those two concepts need to be identified first. A reliable concept recognizer is a prerequisite for the relationship identification to take place. Therefore, we start the methods description first with the description of the data and the concept identification procedure. Next, we discuss the features characterizing both concepts and relationships between them, the relationship representation model and relationship identification procedure, as well as the evaluation measures for comparison.

### Data description

Our participation in the 4<sup>th</sup> i2b2 challenge [3] allowed us to have access to a corpus of fully de-identified medical records manually annotated for concept, assertion, and relationship information. The training data contained discharge summaries from these different hospitals: Partners Healthcare (97 documents), Beth Israel Deaconess Medical Center (73 documents) and University of Pittsburgh Medical Center (98 documents). In addition, a set of progress notes from the University of Pittsburgh Medical Center (81 documents) was also included. The test data contained 477 records, also coming from the same sources. Our particular interest involved the classification of relations between medical problems, tests and treatments.

### Concept identification

The pre-requisite step of relationship identification is the correct identification of the concepts pertaining in the relationship. One particular challenge of our

relationship representation scheme is that the boundaries of the related concepts need to be precisely specified (exact span). Here we describe our concept identification method.

#### Concept identification features

The difference between any two implementations of a named entity recognition task lies in the set of features that are used to represent the entity phrase of interest. These features are traditionally divided into the following groups:

- Word features
- Context features
- Semantic features derived from other sources.

In accordance with those groups we represented each token of a given sentence with the following features: For the word features group we used the current token, its part of speech tag as identified by MEDPOST [25], and a surface feature that identified whether the current token was a number, a stop-word, or a punctuation symbol. For the context features group, we listed the two tokens before and the two tokens after the reference token, when available, as well as their respective part of speech tags, and their surface features. From the semantic features group, we used the priority model prediction class of each token.

Priority model [26], is a statistical method which has been successfully used to identify gene and disease names in biomedical text strings [26,29]. This method, given a phrase representing a named entity, assumes that a word to the right is more likely to be the head word of the phrase—the word more likely to determine the nature of the entity—than a word to the left. This model trains several variable order Markov Models, given a set of strings as training data, for sequences of tokens which represent concept phrases, versus others. While the priority model performance is high on the classification task for all three types of concepts of interest in patient records (problems, tests and treatment), a significant issue with using this approach in a real

production setting is the identification of the boundaries of concept phrases. That is why we decided to incorporate its output as a feature in our concept identification system.

**Concept identification method**

Due to characteristics of the clinical text we decided to use Conditional Random Fields (CRF) [28] trained with the data that we had available. CRFs have been shown to provide state-of-the-art performance in the natural language processing community for named entity recognition. We used the MALLET toolkit [27] to implement concept recognition models for each of our concept classes: medical problem, treatment and test. A common representation for an HMM or CRF based entity recognition model uses three tags (BIO) to label the tokens of a sequence—B to represent the first token of a concept phrase, I to represent any following token part of the concept phrase, and O to represent any other token, outside of the phrase of interest. Due to the modest amount of training data available, we simplified this representation to two tags (IO): I representing any token part of the concept phrase, and O representing any token outside of the concept phrase. We used the above features to train our model. Next, for any unknown piece of text, our system first extracted the features of each token (word, context and semantic), and then it decided whether it was part of the entity phrase of interest (I) or not (O).

**Relationship identification**

The goal of relationship identification is to determine whether two concepts are related, and the type of their relationship. Table 2 summarizes all annotations for the specified relations. These relationships (from our gold standard data) connected concepts appearing in the same sentence. In order to build our machine learning model, first we built the negative examples.

To build the negative dataset, we employed the following procedure: We scanned the text sentence by sentence. For any given sentence, for each identified pair of

concepts, we counted all the possible relationships there could exist between them. The number of relationship candidates that could be generated in this manner is limited. First, there were a limited number of concepts that could be found in a sentence. And second, for any pair of concepts of type (problem, problem), there was only one relationship candidate: *relates*. For any pair of concepts of type (problem, test), there were two relationship candidates: *conducted* and *reveals*. For any pair of concepts of type (problem, treatment), there were five relationship candidates: *improves*, *worsens*, *causes*, *given* and *not given*. After extracting all pairs of concepts in this manner, we found that 14% of the relationship candidates were true (found in the gold standard annotations).

