January 25, 2017

Humboldt Universität zu Berlin
Mathematisch Naturwissenschaftliche Fakultät
Institut für Informatik

Exposé for diploma thesis

Clinical Entity Recognition for ICD-9 Code Prediction in Clinical Discharge Summaries

Author:
Jonathan Bräuer
born 13. October 1988
in Herrenberg, Germany

Supervisors:
Prof. Dr. Ulf LESER
Dr. Johannes STARLINGER

Contents

1 Introduction and Motivation

With the *MIMIC III* Database (Medical Information Mart for Intesive Care) [JAR16], a database of eletronic health records (EHR) of over forty thousand patients, containing over 2 million data points of medical treatment in intensive care units in the Beth Israel Deaconess Medical Center between 2001 and 2012 is made publicly available. This database includes information about demographics, vital sign measurements, laboratory test results, medications, caregiver notes, imaging reports and mortality and therefore offers a wide variety of sources for data analysis. This database has already been proven as a valuable source for different tasks for information extraction, most of which are very specific analyses and concentrate on a single disease or phenomenon [WFF+09, SRSM05, LMC07] or focused on an extracted subset of the database [USSD11, SSV+13].

For this diploma thesis we want to propose to use these EHRs for comparing systems for clinical entity recognition by extracting entities from clinical notes and linking them to the diagnosis given by the caregivers. Evaluations of knowledge extraction tools generally involve testing on texts, which are labelled with mentions by domain experts and then used to evaluate the results of each tool. Fully labelling a corpus as big as the MIMIC III database for this evaluation seems not reasonable, due to the huge time investment required and only using a small subset is prone for over-fitting or not covering all situations. The EHRs already contain a human labelling of its clinical notes in form of ICD-9 (International Identification of Disease) codes, which are tags for performed procedures and final diagnoses and are used for billing in the hospital. We believe, that these codes offer an opportunity to evaluate tools on a much bigger scale, since the used diagnosis codes depend on the available EHR about the patients hospital stay. If a patient is tagged with a diagnosis, it means that either the diagnosis itself, or the disorders responsible for a diagnosis is contained within the clinical notes. We want to focus on the clinical notes to extract clinical entities and investigate the value of this linked information for comparing the performance of different clinical entity extraction methods.

Identifying mentions of biomedical entities within clinical notes is crucial for automated handling of EHRs, as these mentions can be used to index natural language, for frequency-based analyses, to automatically extract relations between entities and many other applications [GdLX11]. Phrases within the text are recognized as known entities and linked to an external database (using a globally unique identifier). These clinical entities can either be references to named entities (entries from a dictionary, defined by a name and synonyms) or references to more complex objects like ontology concepts, which allows to take advantage of the semantic knowledge stored in ontologies like structural information (sub or super classes of concepts etc.).

Named entity recognition (NER) and named entity normalization (NEN, identifying a canonical unambiguous referent for ambiguous names [KJDR08]) have been extensively explored and many tools and methods are available for these tasks, but concept recognition (CR) is more complex, since concepts are not only defined by the term, which was given to them as a name, but by a class of objects, which is described within each concept. To identify the concept "Hypoglycemia" within natural language, it can either be mentioned by its name, synonyms, by describing the effect of this phenotype (a decreased concentration of glucose in the blood), or even only by mentioning a measured value of the glucose concentration. Therefore, simply relying on NER or NEN methods can not cover the full scope of recognising concept mentions.

Another problem of automated handling of EHRs is, that they were written by caregivers and are generally used by other domain experts, which leads to a very domain specific language, filled with technical terms and abbreviations for these, which impedes the transfer of existing

tools to this specific domain. It is necessary to include knowledge about the domain in the knowledge extraction process, even more than within many other domains.

In this exposé we want to introduce a possible approach to use the EHRs to compare the performance of different state-of-the-art systems for clinical entity extraction. We want to use the ICD-9 diagnosis codes, given by responsible caregivers, to evaluate the quality of predicting ICD-9 codes from extracted entities.

