Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2016 December ; 2016: 1081–1088. doi:10.1109/BIBM.2016.7822672.

Towards an Obesity-Cancer Knowledge Base: Biomedical Entity Identification and Relation Detection

Juan Antonio Lossio-Ventura^{*}, William Hogan^{*}, François Modave^{*}, Amanda Hicks^{*}, Josh Hanna^{*}, Yi Guo^{*}, Zhe He[†], and Jiang Bian^{*‡}

^{*}Health Outcomes & Policy, College of Medicine, University of Florida, Gainesville, Florida, USA

[†]School of Information, Florida State University, Tallahassee, Florida, USA

Abstract

Obesity is associated with increased risks of various types of cancer, as well as a wide range of other chronic diseases. On the other hand, access to health information activates patient participation, and improve their health outcomes. However, existing online information on obesity and its relationship to cancer is heterogeneous ranging from pre-clinical models and case studies to mere hypothesis-based scientific arguments. A formal knowledge representation (i.e., a semantic knowledge base) would help better organizing and delivering quality health information related to obesity and cancer that consumers need. Nevertheless, current ontologies describing obesity, cancer and related entities are not designed to guide automatic knowledge base construction from heterogeneous information sources. Thus, in this paper, we present methods for named-entity recognition (NER) to extract biomedical entities from scholarly articles and for detecting if two biomedical entities are related, with the long term goal of building a obesity-cancer knowledge base. We leverage both linguistic and statistical approaches in the NER task, which supersedes the state-of-the-art results. Further, based on statistical features extracted from the sentences, our method for relation detection obtains an accuracy of 99.3% and a f-measure of 0.993.

I. Introduction

Obesity is associated with increased risks of various types of cancer [1], [2] as well as a wide range of other chronic diseases. It accounts for 3–10% of cancer cases and deaths [3], [4]. Interventions that effectively modify excess weight can be used for cancer control, preventing further cancer burden [5]. On the other hand, access to health information activates patient participation, ultimately improves health outcomes [6]. However, existing online information on obesity, especially its relationship to cancer, is heterogeneous ranging from pre-clinical models and case studies to mere hypothesis-based scientific arguments. While the direct causal relationship between obesity and cancer has been difficult to

[‡]Corresponding: bianjiang@ufl.edu.

⁸<http://www.ncbi.nlm.nih.gov/mesh>

¹⁴<http://bioinformatics.ua.pt/becas/>

¹⁶<https://www.ncbi.nlm.nih.gov/bionlp/taggerone/>

definitively prove, stakeholder groups and investigators have generated a tremendous amount of supporting data. But collectively, these data are poorly organized in the public domain. There is an urgent need to organize the vast and increasing amount of information on possible links between obesity and cancer in a way that helps consumers understand and use the information in a meaningful way. This involves collating the information, linking it to evidence in the literature, evaluating its quality, and presenting high quality information relevant to their specific questions in a user-friendly way.

Therefore, we seek to automatically extract and transform obesity and cancer knowledge from heterogeneous, online, free-text sources (e.g., relevant scientific literature in PubMed¹, and online health information from WebMD² site) into a semantic knowledge base (KB) with a formal knowledge representation (i.e., ontology). Thus, the first step is to develop an information extraction (IE) system that can produce an ontology by annotating text with semantic information (e.g., entities and relations).

Many existing ontologies and knowledge bases are curated manually and often by domain experts, which is laborious and time consuming. This has changed in recent years, and a few semi-automatic methods have been proposed. These semi-automatic methods of ontology construction are mostly induced through mining textual data, including *Text2Onto* [7], *OntoLearn* [8], and *Sprat* [9]. Most of these approaches are proposed in general domains, with different techniques to automate the extraction tasks. These studies have shown promising results. Nevertheless, experts are always needed due to the fact that ontology construction and enrichment require a considerable amount of domain knowledge.

There are three fundamental steps in an IE system that extracts ontologically structured information from a free-text corpus [10]: (i) named-entity recognition (NER): find all pairs of named-entities, usually in the same sentence; (ii) relation detection: determine if the two entities are related, and (iii) if yes, classify the specific relation. In this paper, we focus on the **NER** and **relation detection** tasks.

- NER is concerned with identifying the names of entities such as people and locations in a general domain, or diseases and proteins in the biomedical domain. Typically, it is broken down into two main sub-tasks: 1) *find* and then *classify* entities in a piece of text. Despite recent advances in this field, the problem of NER is far from being solved. One of the biggest challenges in constructing a KB is to automate the construction process with more fine-grained entities (e.g., be able to identify “type 2 diabetes” rather than simply “diabetes” from a sentence).
- Relation detection is then a binary classification task (i.e., classify it as *true*, when a two entities are related; or *false* otherwise). This is important for KB construction as related biomedical entities need to be linked in the KB. Nevertheless, relation detection remains a major area in linguistic and semantics research.

¹PubMed is a database of biomedical literature <http://www.ncbi.nlm.nih.gov/pubmed>

²<http://www.webmd.com/>

In this paper, we propose: 1) a novel NER method, focusing on *finding* more fine-grained terms related to obesity and cancer; and 2) a relation detection approach to determine if two biomedical entities are related with novel features, extracted from the text. Through extensive evaluations, we show that our NER approach supersedes state-of-the-art results compared with many existing linguistic and supervised approaches, including TaggerOne, DNorm, BioPortal, MetaMap, and BeCAS. Furthermore, our relation detection method also obtains promising results, with an accuracy of 99.3% and a f-measure of 0.993.

