Interactive Online Learning for Clinical Entity Recognition

Luis Tari
Knowledge Discovery Lab
GE Global Research
Niskayuna, NY, USA
luis.tari@outlook.com

Varish Mulwad
Knowledge Discovery Lab
GE Global Research
Niskayuna, NY, USA
varish.mulwad@ge.com

Anna von Reden
UPMC Enterprises
Pittsburgh, PA, USA
vonredenal@upmc.edu

ABSTRACT
Named entity recognition and entity linking are core natural language processing components that are predominantly solved by supervised machine learning approaches. Such supervised machine learning approaches require manual annotation of training data that can be expensive to compile. The applicability of supervised, machine learning-based entity recognition and linking components in real-world applications can be hindered by the limited availability of training data. In this paper, we propose a novel approach that uses ontologies as a basis for entity recognition and linking, and captures context of neighboring tokens of the entities of interest with vectors based on syntactic and semantic features. Our approach takes user feedback so that the vector-based model can be continuously updated in an online setting. Here we demonstrate our approach in a healthcare context, using it to recognize body part and imaging modality entities within clinical documents, and map these entities to the right concepts in the RadLex and NCIT medical ontologies. Our current evaluation shows promising results on a small set of clinical documents with a precision and recall of 0.841 and 0.966. The evaluation also demonstrates that our approach is capable of continuous performance improvement with increasing size of examples. We believe that our human-in-the-loop, online learning approach to entity recognition and linking shows promise that it is suitable for real-world applications.

CCS Concepts

Information systems • **Computing methodologies → Machine learning → Machine learning approaches → Instance-based learning.**

Keywords
Entity Recognition, Entity Linking, Online Learning, Natural Language Processing, Electronic Medical Records

1. INTRODUCTION
One of the common components in text analytics is identifying entities in text and mapping the entities to the right concepts based on their meaning. These processes are known as entity recognition and entity linking. Performance of a text analytics system can be affected by entities that can have multiple interpretations depending on their context. It is critical to achieve a high performance in identifying entities and their concepts in the clinical domain, as next-generation healthcare systems such as [1][2] are highly dependent on the quality of the extracted entities and their concept mappings. Examples of entities that can be commonly found to have multiple contextual meanings in clinical documents are the word "back", whether it refers to the body part or acts as a preposition, and the word "US", which may refer to ultrasound, the United States, or act as a pronoun.

Clinical documents encountered in patients' electronic medical records include office visits, inpatient progress notes, specialist consults, and other subjective assessments of a clinician's encounter with a patient. These documents provide details about a patient's situation that can greatly impact subsequent care decisions by downstream providers, including specialties like diagnostic imaging and pathology. However, these documents are frequently written in free-form text, and even structured fields in documents can vary by site. To better inform these providers, it is important to leverage automated approaches that can analyze natural language text in clinical documents and present a concise, summarized form of a patient's history and prior assessment. A core component of summarization includes the use of medical concept recognition in clinical documents.

Typical entity recognition and entity linking approaches rely on manually labeled data to train the systems for recognizing entities from text and linking the recognized entities to the corresponding concepts. Such training data can be expensive to compile, and updates to the system are dependent on software developers to repeat the training process and enhance the entity recognition and linking systems. In terms of the development lifecycle of systems, users of these systems play a passive role even when a large group of users are more likely to be the ones who identify unanticipated errors or improvements as compared to a small group of developers. Particularly in the field of medicine, the users themselves, i.e. the healthcare professionals, have the expertise to identify the appropriate meaning of medical terms. The main goal of this proposed method is to identify, extract and disambiguate extracted concepts from unstructured text, and deploy a human-in-the-loop process so that primary software users can actively provide feedback on the extracted entities. Such feedback is leveraged to continuously improve the entity recognition and linking system by our novel online learning method. The core of our proposed approach is the utilization of medical ontologies as a basis for entity recognition and linking. It is then extended with a vector-based, syntactic and semantic representation of the entities in text to capture the neighboring context. Positive and negative examples of entity-to-concept mapping are collected in the form of vectors so that a similarity comparison is performed on-the-fly when new entities are encountered. Such online learning ability supports incremental updates to the systems so that any kind of performance improvement to the model does not require scrapping the existing concept recognition and disambiguation models and restarting the training process.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org. HILDA'16, June 26 2016, San Francisco, CA, USA © 2016 ACM. ISBN 978-1-4503-4207-0/16/06...$15.00 DOI: http://dx.doi.org/10.1145/2939502.2939510