2 Related Work

A wide variety of tools and methods for the NER task in the biomedical domain exists. These tools can generally be divided into 3 categories [KMR12]: dictionary based approaches, which match terms in natural language against a pre defined dictionary; rule-based systems, which use lexical or linguistic rules to identify entities; or machine learning techniques, which treat NER as a classification problem. Many tools also use a combination of the above, as especially machine learning approaches often use dictionaries or lexical rules. One of the most known tools is *MetaMap* [Aro01], which allows the discovery of UMLS Metathesaurus [Bod04] concepts in natural language. MetaMap analyses sentences by splitting them into phrases, which are then searched for named entities (with respect to acronyms, abbreviations and synonyms).

Another well known tool, more specialised on clinical terms is the *C-TAKES* system [SMO$^+$10], which recognises a wide variety of concepts in the SNOMED [SCC97] subset of UMLS.

These tools focus on extracting named entities and normalizing them by linking mentions to UMLS concepts, but can't generally be used directly to recognize complex concepts (like phenotypes), since these concepts are often composed of multiple named entities (or mentioned indirectly). Recent research used NER tools to identify concept candidates and then applied rules to extract ontology concept mentions [KMR11, SCAG15, WA11] or converted the rules into features for machine learning techniques [KMR12, PSO13]. Some of these systems were only tested on biomedical literature and it is yet to be evaluated how well they perform on EHRs.

In the 2013 ShARe-CLEF eHealth challenge [SSV$^+$13], many systems competed in a task to extract disorders (UMLS concept within the semantic group "Disorder"), for which a dataset of 200 labelled clinical notes for training and a test set consisting of 100 labelled clinical notes was provided. A total of 32 teams submitted to the task, and the best system achieved an F1 score of 0.75 [RBN13] by extending Cocoa, a dictionary and rule based entity extraction system. Many of the systems were either based on C-Takes ([LWJS13] with an F1 score of 0.67), MetaMap ([ZHKN13] with an F1 score of 0.58) or a combination of these two ([OGS14] with an F1 score of 0.79). The winning team described the main issues of applying existing tools to the clinical domain as: phrases in the texts are peculiar to clinical notes, many common words are ambiguous and a large number of unknown acronyms is used.

A machine learning approach to extract clinical entities within hospital discharge summaries is presented in [TCW$^+$13]. A corpus, created for the 2010 i2b2 NLP challenge [USSD11], which had a similar task as the 2013 challenge, was used to evaluate their approach. Also a comparison of the performance of structural support vector machines (SSVMs) and conditional random fields (CRFs) was made, using a broad set of features. Their system outperformed all systems, which originally participated in the challenge and therefore seems promising for the NER task in discharge summaries.

In [GKD$^+$15], a comparative study of 3 other systems for HPO concept recognition is presented, evaluated using a gold standard for this task [CRBJ$^+$], consisting of PubMed [fBI] article abstracts. In this study, a new system called Bio-LarK CR was introduced and com-

pared to 2 other systems. Bio-LarK CR was specifically developed for HPO concept recognition and uses a rule based approach, combined with pattern matching methods to recognise non-canonical phenotypes as well as conjunctive terms. The tools for comparison were the NCBO Annotator[JSY+09], which can be used to recognise ontology concepts from over 300 ontologies in unstructured text using a lexical matching and a rule based semantic expansion; and the OBO Annotator [TRM+14], which is based on linguistic patterns. Bio-LarK CR outperformed the other two systems in the experiments on the gold standard.

3 Goal

The MIMIC III database contains 6984 distinct diagnosis codes of 46520 patients, described by 61532 clinical notes, categorized into 15 categories (e.g. discharge summary, nursing, radiology, ECG, and many more). The goal of this thesis is to extract clinical entities within the clinical notes and to predict matching ICD-9 disease codes, based on found entities. We want to compare different state-of-the-art tools and methods to find the best approach. In the following section we will introduce our proposed method in detail.

4 Proposed Method

The idea of our method can be divided into two steps:

1. **Entity Extraction:** use existing tools to annotate the discharge summaries with clinical entities.