The rest of the paper is organized as follows. Section II describes existing work related to NER and relation detection. We detail our approaches in section III. Our evaluation results followed by a short discussion of the work are presented in Section IV. We conclude the work in Section V.

II. Related Work

NER is an important task in natural language processing (NLP). Most existing NER approaches are supervised learning algorithms [11], such as conditional random field (CRF) and deep neural networks [12], based on both local context features [13] and properties of the token itself [14]. However, these supervised approaches require large amount of annotated text as training data. There are fewer studies using semi-supervised and unsupervised approaches in NER, for instance pattern bootstrapping [15], [16], [17] as well as generative approaches [18]. A few NER studies have used a hybrid approach, considering both linguistic and statistical features, which often leads to a more comprehensive set of features.

In the biomedical domain, NER have been used to identify genes and proteins [19], [20], [21], disease names [21], [22], [23], [24] and drug names [25], [26], [27], [28]. In [22], the authors use a CRF model to identify disease names. In [28], a CRF model was also used for identifying adverse drug reactions in social media postings. TaggerOne [24] is another work for joint NER and normalization with a semi-Markov structured linear classifier. In [23], three different bidirectional recurrent neural network (RNN) models were built with word embedding to recognize disease names and classes. In [29], several measures have been proposed to extract technical terms that do not exist in a terminology. In our work, we are interested in extracting meaningful terms regardless whether they are defined in existing terminologies or not.

Further, our work is closely related to the tasks of relation extraction (RE). The goal of RE is to determine semantic relations between entities mentioned in a text. Existing approaches can be grouped into two main categories[30]: (i) unsupervised relation discovery, and (ii) supervised classification. In the unsupervised paradigm, contextual features are often used. The distributional hypothesis theory [31] indicates that words that occur in the same context tend to have similar meanings or be related [32], [33], [34].

In the supervised paradigm, relation classification is considered a multi-classification problem, and existing work concentrates on feature engineering—extracting more complex features. These methods can be categorized into either feature-based [35], [36] or kernel-

based [37], [10]. Further, many recent work has focused on using deep neural networks (DNNs) to automatically learn features [38], [39], [40], [41] from the text.

In the biomedical domain, several approaches have been proposed to extract relations. Some work focused on matching linguistic patterns including those first described in [42] and those used in [43], [44], [42]. Other methods use hierarchical clustering [45], [46], [47] to learn rules for extracting synonyms. In [48], the authors presented MeTAE (Medical Text Annotation and Exploration), which extracts and annotates medical entities and relationships from medical texts and help end-users explore the produced RDF (Resource Description Framework) annotations. SEAM [49] is another system for extracting concepts and relationships from clinical and biomedical documents. Synonym and hierarchical groups are identified using corpus-based semantics and lexico-syntactic patterns in SEAM.

Note that our initial work is only to detect if a two entities are related (binary decision), but not to extract the specific type of relation.

III. Biomedical Named-Entity Recognition and Relation Detection

In this work, we tackle two specific problems with the ultimate goal of automating KB curation: 1) biomedical NER; and then 2) relation detection between two biomedical entities. We detail our approaches below.

A. Biomedical Named-Entity Recognition

We evaluated several existing applications, and formed a set of baseline measures. Our work is then built upon these baseline measures to recognize both biomedical entities that already exist in standard terminology (i.e., Unified Medical Language System) and potential new biomedical entities. We first describes the baseline measures, and then our contribution.

1) Baseline Measure: L-value—Our work is built upon the *L-value* measure [29], which was proposed for the task of Automatic Term Extraction (ATE). The *L-value* measure was designed to enrich existing biomedical ontologies, thus, focuses on extracting new biomedical terms that do not exist in the ontology.

The *L-value* measure (**L**inguistic patterns and **C-value** information, see Equation 1) is a ranking measure based on both linguistic and statistical information. It highlights the more representative terms in a single document. A single document can be considered as a piece of text without delimitation, e.g., a scientific article, a book, a sentence, or a paragraph. The *L-value* is extremely effective when there does not exist a lot of data that contain the term, such as for a emergent concept. For example, the term “Ataxia Neuropathy Spectrum” appeared only in 4 titles/abstracts of scientific articles in PubMed published between 2009 and 2015.