Next, we describe the features that we selected to represent these relationships. These features captured the lexical information, information about the type of concept of each medical entity, and the sentence-context information about the pair of medical concepts.

**Relationship identification features**

We considered all unique token features extracted from our corpus. We experimented both with word stemming and stop-word elimination [30]. Also, for each concept phrase, we used MetaMap [31] to identify all matching UMLS Concept Unique Identifiers (CUI features) and their corresponding Semantic Type categories (SemTyp features) similarly to our previous work [32]. Finally, for each concept, we also used the provided assertion category as annotated in the data: *absent*, *conditional*, *present*, *hypothetical*, *possible*, and *associated-with-someone-else*. All features had binary values, 1 if present and 0 if absent.

**Relationship representation scheme**

We represent a relationship between two concepts as a schema of five, not necessarily consecutive, context-blocks, as shown in Figure 2. This structure—*Introductory Block*, *1<sup>st</sup> Concept Block*, *Connective Block*, *2<sup>nd</sup> Concept Block* and *Conclusive Block*—is naturally marked by the location of the two concepts in the sentence. As an operational decision, the introductory and conclusive blocks contained a maximum of five words. We extracted features to represent each context-block, which all-combined, represented the relationship. This was contrasted with the **Naïve Bag-of-Features** approach, which used all the available features without taking into account the context-block that they were identified from.

**Relationship identification method**

For each relationship, we built a machine learning model that recognized the true relationships (gold standard) from the rest of the candidates (negative examples). The classification algorithm of choice was a linear SVM. We employed a five-fold cross-validation setting

**Table 2 Data description**

Relationship	Number of Positive examples in training set	Number of positive examples in testing set
Improves	107	198
Worsens	56	143
Causes	296	444
Given	1421	2486
Not Given	106	191
Reveals	1733	3033
Conducted	303	588
Relates	1239	1986

The aggregated number of examples extracted from 349 patient records in the training set, and 477 patient records in the test set.

with balanced positive and negative instances for each fold. Our approach was to train our SVM-learner repeatedly, and eliminate a fixed number of lowest-weight features, after each step. Then a new model was learned on the remaining features. We reduced the number of features 500 at a time, until the system's performance did not improve any more. Finally, given a test sentence annotated for concept and assertion, all relevant relationship models were tested for each pair of concepts. Next, each score result was converted to a probability value. The two concepts were predicted to have the relationship which score provided the highest probability. If none of the relationship models provided a probability value higher than 0.5, than the two concepts were not predicted to be related.

Each relationship model was implemented as a context-blocks model, where all available features were organized according to the specific context blocks their appeared in. Each feature had a binary value 0/1 depending on being absent/present in the context block of interest. This representation was contrasted with the naive bag-of-features approach which used all the available features, without distinguishing their position in the sentence, whatsoever. This served as our baseline.

#### Evaluation metrics

We used precision, recall, and F-measure to measure and evaluate the performance of our systems. Precision measures the percentage of correct answers in the result set relative to its complete size and recall measures the percentage of correct answers relative to all true results (gold standard). F-measure is a metric that reflect the overall quality of recall and precision as a harmonic mean on the complete result set,

$$F_{\beta} = \frac{(1 + \beta^2) * p * r}{\beta^2 * (p + r)}$$

Where  $p$  is precision,  $r$  is recall, and  $\beta$  measures the trade-off between precision and recall. In this study, we chose  $\beta = 1$ , as it is commonly chosen for a balanced F-measure.