2. **ICD-9 code classification:** assign ICD-9 codes to discharge summaries based on extracted entities and compare performance of extraction tools

We want to compare a number of possible solutions for both steps. For the first step we want to look into using a number of NER and CR methods and existing tools. Some of the tools we want to compare generate an output, for which a mapping to ICD-9 codes exists. We want to compare the tools based on such mappings, if available. To allow a comparison of tools with different output codings, we will use a cross-fold validated classifier for predicting ICD-9 code labellings.

4.1 Data sets

As mentioned, we want to use the clinical notes contained in the MIMIC III database EHRs. As a proof of concept, to reduce required computation time and to allow for a more thorough evaluation we want to create a small subset by randomly selecting a number of frequently, commonly and rarely diagnosed disease codes and use clinical notes, linked to these codes. For each discharged patient, a note was created to summarize his hospital course. Since these discharge summaries contain a concentrated overview of the treatment and diagnoses, we want to focus on these for the entity extraction. The main data source used for entity coding by many of the proposed tools is the UMLS metathesaurus. It contains definitions, synonyms , relations and abbreviations for many clinical relevant concepts and the SNOMED subset already contains a mapping to ICD-9 codes [oM15], which we will be using for part of the evaluation (see chapter 5). Some of the proposed tools rely on biomedical ontologies as a source for their workflow. Some of the specialised ontologies we will use as background information for these

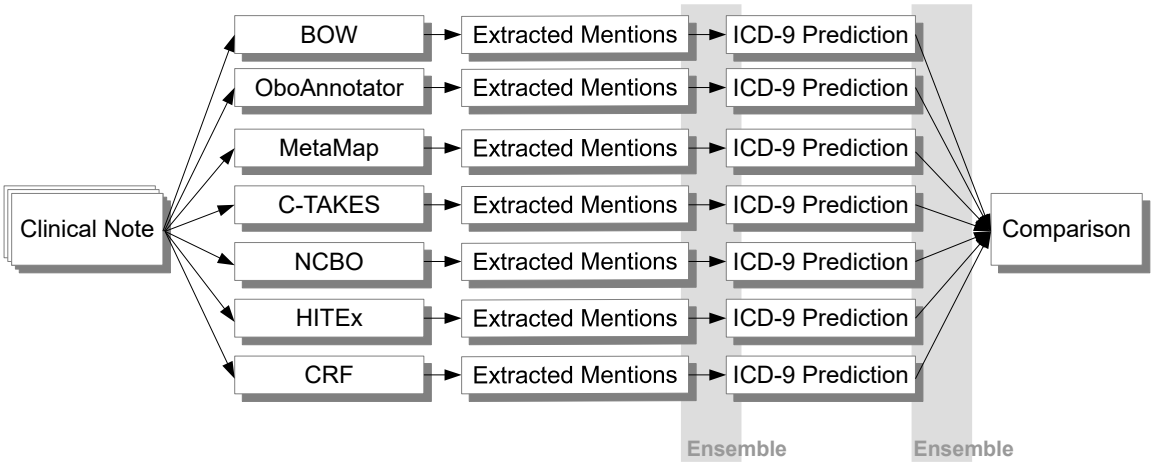

Figure 1: Tool comparison workflow overview.

tools are the HPO (Human Phenotype Ontology) [KDM+13], ChEBI (Chemical Entities of Biological Interest Ontology [HdMD+13]), the PATO (Phenotypic Quality Ontology [Gko10]) and the FMA (Foundational Model of Anatomy Ontolgy [RMJ+03]) ontologies, which all are publicly available. For some of these ontologies (like HPO) a (uncomplete) mapping to UMLS or other ontologies is available, but generally no direct mapping to ICD-9 codes exists, since different domains are represented in the ontologies.

4.2 Clinical Entity Extraction

To extract entities from clinical notes we want to use existing tools and compare their performance. First, we want to run all tools (see next section for detailed listing) and use extracted entities to predict ICD-9 code assignments. Since different tools were created to extract different types of entities we want to include a second step, during which the tool outputs are used to train a CRF for phenotypic concept (HPO) extraction, since we believe that the most significant type for the prediction of tagged diseases are abnormal phenotypes (observable characteristics of an organism).

