Rule-based Approach for Identifying Assertions in Clinical Free-Text Data

Yue Kimi Sun
Faculty of Science & Technology
Queensland University of Technology
QLD 4001 Australia
Kimi.Sun@csiro.au

Anthony Nguyen
The Australian e-Health Research Centre
CSIRO
QLD 4029 Australia
Anthony.Nguyen@csiro.au

Laurianne Sitbon
Faculty of Science & Technology
Queensland University of Technology
QLD 4001 Australia
Laurianne.Sitbon@qut.edu.au

Shlomo Geva
Faculty of Science & Technology
Queensland University of Technology
QLD 4001 Australia
S.Geva@qut.edu.au

Abstract *A rule-based approach for classifying previously identified medical concepts in the clinical free text into an assertion category is presented.There are six different categories of assertions for the task: Present, Absent, Possible, Conditional, Hypothetical and Not associated with the patient. The assertion classification algorithms were largely based on extending the popular NegEx and Context algorithms. In addition, a health based clinical terminology called SNOMED CT and other publicly available dictionaries were used to classify assertions, which did not fit the NegEx/Context model. The data for this task includes discharge summaries from Partners HealthCare and from Beth Israel Deaconess Medical Centre, as well as discharge summaries and progress notes from University of Pittsburgh Medical Centre. The set consists of 349 discharge reports, each with pairs of ground truth concept and assertion files for system development, and 477 reports for evaluation. The system's performance on the evaluation data set was 0.83, 0.83 and 0.83 for recall, precision and F1-measure, respectively. Although the rule-based system shows promise, further improvements can be made by incorporating machine learning approaches.*

Keywords rule-based, medical concept, assertion, NegEx, Context, SNOMED CT.

1 Introduction

A large part of clinical data is recorded in natural language, which makes algorithmic processing by a computer a very hard task. Three sequential tasks defined by the i2b2 NLP Challenge [1] consist of Concept Annotation, Assertion Annotation and Relation Annotation, which are three fundamental steps for processing clinical data. The Concept Annotation task builds toward the Assertion and Relation tasks of the challenge. This means that, the output of the Concept task is used as input to the Assertion task, and the output of both the Concept and Assertion task can be used for the Relation task.

In this paper, only the Assertion Annotation task was studied. In the context of the i2b2 NLP Challenge, an Assertion is defined as a contextual attribute (either 1. Present, 2. Absent, 3. Possible, 4. Conditional, 5. Hypothetical or 6. Not associated with the patient) that is applied to a concept relating to a medical problem.

2 System Description

The system was developed using GATE [1], an open source framework for developing and deploying software components that process natural language. Figure 1 shows the architecture of the assertion classification system. It consists of three stages, namely: 1) Preprocessing, 2) Assertion relevance matching, and 3) Assertion generation.

The system was largely based on a popular regular expression based negation/context algorithm [2, 3], which has been proven to work well with clinical free text data. Additional algorithms were also developed to accommodate assertions that cannot be classified using the NegEx/Context approach.

[1] Fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data. https://www.i2b2.org/NLP/Relations/

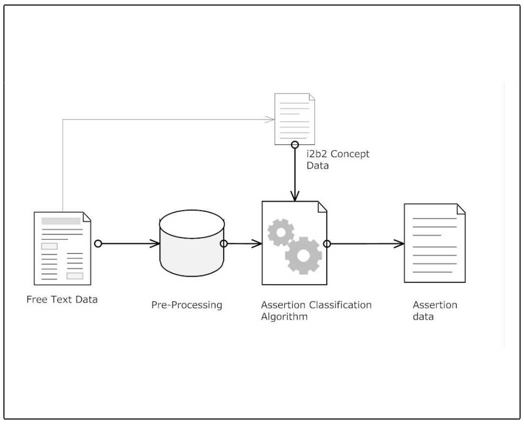

Figure 1: Assertion classification system.

For the Assertion Annotation task, the system is required to generate an assertion category for each concept identified as a medical problem. The input concept data is assumed to be available by the assertion classification system. For the purposes of system development and evaluation, the concept data is provided by human experts for each team. The problem of categorizing concepts into assertion classes is a typical classification task. Figure 2 shows the corpus statistics to the assertion classification task, there were 11968 concepts relating to medical problems in training data for system development, with another 18550 concepts which were used for testing.

	Training	Test	Total
#documents	349	377	726
# annotations	27,837	45,009	72,846
Test	7,369	12,899	20,268
Treatment	8,500	13,560	22,060
Problem	11,968	18,550	30,518
Present	8,052	13,025	21,077
Absent	2,535	3,609	6,144
Possible	535	883	1,418
Hypothetical	651	717	1,368
Conditional	103	171	274
Unassociated	92	145	237

Figure 2: Corpus statistics to assertion classification task

2.1 Preprocessing

The preprocessing step performs the tagging of entities such as tokens, sentences and concepts which were required for the assertion relevance matching stage.