Our proposed approach takes user feedback in an interactive manner in order to improve the underlying model for entity recognition and linking. This mechanism of taking user feedback is different from typical active learning approaches [3][4], in which current active learning methodologies focus on finding the best data sampling strategy to select unlabeled examples and solicit user feedback to provide labels for the examples. The newly labeled examples are then fed into the system to repeat the training process with the existing labeled examples. Such training process usually takes a significant amount of time to compute so that errors that are identified by the users cannot be corrected and reflected to the system instantaneously. The lack of instantaneous update may affect the user confidence of the results over time.

The rest of the paper is organized as follows: in section 2, we provide an overview of prior work related to our proposed approach. Section 3 presents a detailed description of our approach, and section 4 with the evaluation approach and results. We conclude our paper with our findings and elaborate the future direction of the approach in section 5.

2. RELATED WORK

Existing approaches for entity recognition and entity linking can be broadly categorized into: dictionary-based, supervised, semi-supervised and active learning [5]. In the clinical domain, popular systems such as cTakes [6], MedLee [7] and MetaMap [8] perform clinical entity recognition based on the Unified Medical Language System (UMLS) ontology[1] together with syntactic patterns such as noun phrases. More recent approaches such as [9][10][11] leverage the clinical concept extraction task in 2010 i2b2 NLP/VA challenge [12] to develop machine learning approaches for clinical entity recognition and linking. Methods such as conditional random fields (CRF) and support vector machines are popular among the participants for the challenge.

Typical machine learning approaches adopt a batch training methodology that takes in a large amount of training data to create statistical models. To reduce the amount of training data, approaches such as active learning [13] leverage seed examples for initial training and select unlabeled examples for users to give feedback. Recent work in clinical entity recognition adopts such active learning methodologies show promising results in improving the entity recognition and linking performance. The core principle behind the active learning approach in [13] is to leverage CRF to perform training, and the studies focuses on identifying the best strategy in choosing the most effective unlabeled examples that would have the most impact to the performance of the recognition system. In their setting, the newly labeled examples together with the existing labeled examples are fed into the CRF training system to repeat the batch training process. Such model update takes time, which may hinder the user experience when feedback, especially the ones that are identified as false positives, is not instantaneously reflected to the performance of the entity recognition and linking system.

3. APPROACH

Our approach for identifying medical concepts in text corpora such as clinical documents involves the use of medical ontologies as the basis for entity recognition with a user feedback mechanism for online learning. In this paper, we refer an *entity* as word tokens that collectively indicate a concept in text, while a *concept* is referred as a thing in an ontology. An *extracted concept* corresponds to an entity that links to a particular concept in the ontology. The feedback-based online learning mechanism collects positive and negative examples so that these examples can be leveraged to disambiguate concepts used in different context. This mechanism is aimed to reduce the number of false positives when ontologies are utilized for dictionary matching for the input text. Our approach involves a *text processing* step and syntactic and semantic features are generated to form a vector representation of the entities in the *feature vector generation* step. A confidence score is assigned to each of the entities during the *confidence score computation* step, reflecting the likelihood of correctness for an entity to be linked to a particular concept in the ontologies.

3.1 Text Processing

The text processing step involves the use of typical natural language processing components to process the input text. In this case the input to our system is clinical documents such as office notes and progress notes. Apache OpenNLP[2] is applied to perform sentence segmentation, tokenization and parts-of-speech tagging, while ClearNLP [14] is used for lemmatization. To recognize concepts from text, UIMA Concept Mapper [15] is used with RadLex[3] and NCI Thesaurus[4] ontologies as the sources for the UIMA Concept Mapper dictionaries. This is achieved by translating the ontologies in RDF/OWL format into UIMA Concept Mapper dictionaries in XML format. Each of the concepts in the ontologies is translated into its canonical form and the variants for the dictionary format. Canonical form of a concept is represented by a uniformed resource identifier (URI) of the concept, while its variants include the preferred name and synonyms of the concept. When an entity is identified by the UIMA Concept Mapper, the entity is returned with the corresponding URI so that it can be treated as the mapping between the entity to the particular concept in the ontologies.