For all evaluations on the training data, these values were averaged over the five folds of cross validation. For a system balanced both in precision and in recall, we used the F-measure results to select the best models. When the same F-measure was obtained, we broke ties by choosing the model with the smaller number of features. We performed *per-relationship* and *per-record* evaluation of our system on the training data. The first measured the system performance on the relationship candidates, regardless of the patient records they were collected from. The latter measured the system performance on each relationship type, first, for each patient's

record, and next, averaged the results over all patient records. In this case, for each relationship type, only the records that had at least one annotated positive example of that relationship were considered.

For all evaluations on the test data, we measured precision, recall and F-measure, using the 4<sup>th</sup> i2b2 evaluation package in order to have fair comparison between other systems tested on the same dataset.

#### Results

We conducted a wide range of experiments to identify the medical concepts in patient records and the relationships between them. Here we present a summary of our data analysis, concept extraction and relationship identification models.

#### Data analysis

Our training dataset [3] contained 349 fully de-identified medical records from four different hospitals. This corpus was manually annotated for concept, assertion type and relationship information at the sentence level. Clinically relevant concepts are medical problems, treatments and tests. Our training dataset contained more than 27,000 instances of medical problems, divided into 7,073 unique medical problem phrases, 4,844 unique treatment phrases and 4,608 unique test phrases. Table 3 shows sample annotated sentences from the corpus. In addition, each medical problem was annotated with one of the following assertion categories: *absent*, *conditional*, *present*, *hypothetical*, *possible*, and *associated-with-someone-else*. Examples of assertion categories are given in Table 4. Lastly, there were eight different relationship categories between medical problems, treatments and tests. These clinical relationships are illustrated in Figure 1 and detailed with examples in Table 1.

Table 2 shows the number of examples found in training data for each relationship type. These examples correspond to the corpus annotations. We created the negative examples for our machine learning model using all the pairs of annotated concepts for all the sentences in the corpus. Each pair of (problem, problem) concepts

**Table 3 Examples of medical concepts in patient records**

Medical Concept	Example Sentence
Problem	On admission , the patient was found to have a <i>mild fever, myalgias</i> , and <i>arthralgias</i> that were relieved by Tylenol.
Treatment	Infectious Disease was consulted and recommended <u>doxycycline</u> to cover both organisms.
Test	<b>Pending labs</b> included <b>wound , bacterial , and fungal cultures</b> and <b>serologies</b> for Bartonella , Francisella , Yersinia , EBV ,

\*Medical problems are shown in *italics*, tests are shown in bold and treatments are underlined.

**Table 4 Examples of assertion categories of problem concepts in patient records**

Example Sentence	Problem	Assertion
On admission , the patient was found to have a mild fever, myalgias, and arthralgias that were relieved by Tylenol.	"a mild fever" "myalgias" "arthralgias"	Present Present Present
Pending labs included wound , bacterial , and fungal cultures and serologies for Bartonella, Francisella, Yersinia, EBV,	"Bartonella" "Francisella" "Yersinia" "EBV"	Possible Possible Possible Possible
It would be useful to follow this at outpatient if the patient is symptomatic .	"symptomatic"	Hypothetical
He denied ever having any chest pain , chest pressure , shortness of breath , dyspnea on exertion and he was discharged to home .	"chest pain" "chest pressure" "shortness of breath" "dyspnea"	Absent Absent Absent Conditional
Patient has a family history of coronary artery disease	"coronary artery disease"	Associated with someone else

contributed one candidate to the *relates* relationship, each pair of (problem, test) concepts contributed one candidate for each of the *conducted* and *reveals* relationships, and each pair of (problem, treatment) concepts contributed one candidate for each of *improves*, *worsens*, *causes*, *given* and *not given* relationships.

The set of documents used to test the systems at the end of the 4<sup>th</sup> i2b2 challenge consisted of an additional set of 477 discharge summaries, which were accordingly provided labelled and annotated. The number of annotated relationships in the test data is also given in Table 2. These numbers are used to compute the overall weighted average of our system’s performance.