The tokeniser splits the text into simple tokens which were separated by a space. Sentences were separated by line breaks, since this was the general structure in which the reports were formatted. These tokens and sentence annotations were used to annotate the i2b2 input concept data.

Although, the tokeniser and sentence splitter were simplified for this task, in practice more sophisticated algorithms would be required to distinguish sentence boundaries from tokens such as decimal numbers, punctuations and abbreviations. Automatically mapping medical concepts from free text would also be required in practice, since concept annotations are generally not available. A number of medical concept annotators exist, however, their performance may vary [4, 5].

2.2 Contextual analysis

We hypothesized that each assertion category could be largely classified using the methodology adopted in NegEx [2] or more generally the Context [3] algorithm. Context identifies common assertions phrases in the free text, and subsequently applies the respective assertion to a concept (or indexed term) based on a regular expression based template and the type of assertion phrase that was found.

Two types of assertion phrases were defined, namely, pre-assertion and post-assertion phrases. Pre-assertion phrases occur before the term (or concept) they assert, while the post-assertion phrases occur after the term they assert. For example, "pre-assertion" phrases would apply to concepts appearing after the assertion phrase (e.g., the sentence "The patient <pre-negation>denies<pre-negation><concept>chest pain<concept>", would assert the concept "chest pain" as "absent"), and vice versa for "post-assertion" phrases. The scope of search for concepts to apply the assertion was bounded by conjunction phrases and/or sentence boundaries.

The list of assertion phrases used in Context was extended and updated using examples from the i2b2 development data set. This demands a lot of knowledge about the domain language itself to correctly identify assertion phrases.

The algorithm was also extended to incorporate possibility phrases which assert uncertainty between two concepts. An example of a possibility phrase commonly occurring between two concepts is "versus" (or its variants). In such a case, the two concepts appearing before and after the possibility phrase would both be asserted as "possible".

2.3 Self asserted concepts

Although the algorithm above would associate concepts with assertions according to the context surrounding the concept, it cannot classify assertions to concepts when the meaning of morphology of the concept implies the assertion. For example, concepts such as "afebrile" and "nontender" would be considered "self-asserted" concepts and be classified as an absent assertion. To address this limitation, the health based ontology SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) [6] and publicly available dictionaries were incorporated. SNOMED CT is a systematically organized computer processable collection of medical terminology covering diseases, findings, procedures, pharmaceuticals etc. Among these, the concept "Clinical Finding Absent" was used to test if it subsumes (or is an ancestor of) medical concepts found in the free text. If subsumed, then the concepts would be asserted as absent. An in-house ontology server was used to query the subsumption relationships.

In addition, publicly available dictionaries from Internet were incorporated to further identify self-asserted concepts. A public resource from the Internet [8], which consists of 31 English dictionaries (covering 869,228 words or terms), was included in the system. It was conjectured that concepts containing known prefixes representing an absent assertion such as "non" would contain a stem of a word when the prefix was removed. If the stem of the concept is found in the dictionary, then the concept would be considered a "self-asserted" concept and be classified as an absent assertion.

2.4 Post Processing

Post-processing of the assertions was performed to ensure that each concept contains only a single class of assertion. If more than one class of assertion exists for a given concept, the choice of assertion was selected depending on a priority list given by:

1. Not associated with the patient

2. Hypothetical

3. Conditional

4. Possible

5. Absent

6. Present

3 System Evaluation

The i2b2 / VA Challenge data set consists of 349 discharge reports, each with pairs of ground truth concept and assertion files for system development, and 477 reports for evaluation.

The system was evaluated using recall, precision and F1-score measures.

	Annotations		
	Recall	Precision	F1-measure
Absent	0.83	0.89	0.86
Associated with_someone_else	0.52	0.70	0.59
Conditional	0.58	0.23	0.33
Hypothetical	0.74	0.93	0.82
Possible	0.49	0.50	0.49
Present	0.88	0.90	0.89
Summary	0.84	0.87	0.85

Figure 3: Overall System Performance on training data

Overall performance on the 2010 i2b2 /VA Challenge training corpus of 349 discharge reports against a database of ground truth assertion decisions are shown in Figure 3, and resulted in a recall, precision and F1-measure of 0.84,0.87, and 0.85, respectively.

