Identifying Contextual Information in Clinical Texts: A Study of Two Domains

Emilia Apostolova, Noriko Tomuro, Dina Demner-Fushman*
DePaul University, Chicago, IL, USA
*National Library of Medicine, Bethesda, MD, USA
emilia.aposto@gmail.com tomuro@cs.depaul.edu
ddemner@mail.nih.gov

Abstract

Contextual information, such as the affirmation or negation of medical problems, is key in Information Extraction (IE) from clinical texts. As there are very few available annotated clinical datasets, the practical question of training corpora reusability arises. Clinical IE systems face the challenge of disparate clinical sub-domain narratives typically lacking targeted training/testing data. We evaluated the cross-domain reusability of a clinical dataset annotated with contextual information - medical concepts and their assertion status (present, absent, hypothetical, possible, conditional, not associated with the patient). Systems developed on a training dataset consisting of discharge summaries and progress notes were then evaluated on a new sub-domain - radiology reports. We developed a machine learning and a rule-based system and observed that there was a minor performance drop when applied to a new dataset of radiology reports. The rule-based system exhibited stable performance with no statistically significant change. While the machine learning approach had a statistically significant performance drop, it still outperformed the rule-based system. Results suggest that an annotated training dataset could be reused across clinical subdomains for the task of identifying contextual information.

1 Background

Biomedical IE systems and specifically IE systems targeting clinical texts typically involve the extraction of medical problems. Identifying correctly the context of these medical problems is an equally important task. Contextual information refers to concept attributes such as negation status (is the medical problem affirmed, negated, or speculated; e.g. *tumor was felt to be unlikely*), temporality (is the medical problem current or past; e.g. *a prior history of pneumonia 2 years ago*), subject identification (is the medical problem associated with the patient; e.g. *his father had prostate cancer*).

The importance of correctly identifying contextual information is attested by an increasing interest in the task. A number of contextual discovery algorithms have been developed. Most notably, the NegEx[1] (Chapman et al., 2001) negation discovery algorithm was developed and subsequently implemented in a number of Biomedical NLP systems, including MetaMap (Aronson, 2001), CaTIES[2], and Mayo Clinic's Clinical IE system (Savova et al., 2008). Subsequently ConText (Chapman et al., 2007) was developed, as a NegEx extension that identifies additional contextual features from clinical documents, such as temporality and subject identification. Meystre et al. (2008) present a good overview of other systems and algorithms aiming at contextual information extraction from clinical texts.

The availability of annotated clinical text corpora is of crucial importance for the development of contextual information extraction systems. Not surprisingly, few annotated clinical text corpora are publicly available due to patient privacy restrictions and data ownership complications. The BioScope corpus (Szarvas et al., 2008) consists of biomedical texts annotated for negation, speculation and their linguistic scope, and includes 1,954 radiology report excerpts (typically 3 to 4 sentences) that were used in the Computational Medicine Center's 2007 Medical NLP Challenge[3]. The 2010 i2b2 (Informatics for Integrating Bi-

[1] http://www.dbmi.pitt.edu/chapman/NegEx.html
[2] http://caties.cabig.upmc.edu/Wiki.jsp?page=Home
[3] http://www.computationalmedicine.org/challenge

ology and the Bedside[4]) NLP Shared Task released 826 clinical records (discharge summaries and progress notes) annotated with medical problems and their contextual information[5].

As publicly available clinical text corpora are sparse, the acquisition of task-specific annotated datasets remains problematic. Clinical IE systems target various clinical sub-genres (radiology reports, pathology reports, discharge summaries, etc.) and possibly hospital-specific report formating and content characteristics. A very practical concern for the development of such systems is whether or not, and how much, the tools developed using the available corpora annotated with contextual information could be used in other clinical sub-domains. In this paper we explore this question by evaluating the performance of a contextual IE system developed for the clinical sub-genre of discharge summaries on a different sub-genre - radiology reports.

2 Methods

2.1 Dataset

As participants of the 2010 i2b2 NLP Shared Task, we have developed a clinical IE system that extracts contextual information of medical problems. The system was developed using the 2010 i2b2 challenge training dataset consisting of 349 clinical records (discharge summaries and progress notes). The challenge data was annotated for concepts referring to medical problems, tests, and treatments (Table 1). In addition, each medical problem was annotated with its 'assertion status' - one of 6 categories of assertions as described below:

Present - it is asserted that the patient experiences the problem (the default category).

Absent - it is asserted that the problem does not exist in the patient.

Possible - it is asserted that the problem may be present in the patient, but there is uncertainty expressed.

Conditional - it is asserted that the patient experiences the problem only under certain conditions (e.g. allergies).

