Information Extraction In Medical Domain

Lekha

June 12, 2015

Information extraction is an important task of machine learning and natural language processing. It involves extracting meaningful pieces of knowledge from natural language text. Information extraction may involve extracting names of persons, places, organisations; finding temporal information from text; identifying multi-word expressions and many other such applications.

Information extraction tasks present different complexities when applied to different domains and data sources. E.g. information extraction from tweets and social media postings is challenging due to non standard use of language. Specific domains such as legal, technology, medical etc are harder due to abundance of domain specific jargon.

We focus on the challenge of information extraction in the medical domain and specifically form clinical documents.

Information extraction in the medical domain involves handling of a number of vital tasks such as identification of medical terms, identification of attributes such as negation, uncertainty, severity, identification of relationships between entities and mapping terms in the document to concepts in domain specific ontologies. The entire process depends on a number of fundamental NLP processes such as tokenization, part-of-speech tagging, and parsing. There is also a heavy dependence on domain specific resources such as medical dictionaries and ontologies such as the UMLS.

In this survey we aim to highlight the some of the important contributions which exist in the field of information extraction in the medical domain. In particular we will focus on the tasks of clinical entity extraction, modifier detection and relationship detection. We will also discuss resources such as UMLS, ICD 9 / ICD 10, pubMed and Medline.

1 Introduction

The goal of IE in the medical domain is to convert an unstructured medical report into structured information such that the information can then be analysed, aggregated, and mined for insightful patterns. Further, we wish to automate the process of ‘coding’; the mapping of a medical document to a node (or multiple nodes) of hierarchical taxonomy or ontology of diseases.

A medical report contains a large number of medical terms and terminologies. This includes disease names, medicine names, medical procedures, medical devices, laboratory results, patient body measurements etc.

Further each of these medical terms or clinical entities have a number of modifiers attached to them. E.g. a disease may be ‘chronic’, ‘acute’, ‘mild’, ‘atypical’, ‘idiopathic’ etc. Similarly a medicine name may be accompanied by additional information such as frequency, route, quantity of the dose.

The identification of clinical entities along with the modifiers that are associated with them is our primary task. If that is accomplished; each document can be represented by a set of entity-modifier pairs; thus adding structure to the documents.

The next goal is to map each of these entities to a medical concept in an ontology. In particular we use the UMLS semantic network[1]. While this may be straightforward in some cases, multiple senses of the same word, multiple words for the same meaning and the role of context make this task slightly more involved.

Finally using the identified concepts, we must map documents to a taxonomy of diseases. We use the ICD taxonomy[2] for this purpose. Since a doctor-patient interaction focuses on a disease condition; identifying this is vital.

2 Nature of Medical Reports

The dataset used for clinical entity extraction experiments is a collection of clinical documents. Each document is a single clinical report dictated by a doctor (and transcribed later by a third-party) to capture the proceedings of a doctor-patient interaction or to document the results of a medical procedure or test. The reports are typically de-identified using a method such as the Safe Harbour method as per the HIPAA privacy rule [3].

Each document is a few paragraphs long. Sentences may be short phrases or long compound sentences. There are cases of non-grammatical usage of language. Most text is in narrative speech without the use of complicated or stylised constructs. For e.g. phenomena such as double negatives (“not unknown”) are relatively rare.

The documents contains both structured and unstructured data. Header and footer of the document contains patient information, doctor and hospital information, time and date information in structured format. The body of the document may also contain a structured component in the form of a listing of diagnoses, known allergies etc. However, a large component of the body of documents is unstructured. Doctors describe the patient, his condition, diagnosis or proceedings of a procedure in free-form English. Most documents are typically subdivided into sections.

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

[2] http://www.nlm.nih.gov/research/umls/

[3] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#rationale

The documents contain a large number of medical terms such as names of medicines, procedures and anatomical structures. There is also numeric data in the form of measurements and laboratory results. Abbreviations are abundantly used; often ambiguously since the same abbreviation may be medical or non medical; and may expand to different terms based on the context.

What is described above constitutes raw data. This is then usually manually annotated by a set of annotators to identify clinical entities, modifiers, relationships between entities etc.

The documents span a range of medical-sub domains such as cardiology, oncology, psychiatry etc. Documents do not have explicit domain information mentioned with them however, domain can be inferred from the source of the document.