3.2 Feature Vector Generation

As in other vector-based approaches, we associate an entity e that is linked to concept c with a vector representation $x_{e\rightarrow c} = <f_1, \ldots, f_d>$, where f_i corresponds to one of the features among d number of features. A vector represents a positive example or a negative example of entity e linked to concept c. Vectors that correspond to positive examples of c are stored in matrix P_c while negative examples collectively as N_c. Our feature vector generation component takes the output of the text processing components as input to populate the elements in vector $x_{e\rightarrow c}$.

Features can be categorized into syntactic and semantic features. Syntactic features include *isAllUppercase* and part of speech features. Feature *isAllUppercase* is for identifying if all of the letters for an extracted entity are in uppercases. Another type of syntactic features is parts-of-speech features, denoted as *POS features*. Various parts-of-speech tags are considered for POS features: nouns (*NN*), all forms of verbs (*VB*), prepositions (*IN*), to (*TO*), numbers (*CD*), adjectives (*JJ*) and adverbs (*RB*). In addition, the POS features are applied to a window of n word tokens to the left and right of the entity. In this paper, we chose n to be 2. Suppose feature f_i represents the presence of noun for the entity of interest, features f_{i-2} and f_{i-1} refer to occurrences of nouns to the left of the entity, while features f_{i+1} and f_{i+2} correspond to the presence of nouns to the right of the entity. The semantic

[1] UMLS: https://www.nlm.nih.gov/research/umls/

[2] Apache OpenNLP: http://opennlp.apache.org/

[3] Radiology Lexicon (RadLex): http://www.radlex.org/

[4] NCI Thesaurus: https://ncit.nci.nih.gov/ncitbrowser/

features include the identification of the semantic types for the neighboring word tokens of the targeted entity. In particular, the following semantic types are considered: Imaging Modality (*IM*), Body Part (*BP*), Anatomy Modifier (*AM*), Diagnostic Procedure (*DP*), Disease and Disorder (*DD*) and Symptom (*SYM*). The semantic features are applied to the neighboring word tokens that are within a window of *n* word tokens from the entity. Entities are considered to be a member of a particular semantic type if the extracted concept is a subclass of a root concept. Table 1 lists all of the root concepts that we used from RadLex and NCI Thesaurus in identifying entity types.

We illustrate the feature vector generation process with an example shown in Figure 1. The sample sentence

"US thyroid was done through Metabolic Disease Associates, result of which was not available to us at time of completion of this visit."

in Figure 1 shows occurrences of the word tokens "US" and "us". The concept mapping component identifies the word tokens as entities referring to the concept *Ultrasound*. However, only the first occurrence "US" corresponds to the concept. For the sake of brevity, only the features *isAllUppercase*, *NN*, *VB*, *IN*, *BP* are shown in the vector representation for the two entities. The following features are assigned with values of 1 for the entity "US" (denoted as x_1, the short form of $x_{\mathrm{US}\rightarrow Ultrasound}$, in the figure):

- Feature f_1 is assigned as 1 since all letters of the entity are in uppercase letters.
- Features f_2 and f_3 correspond to the occurrence of nouns to the left of the entity, while f_5 and f_6 refer to nouns to the right of entity. Feature f_4 corresponds to the occurrence of noun for the entity, and it is marked as 1 since the word token "US" is recognized as a noun by the parts-of-speech tagger.
- Feature f_{11} corresponds to the verb "was" in the sentence and f_{19} refers to the entity "thyroid" which is recognized as a body part.

For the vector representation for the entity "us" (denoted as x_2, short form of $x_{\mathrm{us}\rightarrow Ultrasound}$ in the figure), the following features are assigned with values of 1:

- Feature f_6 corresponds to the word "time"
- Feature f_{15} refers to the word "at" in the sentence.

By comparing the two vectors x_1 and x_2, we can see that the signatures representing the two word occurrences "US" and "us" are very different. When x_1 is appended into matrix $P_{Ultrasound}$ and x_2 into matrix $N_{Ultrasound}$ as positive and negative examples, the next step is to systematically leverage these matrices to compute the likelihood of correctness for a new entity to be linked to the concept *Ultrasound*.

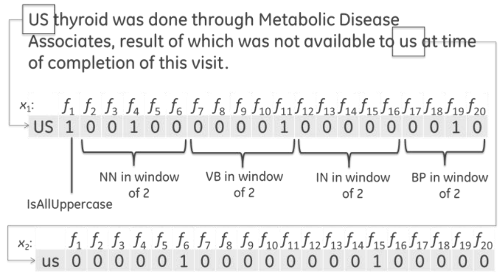

Figure 1. Vector representation of "US" and "us" corresponding to the positive and negative examples of the concept *Ultrasound*.

