

2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text

Özlem Uzuner,¹ Brett R South,^{2,3,4} Shuying Shen,^{2,3,4} Scott L DuVall^{2,3}

► Additional supplements and tables are published online only. To view these files please visit the journal online (www.jamia.org).

¹Department of Information Studies, University at Albany, State University of New York, Albany, New York, USA

²VA Salt Lake City Health Care System, Salt Lake City, Utah, USA

³Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA

⁴Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA

Correspondence to

Dr Özlem Uzuner, College of Computing and Information, University at Albany, SUNY, Draper 114A, 135 Western Ave, Albany, NY 12222, USA; ouzuner@albany.edu

Received 22 February 2011

Accepted 18 April 2011

Published Online First

16 June 2011

ABSTRACT

The 2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records presented three tasks: a concept extraction task focused on the extraction of medical concepts from patient reports; an assertion classification task focused on assigning assertion types for medical problem concepts; and a relation classification task focused on assigning relation types that hold between medical problems, tests, and treatments. i2b2 and the VA provided an annotated reference standard corpus for the three tasks. Using this reference standard, 22 systems were developed for concept extraction, 21 for assertion classification, and 16 for relation classification. These systems showed that machine learning approaches could be augmented with rule-based systems to determine concepts, assertions, and relations. Depending on the task, the rule-based systems can either provide input for machine learning or post-process the output of machine learning. Ensembles of classifiers, information from unlabeled data, and external knowledge sources can help when the training data are inadequate.

INTRODUCTION AND RELATED WORK

Annotated corpora support the development of natural language processing (NLP) systems. In the clinical domain, annotated corpora are not only expensive but also often unavailable for research due to patient privacy and confidentiality requirements. In 2010, i2b2 partnered with VA Salt Lake City Health Care System in manually annotating patient reports from three institutions and created a challenge in which the research community could participate in a head-to-head comparison of their systems. We refer to this challenge as the 2010 i2b2/VA challenge; we refer to the tasks in this challenge as concept extraction, assertion classification, and relation classification.

The 2010 i2b2/VA challenge continued i2b2's efforts to release clinical records to the medical language processing research community. This challenge built on past shared-tasks and challenges^{1–10} (see online supplements at www.jamia.org). It extended previous challenges to new types of data, concepts, assertions, and relations.

DATA

Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center contributed discharge summaries to the 2010 i2b2/VA challenge. In addition, the University of Pittsburgh Medical Center contributed progress reports. A total of 394 training reports, 477 test reports, and 877 unannotated reports were de-identified and released to challenge participants with data use agreements. Table 1 in the online

supplements shows the number of reports from each institution, the division of reports into training and testing sets, and the number of samples in each category of each task in the reference standard. These data will be available to the research community at large in November 2011 from <https://i2b2.org/NLP/DataSets> under data use agreements. An outline of the annotation workflow used for data generation is available online at www.jamia.org.

METHODS

Concept extraction was designed as an information extraction task.^{1–9–10} Given unannotated text of patient reports, systems had to identify and extract the text corresponding to patient medical problems, treatments, and tests.

Assertion classification was run on reports annotated with the reference standard concepts. Its goal was to classify the assertions made on given medical concepts as being present, absent, or possible in the patient, conditionally present in the patient under certain circumstances, hypothetically present in the patient at some future point, and mentioned in the patient report but associated with someone other than the patient. This task extended traditional negation and uncertainty extraction^{11–12} to conditional and hypothetical medical problems and brought in information about the person to whom the medical problem belonged.^{12–13}

Relation classification aimed to classify relations of pairs of given reference standard concepts from a sentence.¹⁴ Box 1 shows the relations annotated for the 2010 i2b2/VA challenge.

Figure 1 shows an excerpt of a patient report and its reference standard for concepts, assertions, and relations.