4.2.1 Named Entity Recognition

The list of tools and methods with generated output codings is shown in table 1 and an overview of the proposed workflow is shown in figure 1. As mentioned earlier, the most commonly used tools for NER in the clinical domain include MetaMap and C-TAKES. Both tools extract a wide variety of semantic types (like diseases or drugs) and have been used successfully for clinical entity extraction. We are also including the OboAnnotator, NCBO and HITEx [GSZ06], an open source natural language processing tool, focused on extracting information from medical documents. To include a purely machine learning based method, we want to include the system proposed by Tang et al. [TCW+13] (see section 2). A simple bag-of-words representation of the text contained in the clinical notes will represent unprocessed features as a baseline. With these we have a total of 7 tools to compare and evaluate.

Many available tools are not usable for our approach, since they are not publicly available

Tool	Output coding	ICD-9 Mapping
Bag-of-words	Word-Vector	No
OboAnnotator	HPO concepts (Phenotypes)	Partial
MetaMap	UMLS concepts	Partial
C-TAKES	SNOMED concepts (subset of clinical UMLS concepts)	Yes
NCBO	Ontology concepts (various ontologies available)	No
HiTEX	UMLS concepts	Partial
CRF	HPO concepts (Phenotypes)	Partial

Table 1: Clinical entity extraction tools with coding schemes of extracted entities.

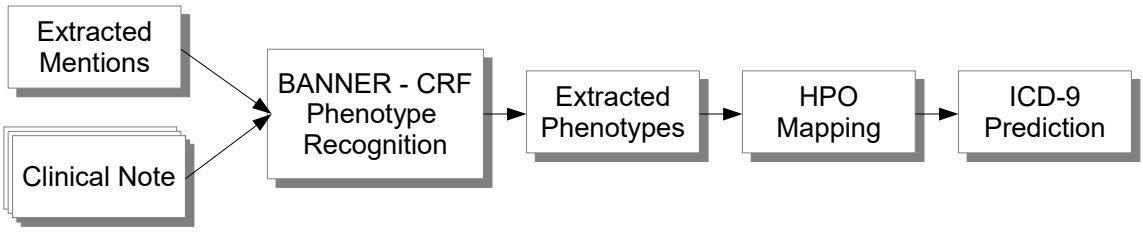

Figure 2: Phenotype extraction workflow overview.

(like Bio-LarK CR) or rely on web APIs (like becas [NCMO13]), which are not suitable for patient data. Other tools, that have proven to be efficient in the clinical domain, like MEDLEE [Fri95] seem not available anymore.

4.2.2 ICD-9 Classification

The selected clinical notes will be annotated using all 7 tools and the results will then be compared for their significance in assigning ICD-9 codes. A mapping for the SNOMED subset of UMLS to ICD-9 codes exists, so tools that generate UMLS output can be evaluated using this mapping.

To allow comparing results between different output codings, we will use classification algorithms for predicting assigned ICD-9 codes. Extracted entities can serve as features for classification methods, using well known algorithms like naive-bayes classification, the k-nearest-neighbours algorithm or classification using linear support vector machines. Which method is suited best will be evaluated in the thesis. Using these classification methods, the most likely ICD-9 code for the extracted clinical entities in each clinical note can be identified and, using cross-fold validation, the predictions generated by the results of each tool can be compared. To find the best approach for ICD-9 diagnosis prediction we also want to look into combining the outputs of different tools by combining the extracted entity mentions of multiple tools for ICD-9 code prediction and by combining the ICD-9 predictions generated by multiple tools.

4.2.3 Phenotype Extraction

The outputs of the mentioned baselines differ, since recognised concepts are contained in different ontologies or dictionaries. In [KMR12] Khordad et al. introduced a method for HPO concept extraction using a baseline system for semantic type tagging. The text is labelled using the

output of the baseline system and the CR task is transformed into a sequential labelling problem, which is solved with BANNER [LG+08], a CRF implementation for Java.

Extracting HPO concepts, generated by the different systems, has the advantage that results are more easily comparable. Also we believe, that using phenotypes to predict the outcome diagnosis is the most meaningful, since diseases are usually diagnosed by observing abnormal phenotypes. Found sequences for phenotype candidates then need to be mapped to HPO concepts using simple rules as mentioned in Khordad's approach and can be evaluated using the test-suite of the HPO goldstandard [GKD+15].