Documents are of different work type. A work type describes the purpose of the document. For example, discharge summary, operation report, history, physical examination report are different types of document work types. Not all documents have work type information explicitly mentioned in them. Around 70% of documents have work type either annotated or easily inferable from the document contents.

Each document is divided in a number of sections. A section is preceded by a section header. Section headers are either all uppercase or camel-case which may or may not be followed by colon. Section names are non-standard. Thus, we may encounter thousands of section headers in reports but typically many are variants of each other and represent the same section type. For e.g. "present history", "presenting history", "history of present illness", "HPI" etc all refer to the same section category, Thus we can group section headers into categories. . Some examples of categories are "History", "Labs", "PE (Physical Examination)", "Complications" etc.

3 Motivation and Applications

The healthcare domain is a constantly evolving critical field of science. The impact of the healthcare industry on day to day patient care and on biomedical research is immense.

Like other industries, for exchange of information and preservation of historic data and to provide accountability and traceability; documentation is a vital component of healthcare. A doctor-patient interaction must be documented in the form of a report.

Such a report has many uses. Some of the needs for documenting medical information are :

- Memory aid for the doctor during subsequent patient visits.
- Knowledge transfer between doctors.
- Ensuring accountability and compliance of hospital procedures.
- Systematic tracking and follow-up of patients.
- Preserving knowledge of current treatments and medical procedures for future reference.

While there is an abundance of medical documentation; all these documents are written in free-form text without any standards or uniformity. Due to this, it is virtually impossible to perform any analysis or data mining on this data.

Hence, conversion of these unstructured documents to structured information is needed. This is the challenge we attempt to solve.

4 Challenges

Medical data such as doctors reports pose many challenges to any information extraction tool. The following constitute some of the major challenges faced:

- **Non Standard Document Structure** : Medical documents have no fixed structure. They may be divided into sections however there is no standardisation on the type of sections or their headings or contents. This depends on hospital to hospital, doctor to doctor.
- **Medical Jargon** : Medical documents contain a large number of medical terms and jargon. NLP tools trained on non-medical domain data perform very poorly on medical data.
- **Non Grammatical Language** : Doctors do not always write their reports using fully grammatical language. Often incomplete phrases or unnaturally long sentences are used. Further, style of writing depends on document source.
- **Abbreviations** : The medical domain experiences an abundant use of abbreviations. Often the same abbreviation can be non medical or medical or can expand to different terms based on the context and intention of the writer. Abbreviations are hard to normalize, classify or resolve.
- **Polysemy and Synonymy** : A single medical term can represent two different ideas based on context. This is known as polysemy. E.g. “inflammation” may refer to a skin problem, a cellular level problem, a non medical activity etc. Further, a single concept can be expressed through many different words. This is known as synonymy. E.g. “foetus” and “baby” mean the same in many medical contexts.
- **Transcription Errors** : Most reports are dictated by doctors and typed by third-party. This introduces a wide array of transcription errors. Inaudible words are left as blanks. Homophones such as “anterior” (front), “interior”(inside) create confusion; similarly spelt words are further mixed up such as “tenia” (band-like structure), “tinea” (fungal infection on the skin). Apart from this the process of transcription also introduces a wide array of grammatical and casual spelling errors.

5 Entity Extraction from Medical Documents

Clinical entity extraction is the most fundamental task in information extraction in the medical domain. It involves the extraction of medical terms and phrases from documents. Medical terms may include disease names, procedures, medical devices, medicine names etc. Clinical entities can be single or multi word units which occur either contiguously or in disjoint spans in the same sentence.

Entity extraction has been widely studied and explored in literature. A number of rule based and statistical approaches with rich feature sets have been used to produce state of the art results.

In this section we discuss popular statistical and rule based models proposed for the task. We also attempt study the similarity between named entity recognition in non medical text to the clinical entity recognition task in medical texts.

5.1 Statistical Methods

Statistical methods are a robust, generalisable choice for many NLP tasks. They depend highly on the need of labelled training data; while producing accurate results.

Hidden Markov Models, MaxEnt systems, Conditional Random Fields are common models used for clinical entity extraction.

5.1.1 Hidden Markov Models

Earlier work such as Collier et al. (2000) used a generative sequence labelling model viz. hidden Markov models for clinical entity detection from text. Transition probabilities between entity types and non entities are used to make predictions along with the output probability of a unigram given its type.