Evaluation metrics

We evaluated systems using precision, recall, and the F₁ measure (equations 1, 2, and 3). These metrics rely on true positives (TP), false positives (FP), and false negatives (FN) which are defined as appropriate in order to provide exact and inexact evaluation of the tasks (see online supplements).

$$\text{Precision(P)} = \text{TP}/(\text{TP} + \text{FP}) \quad 1$$

$$\text{Recall(R)} = \text{TP}/(\text{TP} + \text{FN}) \quad 2$$

$$F_1 = 2 \times P \times R / (P + R) \quad 3$$

Significance tests

We used the z test on two proportions for testing the significance of system performance

Box 1 Relation annotated for the i2b2/VA challenge

1. Medical problem—treatment relations:
 - a. Treatment improves medical problem (TrIP). Includes mention where the treatment improves or cures the problem, for example, *hypertension* was controlled on *hydrochlorothiazide*.
 - b. Treatment worsens medical problem (TrWP). Includes mentions where the treatment is administered for the problem but does not cure the problem, does not improve the problem, or makes the problem worse, for example, the *tumor* was growing despite the available *chemotherapeutic regimen*.
 - c. Treatment causes medical problem (TrCP). The implied context is that the treatment was not administered for the medical problem that it ended up causing, for example, *Bactrim* could be a cause of *these abnormalities*.
 - d. Treatment is administered for medical problem (TrAP). Includes mention where a treatment is given for a problem, but outcome is not mentioned in the sentence, for example, he was given *Lasix* periodically to prevent him from going into *congestive heart failure*.
 - e. Treatment is not administered because of medical problem (TrNAP). Includes mentions where treatment was not given or discontinued because of a medical problem that the treatment did not cause, for example, *Relafen* which is contraindicated because of *ulcers*.
 - f. Treatments and problems that are in the same sentence, but do not fit into one of the above defined relationships, are not assigned a relationship.
2. Medical problem—test relations:
 - a. Test reveals medical problem (TeRP). Includes mentions where a test is conducted and the outcome is not known, for example, *an echocardiogram* revealed a *pericardial effusion*.
 - b. Test conducted to investigate medical problem (TeCP). Includes mentions where a test is conducted and the outcome is not known, for example, a *VQ scan* was performed to investigate *pulmonary embolus*.
 - c. Tests and problems that are in the same sentence, but do not fit into one of the above-defined relationships, are not assigned a relationship.
3. Medical problem—medical problem relations:
 - a. Medical problem indicates medical problem (PIP). Includes medical problems that describe or reveal aspects of the same medical problem and those that cause other medical problem, for example, *Azotemia* presumed secondary to *sepsis*.
 - b. Pairs of medical problems that are in the same sentence, but do not fit into PIP relationship, are not assigned a relationship.

differences.¹⁵ We used a z score of 1.96, which corresponds to $\alpha=0.05$.

Systems

2010 i2b2/VA challenge systems were evaluated on held out test data. These systems included 22 for concept extraction, 21 for assertion classification, and 16 for relation classification. These systems were grouped with respect to their use of external

resources, involvement of medical experts, and methods (see online supplements for definitions).

The most effective concept extraction systems used conditional random fields (CRFs)^{16–24}; the only exception was by deBruijn *et al.*²⁵ Gurulingappa *et al.*¹⁸ trained CRFs on textual features enhanced with the output of a rule-based named entity recognition system.²⁶ Roberts *et al.*²³ broke the concept extraction task into two steps, so that in the first step they trained a CRF on identifying concept boundaries and in the second step they determined the class of the concept. Some others^{16 17 20} utilized CRFs in an ensemble, either of existing named entity recognition systems and chunkers^{27–31} or of different algorithms, with input based on knowledge-rich sources.^{32 33} Jonnalagadda and Gonzalez²¹ applied a semi-supervised CRF that utilized ‘distributional semantics’ features.³⁴

Most effective assertion classification systems used support vector machines (SVMs),^{16 23 25 35–40} either with contextual information and dictionaries that indicate negation, uncertainty, and family history,^{36 39} or with the output of rule-based systems.^{16 35 37} Roberts *et al.*²³ and Chang *et al.*⁴⁰ used both dictionaries and rule-based systems. Chang *et al.* complemented SVMs with logistic regression, multi-logistic regression, and boosting, which they combined using voting. deBruijn *et al.*²⁵ created an ensemble whose final output was determined by a multi-class SVM. Clark *et al.*⁴¹ used a CRF to determine negation and uncertainty with their scope, and added sets of rules to separate documents into zones, to identify cue phrases, to scope cue phrases, and to determine phrase status. They combined the results from the found cues and the phrase status module with a maximum entropy classifier that also used concept and contextual features.