We want to evaluate the feasibility of this process for the different baseline systems and compare the results of raw baseline output with identified phenotype concepts. This also enables usage of structural distance measures, based on the concept structure in the HPO. An overview of this approach is shown in figure 2. The entities, extracted by the earlier presented tools are used to label the clinical notes. A CRF is trained on a phenotype labelled corpus and then used to sequentially label clinical notes with phenotypes. These phenotypes are mapped to HPO concepts and used for the ICD-9 code prediction (using similar classification methods as mentioned in section 4.2.2).

Due to the necessity of a training corpus, it is yet unknown, if this evaluation is applicable. The earlier mentioned gold standard seems as a good starting point, but does not consist of discharge summaries or similar text types (it was created from biomedical article abstracts). Another possible corpus is the 2013 shARe/CLEF corpus, but we are still awaiting access rights for this data. At this point it is unclear, whether this phenotype extraction step will be feasible, since this approach relies on a high quality baseline method for semantic type tagging and having access to a training corpus. We reserve excluding this step from the diploma thesis.

5 Evaluation

After identifying entities using the presented methods and mapping or predicting ICD-9 codes to discharge summaries, we want to compare the results for all methods to find the best tool or combination of tools for ICD-9 code assignment. The correctness for predictions generated by a direct mapping can be checked by simply comparing manually tagged ICD-9 codes with the extracted mentions and calculating usual measures (precision, recall and F1 score).

The correctness of ICD-9 code predictions generated by a classifier can be evaluated through cross-fold validation. After removing a small subset of the selected document, the classifiers are trained on the remaining documents. For each removed document, the classifiers then calculates a probability for each code assignment. This allows to order assigned codes for each document by probability and checking, whether the most likely code is the correct one, or if the correct code is contained in the most likely predicted codes.

Since ICD-9 codes are defined in a hierarchical structure and ontologies contain structural information about relations between concepts, we want to look into using distance based measures to improve the evaluation estimates. If a code is predicted, which is structural close to the manually assigned code, the result is "less wrong" then one, which is structural distant.

We hope that this evaluation can show, which tools are suited well for usage in clinical notes and identify causes for shortcomings.

References

[Aro01] 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.

[Bod04] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[CRBJ+] K Bretonnel Cohen, Christophe Roeder, William A Baumgartner Jr, Lawrence E Hunter, and Karin Verspoor. Test suite design for ontology concept recognition systems.

[fBI] US National Center for Biotechnology Information. Pubmed. https://www.ncbi.nlm.nih.gov/pubmed, accessed: 31.10.2016.

[Fri95] Carol Friedman. Medlee-a medical language extraction and encoding system. *Columbia University, and Queens College of CUNY*, 1995.

[GdLX11] Rodrigo Rafael Villarreal Goulart, Vera Lúcia Strube de Lima, and Clarissa Castellã Xavier. A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, 17(2):103–116, 2011.

[GKD+15] Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M Couto, Gareth Baynam, Andreas Zankl, and Peter N Robinson. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, 2015:bav005, 2015.

[Gko10] George Gkoutos. Phenotype, attribute and trait ontology, 2010. http://obofoundry.org/ontology/pato.html, accessed: 31.10.2016.

[GSZ06] Sergey Goryachev, Margarita Sordo, and Qing T Zeng. A suite of natural language processing tools developed for the i2b2 project. In *AMIA Annual Symposium Proceedings*, volume 2006, page 931. American Medical Informatics Association, 2006.

[HdMD+13] Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, et al. The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463, 2013.

[JAR16] Shen L Lehman L Feng M Ghassemi M Moody B Szolovits P Celi LA Johnson AEW, Pollard TJ and Mark RG. Mimic-iii, a freely accessible critical care database. *Scientific Data*, page DOI: 10.1038/sdata.2016.35., 2016.

[JSY+09] Clement Jonquet, Nigam Shah, Cherie Youn, Chris Callendar, Margaret-Anne Storey, and M Musen. Ncbo annotator: semantic annotation of biomedical data. In *International Semantic Web Conference, Poster and Demo session*, volume 110, 2009.