**Concept identification**

Our results of concept identification for the three medical concept classes are listed in Table 5. These results are expressed in terms of precision, recall and F-measure and measurements are produced for exact matching of the phrase to the annotated phrase, and partial or inexact matching to the annotated concept phrase. They

**Table 5 Concept identification**

Concept	Exact span evaluation			Inexact span evaluation		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Problem	0.802	0.818	0.810	0.911	0.914	0.912
Treatment	0.929	0.891	0.909	0.975	0.946	0.960
Test	0.943	0.893	0.917	0.975	0.934	0.954
<b>Overall</b>	<b>0.878</b>	<b>0.861</b>	<b>0.870</b>	<b>0.948</b>	<b>0.930</b>	<b>0.939</b>
Best i2b2 system	0.869	0.836	0.852	0.932	0.917	0.924

are computed using the i2b2 evaluation tool for compatible comparison with other systems. Table 5 also contrasts our results for concept extraction with the best reported result at the 4<sup>th</sup> i2b2 challenge (Ozlem Uzuner, personal communication).

**Relationship identification**

We approached the relationship identification task as a classification task. Similarly to other participants of the i2b2 challenge we also utilized support vector machines, and we modeled each relationship separately. Differently from other approaches though, we conceptualized relationships as a composite structure of consecutive context blocks.

**Concept-blocks relationship model performs best**

Table 6 shows detailed evaluation results for the *relates* relationship identification using the string matching model, the naive bag-of-features model and the context-blocks relationship representation model. In addition, we experimented with other features for each of the Concept blocks, such as assertion category, and mappings to UMLS Concept Identifiers (CUI) and UMLS Semantic Types (SemTyp). Our exploration of feature space in building a relationship identification model is also shown in Table 6.

**Assertions, Concept identifiers and Semantic Types are important for different relationships**

Table 7 shows performance evaluation for all eight clinical relationships. For each relationship, we used the context-blocks representation to identify the best model combining the word features with the assertion, CUI and SemTyp features. We selected the best model based on the F-measure values. These results illustrate that different relationships benefited from different additional concept features. In addition, we used the same features in the non-context-blocks setting, or naive bag-of-features, with the same SVM classifier, and those results are also listed in Table 7.

**Feature selection refined relationship identification**

We applied the SVM iterative feature selection to each context-blocks relationship model selected in Table 7.

**Table 6 Performance evaluation for the relates relationship, using string matching and SVM models**

Relationship Model	Precision	Recall	F-measure
String Matching	0.177	0.511	0.263
Naive bag-of-words SVM	0.254	0.960	0.402
Context-blocks SVM (words)	0.601	0.796	0.685
Context-blocks SVM (Words + Assertion)	0.598	0.784	0.679
Context-blocks SVM (Words + CUI)	<b>0.646</b>	0.746	<b>0.692</b>
Context-blocks SVM (Words + SemType)	0.590	<b>0.797</b>	0.678

**Table 7 Performance evaluation for the best models of all relationships**

Relationship	Naive bag-of-features SVM F-measure	Context-blocks SVM			Features in the best model(Words +)
		Precision	Recall	F-measure	
Improves	0.489	0.619	0.607	<b>0.613</b>	SemTyp
Worsens	0.481	0.800	0.358	<b>0.494</b>	Assertion
Causes	0.470	0.644	0.588	<b>0.615</b>	CUI, Assertion, SemTyp
Given	0.713	0.727	0.872	<b>0.793</b>	CUI, Assertion, SemTyp
Not Given	0.618	0.800	0.604	<b>0.688</b>	CUI
Reveals	0.772	0.805	0.932	<b>0.864</b>	-
Conducted	0.533	0.584	0.742	<b>0.654</b>	Assertion
Relates	0.543	0.646	0.746	<b>0.692</b>	CUI

After feature selection we identified 1000 features for each relationship. Table 8 presents the F-measures obtained, both before and after feature selection, for each relationship using five-fold cross validation. Metrics are computed using both *per-relationship* and *per-record* evaluation.