The *linguistic* information in the *L-value* measure serves to detect low frequency terms. A pattern weight (see Table I) is associated with each candidate term, which represents the *probability* of a candidate term being a relevant biomedical term. The *probability* is only assigned if the syntactic structure of the term appears in the linguistic pattern list. This

probability is designed to tackle the silence problem, allowing the extraction of discriminant terms. For example, in a biomedical corpus, the *probability* would favor a low frequency term “*virus production*” than a high frequency “*human monocytic cell*”, since the linguistic pattern weight of *NN* is higher than *NNN*. The *C-value* [50] in *L-value* measure is based on term frequency. The *C-value* favors a longer candidate term. For instance, in a specialized corpus (Ophthalmology), the authors [50] were able to identify “*soft contact lens*” using *C-value* rather than the more frequent and shorter “*soft contact*”.

$$L-value(A) = P(A_{LP}) \times C-value(A) \quad (1)$$

The authors [29] constructed a list of the most common linguistic patterns according to the syntactic structure of terms present in the UMLS¹. They carried out automatic part-of-speech (POS) tagging² of the biomedical terms using Tree-Tagger³, and then computed the frequency of the syntactic structures. The top 200 most frequent patterns were selected, and a weight (the probability of the pattern) was computed for each pattern, i.e. the frequency of the pattern over the sum of all pattern frequencies. Overall, they examined 3,000,000 terms to build the list of patterns. Table I illustrates the computation of the linguistic patterns and their weights.

2) Pattern Grouping—The patterns constructed in [29] take into account the difference between the specific classes of a grammatical category. For example, for a noun, the baseline measure notes the difference between the common and plural nouns. For our purpose, such difference may not be important. Thus, we recreated the list of patterns using the English UMLS terms, where we grouped related classes in a grammatical category into one category.

Furthermore, we evaluated two POS tools: the TreeTagger and Stanford Tagger⁶). Based on our evaluation, we chose TreeTagger, which shows better POS results, particularly for nouns and verbs [51].

Specifically:

- **N** represents all nouns: *NN* (common nouns), *NNS* (proper and singular nouns), *NNP* (proper and singular nouns), and *NNPS* (common and plural nouns).
- **J** represents all adjectives: *JJ* (ordinal numerals or adjectives), *JJS* (comparative adjectives), and *JJR* (superlative adjectives)

Through this process, we merged 1,665 patterns initially identified in UMLS into 200 patterns. Table II shows a sample of the merged patterns and how the final patterns (left side) are built according to the initial list of patterns (right side). We thus effectively expanded the original pattern list (i.e., from 200 to 1,665). We denote this new measure as ***L-value-FP***.

³<http://www.nlm.nih.gov/research/umls>

⁶<http://nlp.stanford.edu/software/tagger.shtml>

To illustrate the effect of grouping, Table III shows the NER results of two sentences with the same linguistic structure, i.e., the same POS structure. The baseline *L-value* measure would have ignored the evaluation of these two sentences due to the ranking of their linguistic pattern (i.e., 624 in the original list, where the original list only considers the top 200 patterns). After pattern grouping, the same pattern is ranked 98 in the list.

With the improved *L-value* measure, we can now identify more potential biomedical entities in each sentence.

3) RAKE (RAnKing of Entities)—However, not all the extracted entities are relevant to our study domain (i.e., obesity and cancer). Thus, we propose, *RAKE*, a ranking measure to select the appropriate biomedical entities. The *RAKE* measure is based on three statistics: the number of words contained in each entity, whether the entity belongs to an existing terminology and if the entity is part of a longer entity. We consider four biomedical terminologies relevant to obesity and cancer: UMLS⁴, MeSH⁵, NCI Thesaurus⁹, and SNOMED¹⁰. The formula of the *RAKE* measure is as follows:

$$RAKE(A) = (\log_2(nW + 1) + notN) \frac{U + M + N + S}{4} \quad (2)$$

where A is the entity, nW the number of words composing the entity, $notN$ is 1 if the entity is not nested and 0 otherwise, and U, M, N, S are UMLS, MeSH, NCI, and SNOMED respectively (i.e., the value is 1 if at least one word of the term belongs to that terminology). Through experiments, we determined the threshold of the *RAKE* is 2.5 ($RAKE(A) \geq 2.5$) to select entities relevant to our study domain. Note that MeSH, NCI Thesaurus, and SNOMED are part of UMLS. Thus, effectively we purposely ranked terms that contain MeSH, NCI Thesaurus, and SNOMED terms higher than terms in biomedical domain (i.e., UMLS), and even higher than general terms (i.e., those that are not in UMLS). This is because, through evaluation of these terminologies, we believe MeSH, NCI Thesaurus, and SNOMED terms are more relevant to obesity and cancer (i.e., contain more obesity and cancer related terms).

B. Relation Detection

After selecting the appropriate entities, we use a set of statistical, lexical, and semantic features to predict if two biomedical entities are related. Besides basic statistical measures, we derive a set of semantic features based on four biomedical terminologies relevant to obesity and cancer: UMLS, MeSH, NCI, and SNOMED. Table IV shows the 51 features, where $Dict = \{UMLS, MeSH, NCI, SNOMED\}$. We then tested a wide range of supervised learning algorithms following machine learning best practice to determine if two entities are related based on this set of 51 features.

⁴Part-of-Speech (POS) tagging is the process of assigning each word in a text to its grammatical category (e.g. noun, adjective).

⁵<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

⁹<https://ncit.nci.nih.gov/>

¹⁰<http://www.ihtsdo.org/snomed-ct>

IV. Results and Discussion

A. Dataset

Our first evaluation dataset is composed of 214 sentences extracted from 23 PubMed abstracts based on the search keywords “obesity” and “cancer”. Two domain experts manually annotated this corpus to extract the appropriate biomedical entities and relations. This dataset is publicly available online ¹¹. Our second dataset is the NCBI Disease corpus¹²[52], which is fully annotated at the mention and concept level. Table V summarizes the details of the two gold-standard datasets.