Hypothetical - it is asserted that the patient may develop the medical problem.

Not associated with the patient - the medical problem is associated with someone who is not the patient.

Assertion category examples are shown in Table 2. The distribution of the 6 assertion categories within the i2b2 training dataset is shown in Figure 1.

Concept Category	Example
Medical Problem (any abnormality observed in patient)	*She developed **diabetes**.*
Test (procedures, panels, and measures)	***Chest x-ray** revealed* clear lungs.
Treatment (procedures, interventions, and substances)	*He was placed on **a morphine drip**.*

Table 1: Concept categories in the 2010 i2b2 NLP Shared Task.

Concept Category	Example
Present	*He has **pneumonia**.*
Absent	***No pneumonia** was suspected.*
Hypothetical	*If you experience **wheezing** or **sob**.*
Possible	***Pneumonia** is possible/probable.*
Conditional	*Penicillin causes **a rash**.*
Not associated with the patient	*Brother had **asthma**.*

Table 2: Medical concept 'assertion' categories in the 2010 i2b2 NLP Shared Task.

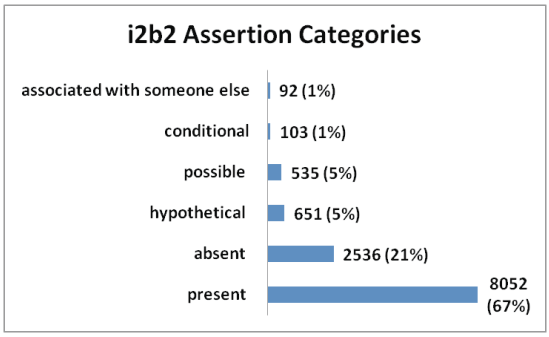

Figure 1: Distribution of the assertion categories across the 11,969 medical problems in the i2b2 training dataset of 349 documents.

Systems developed with the training dateset described above were tested against two datasets. One was provided by the i2b2 challenge - 477 discharge summaries and progress notes similar to the training dataset. The second dataset was created by annotating 70 deidentified Lung CT radiology reports[6]. The radiology reports were anno-

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

[5]The 2010 i2b2 Shared Task data (currently available to challenge participants) will be made available to the research community at large one year after the evaluation.

[6]The reports were randomly selected from a proprietary dataset of deidentified radiology reports spanning a period of 3 years from the Department of Radiology at Northwestern University Medical School.

tated by a single annotator for tests, treatments, medical problems and their assertion types following the 2010 i2b2 challenge annotation guidelines. The dataset comprised of 2,322 sentences and 32, 592 tokens, with an average report length of 33 sentences. Lung CT studies were chosen as the nature of the procedure usually results in more verbose reports with an abundance of findings (e.g. as compared to routine exams such as mammography). The annotator identified a total of 1,564 medical problems, 431 tests, and 92 treatments. The distribution of the assertion categories across the annotated medical problems is shown in Figure 2. Unlike discharge summaries and progress notes, radiology reports do not list allergies, 'as needed' medication prescriptions, or narrate family history. As a result, there were no instances of the categories *conditional, hypothetical,* and *associated with someone else*. Also notable is that the percentage of *present* medical problems is relatively lower (60%) compared to the dataset of discharge summaries/progress notes (67%), while the percentages of *possible* and *absent* medical problems are higher (5% vs. 14% and 21% vs. 26% respectively). The difference is again due to the nature of radiology reports, they are often used to rule out conditions and to suggest further investigation of possible medical problems.

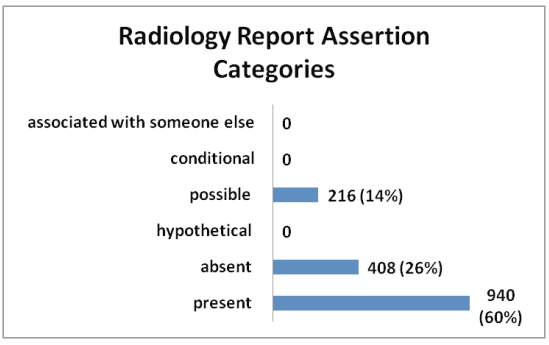

Figure 2: Distribution of the assertion categories across the 1,564 medical problems in the test dataset of 70 radiology reports.

2.2 System Description

We evaluated the performance of a rule-based approach and a machine learning approach against the two datasets. Both systems were developed using the i2b2 training dataset consisting of 349 discharge summaries and progress notes.