5.1.2 Maximum Entropy Markov Models

Finkel et al. (2005), Saha et al. (2009) and Finkel et al. (2004) use a discriminative framework through maximum entropy Markov model for the task. This method allows the use of a wider variety of features. Lexical features such as unigrams, suffixes, lemma are found to be influential. Further, linguistic features such as part of speech also play a role. Further, lexicon based approaches are used to create additional features.

5.1.3 Conditional Random Fields

Since the introduction of Conditional Random Fields (Lafferty et al., 2001), they have been a popular choice for sequence labelling tasks. CRF has been used for clinical entity extraction in Settles (2004) , McDonald and Pereira (2005), Bodnari et al. (2013) , Grouin (2014) , Tang et al. (2014) etc. Features used with MEMM are also found suitable with CRF. A number of orthographic features such as case information, presence of punctuation etc. is also found to provide additional cues. CRFs overcome the label-bias problem faced by MEMMs and have been theoretically and empirically proven to be more robust and accurate in sequence labelling tasks.

5.1.4 Support Vector Machines

SVM and MaxEnt classifiers have also been employed in entity detection tasks in Doan and Xu (2010) and Saha et al. (2009). Further, SVMs have been modified for sequence labelling tasks in the form of structured SVMs used in Cogley et al. (2013) and Yamamoto et al. (2003). Comparable results are produced with structured SVM as well.

5.1.5 Combining and Comparing Models

A number of works attempt to combine multiple statistical models or statistical and rule based models either in a pipeline or in a parallel architecture combined using majority voting. Dehghan (2013) post processes the CRF output to correct boundary identification errors. Wang and Patrick (2009) combines CRF and MaxEnt outputs using majority voting. They also attempt to use CRF for only boundary identification of entities which is post processed by a MaxEnt system for entity type classification.

Comparative studies such as those made by Abacha and Zweigenbaum (2011) reveal that statistical systems such as CRF perform better than pure rule based methods. Further, CRF outperforms a rule based boundary identifier followed by a SVM based entity type classifier also.

5.2 Rule Based Approaches

A number of rule based methods have also been proposed for clinical named entity recognition in medical texts. The methods fall broadly in two categories. The first exploits linguistic principles to identify named entities. The second popular approach uses semantic ontologies, lexicons and lookup based approaches to identify clinical entities.

5.2.1 Linguistic Approaches

Language based approaches usually rely on parsing. Syntactical parsing is performed and its output is post processed using a number of hand-crafted rules to identify named entities. In particular, named entities tend to be noun phrases occurring at the subject (or sometimes object) position of sentences. Proux et al. (1998) performs a number of rule based filtering steps to identify clinical entities. Wilbur et al. (1999) performs sentence segmentation using rule based methods. They also perform a 2 stage approach where a rule based system is followed by a classifier which identifies entity type. Similarly Rebholz-Schuhmann et al. (2006) employs a number of stages of filtering using both rule based and statistical principles. Jimeno et al. (2008) includes statistical model in the rule based system by making use of word frequency and co-occurence counts.

5.2.2 Ontology Based Approaches

Ontology and lexicon based approaches make use of UMLS[4], SNOMED[5] and other popular medical lexicons and semantic networks to perform lookup of tokens and to identity their type. For example Fan et al. (2013) uses concepts in SNOMED lexicon. Similarly MetaMap (Aronson, 2001) is a popular rule based tool relying on UMLS. MedEx (Xu et al., 2010) is a more recent state of the art clinical entity extraction tool which combines parsing, lexicon lookup and regular expressions to extract clinical entities from text.

Other such tools include dNorm [6], cTakes[7] and yTex[8]

5.3 Similarity With Named Entity Extraction From Non Medical Text

The Clinical Entity Extraction task has close similarity with the task of named entity extraction from normal text. Named entities (Bikel et al., 1999) include person, location, organization names and their identification is an important first step in the field of information extraction (Nadeau and Sekine, 2007, Arora)

Named entity recognition has been a widely researched and experimented domain. Rule based approaches, syntax parsing, use of web based and hand crafted lexicons , use of statistical tools such as CRF (McCallum and Li, 2003, Tkachenko and Simanovsky, 2012), HMM (Zhou and Su, 2002), and MEMM (Bender et al., 2003) have been explored for NER.