B. Experiments for Biomedical NER

We perform two sets of experiments based on the two gold-standard datasets.

1) Experiments on the Obesity and Cancer PubMed Abstracts—We compared the proposed *L-value-FP* measure for biomedical NER with a baseline (i.e., simply matching the words with UMLS terms) algorithm and four state-of-the-art biomedical NER applications: **BioPortal**¹³ [53], **BeCAS**⁶ [54], **MetaMap**¹⁵ [55], **L-value** [29].

Table VI shows two examples of the entities extracted by the six evaluated methods. Note that, we aim to recognize all possible entities including potential new biomedical entities, i.e., terms that are not in existing biomedical terminologies and ontologies. As shown in Table VI, Simple Match, BioPortal, BeCAS, and MetaMap can only recognize biomedical entities in existing terminologies and ontologies.

A second test is to select the most appropriate biomedical entities from all the entities being extracted from each sentence. Table VII shows a sample of preferred biomedical entities based on manual annotations. Table VIII compares the performance of the six evaluated methods, where *P* is the average precision, and *N* is the number of entities that the applications have not extracted. As shown in Table VIII, *L-value-FP* outperforms all other methods.

2) Experiments on the NCBI disease corpus—The second set of experiments are done with the NCBI disease corpus, comparing the proposed method with two other state-of-the-art NER tools: **TaggerOne**⁷ [24] and **DNorm** [56]. **DNorm** until 2015, had reached the highest published performance on the NCBI disease corpus and also achieved the highest performance in a number of other disease recognition challenges. **TaggerOne** later outperformed **DNorm**.

Table IX presents the results of our experiments, where *P* is precision, *R* is recall, and *F* is the f-measure. We see that the proposed **L-value-FP** achieved better performance than the state-of-the-art applications, with a *f-measure* of 0.901.

¹¹<https://github.com/bianjiang/obesity-cancer-kb>

¹²<http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/>

¹³<http://bioportal.bioontology.org/annotator>

¹⁵<https://metamap.nlm.nih.gov/>

⁷<http://www.nlm.nih.gov/research/umls/>

3) Discussion—The experiment results demonstrate that our approach is effective. Our approach achieved much better performance than the state-of-the-art (i.e., TaggerOne and DNorm) applications. Our approach is linguistic- and statistic- based, while DNorm and TaggerOne are based on supervised learning algorithms. We also manually analyzed the entities that were not extracted by our application. We observed that most of the not extracted entities are composed of signs and symbols, e.g., “*breast/ovarian cancer*”, and “*infantile form of g (m2) gangliosidosis, type 1*”. Our current method did not take into account symbols in the extraction process, which will be an interesting future work.

Further, the NER results are influenced by the performance of the POS tagging tool. For example, the sentence “*Red blood cells increase with ...*” was tagged by the Stanford Tagger as “*adjective noun noun verb preposition ...*”, whereas the TreeTagger tagged it as “*adjective noun noun noun preposition ...*”.

C. Experiments for Relation Detection

In this section, we report our experiments on relation detection. Results are presented in terms of *accuracy (A)*, *precision (P)*, *recall (R)*, and *f-measure (F)*. We used the dataset we manually curated based on 214 sentences from 23 obesity and cancer PubMed abstracts (i.e., the first dataset in our previous experiment for biomedical NER). The datasets contain 154 positive samples (the two identified entities are related) and 150 negative samples (the two identified entities are not related).

We experimented with a wide range of well-known supervised learning algorithms using Weka¹⁷, including zero rule (**ZeroR**), support vector machine (**SVM**), one rule (**OneR**), meta bagging (**MB**), naives bayes (**NB**), multilayer perceptron (**NN**), AdaBoost (**AB**), random forest (**RF**), decision tree (**TD**), and multi-class logistic classifier (**MCC**). We followed best practice in machine learning studies, including using 10-fold cross-valuation, and parameter tuning.

Table X shows the results of our experiments for relation detection, where the random forest (RF) model achieved the best results, with an *accuracy (A)* of 0.993. Most of the other learning algorithms except ZeroR and OneR achieved good performance.

1) Discussion—We also explored in detail the informativeness of the features we used. We evaluated the decision tree created from the TD model (as shown in Table X, the TD model has achieved reasonably good performance), shown in Figure 1. Only 14 of the 51 features are informative for the classification task; a large portion of these informative features are semantic features related to the four standard terminologies, which indicates that the biomedical entities in these terminologies are more likely to be related. Moreover, as shown in Figure 1, “is-*S_{UMLS}*” is the most informative feature separating related and not related biomedical entities. Table XI shows the confusion matrix of the TD, corresponding to an *accuracy (A)* of 0.96 (see Table X, column *A*, row *TD*).

¹⁷<http://www.cs.waikato.ac.nz/ml/weka/>

V. Conclusions

Biomedical named-entity recognition and relation detection are important building blocks towards automating ontologically structured knowledge base constructions from large free-text datasets.

In this paper, built upon the *L-value* measure, we first proposed a improved biomedical NER model with an expanded list of patterns incorporating both linguistic and statistic characteristics of existing biomedical terminologies. We also proposed a new measure to rank and select the most appropriate biomedical entities in a sentence for the downstream relation detection task. Through extensive experiments we show that the proposed method outperforms existing well-known systems for biomedical NER tasks.

