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

Yue K. Sun[1], Anthony N. Nguyen, PhD[2], Shlomo Geva, PhD[1], Laurianne Sitbon, PhD[3]
[1]Faculty of Science & Technology, Queensland University of Technology, Brisbane, Australia; [2]The Australian e-Health Research Centre, CSIRO, Brisbane, Australia; [3]The School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

Abstract

A rule-based approach for classifying previously identified medical concepts in the clinical free text into an assertion category for the 2010 i2b2 / VA Challenge 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. Overall performance 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.

1 Introduction

A large part of clinical data is recorded in natural language, which makes algorithmic processing by a computer a very hard task. This year is the 4th year of the i2b2 / VA Challenge for clinical text data processing. Three sequential tasks were defined for the challenge this year and consist of Concept Annotation, Assertion Annotation and Relation Annotation. 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 that is applied to a concept relating to a medical problem. According to 2010 i2b2 / VA Challenge Evaluation Assertion Annotation Guidelines[1], concepts identified as medical problems are assigned to one of the six pre-defined assertion categories, namely 1) Present, 2) Absent, 3) Possible, 4) Conditional, 5) Hypothetical or 6) Not associated with the patient.

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.

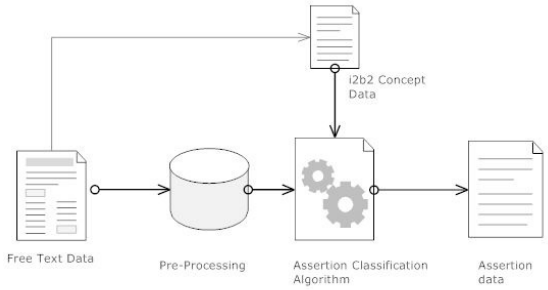

Figure 1. Assertion classification system

For the Assertion Annotation task, the system is required to generate an assertion category for each concept identified as a medical problem. The input concepts data is assumed to be available by the assertion classification system. For the purposes of system development and evaluation, the concepts data is provided by the challenge. The problem of categorizing concepts into assertion classes is a typical classification task.

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

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 concepts data.

Although, the tokeniser and sentence splitter were simplified for the i2b2 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 concept annotators exist, however, their performance may vary [4, 5]

2.2 Assertion Relevance Matching

In this part, the aim is to assign one of the six categories of assertions (Present, Absent, Possible, Conditional, Hypothetical or Not associated with the patient) to concepts relating to medical problems.

2.2.1 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 has been 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".

2.2.2 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 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[2], 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.

3 System Evaluation

The i2b2 / VA Challenge data set consists of 348 discharge reports, each with pairs of ground truth

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

concept and assertion files for system development, and 477 reports for evaluation.

The system was evaluated using recall, precision and F1-score measures. The system performance on the development and test data are shown in Tables 1 and 2.

	Annotations		
	Recall	Precision	F1-measure
Absent	0.81	0.84	0.82
Associated with_someone_else	0.51	0.66	0.57
Conditional	0.57	0.22	0.32
Hypothetical	0.66	0.83	0.74
Possible	0.46	0.45	0.45
Present	0.78	0.77	0.77
Summary	0.76	0.76	0.76

Table 1 System Performance on training data

	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

Table 2 System Performance on testing data

Overall performance 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. While the performance of the system shows promise, the methodology could be much improved to enhance the performance of the less prevalent assertion classes.

Compared with training data, there is a drop in Recall. This is because the system itself was tuned to achieve best possible overall Recall, Precision and F1 measure by the given ground truth Assertions during development period. As a result, new assertion phrases, for example, would be unknown to the system.

5 Possible Improvements

The NegEx/Context based algorithm is limited to the list of assertion phrases known to the system. New unseen phrases will therefore be overlooked and result in misclassifications. The assertion phrases themselves are also subject to a tradeoff between recall and precision. Significant knowledge about the domain language itself to correctly identify assertion phrases is thus necessary.

The algorithm can also be extended to take into account of the low-level 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.

The assertion phrases used for detection are also subject to a recall-precision trade-off, The correct understanding of an assertion requires the complete knowledge of the intended language and domain. 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.")

So searching for the best negation algorithm to use is critical. Additional information like POS (Part of Speech) tagging or sentence structure may help to achieve this goal.

References

1. Hamish Cunningham, Diana Maynard, Kalina Bontcheva 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.