SVMs were also the common theme among the most effective relation extraction systems.^{19 23 35–39 42} Given that our corpus contained an abundance of concept pairs with no relations, some of the relation extraction systems chose to first separate those pairs with relations from those with no relation and as a second step identified the nature of the relation.^{21 25 39} Anick *et al.*³⁹ used lists of n-grams with specific semantics, Divita *et al.*³⁷ augmented the reference standard for the least prevalent relation classes, Demner-Fushman *et al.*³⁵ utilized UMLS CUIs⁴³ and exercised feature reduction through cross-validation, and Grouin *et al.*³⁶ complemented their machine learning component with hand-built linguistic patterns and made use of simplified representations of text. Last but not least, deBruijn *et al.*²⁵ corrected for the label imbalance in the training data, calculated the ‘relatedness’ of two concepts using pointwise mutual information in Medline, and bootstrapped with unlabeled examples.

In comparison to systems developed for previous challenges, the 2010 challenge systems showed novel uses of combinations and ensembles as applied to concept, assertion, and relation tasks. These combinations could have multiple layers, for example, the output of one system is a direct input to another which in turn participates in a voting scheme. The combinations and ensembles leveraged the complementary strengths of systems that by themselves could address (portions of the) concept, assertion, and relation tasks. When used in a combination/ensemble, these systems gave state of the art results.

Challenge 2010 systems adapted open-domain NLP methods in concept and relation tasks to the clinical NLP domain. However, challenge 2010 tasks are different from their counterparts in open-domain NLP. For example, the concept of a person is very narrow in open-domain NLP however, the concept of a treatment is very broad in challenge 2010 as it

Figure 1 Sample text excerpt, its concepts, assertions, and relations.

Line #	Input Text
25	The patient is a 63-year-old female with a three-year history of bilateral hand numbness and occasional weakness .
26	Within the past year , these symptoms have progressively gotten worse , to encompass also her feet .
27	She had a workup by her neurologist and an MRI revealed a C5-6 disc herniation with cord compression and a T2 signal change at that level .
Concepts	
c="bilateral hand numbness" 25:11 25:13 t="problem"	
c="occasional weakness" 25:15 25:16 t="problem"	
c="these symptoms" 26:5 26:6 t="problem"	
c="a workup" 27:2 27:3 t="test"	
c="an mri" 27:8 27:9 t="test"	
c="a c5-6 disc herniation" 27:11 27:14 t="problem"	
c="cord compression" 27:16 27:17 t="problem"	
c="a t2 signal change" 27:19 27:22 t="problem"	
Assertions	
c="bilateral hand numbness" 25:11 25:13 t="problem" a="present"	
c="occasional weakness" 25:15 25:16 t="problem" a="present"	
c="these symptoms" 26:5 26:6 t="problem" a="present"	
c="a c5-6 disc herniation" 27:11 27:14 t="problem" a="present"	
c="cord compression" 27:16 27:17 t="problem" a="present"	
c="a t2 signal change" 27:19 27:22 t="problem" a="present"	
Relations	
c="an mri" 27:8 27:9 r="TeRP" c="a c5-6 disc herniation" 27:11 27:14	
c="an mri" 27:8 27:9 r="TeRP" c="cord compression" 27:16 27:17	
c="an mri" 27:8 27:9 r="TeRP" c="a t2 signal change" 27:19 27:22	
c="a c5-6 disc herniation" 27:11 27:14 r="PIP" c="cord compression" 27:16 27:17	

includes medications as well as procedures that can be given in response to a medical problem. Challenge 2010 relations are very fine-grained; for example, a number of relations can hold between a treatment and a medical problem. The number of relations that can link two concepts in most open-domain NLP tasks is usually much smaller.

Challenge 2010 systems required extensive feature engineering. The novelty of systems in this regard came not just from the specific features but also from the volume and variety of features they employed. The most successful systems performed feature selection from a vast volume of engineered features.

RESULTS

The most effective concept extraction system achieved an exact F measure of 0.852 (see table 2) and was significantly different from the rest of the concept extraction systems (see table 3 in the online supplements). Despite the diversity of mentions of treatments versus tests and problems, tables 4 and 5 in the

online supplements show that the system performances on the three categories were comparable.

In general, all concept extraction systems performed better in inexact evaluation than exact evaluation. The concept extraction systems benefited the most from textual features and reported disappointing gains from the inclusion of knowledge-rich resources such as the UMLS. Consequently, the best system in this task used a very high dimensional feature space with millions of textual features.²⁵ The most challenging examples for the concept extraction systems were abbreviations, for example, CXR for chest x-ray, and descriptive concept phrases, for example, *subtle decreased flow signal within the sylvian branches*.