Further, we proposed a supervised learning method to predict if two biomedical entities in a sentence are related. We explored 51 statistic, lexical and semantic features, which allowed good characterization of our dataset. We experimented with ten most well-known supervised learning algorithms implemented in Weka. Our results show that the best model obtained 99.3% of accuracy, with a 0.993 f-measure.

In future work, we will explore methods to predict the specific kind of relation between two biomedical entities (i.e., relation extraction) using supervised learning algorithms to complete the process of automatic KB construction.

Acknowledgments

This work was supported in part by NIH grants UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Calle EE, Thun MJ. Obesity and cancer. *Oncogene*. 2004; 23(38):6365–6378. [PubMed: 15322511]
2. Calle EE, Kaaks R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nature Reviews Cancer*. 2004; 4(8):579–591. [PubMed: 15286738]
3. Arnold M, Pandeya N, Byrnes G, Renehan AG, Stevens GA, Ezzati M, Ferlay J, Miranda JJ, Romieu I, Dikshit R, et al. Global burden of cancer attributable to high body-mass index in 2012: a population-based study. *The Lancet Oncology*. 2015; 16(1):36–46. [PubMed: 25467404]
4. Katzke VA, Kaaks R, Kühn T. Lifestyle and cancer risk. *The Cancer Journal*. 2015; 21(2):104–110. [PubMed: 25815850]
5. Nimptsch K, Pischon T. Body fatness, related biomarkers and cancer risk: an epidemiological perspective. *Hormone molecular biology and clinical investigation*. 2015; 22(2):39–51. [PubMed: 25781710]
6. Matheson GO, Klügl M, Engebretsen L, Bendiksen F, Blair SN, Börjesson M, Budgett R, Derman W, Erdener U, Ioannidis JP, et al. Prevention and management of non-communicable disease: the ioc consensus statement, lausanne 2013. *British journal of sports medicine*. 2013; 47(16):1003–1011. [PubMed: 24115479]
7. Cimiano, P., Völker, J. Text2onto: A framework for ontology learning and data-driven change discovery. *Proceedings of the 10th International Conference on Natural Language Processing and Information Systems*; Berlin, Heidelberg: Springer-Verlag; 2005. p. 227-238.ser. NLDB'05
8. Velardi P, Cucchiarelli A, Petit M. A taxonomy learning method and its application to characterize a scientific web community. *Knowledge and Data Engineering, IEEE Transactions on*. 2007; 19(2): 180–191.

9. Maynard, D., Funk, A., Peters, W. SPRAT: a tool for automatic semantic pattern-based ontology population. International Conference for Digital Libraries and the Semantic Web; Trento, Italy: Citeseer; 2009.
10. Mintz, M., Bills, S., Snow, R., Jurafsky, D. Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2; Stroudsburg, PA, USA: Association for Computational Linguistics; 2009. p. 1003-1011.ser. ACL'09
11. Luo, G., Huang, X., Lin, C-Y., Nie, Z. Joint entity recognition and disambiguation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; Lisbon, Portugal: Association for Computational Linguistics; Sep. 2015 p. 879-888.ser. EMNLP'15[Online]. Available: <http://aclweb.org/anthology/D15-1104>
12. Tomori, S., Ninomiya, T., Mori, S. Domain specific named entity recognition referring to the real world by deep neural networks. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Berlin, Germany: Association for Computational Linguistics; Aug. 2016 p. 236-242.ser. ACL'16[Online]. Available: <http://anthology.aclweb.org/P16-2039>
13. Finkel, JR., Grenager, T., Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; Stroudsburg, PA, USA: Association for Computational Linguistics; 2005. p. 363-370.ser. ACL'05
14. Ratnikov, L., Roth, D. Design challenges and misconceptions in named entity recognition. Proceedings of the Thirteenth Conference on Computational Natural Language Learning; Stroudsburg, PA, USA: Association for Computational Linguistics; 2009. p. 147-155.ser. CoNLL'09
15. Nadeau, D., Turney, PD., Matwin, S. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. Proceedings of the 19th International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence; Berlin, Heidelberg: Springer-Verlag; 2006. p. 266-277.ser. AI'06
16. Thelen, M., Riloff, E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. Proceedings of the ACL-2 Conference on Empirical Methods in Natural Language Processing - Volume 10; Stroudsburg, PA, USA: Association for Computational Linguistics; 2002. p. 214-221.ser. EMNLP'02
17. Brooke, J., Hammond, A., Baldwin, T. Bootstrapped text-level named entity recognition for literature. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Berlin, Germany: Association for Computational Linguistics; Aug. 2016 p. 344-350.ser. ACL'16[Online]. Available: <http://anthology.aclweb.org/P16-2056>
18. Elsner, M., Charniak, E., Johnson, M. Structured generative models for unsupervised named-entity clustering. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics; Stroudsburg, PA, USA: Association for Computational Linguistics; 2009. p. 164-172.ser. NAACL'09
19. Settles B. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005; 21(14):3191-3192. [PubMed: 15860559]
20. Leaman R, Gonzalez G. Banner: an executable survey of advances in biomedical named entity recognition. Pacific symposium on biocomputing. 2008; 13:652-663.
21. Leaman, R., Miller, C., Gonzalez, G. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. Proceedings of the 2009 Symposium on Languages in Biology and Medicine; 2009.
22. Mahbub Chowdhury, MF., Lavelli, A. Disease mention recognition with specific features. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing; Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 83-90.ser. BioNLP'10
23. Sahu, S., Anand, A. Recurrent neural network models for disease name recognition using domain invariant features. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Berlin, Germany: Association for Computational