Table 1. List of ontology terms and their subclasses used in identifying entity types. "radlex" corresponds to the namespace for RadLex ontology and "ncit" for NCIT ontology.

URI	Label	Entity Type
radlex:RID10311	Imaging Modality	Imaging Modality (*IM*)
radlex:RID3	Anatomical Entity	Body Part (BP)
radlex:RID39102	Anatomy Modifier	Anatomy Modifier (*AM*)
radlex:RID13060	Imaging Procedure	Diagnostic Procedure (DP)
radlex:RID39050	Symptom	Symptom (SYM)
radlex:RID4736	Pathophysiologic Finding	Disease and Disorder (*DD*)
ncit:Anatomy_Modifier	Anatomy Modifier	Anatomy Modifier (*AM*)
ncit:Body_Cavity	Body Cavity	Body Part (BP)
ncit:Cardiac_Diagnostic_Procedure	Cardiac Diagnostic Procedure	Diagnostic Procedure (DP)
ncit:Diseases_Disorders_and_Findings	Diseases, Disorders and Findings	Disease and Disorder (*DD*)

3.3 Confidence Score Computation

The goal of confidence score computation is to determine the likelihood for an entity *e* to be linked correctly to concept *c*, i.e. the likelihood of correctness for an extracted concept denoted as $e\rightarrow c$. This is achieved by comparing the vector representation of $e\rightarrow c$ (denoted as $x_{e\rightarrow c}$) with the positive examples P_c and the negative examples N_c that have been collected for concept *c* in the disambiguation model. Intuitively, the confidence score for $e\rightarrow c$ is computed based on the level of similarity between $e\rightarrow c$ and the positive examples in P_c and the level of dissimilarity between $e\rightarrow c$ and the negative examples in N_c. The confidence score computation $score(e\rightarrow c)$ is characterized by the following formulas:

$$score(e \rightarrow c) = \begin{cases} 1 \; if \; N_c = \phi \\ conf(e \rightarrow c) \end{cases}$$

$$conf(e \rightarrow c) = w_p \cdot \max_{p \in P_c} similarity(x_{e\rightarrow c}, p) + w_n \left(1 - \max_{n \in N_c} similarity(x_{e\rightarrow c}, n)\right)$$

w_p and w_n corresponds to the weights, which are both assigned as 0.5 in our experiments. It is important to note and justify the rationale behind the assignment of $score(e \rightarrow c)$ as 1 when $N_c = \phi$, i.e. when there are no negative examples associated with concept c. The basis of our approach is to rely on the ontology to identify concepts, and disambiguation is needed only for concepts that can be used in different context. This can be reflected by the existence of negative examples for the concepts of interest, in which a concept with negative examples indicate that such concept can be used in multiple contexts in text. Such behavior is captured in the formulas above so that $conf(e \rightarrow c)$ is only computed for an extracted concept if negative examples of c have been collected. A high $score(e \rightarrow c)$ implies that $x_{e \rightarrow c}$ is deemed to be highly similar to one of the positive examples in P_c and highly dissimilar to one of the negative examples in N_c. Similarity between the extracted concept and an example is computed based on their corresponding vector representation with the use of cosine similarity defined as follows:

$$similarity(x, y) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}||\ ||\vec{y}||}$$

3.4 User Feedback

Concept recognition module, when deployed to the users, is usually treated as a black box so that any kind of errors that the users might observe from the output may only be kept in a log. Such errors might not be addressed until the developers resolve the issues and release them in the next software release cycle. This can affect the user's confidence in using the system, and results accuracy matters to healthcare professionals who depend on the results in making diagnosis decision.

The proposed approach described in this section supports the idea of online learning. Typical machine learning approaches perform batch learning so that training data is taken as input to generate statistical models. The batch learning process generally assumes that there is sufficient labeled data available for training, and improvement of the generated models would require a repeat of model training. On the other hand, online learning takes examples in increments so that models can be improved over time. In our case, positive and negative examples collected from user feedback are treated as increments to append the models. This is different from active learning approaches, as active learning mainly focuses on selecting the best unlabeled examples and presenting them to the users to give feedback. The newly labeled examples are then leveraged to perform training with the existing labeled examples.