Ratinov and Roth (2009) discusses various design challenges for NER using representation of named entities, tag schemes, models and feature sets; all of which are relevant for clinical entity extraction also.

Jiang and Zhai (2006) discusses the importance of a domain ontology to improve results of NER when dealing with a specific target domain. This is especially relevant for us since the medical domain is supplemented with many ontologies such as UMLS.

Gazetteers or lookup engines which use the web as a resource or any other major semantic ontology are also vital for NER. Mikheev et al. (1999) presents various rule based techniques with and without the use of a gazetteer.

[4] http://www.nlm.nih.gov/research/umls/

[5] http://www.ihtsdo.org/snomed-ct/

[6] http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/DNorm/

[7] http://ctakes.apache.org/

[8] https://code.google.com/p/ytex/

6 Modifier Detection

Modifiers are tokens which provide additional vital information about entities. A modifier may negate, quantify or describe an entity. The semantic role of an entity in a sentence can only be discovered after combining it with its modifiers.

6.1 Rule Based Approaches

An important subproblem of modifier detection is the problem of negation detection, which has been widely studied in literature. A simple yet popular rule based negation detection system is the NegEx algorithm proposed by Chapman et al. (2001) which uses a lexicon of 35 negation phrases used for negation detection along with a context window of 5 words to detect the negated entity. An F score of 81.03% is achieved in this simple scheme. This is extended by Harkema et al. (2009) where along with trigger terms, a second list of termination terms is used and they perform detection of negation, uncertainty and subject modifiers.

Mutalik et al. (2001) supplement the idea of the NegEx algorithm by also using syntax parsing to discover the scope of the negation.

Elkin et al. (2005) proposed a negation assignment grammar, a set of reduction rules which can be applied to language to discover negation terms and their scope. They achieve an F measure of 94.1%. The paper also explores the concept that all text is divided into four categories; kernel concepts, modifiers, quantifiers, and negation terms.In the context of negation of clinical entities Patrick et al. (2006) mention an important classification of negations. Some negations are included as part of the clinical entity in the UMLS ontology, while others are instances of classic negation wherein the cue for negation is disjoint from the clinical entity phrase.

Tolentino et al. (2006) make use of a finite state machine along with a list of negation phrases to achieve an F score of 91.43%.

Rule based methods which rely heavily on lexical and linguistic clues dominate the negation detection efforts. Many ideas can be mapped to identify other categories of modifiers as well. For example, a syntax parsing approach such as that of Gindl et al. (2008) can be used to detect the scope of any modifier phrase. Modifiers also tend to come from a limited vocabulary of adjective or adjective-like phrases. The terms they modify determine whether the modifier is medical or non medical. E.g. ‘chronic’ can be a medical modifier in the phrase ‘chronic diabetes’ but a non-medical phrase in the context ‘chronic shopper’. Similarly ‘mild’ can be medical in ‘mild pain’ but non medical in ‘mild soap’.

6.2 Statistical Approaches

Uzuner et al. (2011) as well as de Bruijn et al. (2011) proposed a classifier based approach using SVM classifier and a range of lexical as well as syntactic features for assertion classification.

Clark et al. (2011) divide the modifier detection task into two steps viz. cue detection and cue scope detection. They model each of these steps as sequence labelling tasks using CRFs.

6.3 Use of Dependency Parsing for Modifier Detection

Sohn et al. (2012) demonstrate the usefulness of dependency parsing for negation detection through a rule-based approach where manually created patterns on the dependency parse tree are used to identify instances of negation. However, any approach involving dependency parse output in the form of generalisable features to a statistical system has not been used before for modifier detection to the best of our knowledge.

6.4 Dependency Parsing used with Statistical Approaches

Using parse structure along with classifiers have been largely implemented through the design of specialised classifier kernels that measure parse tree similarity. Convolution kernels (Collins and Duffy, 2002, Moschitti, 2004) are one such approach.

Sidorov et al. (2013) makes use of dependency parse based features to a classifier for the task of author attribution.

Joshi and Penstein-Rosé (2009) also use flat features extracted from dependency parse output for opinion mining.