Linguistics; Aug. 2016 p. 2216-2225.ser. ACL'16[Online]. Available: <http://www.aclweb.org/anthology/P16-1209>

24. Leaman R, Lu Z. Taggerone: Joint named entity recognition and normalization with semi-markov models. *Bioinformatics*. 2016:btw343.
25. Sánchez Cisneros, D., Aparicio Gali, F. Uem-uc3m: an ontology-based named entity recognition system for biomedical texts. *Proceedings of the 7th International Workshop on Semantic Evaluation; Association for Computational Linguistics; 2013. ser. SemEval'13*
26. He L, Yang Z, Lin H, Li Y. Drug name recognition in biomedical texts: a machine-learning-based method. *Drug discovery today*. 2014; 19(5):610–617. [PubMed: 24140287]
27. Korkontzelos I, Piliouras D, Dowsey AW, Ananiadou S. Boosting drug named entity recognition using an aggregate classifier. *Artificial intelligence in medicine*. 2015; 65(2):145–153. [PubMed: 26116947]
28. Nikfarjam A, Sarker A, OConnor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*. 2015:671–681. [PubMed: 25755127]
29. Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*. 2016; 19(1–2):59–99.
30. Bach N, Badaskar S. A review of relation extraction. 2007
31. Harris ZS. Distributional structure. *Word*. 1954
32. Hasegawa, T., Sekine, S., Grishman, R. Discovering relations among named entities from large corpora. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics; Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. ser. ACL'04*
33. Bollegala, DT., Matsuo, Y., Ishizuka, M. Relational duality: Unsupervised extraction of semantic relations between entities on the web. *Proceedings of the 19th International Conference on World Wide Web; New York, NY, USA: ACM; 2010. p. 151-160.ser. WWW'10*
34. Yao, L., Haghighi, A., Riedel, S., McCallum, A. Structured relation discovery using generative models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing; Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 1456-1466.ser. EMNLP'11*
35. Kambhatla, N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions; Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. ser. ACLdemo'04*
36. Suchanek, FM., Ifrim, G., Weikum, G. Combining linguistic and statistical analysis to extract relations from web documents. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; New York, NY, USA: ACM; 2006. p. 712-717.ser. KDD'06*
37. Qian, L., Zhou, G., Kong, F., Zhu, Q., Qian, P. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1; Association for Computational Linguistics; 2008. p. 697-704.ser. COLING'08*
38. Turian, J., Ratniov, L-A., Bengio, Y. Word representations: A simple and general method for semi-supervised learning; *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics; Stroudsburg, PA, USA: Association for Computational Linguistics; Jul. 2010 p. 384-394.*
39. Socher, R., Huval, B., Manning, CD., Ng, AY. Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 1201-1211.ser. EMNLP-CoNLL '12*
40. Hashimoto, K., Miwa, M., Tsuruoka, Y., Chikayama, T. Simple customization of recursive neural networks for semantic relation classification. *Proceedings of the 2013 International Conference on Empirical Methods in Natural Language Processing; Stroudsburg, PA, USA: Association for Computational Linguistics; 2013. ser. EMNLP'13*

41. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J. Relation classification via convolutional deep neural network. *Proceedings of the 25th International Conference on Computational Linguistics*; 2014. p. 2237-2248.ser. COLING' 14
42. Liu K, Chapman W, Savova G, Chute C, Sioutos N, Crowley RS, et al. Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. *Methods of information in medicine*. 2011; 50(5):397. [PubMed: 21057720]
43. Charlet J, Bachimont B, Jaulent MC. Building medical ontologies by terminology extraction from texts: an experiment for the intensive care units. *Computers in biology and medicine*. 2006; 36(7): 857–870. [PubMed: 16198328]
44. Baneyx A, Charlet J, Jaulent MC. Building an ontology of pulmonary diseases with natural language processing tools using textual corpora. *International Journal of Medical Informatics*. 2007; 76(2):208–215. [PubMed: 16797227]
45. Kuo, Y-T., Lonie, A., Sonenberg, L., Paizis, K. Domain ontology driven data mining: A medical case study. *Proceedings of the 2007 International Workshop on Domain Driven Data Mining*; New York, NY, USA: ACM; 2007. p. 11-17.ser. DDDM'07
46. Henriksson A, Moen H, Skeppstedt M, Daudaravicius Y, Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*. 2014; 5(6)
47. Wang C, Cao L, Zhou B. Medical synonym extraction with concept space models. 2015 arXiv preprint.
48. Abacha AB, Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of Biomedical Semantics*. 2011; 2
49. Doing-Harris K, Livnat Y, Meystre S. Automated concept and relationship extraction for the semi-automated ontology management (seam) system. *Journal of biomedical semantics*. 2015; 6(1):15. [PubMed: 25874077]
50. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*. 2000; 3(2):115–130.
51. Tian, Y., Lo, D. A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports. *Proceedings of the 22nd International IEEE Conference on Software Analysis, Evolution, and Reengineering*; Montreal, Canada: IEEE; Mar. 2015 p. 570-574.ser. SANER'15
52. Do an RI, Leaman R, Lu Z. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*. 2014; 47:1–10. [PubMed: 24393765]
53. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith NB, Jonquet C, Rubin DL, Storey M-A, Chute CG, Musen MA. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*. 2009; 37:170–173.
54. Nunes T, Campos D, Matos S, Oliveira JL. Becas: biomedical concept recognition services and visualization. *Bioinformatics*. 2013; 29(15):1915–1916. [PubMed: 23736528]
55. Aronson AR, Lang FM. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*. 2010; 17(3):229–236. [PubMed: 20442139]
56. Leaman R, Do an RI, Lu Z. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*. 2013:btt474.

```

is-SUMLS <= 0
| nW-S <= 1: Not Related (4.0/1.0)
| nW-S > 1: Related (76.0)
is-SUMLS > 0
| nBE-SMeSH <= 1
| | nJJ <= 5
| | | nBE-SenSNOMED <= 0
| | | | nBE-SenNCI <= 20: Not Related (98.0)
| | | | nBE-SenNCI > 20
| | | | | nBE-SenUMLS <= 28: Related (2.0)
| | | | | nBE-SenUMLS > 28: Not Related (7.0)
| | | nBE-SenSNOMED > 0
| | | | nNNS <= 0: Related (13.0)
| | | | nNNS > 0
| | | | | nNoun <= 11
| | | | | | is-PUMLS <= 0: Not Related (22.0)
| | | | | | is-PUMLS > 0
| | | | | | | is-PNCI <= 0: Related (6.0)
| | | | | | | is-PNCI > 0
| | | | | | | | nBE-OSNOMED <= 0: Not Related (12.0)
| | | | | | | | nBE-OSNOMED > 0
| | | | | | | | nBE-OMeSH <= 0: Not Related (6.0)
| | | | | | | | nBE-OMeSH > 0
| | | | | | | | | nBE-SMeSH <= 0
| | | | | | | | | | is-OMeSH <= 0: Related (3.0)
| | | | | | | | | | is-OMeSH > 0: Not Related (2.0)
| | | | | | | | | nBE-SMeSH > 0: Related (6.0)
| | | | | nNoun > 11: Related (18.0)
| | nJJ > 5: Related (15.0)
| nBE-SMeSH > 1: Related (14.0)

```

Fig. 1.
The generated decision tree of the TD model.

TABLE I

Example of pattern constructions (where *NN* is a noun, *IN* is a preposition or subordinating conjunction, *JJ* is an adjective, and *CD* is a cardinal number)

Pattern	Frequency	Probability
NN IN JJ NN IN JJ NN	3006	$3006/4113 = 0.73$
NN CD NN NN NN	1107	$1107/4113 = 0.27$
	4113	1.00

Pattern grouping (where *N* is a noun, *IN* is a preposition or subordinating conjunction, *J* is an adjective, and *CD* is a cardinal number)

TABLE II

Final List of Patterns			Initial List of Patterns		
Pattern	Freq	Position	Pattern	Freq	Position
N J CD	58	N° 84	NN JJ CD	31	N° 246
			NNS JJ CD	23	N° 300
			NP JJ CD	4	N° 916
N IN J N	43	N° 98	NN IN JJ NN	17	N° 369
			NNS IN JJ NN	9	N° 533
			NN IN JJ NNS	7	N° 624
			NNS IN JJ NNS	4	N° 962
			NNS IN JJ NP	2	N° 1423
			NN IN JJ NP	2	N° 1584
			NN IN JJ NPS	1	N° 2559
			NP IN JJ NN	1	N° 2927
			JJ NN	22006	N° 5
			JJ NNS	8147	N° 10
J N	37055	N° 3	JJ NP	6562	N° 11
			JJ NPS	321	N° 60
			JJR NNS	5	N° 858
			JJR NN	5	N° 873
			JJR NP	3	N° 1221
			JJR NPS	2	N° 1540
			JJS NN	2	N° 1783
			JJS NNS	1	N° 2440
			JJS NPS	1	N° 3711

TABLE III

The effect of pattern grouping on two sample sentences.

risk of Western cancers				
regression of coronary stenoses				
Ranked 624 in Initial LP	<i>NN</i>	<i>IN</i>	<i>JJ</i>	<i>NNS</i>
Ranked 98 after Pattern Grouping	<i>N</i>	<i>IN</i>	<i>J</i>	<i>N</i>