The rule-based system is a slightly modified implementation of the ConText algorithm (Chapman et al., 2007). The ConText algorithm relies on hand-crafted sets of trigger terms in proximity of clinical conditions to discover if the conditions are affirmed, negated, or possible; recent, historical, or hypothetical; experienced by the patient or other. Trigger terms are phrases preceding or following medical problems such as *no evidence to suggest*, *negative for*, *may be ruled out*, etc. We slightly modified the algorithm by extending the list of 'possible' trigger-terms (118 additional expressions including morphological variants), extending the list of 'absent' and 'hypothetical' trigger terms (4 and 2 additional expressions respectively), and introducing a small set of 'conditional' trigger terms (12 expressions). We also disregarded 'historical' cues as the challenge task does not differentiate between historical and recent medical problems.

The Machine Learning system models the problem as a classification task that assigns each of the annotated medical problems into one of the 6 categories. We trained a one-against-all SVM (Chang and Lin, 2001) classifier - a series of binary classifiers for each assertion category against all other categories. Empirically, we identified an optimum set of features as described below. The GATE framework (Cunningham et al., 2002) was used to generate and experiment with features sets.

Feature Set:

1. *Token window of size 5*: Tokens surrounding the medical problem (within sentence boundaries). Numbers were normalized (converted to the string $number). Tokens belonging to concepts were converted to their corresponding concept types (e.g. 'coronary bypass surgery' was substituted by the concept's category - *treatment*).

2. *Negative prefix*: This feature targets the discovery of absent medical problems identified as such by the presence of a morphological prefix, as in 'afebrile' or 'nontender'. Possible values are a-, ab-, un-, an-, anti-, dis-, non-, in-, il-, ir-, or im-.

3. *Section heading preceding the problem concept*: Section headings could be helpful in identifying most assertion categories. For example, problems that fall under the heading 'Family History' typically fall into the 'not associated with patient' assertion category. Headings were identified as the last string preceding the problem concept that matches the regular expression 'Beginning of line, One or more characters, Colon, White space, End of line'.

4. *ConText Cues*: Occasionally cues or trigger

terms outside the limitations of the 5-token window were necessary for a human reader to identify the assertion category. This feature was used to identify ConText cues preceding or following the medical concept outside the token window size.

5. *Semantic Type*: Conditional medical problems are typically related to allergy symptoms and other temporary conditions (e.g. *penicillin causes a rash; dyspnea on exertion*). A hand-crafted dictionary was created to map problem concepts to such semantic types. For example, expressions such as *rash, hive, itching, dyspepsia*, etc. were mapped to the semantic type *allergy symptom*.

In addition, token-window Part-of-Speech, UMLS[7] term and semantic type features were also considered. However, these features had no positive effect on the system performance and were excluded from the final system.

3 Results

Results from applying the original and the modified versions of the ConText algorithm on the i2b2 dataset of 477 discharge summaries/progress notes are shown in Tables 3 and 4 respectively. As shown, the addition of hand-crafted trigger terms improved performance across most assertion categories. Table 5 shows results obtained on the same dataset by the SVM-based system.

Category	TP	FN	FP	R	P	F1
Present	12663	362	2061	97.22	86.0	91.27
Absent	2877	732	461	79.72	86.19	82.83
Hypothetical	327	390	42	45.61	88.62	60.22
Possible	53	830	15	6.0	77.94	11.14
Conditional	0	171	0	0.0	0	0
Not patient	84	61	13	57.93	86.6	69.42
Overall	16004	2546	2592	86.27	86.06	86.17

Table 3: Results from applying the ConText algorithm (unmodified) to the 2010 i2b2 NLP Shared Task test dataset containing 11,969 medical problems (TP=True Positive, FN=False Negative, FP=False Positive, R=Recall, P=Precision, F1=F1-score).

Results from applying the original and the modified versions of the ConText algorithm on the test dataset of 70 CT Lung radiology reports are shown in Tables 6 and 7 respectively. The performance gain from the addition of hand-crafted trigger terms is more notable in the radiology dataset. The i2b2 dataset of discharge summaries contained a very low portion of 'possible' assertions (5%), and even though the new rules improved the F1-score from 11.14 to 50.57, the overall performance gain was negligible. However, the radiology report dataset contained a larger portion of 'possible' assertions (14%) that account for the improved performance over the base-line. Table 8 shows results obtained on the same dataset by the SVM-based system.

[7]Unified Medical Language System ©The National Library of Medicine

Category	TP	FN	FP	R	P	F1
Present	12338	687	1559	94.73	88.78	91.66
Absent	2876	733	452	79.69	86.42	82.92
Hypothetical	453	264	64	63.18	87.62	73.42
Possible	398	485	293	45.07	57.6	50.57
Conditional	42	129	146	24.56	22.34	23.4
Not patient	89	56	15	61.38	85.58	71.49
Overall	16196	2354	2529	87.31	86.49	86.90

Table 4: Results from applying a modified version of the ConText algorithm to the 2010 i2b2 NLP Shared Task test dataset containing 11,969 medical problems.