**Context-blocks model is important for relationship identification**

We studied the feature composition of the selected models for each relationship category. We found that specific words were selected in specific context blocks. Consider, for example, the *conducted* and *relates* relationships, as illustrated in Figure 3. The word “revealed” was weighted positively in the Connective block of the *relates* relationship, but it was weighted negatively in the Connective block of the *conducted* relationship. Stop-words were also highly weighted features, both positively and negatively, in all relationship models.

**Relationship identification model robust after concept extraction**

One of the main goals of this study was to demonstrate that a realistic application setting is possible. In a realistic application test, one would start with the concept extraction, and precisely identify the concept

boundaries, so that relationship identification may be performed. In order to address this, we ran our concept extraction model on the i2b2 test dataset, and marked the predicted concept phrases and their type. Next, for each test sentence in the test records, all relevant relationship models were tested for each pair of concepts. In the end, each score result was converted to a probability value. Two concepts were predicted to have a relationship, if the probability was higher than the threshold (0.5). The relationship type, however, was assigned to the one with highest probability value amongst competing relationships.

Table 9 lists these results. The first two columns show the results of the relationship identification models using the test data annotated concepts, and the last two columns show the results using the automatically extracted concepts. The first and the third columns show results when the original model is applied, and the other two columns show results when the feature-selection refined model is applied. The results of column three and four are statistically different (T-test,  $p=0.005$ ), but the results presented in column four are not statistically different from those presented in the first two columns. Overall averages are computed by weighting each F-measure with the number of examples of that particular relationship type (shown in Table 2).

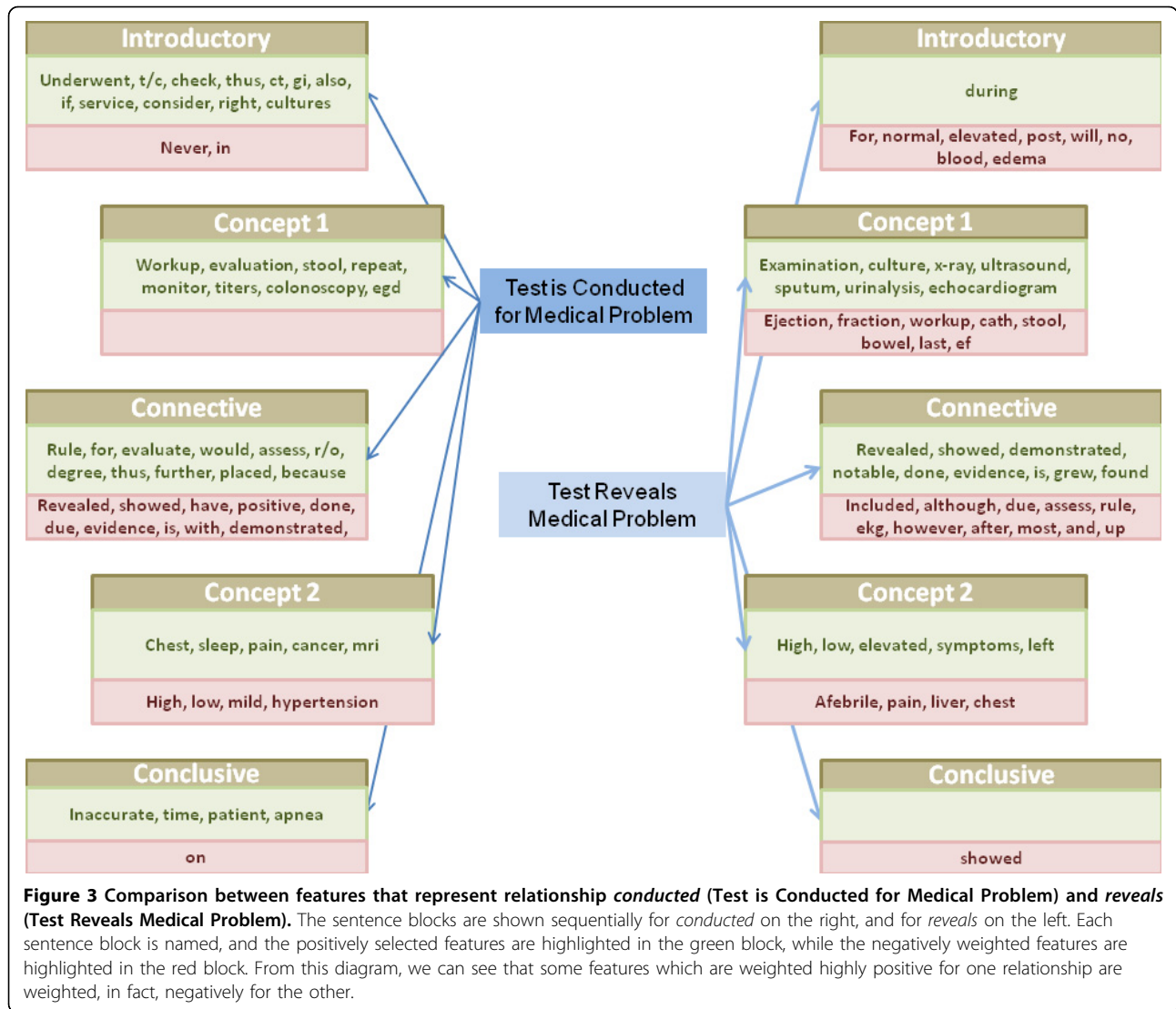
**Table 8 Per-relationship and per-record f-measures computed prior to and after feature selection**