7 Semantic Networks and Ontologies

7.1 UMLS

The UMLS(Unified Medical Language System)[9] is a large medical domain ontology created by the U.S. National Library of Medicine. The UMLS contains the following components:

- **Metathesaurus** : This is a comprehensive medical vocabulary i.e. an extensive list of medical terms and terminologies from various medical sub-domains. The UMLS metathesaurus consists of many medical vocabularies such as SNOMED CT[10], RxNorm[11], ICD-10-CM[12], etc

- **Semantic Network** : The terms in the metathesaurus vocabulary are combined in the semantic network. Relationships and term hierarchies are defined.

[9] http://www.nlm.nih.gov/research/umls/

[10] http://www.ihtsdo.org/snomed-ct/

[11] www.nlm.nih.gov/research/umls/rxnorm/

[12] http://www.cdc.gov/nchs/icd/icd10cm.htm

- **SPECIALIST lexicon and Lexical Tools** : Lexical variants, morphology analysers, and other such lexical and linguistic knowledge of terms in the UMLS network are made available through this this suite of tools.

  The UMLS is a fundamental and comprehensive ontology for the medical domain. Information extraction relies heavily on its use as a lexicon, terminology store, and as an ontology of relationships and linkages of terms.

7.2 Medline and PubMed

Medline (Medical Literature Analysis and Retrieval System Online)[13] is a large repository of citations, abstracts and publications from the biosciences and biomedical domain. PubMed is a freely available access point to the Medline repository. Full papers are also provided wherever available. Medline is a valuable resource for medical domain natural language data. Citations and abstracts serve as title keyword-pool whereas content of papers provide millions of documents of medical text. Medical vocabularies can be built based on this corpus. Medline is widely used in Biomedical Natural Language Processing research. Subsets of the Medline dataset have been annotated at sentence and token level to identify names of clinical entities, biomolecules, gene and protein names etc.

7.3 ICD-10

International Statistical Classification of Diseases [14] is a hierarchical classification of diseases created by the World Health Organization. While it has many versions ICD-10 is the latest version which is used by hospitals, insurance agencies and medical practitioners for unambiguous record and exchange of medical information. ICD-10 divides diseases based on their location (diseases of the heart, diseases of the digestive system) and their nature (autoimmune diseases, infectious and parasitic diseases, congenital diseases).

The classification scheme provides a hierarchy such that the lowermost level containing leaf nodes refers to a very specific disease whereas higher levels in the hierarchy refer to categories or groupings of diseases.

E.g. “chronic conjunctivitis” is a leaf node with ICD-10 code H10.4 which has the following nodes as ancestors:

H10.4 Chronic conjunctivitis
H10 Conjunctivitis
VII Disease of the eye and adnexa

Each term (internal node or leaf node in the hierarchy) is supplemented by a name and a short description of the disease. Alternative names of the same condition are also mentioned along with the hierarchy.

[13] http://www.nlm.nih.gov/bsd/pmresources.html

[14] http://apps.who.int/classifications/icd10/browse/2010/en

References

Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 201–207. Association for Computational Linguistics, 2000.

Jenny Finkel, Shipra Dingare, Christopher D Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. Exploring the boundaries: gene and protein identification in biomedical text. *BMC bioinformatics*, 6(Suppl 1):S5, 2005.

Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. Feature selection techniques for maximum entropy based biomedical named entity recognition. *Journal of biomedical informatics*, 42(5):905–911, 2009.

Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair. Exploiting context for biomedical entity recognition: from syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 88–91. Association for Computational Linguistics, 2004.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann, 2001. ISBN 1-55860-778-1.

Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics, 2004.

Ryan McDonald and Fernando Pereira. Identifying gene and protein mentions in text using conditional random fields. *BMC bioinformatics*, 6(Suppl 1):S6, 2005.

Andreea Bodnari, Louise Deléger, Thomas Lavergne, Aurélie Névéol, and Pierre Zweigenbaum. A supervised named-entity extraction system for medical text. In Forner et al. (2013). URL `http://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-BodnariEt2013.pdf`.

Cyril Grouin. Biomedical entity extraction using machine-learning based approaches. *LREC*, 6: 1–611, 2014.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014, 2014.

Son Doan and Hua Xu. Recognizing medication related entities in hospital discharge summaries using support vector machine. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 259–266. Association for Computational Linguistics, 2010.

James Cogley, Nicola Stokes, and Joe Carthy. Medical disorder recognition with structural support vector machines. In Forner et al. (2013). URL http://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-CogleyEt2013.pdf.