	Annotations		
	Recall	Precision	F1-measure
Absent	0.77	0.87	0.82
Associated with_someone_else	0.56	0.68	0.61
Conditional	0.19	0.11	0.14
Hypothetical	0.58	0.80	0.68
Possible	0.44	0.49	0.47
Present	0.76	0.90	0.82
Summary	0.73	0.85	0.79

Figure 4: System Performance on testing data by only use Contextual Analysis

	Annotations		
	Recall	Precision	F1-measure
Absent	0.85	0.81	0.83
Associated with_someone_else	0.61	0.69	0.65
Conditional	0.24	0.13	0.16
Hypothetical	0.65	0.83	0.73
Possible	0.49	0.48	0.48
Present	0.86	0.88	0.87
Summary	0.83	0.83	0.83

Figure 5: Overall System Performance on testing data

The performance on the testing data are shown in Figures 4 (Contextual analysis only) and Figure 5 (Contextual analysis and self-assertions). The performance of Contextual analysis algorithm on the 2010 i2b2 /VA Challenge test corpus of 477 discharge reports against a database of ground truth assertion decisions were 0.73, 0.85, and 0.79 for recall, precision and F1-measure, respectively. This is the baseline performance for the core NegEx and Context algorithms, which didn't include the processing of self asserted concepts as described in section 2.3.

The performance improves further when self-asserted concepts are incorporated. Overall performance on the test corpus were 0.83, 0.83, and 0.83 for recall, precision and F1-measure, respectively. While the performance of the system shows promise, the methodology could be much improved to enhance the performance of the less prevalent assertion classes.

4 Possible Improvements

The proposed rule-based system shows promise but is limited in performance compared with the best performing Supervised or Hybrid systems, which can perform up to 0.93 for recall, precision and F1-measure.

The contextual analysis based algorithm is limited to the list of assertion phrases known to the system and unable to always make linguistic sense or are consistent with various types of semantic constraints. New unseen phrases will therefore be overlooked and result in misclassifications. The assertion phrases themselves are also subject to a trade-off between recall and precision. Significant knowledge about the domain language itself to correctly identify assertion phrases is thus necessary. For example, one word could completely change the sense of a statement. The statement could then be inverted, weakened or amplified. The following simple example by Horn [7] shows this effect in negated sentences:

1. I'm not tired.

2. I'm not a bit tired. (which equals "I'm not at all tired.")

3. I'm not a little tired. (which equals "I'm quite tired.")

The algorithm can also be extended to take into account of the low-level POS (Part of Speech) and grammatical sentence structure and/or use machine learning based approaches such as Conditional Random Fields (CRF) to learn the association between the phrases in the free text and the possible assertions that they represent. One the other hand, active learning methods maybe useful for selectively sampling (as opposed to randomly sampling) from a large corpus for tagging using various entropy-based scores [9].

5 Conclusion

A simple rule-based approach for classifying previously identified medical concepts in the clinical free text into an assertion category was proposed and shows promise. Further improvements can be made by incorporating machine learning approaches to learn the associations between concepts and assertions that are difficult to achieve with rule-based approaches.

Acknowledgements The data for this project includes discharge summaries from Partners HealthCare and from Beth Israel Deaconess Medical Center (MIMIC II Database), as well as discharge summaries and progress notes from University of Pittsburgh Medical Center. We would like to thanks for their contributions to this task.

References

[1] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Marin Dimitrov et al. Developing Language Processing Components with GATE Version 5.2009;1:17-25.

[2] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001, 34:301-10.

[3] Henk Harkema, John N. Dowling, Tyler Thornblade, Wendy W. Chapman, ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports, Journal of Biomedical Informatics, Volume 42, Issue 5, Biomedical Natural Language Processing, October 2009, Pages 839-851.

[4] Aronson, AR, Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program, American Medical Informatics Association (AMIA) Annual Symposium, pp. 17-21, 2001.

[5] Nguyen AN, Lawley MJ, Hansen DP, Colquist S., A Simple Pipeline Application for Identifying and Negating SNOMED Clinical Terminology in Free Text. Health Informatics Conference, Canberra, Australia. pp. 188-193, August 2009.

[6] International Health Terminology Standards Development Organisation , SNOMED CT, http://www.ihtsdo.org/snomed-ct/

[7] Laurence R. Horn. A natural history of negation. University of Chicago Press, Chicago, Illinois, June 15 1989.

[8] http://www.linux-pour-lesnuls.com/traduc/Dictionnaires/

[9] Tang, M, Luo, X, and Roukos, S. Active learning for statistical natural language parsing, in Proc. of the 40th Annual Meeting of the Association for Computational Linguistics pp 120-127 July 7-12 Philadelphia, PA 2002.