Feature Selection	Prior	After	Prior	After
Evaluation	Per Relation	Per Relation	Per Record	Per Record
Improves	0.613	0.683	0.703	0.814
Worsens	0.494	0.673	0.569	0.621
Causes	0.615	0.735	0.695	0.793
Given	0.793	0.866	0.768	0.831
Not Given	0.688	0.754	0.685	0.743
Reveals	0.864	0.912	0.819	0.855
Conducted	0.654	0.759	0.803	0.883
Relates	0.692	0.775	0.761	0.823

**Discussion**

In this study, we defined the relationship between two concepts as a structure of five distinct context-blocks: the *introductory* block, the *first concept* block, the *connective* block, the *second concept* block, and the *conclusive* block. Such a representation was successful in identifying eight relationships between medical problems, treatments and tests in patient records. The performance degraded considerably when the context-blocks structure was removed for the same relationships, with the same set of features and the same classification algorithm (the naïve bag-of-features model). The context-blocks representation captured the individual word





**Table 9 F-measures computed prior to and after feature selection for the test dataset relationship prediction. Results are computed using the annotated concepts (columns 1 and 2), and using the predicted concepts as identified in the concept recognition step (columns 3 and 4)**

Results of relationship identification	Annotated concepts		Predicted concepts	
	Before feature selection	After feature selection	Before feature selection	After feature selection
Improves	0.735	0.727	0.558	0.728
Worsens	0.598	0.600	0.398	0.541
Causes	0.655	0.664	0.401	0.632
Given	0.735	0.738	0.536	0.735
Not Given	0.628	0.427	0.494	0.458
Reveals	0.814	0.823	0.625	0.821
Conducted	0.593	0.583	0.404	0.531
Relates	0.591	0.588	0.390	0.588
<b>Overall weighted average</b>	0.712	0.711	0.516	0.704

The F-measures are computed using the i2b2 evaluation package, and the Average is computed weighting each individual F-measure by the number of all positive examples in that particular relationship category.

positions, and treated them accordingly. For example, for the *conducted* relationship the word “without” was a highly weighted *negative* feature in the introductory block and a highly weighted *positive* feature in the connective block.

Also, stop-words were very valuable in this study as also reported in [30]. For example, the word “no” was a highly weighted negative feature in the introductory block of the *given* relationship, while being a highly weighted positive feature in the same block of the *not given* relationship. Similarly, the words “for”, “but”, “because”, and other stop-words, were observed to fulfill analogous roles.