Table 6 shows the systems' performance in the assertion and relation classification tasks. Significance tests in tables 7 and 8 in the online supplements show that the top four assertion classification systems were not significantly different from each other; similarly, the top two relation classification systems were not significantly different from each other. Tables 9–13 in the online supplements show the performance of systems on individual assertion and relation classes.

Assertion classification data contained ample examples of some assertion classes and scarce examples of others. In general, systems recognized the larger classes but even the input from dictionaries and rule-based systems did not help machine learning systems recognize the less prevalent classes.

The relation extraction task included a variety of relations with varying class sizes as well. The classifiers could capture the larger classes accurately by using basic textual features. They benefited from down-sampling the larger classes and were augmented with hand-built rules in order to recognize the less prevalent classes. We observed the lack of context in some of the relations found in the reference standard, indicating the possible use of domain knowledge in the annotation of these examples. In some other cases, the complexity of the language got in the way of relation extraction via machine-learning systems.

Table 2 Exact and inexact evaluation on the concept extraction task

Concept extraction					
System by	Medical experts	Method	External?	Exact F measure	Inexact F measure
deBruijn <i>et al</i> ²⁵	N	Semi-supervised	N	0.852	0.924
Jiang <i>et al</i> ¹⁶	Y	Hybrid	Y	0.839	0.913
Kang <i>et al</i> ¹⁷	N	Hybrid	Y	0.821	0.904
Gurulingappa <i>et al</i> ¹⁸	N	Supervised	Y	0.818	0.905
Patrick <i>et al</i> ¹⁹	N	Supervised	Y	0.818	0.898
Torii and Liu ²⁰	N	Supervised	N	0.813	0.898
Jonnalagadda and Gonzalez ²¹	N	Semi-supervised	N	0.809	0.901
Sasaki <i>et al</i> ²²	N	Supervised	N	0.802	0.887
Roberts <i>et al</i> ²³	N	Supervised	N	0.796	0.893
Pai <i>et al</i> ²⁴	Y	Hybrid	N	0.788	0.884

Table 6 Performance on the assertion and relation classification tasks

System by	Medical experts	Method	External?	F measure
Assertion classification				
deBruijn <i>et al</i> ²⁵	N	Semi-supervised	N	0.936
Clark <i>et al</i> ⁴¹	N	Hybrid	Y	0.934
Demner-Fushman <i>et al</i> ³⁵	N	Supervised	Y	0.933
Jiang <i>et al</i> ¹⁶	N	Hybrid	Y	0.931
Grouin <i>et al</i> ³⁶	N	Supervised	Y	0.931
Divita <i>et al</i> ³⁷	N	Supervised	Y	0.930
Cohen <i>et al</i> ³⁸	N	Supervised	N	0.928
Roberts <i>et al</i> ²³	N	Supervised	N	0.928
Anick <i>et al</i> ³⁹	Y	Supervised	Y	0.923
Chang <i>et al</i> ⁴⁰	N	Hybrid	Y	0.921
Relation classification				
Roberts <i>et al</i> ²³	N	Supervised	N	0.737
deBruijn <i>et al</i> ²⁵	N	Semi-supervised	N	0.731
Grouin <i>et al</i> ³⁶	N	Hybrid	Y	0.709
Patrick <i>et al</i> ¹⁹	N	Supervised	Y	0.702
Jonnalagadda and Gonzalez ²¹	N	Supervised	N	0.697
Divita <i>et al</i> ³⁷	Y	Supervised	Y	0.695
Solt <i>et al</i> ⁴²	N	Supervised	Y	0.671
Demner-Fushman <i>et al</i> ³⁵	N	Supervised	N	0.666
Anick <i>et al</i> ³⁹	Y	Supervised	Y	0.663
Cohen <i>et al</i> ³⁸	N	Supervised	N	0.656

DISCUSSION

The 2010 i2b2/VA challenge systems exhibit a trend toward ensembles of complementary approaches for improved performance. This trend is a result of the nature of the tasks: concept, assertion, and relation tasks are all disease-agnostic and linguistic. Although the concepts, assertions, and relations studied in the 2010 i2b2/VA challenge are unique and tailored toward clinical applications, the disease-agnostic and linguistic nature of the tasks allowed the systems developed for the open-domain NLP to be adapted and combined in order to respond to the 2010 challenge. In return, the methods developed for the 2010 challenge can be transferred back to open domain.