Figure 2 describes the process in taking user feedback to append to an existing model so that the entity recognition and linking performance can be improved over time. Users provide text paragraphs for the system to process, and this process involves text processing, feature vector generation and confidence score computation. Extracted concepts with a confidence score > 0.5 are presented to the users for feedback. Through the user interface, users can mark an extracted concept as either positive or negative. Once the feedback is submitted, the system appends the examples and stores the corresponding vectors to the disambiguation model. Figure 3 shows a screenshot for our prototype in soliciting user feedback on extracted concepts.

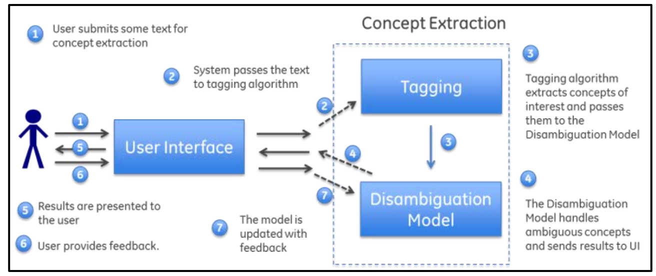

Figure 2. Workflow for the user feedback and model update processes.

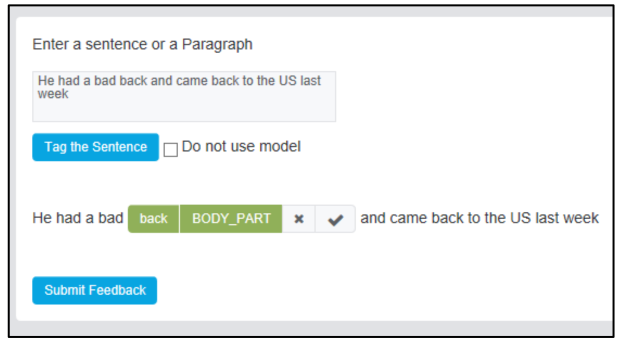

Figure 3. Screenshot of our prototype that takes user feedback to validate correctness of the extracted entities.

4. RESULTS

We evaluated the impact of user feedback by iteratively training the disambiguation model using examples from a training dataset of clinical documents. We manually labeled and generated a set of 228 positive examples and 296 negative examples from a corpus of clinical documents as our training dataset. We used a separate set of 20 clinical documents as our test dataset. Medical experts annotated the test dataset with expected Body Part and Imaging Modality concepts, which forms the basis for computing precision and recall. The test dataset composes of 354 body part and 27 imaging modality mentions.

We iteratively trained our disambiguation model by randomly selecting equal number of positive and negative examples from the training dataset. The selected examples go through text processing, feature vector generation and confidence score computation. Using the available feedback (positive or negative) extracted concepts with a confidence score > 0.5 along with their vector representation are appended to appropriate positive and negative matrices. The baseline was to only use the medical ontologies for concept recognition and linking. Then we examined the performance with the disambiguation model by iterating selecting examples for training. In iteration 1, we randomly sampled 12 positive and 12 negative examples to update the disambiguation model so that the concept extraction was applied to the test dataset to compute precision and recall. In iteration 2, we appended the model from iteration 1 with an additional 25 randomly selected positive examples and 25 negative examples. We re-computed precision and recall using the model updated at the end of iteration 2. For every other iteration, we sampled an additional 25 positive and 25 negative examples. Each iteration was attempted 3 times, and we reported the resulting averaged precision and recall.

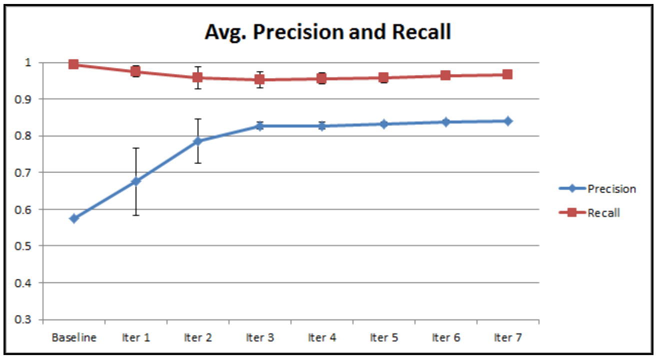

Figure 4. Comparison between baseline and disambiguation model for concepts belonging to Imaging Modality and Body Part. Average Precision and Recall is computed over three separate runs.