Kaoru Yamamoto, Taku Kudo, Akihiko Konagaya, and Yuji Matsumoto. Protein name tagging for biomedical annotation in text. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 65–72. Association for Computational Linguistics, 2003.

Azad Dehghan. Boundary adjustment of events in clinical named entity recognition. *CoRR*, abs/1308.1004, 2013. URL http://arxiv.org/abs/1308.1004.

Yefeng Wang and Jon Patrick. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the workshop on biomedical information extraction*, pages 42–49. Association for Computational Linguistics, 2009.

Asma Ben Abacha and Pierre Zweigenbaum. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, pages 56–64. Association for Computational Linguistics, 2011.

Denys Proux, François Rechenmann, Laurent Julliard, Violaine Pillet, Bernard Jacq, et al. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. *Genome informatics series*, pages 72–80, 1998.

W John Wilbur, George F Hazard Jr, Guy Divita, James G Mork, Alan R Aronson, and Allen C Browne. Analysis of biomedical text for chemical names: a comparison of three methods. In *Proceedings of the AMIA Symposium*, page 176. American Medical Informatics Association, 1999.

Dietrich Rebholz-Schuhmann, Harald Kirsch, Sylvain Gaudan, Miguel Arregui, and Goran Nenadic. Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, pages 11–18. Association for Computational Linguistics, 2006.

Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC bioinformatics*, 9(Suppl 3):S3, 2008.

Jung-Wei Fan, Navdeep Sood, and Yang Huang. Disorder concept identification from clinical notes: an experience with the share/clef 2013 challenge. In Forner et al. (2013). URL http://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-FanEt2013.pdf.

Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C

Denny. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.

Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. An algorithm that learns what's in a name. *Machine learning*, 34(1-3):211–231, 1999.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Satpreet Arora. *Named Entity Recognition - A Survey.* PhD thesis, Indian Institute of Technology, Bombay.

Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics, 2003.

Maksim Tkachenko and Andrey Simanovsky. Named entity recognition: Exploring features. In *Proceedings of KONVENS*, volume 2012, pages 118–127, 2012.

GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics, 2002.

Oliver Bender, Franz Josef Och, and Hermann Ney. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 148–151. Association for Computational Linguistics, 2003.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

Jing Jiang and ChengXiang Zhai. Exploiting domain structure for named entity recognition. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 74–81. Association for Computational Linguistics, 2006.

Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 1999.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851, 2009.

Pradeep G Mutalik, Aniruddha Deshpande, and Prakash M Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents a quantitative study using the umls. *Journal of the American Medical Informatics Association*, 8(6):598–609, 2001.

Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13, 2005.

Jon Patrick, Yefeng Wang, and Peter Budd. Automatic mapping clinical notes to medical terminologies. In *Proc. Of the 2006 Australian Language Technology Workshop*, pages 75–82, 2006.

Herman Tolentino, Michael Matters, Wikke Walop, Barbara Law, Wesley Tong, Fang Liu, Paul Fontelo, Katrin Kohl, and Daniel Payne. Concept negation in free text components of vaccine safety reports. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1122. American Medical Informatics Association, 2006.

Stefan Gindl, Katharina Kaiser, and Silvia Miksch. Syntactical negation detection in clinical practice guidelines. *Studies in health technology and informatics*, 136:187, 2008.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 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, 2011.

Cheryl Clark, John Aberdeen, Matt Coarr, David Tresner-Kirsch, Ben Wellner, Alexander Yeh, and Lynette Hirschman. Mitre system for clinical assertion status classification. *Journal of the American Medical Informatics Association*, pages amiajnl–2011, 2011.

Sunghwan Sohn, Stephen Wu, and Christopher G Chute. Dependency parser-based negation detection in clinical narratives. *AMIA Summits on Translational Science Proceedings*, 2012: 1, 2012.

Michael Collins and Nigel Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 263–270. Association for Computational Linguistics, 2002.

Alessandro Moschitti. A study on convolution kernels for shallow semantic parsing. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 335. Association for Computational Linguistics, 2004.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic dependency-based n-grams as classification features. In *Advances in Computational Intelligence*, pages 1–11. Springer, 2013.

Mahesh Joshi and Carolyn Penstein-Rosé. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316. Association for Computational Linguistics, 2009.

Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors. *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*, 2013. CEUR-WS.org. URL `http://ceur-ws.org/Vol-1179`.