In order to continue the collaboration between clinical and open-domain NLP we hope to continue the disease-agnostic and linguistic tasks in the future. The 2011 i2b2/VA challenge will focus on co-reference resolution, a specific case of relation extraction where the relation is the 'equivalence' relation between two concepts. Again, the disease-agnostic and linguistic nature of this task will allow collaborations with the open-domain NLP field.

The 2010 i2b2/VA challenge overcame one large hurdle in making de-identified clinical records available to the research community. In 2010, for the first time, *multiple* institutions and multiple report types were shared with the community under data use agreements. As a result, the systems developed are not biased by the idiosyncrasies of individual data sets and can address the challenge tasks on two types of reports that come from multiple independent institutions. We expect that these systems also have a better chance of generalizing to other institutions' data. The diversification of the data sets is continuing in 2011, with data from more institutions and more report types becoming available to the research community.

Past i2b2 challenge data sets are now in use by more than 200 individuals, in addition to challenge participants, from all around the world. The data sets generated for these challenges have not only provided the essential elements for improvement

of the state of the art, but they also support education by being a major building block for NLP coursework in many academic institutions.

CONCLUSIONS

The 2010 i2b2/VA challenge evaluated systems on three tasks: concept extraction, assertion classification, and relation classification. The results of the challenge showed that of the three tasks, assertion classification was the easiest and best studied, concept extraction was relatively complex because of the difficulty of boundary detection for concepts, and relation classification, being the most novel task in the 2010 i2b2/VA challenge, was the most difficult. The best performance in relation extraction was 0.737, leaving about a quarter of the relations in the corpus incorrectly classified. The difficulty of classifying these relations comes from lack of explicit contextual information that describes the relations and/or the complexity of the language used in presenting the relations. While deeper syntactic analysis may help with the complex language, in the absence of context, domain knowledge may provide a good starting point.

Acknowledgments We thank all participating teams for their contributions to the challenge, Harvard Medical School for their technical support, and AMIA and Mediquist for co-sponsoring the workshop.

Funding This work was supported in part by the NIH Roadmap for Medical Research, Grant U54LM008748 from the NIH/National Library of Medicine (NLM). These efforts were undertaken as part of the 2010 i2b2/VA challenge that was in partnership with the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374. The views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs, NLM, or NIH. IRB approval has been granted for the studies presented in this manuscript.

Competing interests None.

Ethics approval This study was conducted with the approval of Partners HealthCare, the University of Pittsburgh, and Beth Israel Deaconess Medical Center.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Data are available from <https://i2b2.org/NLP> under a data use agreement.

REFERENCES

1. Uzun O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514–18.
2. Uzun O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;**14**:550–63.
3. Uzun O, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge summaries. *J Am Med Inform Assoc* 2008;**15**:14–24.
4. Uzun O. Recognizing obesity and co-morbidities in sparse data. *J Am Med Inform Assoc* 2009;**16**:561–70.
5. Uzun O, Solti I, Xia F, *et al*. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc* 2010;**17**:519–23.
6. Pestian JP, Brew C, Matykievicz P, *et al*. A shared task involving multi-label classification of clinical free text. In: *Proceedings of ACL*. Stroudsburg, PA, USA: Association for Computational Linguistics, BioNLP, 2007:97–104.
7. Hirschman L, Yeh A, Blaschke C, *et al*. Overview of BioCreAtivE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005;**6** (Suppl 1):S1.
8. Hersh W, Bhupatiraju RT, Corley S. Enhancing access to the bibliome: the TREC genomics track. *Stud Health Technol Inform* 2004;**11**:773–7.
9. Grishman R, Sundheim B. Message understanding conference-6: a brief history. *16th Conference on Computational Linguistics (COLING)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Stroudsburg, 1996:466–71.
10. Sparck Jones K. Reflections on TREC. *Inform Process Manag* 1995;**31**:291–314.
11. Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;**34**:301–10.
12. Uzun O, Zhang X, Sibanda T. Machine Learning and Rule-Based Approaches to Assertion Classification. *J Am Med Inform Assoc* 2009;**16**:109–15.
13. Chapman WW, Chu D, Dowling JN. *ConText: An Algorithm for Identifying Contextual Features from Clinical Text*. Prague: BioNLP 2007: Biological, Translational, and Clinical Language Processing, 2007:81–8.