We compared the performance of the iterative training of the disambiguation model against a baseline model that simply used dictionary matching based on RadLex and NCIT ontologies. Figure 3 shows the averaged precision and recall with standard deviations over 7 iterations and each with 3 runs. Our results showed that our entity recognition with disambiguation model at iteration 7, in which a total of 324 examples was used, outperformed the baseline with an average precision of 0.84 as compared to 0.57. Figure 3 also demonstrates that the benefit of user feedback. The performance of the disambiguation model steadily improves as the number of training examples increases. Even with only 12 positive and 12 negative examples, the precision of 0.67 for our approach is significantly better than 0.57 for the baseline, while the recall across all the iterations remains competitive compared with the baseline. The high recall for the baseline comes at the cost of low precision, whereas the disambiguation model maintains both a high recall and high precision. The standard deviation for precision is slightly high for the first two iterations (0.09 and 0.05 respectively), but it quickly became stabilized, demonstrating the minimal impact of which examples were chosen to train the model.

5. DISCUSSION AND CONCLUSION

We described a novel online learning approach that is capable of continuous update to the entity recognition and linking model. We believe that an interactive and online learning approach is suitable to be applied to real-world applications such as clinical domain for the following reasons: (1) online learning allows updates to the model regardless to the software development and deployment cycle; (2) terminologies and acronyms may vary from one site to another so that an online learning approach has the opportunity to be adapted to the designated environment; (3) feedback taken via the user interaction can be leveraged to update and reflect changes instantaneously to improve user experience; (4) users in this case are expert users for the clinical domain such as healthcare professionals so that their feedback can be trustworthy.

This approach is a work in progress and hence we are only able to perform our evaluation with a small dataset of clinical documents. We plan to perform further extensive evaluation using the 2010 i2b2/VA NLP challenge dataset [12]. We will also explore different mechanisms to incorporate such human-in-the-loop online learning feature into a healthcare system. One possible mechanism is to use an approach similar to CAPCHA [16] that instead of presenting images to the users, the system can solicit validation from the users by presenting entities and their originating sentences.

6. REFERENCES

[1] Lalithsena, S., *et al*. 2015. Feedback-Driven Radiology Exam Report Retrieval with Semantics. International Conference on Healthcare Informatics (ICHI), 233-242.

[2] Mabotuwana, T., Lee, M. C. and Cohen-Solal, E. V. 2013. An ontology-based similarity measure for biomedical data application to radiology reports. Journal of Biomedical Informatics, 46(5), 857-868.

[3] Settles, B. 2010. Active learning literature survey. University of Wisconsin, Madison, 52.55-66: 11.

[4] Fu, Y., Zhu, X., and Li, B. 2013. A survey on instance selection for active learning. Knowledge and information systems, 35.2: 249-283.

[5] Shen, W., Wang, J., and Han, J. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering, 27.2: 443-460.

[6] Savova, G. K., *et al.* 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17.5, 507-513.

[7] Hripcsak, G., *et al*. 1995. Unlocking clinical data from narrative reports: a study of natural language processing. Annals of internal medicine, 122.9, 681-688.

[8] Aronson, A. R., and Lang, F. M. 2010. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17.3, 229-236.

[9] Patrick, J., and Li, M. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. Journal of the American Medical Informatics Association, 17.5, 524-527.

[10] de Bruijn, B., *et al*. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. Journal of the American Medical Informatics Association, 18.5, 557-562.

[11] Jiang, M., *et al*. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. Journal of the American Medical Informatics Association, 18.5, 601-606.

[12] Uzuner, Ö., *et al*. 2010. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18.5, 552-556.

[13] Chen, Y., Lasko, T.A., Mei, Q., Denny, J.C. and Xu, H. 2015. A study of active learning methods for named entity recognition in clinical text. Journal of Biomedical Informatics, 58, 11-18.

[14] Choi, J. D. and McCallum, A. 2013. Transition-based Dependency Parsing with Selectional Branching. Association for Computational Linguistics (ACL).

[15] Tanenblatt, M. A., Coden, A. and Sominsky, I. L. 2010. The ConceptMapper Approach to Named Entity Recognition. In Language Resources and Evaluation Conference (LREC).

[16] Von Ahn, L., et al. 2003. CAPTCHA: Using hard AI problems for security. Advances in Cryptology — EUROCRYPT 2003, Springer Berlin Heidelberg, 294-311.