In this study, we also addressed its natural pre-requisite problem that, in order for a relationship to be identified between two co-occurring concepts; those two concepts need to be identified first. We built a reliable concept recognizer that exhibited high accuracy at identifying concept boundaries; critical for the context-blocks relationship model. While feature selection did not have a significant impact on relationship extraction based on manually annotated concepts, it significantly improved the performance of relationship extraction based on automatically extracted concepts (T-test,  $p=0.005$ ). Overall, as can be seen from Table 9, feature selection was a key step allowing us to obtain similar relationship extraction performance as high for automatically extracted concepts as for manually annotated concepts.

Naturally, a careful study of the clinical texts may define other types of relationships between medical concepts. In that case, the context-blocks model could be easily adapted. Finally, this model only considered the text within a sentence. Such a simplification, by definition, puts a limitation on the sensitivity of the produced results. Future work should include natural language techniques in order to obtain a better understanding of the text, as well as resolve pronouns and inference.

## Conclusions

In this work, we present a successful end-to-end method for relationship extraction from clinical documents. Automatic recognition of medical concepts in clinical records is a challenging first step towards semantically relating the concepts and more advanced reasoning applications of text mining in the patient records. To address this, we built a reliable concept recognizer that exhibited high accuracy (F-measure = 0.870) at identifying concept boundaries; critical for the context-blocks relationship model. We defined a relationship identification schema between two concepts in text. In this scheme, the relationship is represented as a structure of five context-blocks: the *introductory*, *first concept*, *connective*, *second concept*, and *conclusive* block.

This scheme automatically captured the word positions information; critical in certain relationships. We found that assertion information was useful in detecting clinical tests conducted to investigate medical problems, and treatments which cause medical problems to get worse. Semantic types were useful in identifying treatments that improved a medical problem and UMLS concept identifiers were relevant in identifying two medical problems that were related to each other. Our system benefited from inclusion of stop-words, especially when found in the introductory and connective blocks of the relationship representation. Our results may serve as benchmark for comparison to other systems developed on i2b2 challenge data. Finally, our system may serve as a preliminary step for other discovery tasks in medical informatics.

## Acknowledgements

*Funding:* This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 3, 2011: Machine Learning for Biomedical Literature Analysis and Text Retrieval. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/12?issue=S3>.

## Authors' contributions

RID designed the study, developed the concept and relationship models, performed the evaluation and wrote the draft of the manuscript. AN and ZL contributed to the study design, data preparation and evaluation. All authors have read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

Published: 9 June 2011

## References

1. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH: **Automatic de-identification of textual documents in the electronic health record: a review of recent research.** *BMC Med Res Methodol* 2010, **10**:70.
2. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, Malin B, Hirschman L: **The MITRE Identification Scrubber Toolkit: design, training, and assessment.** *Int J Med Inform* 2010, **79**(12):849-59.
3. **Fourth i2b2/VA Shared-Task and Workshop.** [<https://www.i2b2.org/NLP/Relations>].
4. Friedman C, Shagina L, Lussier Y, Hripcsak G: **Automated encoding of clinical documents based on natural language processing.** *J Am Med Inform Assoc* 2004, **11**(5):392-402.
5. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC: **MedEx: a medication information extraction system for clinical narratives.** *J Am Med Inform Assoc* 2010, **17**(1):19-24.
6. Ponomareva N, Rosso P, Pla F, Molina A: **Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task.** *Proc RANLP'07* 2007.
7. Uzuner O, Goldstein I, Luo Y, Kohane I: **Identifying patient smoking status from medical discharge records.** *J. Am. Med. Inform. Assoc.* 2008, **15**(1):14-24.
8. Pakhomov SV, Ruggieri A, Chute CG: **Maximum entropy modeling for mining patient medication status from free text.** *Proc AMIA Symp* 2002, 587-91.
9. Cohen AM, Hersh WR: **A survey of current work in biomedical text mining.** *Brief Bioinform* 2005, **6**(1):57-71.
10. Craven M: **Learning to Extract Relations from MEDLINE.** *AAAI-99 Workshop on Machine Learning for Information Extraction* 1999, 25-30.

11. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, Valencia A: **An overview of BioCreative II.5.** *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2010, **7**(3):385-99.
12. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** *Genome Biol.* 2008, **9**(Suppl 2):S4.
13. Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T: **Complex event extraction at PubMed scale.** *Bioinformatics* 2010, **15**;26(12):i382-90.
14. Bundschuh M, Dejori M, Stetter M, Tresp V, Kriegel HP: **Extraction of semantic biomedical relations from text using conditional random fields.** *BMC Bioinformatics* 2008, **9**:207.
15. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C: **Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study.** *J. Am. Med. Inform. Assoc.* 2008, **15**(1):87-98.
16. Wang X, Chused A, Elhadad N, Friedman C, Markatou M: **Automated knowledge acquisition from clinical narrative reports.** *AMIA Annu Symp Proc.* 2008, **6**:783-7.
17. Wang X, Chase H, Markatou M, Hripcsak G, Friedman C: **Selecting information in electronic health records for knowledge acquisition.** *J. Biomed. Inform.* 2010, **43**(4):595-601.
18. Harpaz R, Haerian K, Chase H, Friedman C: **Mining Electronic Health Records For Adverse Drug Effects Using Regression.** *Proc ACM IHI* 2010.
19. Rindflesch TC, Fiszman M: **The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.** *J. Biomed. Inform.* 2003, **36**(6):462-77.
20. Rindflesch TC, Pakhomov SV, Fiszman M, Kilicoglu H, Sanchez VR: **Medical facts to support inferencing in natural language processing.** *AMIA Annu. Symp. Proc.* 2005, 634-8.
21. Wang X, Chase HS, Li H, Hripcsak G, Friedman C: **Integrating Heterogeneous Knowledge Sources to Acquire Executable Drug-Related Knowledge.** *Proc AMIA Symp* 2010, 852-6.
22. Uzuner O, Mailoa J, Ryan R, Sibanda T: **Semantic relations for problem-oriented medical records.** *Artificial Intelligence in Medicine* 2010, **50**(2):63-73.
23. Roberts A, Gaizauskas R, Hepple M, Guo Y: **Mining clinical relationships from patient narratives.** *BMC Bioinformatics* 2008, **9**(11):S3.
24. Islamaj Doğan R, Névél A, Lu Z: **A textual representation scheme for identifying clinical relationships in patient records.** *IEEE Proceedings of the International Conference on Machine Learning Applications* 2010.
25. Smith L, Rindflesch T, Wilbur WJ: **MedPost: a part-of-speech tagger for bioMedical text.** *Bioinformatics* 2004, **20**(14):2320-1.
26. Tanabe L, Wilbur WJ: **A Priority Model for Named Entities.** *Proc HLT-NAACL BioNLP Workshop* 2006, 33-40.
27. McCallum AK: **MALLET: A Machine Learning for Language Toolkit.** 2002 [<http://mallet.cs.umass.edu>].
28. Lafferty J, McCallum A, Pereira F: **Conditional random fields: probabilistic models for segmenting and labeling sequence data.** *Proc. 18th Int Conf on Machine Learning* 2001, 282-289.
29. Névél A, Kim W, Wilbur WJ, Lu Z: **Exploring Two Biomedical Text Genres for Disease Recognition.** *NAACL BioNLP Workshop* 2009.
30. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ: **Multi-dimensional classification of biomedical text: Toward automated practical provision of high-utility text to diverse users.** *Bioinformatics* 2008, **24**(18):2086-2093.
31. Aronson AR, Lang FM: **An overview of MetaMap: historical perspective and recent advances.** *J Am Med Inform Assoc* 2010, **17**(3):229-36.
32. Islamaj Doğan R, Lu Z: **Click-words: learning to predict document keywords from a user perspective.** *Bioinformatics* 2010, **26**(21):2767-75.

doi:10.1186/1471-2105-12-S3-S3

**Cite this article as:** Islamaj Doğan *et al.*: A context-blocks model for identifying clinical relationships in patient records. *BMC Bioinformatics* 2011 **12**(Suppl 3):S3.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