14. **Uzuner Ö**, Mailoa J, Ryan RJ, *et al*. Semantic Relations for Problem-Oriented Medical Records. *Artif Intell Med* 2010;**50**:63–73.
15. **Osborn CE**. *Statistical Applications for Health Information Management*. 2nd edn. Boston, MA, USA: Jones & Bartlett Publishers, 2005.
16. **Jiang M**, Chen Y, Liu M, *et al*. Hybrid approaches to concept extraction and assertion classification - vanderbilt's systems for 2010 i2b2 NLP Challenge. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
17. **Kang N**, Barendse RJ, Afzal Z, *et al*. Erasmus MC approaches to the i2b2 Challenge. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
18. **Gurulingappa H**, Hofmann-Apitius M, Fluck J. Concept identification and assertion classification in patient health records. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
19. **Patrick JD**, Nguyen DHM, Wang Y, *et al*. i2b2 challenges in clinical natural language processing 2010. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
20. **Torii M**, Liu H. BioTagger-GM for detecting clinical concepts in electronic medical reports. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
21. **Jonnalagadda S**, Gonzalez G. Can distributional statistics aid clinical concept extraction? *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
22. **Sasaki Y**, Ishihara K, Yamamoto Y, *et al*. TTI's systems for 2010 i2b2/VA challenge. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
23. **Roberts K**, Rink B, Harabagiu S. Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
24. **Pai AK**, Agichtein E, Post AR, *et al*. The emory system for extracting medical concepts at 2010 i2b2 challenge: integrating natural language processing and machine learning techniques. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
25. **deBruijn B**, Cherry C, Kiritchenko S, *et al*. NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
26. **Hanisch D**, Fundel K, Mevissen HT, *et al*. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics* 2005;**6**(Suppl 1):S14.
27. **Settles B**. Biomedical named entity recognition using conditional random fields and rich feature sets. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*. Geneva, Switzerland: Association for Computational Linguistics, Stroudsburg, PA, USA. 2004:104–7.
28. **LingPipe**. <http://alias-i.com/lingpipe> (accessed 29 Nov 2010).
29. **Buyko E**, Wermter J, Poprat M, *et al*. Automatically adapting an NLP core engine to the biology domain. *Proceedings of the Joint BioLINK/Bio-Ontologies Meeting 2006*. Fortaleza, Brasil: 2006:65–8.
30. **Schuemie MJ**, Jelier R, Kors JA. Peregrine: lightweight gene name normalization by dictionary lookup. *Proceedings of the BioCreAtivE II Workshop*. Madrid, Spain: 2007:131–3.
31. **StanfordNer**. <http://nlp.stanford.edu/software/CRF-NER.shtml> (accessed 29 Nov 2010).
32. **Denny JC**, Smithers JD, Miller RA, *et al*. "Understanding" medical school curriculum content using knowledgeMap. *J Am Med Inform Assoc* 2003;**10**:351–62.
33. **Friedman C**, Alderson PO, Austin J, *et al*. A general natural language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;**1**:161–74.
34. **Jonnalagadda S**, Leaman R, Cohen T, *et al*. A distributional semantics approach to simultaneous recognition of multiple classes of named entities. *Computational Linguistics and Intelligent Text Processing (CICLing)*. LNCS Vol. **6008**. Springer-Verlag Berlin Heidelberg 2010.
35. **Demner-Fushman D**, Apostolova E, Islamaj Doğan R, *et al*. NLM's system description for the fourth i2b2/VA challenge. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
36. **Grouin C**, Abacha AB, Bernhard D, *et al*. CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
37. **Divita G**, Treitler OZ, Kim YJ, *et al*. Salt lake city VA's challenge submissions. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
38. **Cohen AM**, Ambert K, Yang J, *et al*. OHSU/portland VAMC team participation in the 2010 i2b2/VA challenge tasks. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
39. **Anick P**, Hong P, Xue N, *et al*. i2B2 2010 challenge: machine learning for information extraction from patient records. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
40. **Chang E**, Xu Y, Hong K, *et al*. A hybrid approach to extract structured information from narrative clinical discharge summaries. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
41. **Clark C**, Aberdeen J, Coarr M, *et al*. Determining assertion status for medical problems in clinical records. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
42. **Solt I**, Szidarovszky FP, Tikk D. Concept, assertion and relation extraction at the 2010 i2b2 relation extraction challenge using parsing information and dictionaries. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*. Boston, MA, USA: i2b2, 2010.
43. **Aronson A**. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. *AMIA Annu Symp Proc* 2001;**2001**:17–21.