Category	TP	FN	FP	R	P	F1
Present	12808	217	930	98.33	93.23	95.71
Absent	3331	278	145	92.29	95.82	94.02
Hypothetical	568	149	38	79.21	93.72	85.86
Possible	449	434	102	50.84	81.48	62.62
Conditional	44	127	14	25.73	75.86	38.42
Not patient	111	34	10	76.55	91.73	83.45
Overall	17311	1239	1239	93.32	93.32	93.32

Table 5: Results from applying an SVM classifier to the 2010 i2b2 NLP Shared Task test dataset containing 11,969 medical problems.

Category	TP	FN	FP	R	P	F1
Present	936	4	241	99.57	79.52	88.42
Absent	362	46	23	88.73	94.03	91.3
Hypothetical	0	0	0	0	0	0
Possible	2	214	0	0.93	100.0	1.84
Conditional	0	0	0	0	0	0
Not patient	0	0	0	0	0	0
Overall	1300	264	264	83.12	83.12	83.12

Table 6: Results from applying the ConText algorithm (unmodified) to a dataset of 70 CT Lung radiology reports containing 1,564 medical problems.

As could be seen by comparing Tables 4 and 7, the overall performance of the rule-based approach decreased from an F1-score of 86.90 to 86.32 when applied to a new clinical sub-domain - radiology. The performance drop was not statistically significant (we used a two-tailed Z-test on two proportions with a confidence level of 95%). Introducing new heuristics to the ConText rule-based system proved beneficial as the original

Category	TP	FN	FP	R	P	F1
Present	924	16	184	98.3	83.39	90.23
Absent	362	46	18	88.73	95.26	91.88
Hypothetical	0	0	0	0	0	0
Possible	64	152	12	29.63	84.21	43.84
Conditional	0	0	0	0	0	0
Not patient	0	0	0	0	0	0
Overall	1350	214	214	86.32	86.32	86.32

Table 7: Results from applying a modified version of the ConText algorithm to a dataset of 70 CT Lung radiology reports containing 1,564 medical problems.

Category	TP	FN	FP	R	P	F1
Present	929	11	125	98.82	88.14	93.17
Absent	392	16	9	96.07	97.75	96.90
Hypothetical	0	0	1	0	0	0
Possible	101	115	4	46.75	96.19	62.92
Conditional	0	0	0	0	0	0
Not patient	0	0	3	0	0	0
Overall	1422	142	142	90.92	90.92	90.92

Table 8: Results from applying an SVM classifier to a dataset of 70 CT Lung radiology reports containing 1,564 medical problems.

ConText algorithm developed for discharge summaries dropped from an F1-score of 86.17 to 83.12 (Tables 3 and 6) on the new domain (statistically significant with a confidence level of 99%).

As shown in Tables 5 and 8, the machine learning system dropped in performance from an F1-score of 93.32 to 90.92 (statistically significant with a confidence level of 99%). The performance of the SVM classifier was hindered by the new dataset as the system was trained on 'genre-specific' features such as Section Headings. Even though performance dropped, the machine learning system still significantly outperformed the rule-based approach.

4 Conclusions

Clinical IE systems face the challenge of content and format differences across narrative sub-genres and environments. As publicly available clinical corpora are sparse, the practical question of re-use of existing training corpora arises. While not all clinical IE tasks would render themselves to annotated corpora re-use, the task of contextual information extraction is common across clinical sub-domains and expectations were that tools developed using the existing corpora could be ported to different types of clinical texts.

We developed two systems using the recently released i2b2 corpus containing medical problems annotated with contextual information. The performance of the two systems was evaluated against an independent dataset of clinical records from a different domain - radiology. Performance of the SVM-based machine learning system deteriorated, while performance the rule-based system proved more robust. However, the machine learning system still significantly outperformed the rule-based approach. Results suggest that adapting systems for identifying contextual information can be avoided as they can be successfully ported to new clinical domains.

References

A.R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

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

W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

W.W. Chapman, D. Chu, and J.N. Dowling. 2007. ConText: An algorithm for identifying contextual features from clinical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 81–88. Association for Computational Linguistics.

D.H. Cunningham, D.D. Maynard, D.K. Bontcheva, and M.V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications.

S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and JF Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, page 128.

G.K. Savova, K. Kipper-Schuler, J.D. Buntrock, and C.G. Chute. 2008. UIMA-based Clinical Information Extraction System. In *Proc. UIMA for NLP Workshop. LREC*.

G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45. Association for Computational Linguistics.