[KDM+13] Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme CM Black, Danielle L Brown, Michael Brudno, Jennifer Campbell, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, page gkt1026, 2013.

[KJDR08] Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval*, pages 705–710. Springer, 2008.

[KMR11] Maryam Khordad, Robert E Mercer, and Peter Rogan. Improving phenotype name recognition. In *Canadian Conference on Artificial Intelligence*, pages 246–257. Springer, 2011.

[KMR12] Maryam Khordad, Robert E Mercer, and Peter Rogan. A machine learning approach for phenotype name recognition. In *Proceedings of Coling*, pages 1425–1440, 2012.

[LG+08] Robert Leaman, Graciela Gonzalez, et al. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific symposium on biocomputing*, volume 13, pages 652–663, 2008.

[LMC07] Qiao Li, Roger G Mark, and Gari D Clifford. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. *Physiological measurement*, 29(1):15, 2007.

[LWJS13] Hongfang Liu, Kavishwar B Wagholikar, Siddhartha Jonnalagadda, and Sunghwan Sohn. Integrated ctakes for concept mention detection and normalization. In *CLEF (Working Notes)*, 2013.

[NCMO13] Tiago Nunes, David Campos, Sérgio Matos, and José Luís Oliveira. Becas: biomedical concept recognition services and visualization. *Bioinformatics*, 29(15):1915–1916, 2013.

[OGS14] John David Osborne, Binod Gyawali, and Thamar Solorio. Evaluation of ytex and metamap for clinical concept recognition. *arXiv preprint arXiv:1402.1668*, 2014.

[oM15] U.S. National Library of Medicine. Icd-9-cm diagnostic codes to snomed ct map, 2015. https://www.nlm.nih.gov/research/umls, accessed: 12.12.2016.

[PSO13] Jon D Patrick, Leila Safari, and Ying Ou. Share/clef ehealth 2013 named entity recognition and normalization of disorders challenge. In *CLEF (Working Notes)*, 2013.

[RBN13] SV Ramanan, Shereen Broido, and P Senthil Nathan. Performance of a multi-class biomedical tagger on clinical records. In *CLEF (Working Notes)*, 2013.

[RMJ+03] Cornelius Rosse, José LV Mejino Jr, et al. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500, 2003.

[SCAG15] Daniel Sánchez Cisneros and Fernando Aparicio Gali. Uem-uc3m: an ontology-based named entity recognition system for biomedical texts. Association for Computational Linguistics, 2015.

[SCC97] Kent A Spackman, Keith E Campbell, and Roger A Côté. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association, 1997.

[SMO+10] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 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, 2010.

[SRSM05] JX Sun, AT Reisner, M Saeed, and RG Mark. Estimating cardiac output from arterial blood pressure waveforms: a critical evaluation using the mimic ii database. In *Computers in Cardiology, 2005*, pages 295–298. IEEE, 2005.

[SSV+13] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer, 2013.

[TCW+13] Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. *BMC medical informatics and decision making*, 13(1):1, 2013.

[TRM+14] Maria Taboada, Hadriana Rodríguez, Diego Martínez, María Pardo, and María Jesús Sobrido. Automated semantic annotation of rare disease cases: a case study. *Database*, 2014:bau045, 2014.

[USSD11] 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.

[WA11] Chunye Wang and Ramakrishna Akella. Ucsc's system for clef ehealth 2013 task. 2011.

[WFF+09] David T Winkler, Felix Fluri, Peter Fuhr, Stephan G Wetzel, Philippe A Lyrer, Stephan Ruegg, and Stefan T Engelter. Thrombolysis in stroke mimics frequency, clinical characteristics, and outcome. *Stroke*, 40(4):1522–1525, 2009.

[ZHKN13] Guido Zuccon, Alexander Holloway, Bevan Koopman, and Anthony Nguyen. Identify disorders in health records using conditional random fields and metamap: Aehrc at share/clef 2013 ehealth evaluation lab task 1. In *Proceedings of CLEF Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*, 2013.