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE IV

Features considered for relation detection

Features
nW-Sen : number of words in the sentence. nW-S : number of words contained in the subject. nW-O : number of words contained in the object. nW-P : number of words contained in the predicate.
nBE-Sen_{Dict} : number of biomedical entities in the sentence that are in the four terminologies <i>Dict</i> . nBE-S_{Dict} : number of biomedical entities appearing in the subject that are in <i>Dict</i> , it could be zero or more. nBE-O_{Dict} : number of biomedical entities appearing in the object that are in <i>Dict</i> , it could be zero or more. nBE-P_{Dict} : number of biomedical entities contained in the predicate that are in <i>Dict</i> , it could be zero or more.
is-S_{Dict} : <i>true</i> if the subject is included in the four terminologies <i>Dict</i> , <i>false</i> otherwise. is-O_{Dict} : <i>true</i> if the object is included in the four terminologies <i>Dict</i> , <i>false</i> otherwise. is-P_{Dict} : <i>true</i> if the predicate is included in the four terminologies <i>Dict</i> , <i>false</i> otherwise.
dist-SO : is the distance in terms of the number of tokens between the subject and object. dist-SP : is the distance in terms of the number of tokens between the subject and predicate. dist-PO : is the distance in terms of the number of tokens between the predicate and object.
nNoun : is the total number of nouns in the sentence. nVerb : is the total number of verbs in the sentence. nAdj : is the total number of adjectives.
nNN : is the total number of nouns, common, singular or mass, in the sentence. nNNS : is the total number of nouns, proper, singular, in the sentence. nNNP : is the total number of nouns, proper, plural, in the sentence. nNNPS : is the total number of nouns, common, plural, in the sentence.
nVB : is the total number of verbs, base form, in the sentence. nVBD : is the total number of verbs, past tense, in the sentence. nVBG : is the total number of verbs, present participle or gerund, in the sentence. nVBN : is the total number of verbs, past participle in the sentence. nVBP : is the total number of verbs, present tense, not 3rd person singular, in the sentence. nVBZ : is the total number of verbs, present tense, 3rd person singular, in the sentence.
nJJ : is the total number of adjectives or numerals, ordinal, in the sentence. nJJR : is the total number of adjectives, comparative, in the sentence. nJJS : is the total number of adjectives, superlative, in the sentence.

TABLE V

Summary of the two gold-standard datasets.

Corpus & Entity	Annotations	Unique Mentions
Pubmed Abstracts	462	172
NCBI Disease	6892	2135

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE VI

Comparing the proposed *L-value-FP* method with state-of-the-art biomedical Named-Entity Recognition applications.

Sentence	Simple Match	BioPortal	BeCAS	MetaMap	L-value	L-value-FP
Associated liver steatosis is a significant risk factor for colorectal cancer.	liver liver steatosis is a significant risk factor colorectal	colorectal cancer liver colorectal associated liver steatosis is a risk cancer significant factor steatosis risk factor	colorectal cancer liver liver steatosis	associated liver steatosis colorectal cancer associated significant risk factor colo rectal cancer	cancer factor liver risk steatosis liver steatosis risk factor colorectal cancer significant risk factor	liver steatosis risk factor cancer factor liver risk steatosis colorectal cancer significant risk factor risk factor for colorectal cancer factor for colorectal risk factor for colorectal
Paradoxically, mTORC1 has also been identified as a mediator of the Warburg effect that allows cell survival under hypoxia.	mtorc1 identified mediator warburg effect effect cell cell survival survival	hypoxia identified cell survival warburg effect effect survival mediator cell mtorc1	hypoxia cell survival warburg effect cell mtorc1	allow hypoxia identified cell survival warburg effect mediator mtorc1	warburg cell effect hypoxia mediator mtorc1 survival cell survival warburg effect cell survival under hypoxia survival under hypoxia mediator of the warburg mediator of the warburg	cell survival warburg effect cell effect hypoxia mediator mtorc1 survival warburg cell survival under hypoxia survival under hypoxia mediator of the warburg mediator of the warburg effect

TABLE VII

Samples of biomedical entities manually extracted.

Associated liver steatosis is a significant risk factor for colorectal cancer.	risk factor for colorectal cancer liver steatosis
Paradoxically, mTORC1 has also been identified as a mediator of the Warburg effect that allows cell survival under hypoxia.	mtorc1 cell survival under hypoxia warburg effect

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE VIII

Benchmarks of six biomedical NER methods in terms of selecting the most appropriate terms.

	<i>P</i>	<i>N</i>
Simple	0.286	135
BioPortal	0.428	101
BeCAS	0.284	124
MetaMap	0.496	88
L-value	0.693	54
L-value-FP	0.823	32

TABLE IX

Results of NER on NCBI disease dataset

	<i>P</i>	<i>R</i>	<i>F</i>
DNorm	0.822	0.775	0.798
TaggerOne	0.851	0.808	0.829
L-value-FP+RAKE	0.905	0.897	0.901

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE X

Results of relation detection.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F</i>
ZeroR	0.507	0.257	0.507	0.341
OneR	0.849	0.851	0.849	0.848
AB	0.921	0.921	0.921	0.921
NB	0.928	0.928	0.928	0.928
SVM	0.947	0.947	0.947	0.947
MCC	0.947	0.948	0.947	0.947
NN	0.954	0.954	0.954	0.954
TD	0.961	0.961	0.961	0.961
MB	0.964	0.964	0.964	0.964
RF	0.993	0.993	0.993	0.993

TABLE XI

Confusion matrix of the trained TD model.

	<i>Rel</i>	<i>Not Rel</i>	← Classified as
<i>Rel</i>	153	1	154
<i>Not Rel</i>	1	149	150

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript