Improving the Extraction of Clinical Concepts from Clinical Records

Xiao Fu and Sophia Ananiadou
National Centre for Text Mining
School of Computer Science, University of Manchester
Manchester Institute of Biotechnology,
131 Princess Street, M1 7DN, Manchester, UK
E-mail: fux@cs.man.ac.uk, sophia.ananiadou@manchester.ac.uk

Abstract

Essential information relevant to medical problems, tests, and treatments is often expressed in patient clinical records with natural language, making their processing a daunting task for automated systems. One of the steps towards alleviating this problem is concept extraction. In this work, we proposed a machine learning-based named entity recognition system to extract clinical concepts from patient discharge summaries and progress notes without the need for any external knowledge resources. Three pre- and post-processing methods were investigated, i.e. truecasing, abbreviation disambiguation, and distributional thesaurus lookup, the individual annotation results of which were combined into a final annotation set using two refinement schemes. While truecasing and abbreviation disambiguation capture the inflectional morphology of words, the distributional thesaurus lookup allows for statistics-based similarity matching. We achieved a maximum F-score of 0.7586 and 0.8444 for exact and inexact matching, respectively. Our results show that truecasing and annotation combination are the enhancements which best increase the system performance, whereas abbreviation disambiguation and distributional thesaurus lookup bring about insignificant improvements.

Keywords: concept extraction, clinical records, truecasing, abbreviation disambiguation, distributional thesaurus

1. Introduction

The use of electronic medical records (EMRs) that contain clinical information on individual patients has been shown to reduce healthcare costs as well as improve the quality of healthcare provided (Holroyd-Leduc et al., 2011). It has become one of the most important novel technologies in the clinical domain (Murphy et al., 2007). However, clinicians and healthcare providers often find difficulties in filtering and retrieving useful knowledge from those clinical documents since the majority of EMR data is still expressed in narrative form (Pedersen, 2006; Xu et al., 2012a; Byrd et al., 2013). Therefore, concept extraction which has been actively employed to unlock information from free-text content (Denny et al., 2010; Gonzalez et al., 2012; Xu et al., 2012b) could be used to address this problem and consequently, to improve clinical care.

In this paper, we investigated a machine learning (ML)-based system to identify clinically relevant entities from patient discharge summaries and progress notes, and assign semantic types (i.e., problem, test, and treatment) to them, as specified by the concept extraction task of the 2010 i2b2/VA challenge (Uzuner et al., 2011).

2. Related Work

Recently, several studies have been conducted to apply rule- and/or ML-based approaches on EMRs to assist scientists in the extraction of valuable information. deBruijn et al. (2011) developed a discriminative semi-Markov hidden Markov model (HMM) based on a wide range of features generated from both training texts and external knowledge sources to identify clinical concepts in discharge summaries and progress reports. They observed that projecting textual features onto a high-dimensional feature space as well as the utilisation of external sources for semantic and syntactic tagging are beneficial. Jiang et al. (2011) proposed a hybrid clinical entity extraction system combining an ML-based named entity recogniser with rule-based methods for post-processing. Two ML algorithms, conditional random fields (CRF) and support vector machines (SVM) were applied. Their results suggest that CRF outperformed SVM, and the semantic features derived from existing medical knowledge bases can enhance the performance of clinical named entity recognition (NER) significantly. Similarly, Patrick et al. (2010) developed a hybrid medication extraction model which was based on a cascaded approach, incorporating two ML classifiers (i.e., CRF and SVM) and several pattern matching rules. In order to effectively use the two ML methods, they manually annotated another 145 records to augment the training set since they had access to only 17 annotated records initially. The performance of their model is better than those of the rule-based approaches adopted for the same task.

Instead of developing new ML- or rule-based methods, some researchers have focussed on combining existing approaches. Kang et al. (2011) integrated six named entity recognisers and chunkers, including both ML- and thesaurus-based methods, to annotate clinical records. Majority voting (Penrose, 1946), a method that validates if majority of the systems have given identical results, was then applied to generate composite annotations. They conclude that a combined annotation system for clinical records performs substantially better than any of the individual systems.

In our study, on the other hand, we did not leverage any external medical ontologies, such as UMLS (Bodenreider

et al., 2004) or SNOMED-CT (Lee et al., 2014). A relevant work is performed by Yang (2010), in which a rule-based medication extraction system for textual patient reports without employing any external medical knowledge resources was constructed. Seven types of medication information including drug names, dosages, modes of administration, frequencies, durations, reasons and contexts were extracted. Based on his findings, the rule-based approach, built based on a small, annotated development corpus and without utilising any knowledge bases, obtained satisfactory performance in the concept extraction task. In contrast, this paper investigates the recognition of different concept types and in clinical notes using a ML-based concept extraction system.

3. Methods

We initially constructed a CRF-based NER system to tag clinical records, which was considered as the baseline in our experiments. Several pre- and post-processing methods were then explored to improve the annotation performance.

3.1 Dataset

This study is performed on patient discharge summaries and progress notes in the 2010 i2b2/VA challenge data set (Uzuner et al., 2011). Seventy-three (73) human annotated records were used for system training, while the test data set was comprised of 256 annotated records. For each record, there are two types of files. One is the report file, which has already been split into sentences, and the other is the annotation file, in which annotations are specified by means of line and word numbers that indicate text spans corresponding to concepts. Table 1 shows a sentence in report files and its corresponding annotations in annotation files. In this corpus, there are 16,779 entities which were assigned the problem label, 12,261 assigned the test label and 12,417 annotated as treatment.

Sentences (Report File)	Trauma series demonstrated no evidence of a pelvic fracture.
Concept Annotations (Annotation File)	c='trauma series' 44:0 44:1||t='test' c='a pelvic fracture' 44:6 44:8||t='problem'

Table 1: Examples from a pair of report and annotation files

3.2 Baseline

We selected NERsuite[1], which is a freely available NER tagger based on the CRFsuite implementation of CRFs (Okazaki, 2007), to generate the concept annotations. NERsuite consists of three modules, a tokeniser, a modified version of GENIA tagger, and a NER. The procedure is as follows: First, the sentence-split report files were processed by the tokeniser which was used to segment each sentence into tokens, and compute the position of each token in the sentence. The modified GENIA tagger was then applied to produce three token features, i.e., the part-of-speech (POS) tags, lemmas, and chunk tags. In order to train the NER model, the annotated records were converted into a Begin, Inside, Outside (BIO) format to obtain the correct named entity (NE) label for individual tokens. Consequently, we used a total of seven possible NE labels, i.e., 'B-/I- problem, test, treatment', and 'O'. The example sentence in Table 1 will be transformed to the token-level annotation shown in Table 2. A CRF model was then trained to assign the label to each token in the test corpus.

3.3 Truecasing

Clinicians are used to writing short and terse sentences with limited use of full sentence syntax in clinical records, examples of which include *'Percocet for pain as needed. Aspirin 81 mg daily.'* and *'Patient may shower, no baths. No driving for at least one month.'*. While capitalisation is typically used only to begin sentences, we have observed that several full words are also capitalised for the purpose of emphasis. Given these considerations, we considered the application of the truecasing method, generally used to restore the correct case of tokens in raw texts to consistently transform token expressions to their canonical forms (Lita et al., 2003; Pyysalo and Ananiadou, 2013). The Truecase Asciifier module in Argo[2] (Rak et al., 2012a; Rak et al., 2012b) was employed to generate tokens in their normalised case form. Truecasing was performed before tokenisation (i.e., on input sentences) as it takes into consideration the context surrounding any given token.

3.4 Abbreviation Disambiguation

Acronyms and abbreviations also appear widely in clinical records, some of which follow certain conventions whereas others are ambiguous (Carroll et al., 2012). For instance, *q.d.* is the standard acronym for *'quaque die'*, which means once a day. *MI*, having more than 80 possible full forms, can stand for *myocardial infarction*, *mitotic index*, and *myo-inositol*, while *CAD* may mean any of *coronary artery disease*, *computer-aided diagnosis*, and *caldesmon* depending on the context. Therefore, a disambiguation process is crucial to ensure the correct interpretation of the records (Uzuner et al., 2011). In our study, we examined the impact of abbreviation disambiguation (AD) on NER by transforming the acronyms and abbreviations or short forms in the report files to their suitable expanded full forms estimated by Acromine Disambiguator[3], a word sense disambiguation classifier trained on MEDLINE abstracts (Okazaki et al., 2010).

3.5 Distributional Similarity

After we automatically annotated the test corpus using the newly trained NERsuite model, a distributional thesaurus,

[1] http://nersuite.nlplab.org/

[2] http://argo.nactem.ac.uk/

[3] http://www.nactem.ac.uk/software/acromine_disambiguation/

NE Label	Beginning Position	Past-the-End Position	Token	Lemma	POS	Chunk
B-test	0	6	Trauma	Trauma	NN	B-NP
I-test	7	13	series	series	NN	I-NP
O	14	26	demonstrated	demonstrate	VBD	B-VP
O	27	29	no	no	DT	B-NP
O	30	38	evidence	evidence	NN	I-NP
O	39	41	of	of	IN	B-PP
B-problem	42	43	a	a	DT	B-NP
I-problem	44	50	pelvic	pelvic	JJ	I-NP
I-problem	51	59	fracture	fracture	NN	I-NP
O	60	61	.	.	.	O

Table 2: The transformed annotations of 'Trauma series demonstrated no evidence of a pelvic fracture.'

in which each word is associated with a list of other words according to their distributional similarity (DS) scores (Carroll et al., 2012), was constructed using the training documents. The thesaurus was then applied to reassign concept types to tokens in the initial results for improving recall, based on the intuition that similar words tend to have the same concept type. In Figure 1 is a list of six words which were identified as most similar to 'artery' in the thesaurus. The numbers in the second column are DS scores that indicate the degree of similarity.

artery	artery	1.0
	embolism	0.1701
	insufficiency	0.1585
	angioplasty	0.1496
	coronary	0.1493
	percutaneous	0.1033
	⋮	

Figure 1: Distributional thesaurus entries for 'artery'

Pair-wise DS scores between words in the documents were computed using Lin's measurement (Lin, 1998), as shown in Equation 1.

$$sim(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1,r,w) + I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w) + \sum_{(r,w)\in T(w_2)} I(w_2,r,w))} \quad (1)$$

where w and r represent words and the relationship between two words, respectively. $I(w,r,w')$ is equal to the mutual information between w and w' (see Equation 2).

$$I(w,r,w') = \log \frac{||w,r,w'|| \times ||*,r,*||}{||w,r,*|| \times ||*,r,w'||} \quad (2)$$

where $||w,r,w'||$ denotes the frequency of the triple (w,r,w') in the corpus, and $*$ means there are no specific words. $T(w)$ means the set of pairs (r,w') such that $I(w,r,w')$ is positive. Here we used the 4 types of proximity relationships as in the work of Carroll et al. (2012), shown in Table 3.

Relation Name	Explanation
prev	Previous word
prev_window	Word within a distance of 2-5 words to the left
next	Next word
next_window	Word within a distance of 2-5 words to the right

Table 3: Proximity relationships used to calculate DS

3.6 Hybrid Methods

Since a hybrid annotation system has been demonstrated to have a better performance than any of the individual systems (Kang et al., 2010; Uzuner et al., 2011), we combined the three aforementioned techniques (i.e., truecasing, AD, and DS) into a final annotation system by two schemes . Each of the schemes is illustrated in Figures 2 and 3, respectively.

The first is a sequential scheme in which the training and test documents were processed by the truecasing and the AD modules consecutively. In this way, we prevent the AD module from treating words in all uppercase letters as short forms and incorrectly expanding them. For example, AD will typically expand *'END'* to *'endurance'*, but with the application of the truecasing module, the former will be first transformed to *'end'*, hence avoiding its unnecessary expansion. Next, the texts with the correct case information and expanded forms were used to train and test the NERsuite model to obtain the concept extraction results. A distributional thesaurus built on the new training documents was then employed to assign new annotations to the test documents.

In the second scheme, a parallel one, we simply compute the union of the annotation results of these three systems. If the concept annotations provided by any two of the three systems are identical, they are generated as final, combined annotations. Otherwise, the result of truecasing is considered as the final annotation, since this technique performed best on the training texts.

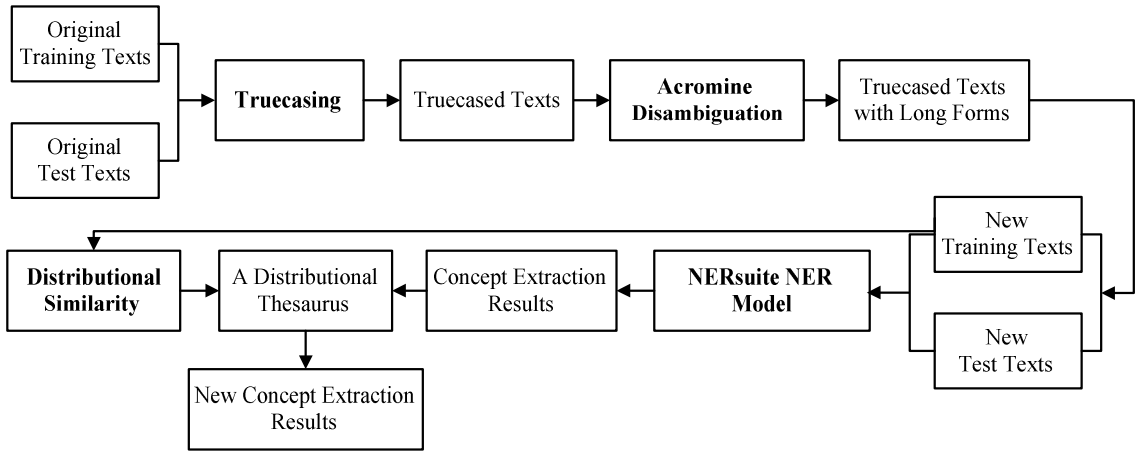

Figure 2: The framework of the sequential hybrid annotation system

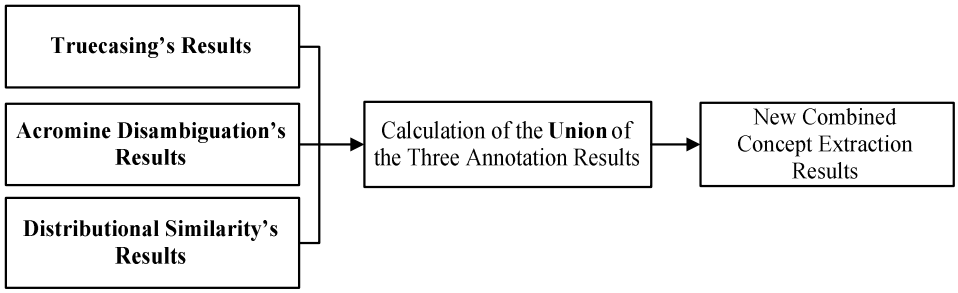

Figure 3: The framework of the parallel hybrid annotation system

4. Results and Discussion

4.1 Results

We explored the performance of truecasing, AD and DS alone, and various combinations of these three methods for clinical concept annotation. Table 4 summarises the performance of our systems on the 256 test records based on exact and inexact matching. The former requires that automatically generated annotations are exactly the same as those in the gold standard, while the latter additionally takes into account annotations with inexact boundaries as long as they have overlapping spans (Kang et al., 2010; Uzuner et al., 2011).

The truecasing method obtained an F-score of 0.7586 according to exact matching, making it our best performing concept annotation system. In contrast, the AD method did not show any improvements. While the DS method produced optimal recall (0.7225), precision was compromised (0.6466), resulting in a reduction in F-score. The F-score of the sequential combination of the truecasing and AD methods is 0.7556 which is worse than that of the purely truecasing-based system and even that of the baseline. The addition of the DS method further reduced the F-score to 0.6877. Nevertheless, integrating those three methods using the parallel scheme achieved better performance than the baseline and the parallel combination of truecasing and AD.

The best F-score measurements using inexact matching is 0.8444, achieved by the parallel combination of truecasing and AD. The parallel combination of truecasing, AD, and DS, truecasing on its own, and the sequential combination of truecasing and AD models outperformed the baseline as well.

Systems	Recall	Precision	F-score
Exact Matching			
Baseline	0.7132	0.8054	0.7565
T	**0.7147**	**0.8083**	**0.7586**
AD	0.7099	0.8014	0.7529
DS	0.7225	0.6466	0.6825
T + AD (S)	0.7115	0.8055	0.7556
T + AD + DS (S)	0.7210	0.6573	0.6877
T + AD (P)	0.7283	0.7849	0.7556
T + AD + DS (P)	0.8047	0.7140	0.7567
Inexact Matching			
Baseline	0.7927	0.8951	0.8408
T	0.7931	0.8970	0.8419
AD	0.7925	0.8947	0.8405
DS	0.8133	0.7280	0.7683
T + AD (S)	0.7922	0.8969	0.8413
T + AD + DS (S)	0.8138	0.7420	0.7762
T + AD (P)	**0.8140**	**0.8772**	**0.8444**
T + AD + DS (P)	0.7944	0.8953	0.8419

T, truecasing; S, the sequential scheme; P, the parallel scheme;

Table 4: The performance for clinical concept extraction

Concept Types	Problem			Treatment			Test		
	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
Exact Matching									
Baseline	0.7094	0.7757	0.7411	0.6846	0.8237	0.7477	0.7473	0.8292	0.7861
T	0.7138	0.7803	0.7456	0.6890	0.8271	0.7517	0.7419	0.8298	0.7834
AD	0.7071	0.7742	0.7392	0.6854	0.8196	0.7465	0.7385	0.8213	0.7780
DS	0.6562	0.7175	0.6855	0.6304	0.7527	0.6862	0.5110	0.5685	0.5382
T + AD (S)	0.7102	0.7793	0.7432	0.6881	0.8232	0.7496	0.7368	0.8252	0.7785
T + AD + DS (S)	0.7181	0.5596	0.6290	0.6974	0.7827	0.7376	0.7487	0.7124	0.7301
T + AD (P)	0.7245	0.7559	0.7399	0.7032	0.8063	0.7512	0.7589	0.8051	0.7813
T + AD + DS (P)	0.7099	0.7757	0.7413	0.6849	0.8251	0.7485	0.7493	0.8262	0.7859
Inexact Matching									
Baseline	0.8202	0.8970	0.8569	0.7450	0.8963	0.8137	0.8035	0.8915	0.8452
T	0.8237	0.9005	0.8604	0.7470	0.8967	0.8150	0.7982	0.8927	0.8428
AD	0.8188	0.8966	0.8559	0.7459	0.8919	0.8124	0.8039	0.8947	0.8469
DS	0.7851	0.8584	0.8201	0.8169	0.8560	0.7803	0.6178	0.6873	0.6507
T + AD (S)	0.8185	0.8982	0.8565	0.7471	0.8937	0.8139	0.8018	0.8980	0.8472
T + AD + DS (S)	0.8444	0.6579	0.7396	0.7626	0.8559	0.8065	0.8241	0.7841	0.8036
T + AD (P)	0.8412	0.8776	0.8590	0.7656	0.8779	0.8179	0.8259	0.8761	0.8503
T + AD + DS (P)	0.8219	0.8981	0.8583	0.7452	0.8977	0.8144	0.8066	0.8894	0.8460

Table 5: Concept extraction results for each concept type

Systems	Exact Matching				Inexact Matching			
	FN	(%)	FP	(%)	FN	(%)	FP	(%)
Baseline	711	2.14	379	1.14	579	1.75	298	0.90
T	706	2.13	372	1.12	586	1.77	294	0.89
AD	724	2.18	472	1.42	565	1.70	372	1.12
DS	647	1.95	1438	4.34	513	1.55	1270	3.83
T + AD (S)	707	2.13	477	1.44	558	1.68	377	1.14
T + AD + DS (S)	650	1.96	1111	3.35	501	1.51	909	2.74
T + AD (U)	621	1.87	541	1.63	464	1.40	443	1.34
T + AD + DS (U)	693	2.09	406	1.22	559	1.69	317	0.96

%: The number of unrecognised abbreviations/The total number of abbreviations in the test set*100

Table 6: The number of unrecognised abbreviations

We also assessed the extraction performance for each entity type (i.e., problem, treatment, and test), as shown in Table 5. Results for the test type indicate the highest F-scores, with around 0.77 using exact matching for all the systems except DS. Based on inexact matching, the systems excluding the sequential combination of truecasing, AD and DS achieved the best results in problem extraction with F-scores over 0.82.

In our system-level evaluation, truecasing obtained the best performance in recognising problems and treatments on the test records using exact matching, while none of the methods improved extraction performance for tests.

The highest F-score using inexact matching was also achieved by truecasing; adding AD to it employing the parallel combination scheme resulted to the highest improvements in both treatment and test extraction.

4.2 Discussion

In our study, several NLP methods were investigated to construct NER systems for clinical concept extraction. Instead of using large-dimensional bags of complex features and rules derived from the text itself and external sources (deBruijn et al., 2011; Jiang et al., 2011), we used only lemmas, POS tags and chunk tags generated by the modified GENIA tagger. Our approach offers the possibility to construct effective clinical concept annotation systems on a simple feature set, without using dictionaries or ontologies.

Recent studies have shown that the performance of combined or hybrid annotation systems is better than that of any individual systems (Kang et al., 2011). However, our experiment results are not able to support their findings. Only the parallel combination of truecasing and AD outperformed truecasing and AD individually; the performance of the other hybrid NER models fell in between that of the best and worst performing individual methods. This can be attributed to the possibly conflicting contributions of those three pre- or post-processing methods.

The truecasing method improved the concept extraction performance based on both exact and inexact matching,

demonstrating that correct case information is beneficial for entity recognition in clinical records. The expansion and disambiguation of abbreviations were supposed to decrease the size of acronyms and abbreviations in the set of false negatives. Unexpectedly, the false negatives generated by AD are even slightly greater than that from truecasing and the baseline methods (see Table 6). The low performance of AD can partly be explained by the fact that the Acromine Disambiguation tool which we used was trained on MEDLINE abstracts (Okazaki et al., 2010). Training on a suitable abbreviation disambiguation dictionary specifically geared towards terms in clinical records would likely enhance the performance of Acromine Disambiguation.

The DS method did not add much value to the performance of the NER models, and in certain cases, even reduced their performance significantly. For example, from Table 4 we found an 8.99 and 7.74 percentage points drop in F-score based on exact and inexact matching, respectively, when we added DS to the sequential combination of truecasing and AD. A possible explanation is that the amount of the training records used to build the distributional thesaurus is too small to cover the full breadth of term occurrence, which limited the thesaurus' efficacy. Thus, more annotated records need to be involved to build a high-quality distributional thesaurus.

5. Conclusion

In this study, we developed an ML-based model for clinical entity recognition and systematically evaluated the effects of three pre- and post-processing methods (i.e., truecasing, AD, and DS) which were combined using two schemes (i.e., sequential and parallel). Based on our results, the original model with the addition of truecasing achieved the best performance using exact matching with an F-score of 0.7586. Using inexact matching, the maximum F-score of 0.8444 was obtained by the parallel combination of the truecasing and AD methods.

The utilisation of a more sizeable clinical record data set for training the models can potentially improve the performance of the system. We plan to continue with the development of our system upon gaining access to such data sets; the MIMIC II database that contains clinical records for 32,077 patients (Saeed et al., 2002; Scott et al., 2013) could be considered as an alternative option.

6. References

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1), pp. D267--D270.

Byrd, R.J., Steinhubl, S.R., Sun J., Ebadollahi, S., and Stewart, W.F. (2013). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International Journal of Medical Informatics*.

Carroll, J., Koeling, R., and Puri, S. (2012). Lexical acquisition for clinical text mining using distributional similarity. In *Computational Linguistics and Intelligent Text Processing*, pp. 232--246.

deBruijn, B., Cherry, C., Kiritchenko, S., Martin, J., and Zhu, X. (2011). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5), pp. 557--562.

Denny, J.C., Peterson, J.F., Choma, N.N., Xu, H., Miller, R.A., Bastarache, L., and Peterson, N.B. (2010). Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *Journal of the American Medical Informatics Association*, 17(4), pp. 383--388.

Holroyd-Leduc, J.M., Lorenzetti, D., Straus, S.E., Sykes, L., and Quan, H. (2011). The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence. *Journal of the American Medical Informatics Association*, 18(6), pp. 732--737.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S.T., Mani, S., Denny, J.C., and Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5), pp. 601--606.

Jonnalagadda, S., Cohen, T., Wu, S., and Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1), pp. 129--140.

Kang, N., Barendse, R.J., Afzal, Z., Singh, B., Schuemie, M.J., van Mulligen, E.M., and Kors, J A. (2010). Erasmus MC approaches to the i2b2 Challenge. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*.

Lee, D., de Keizer, N., Lau, F., and Cornet, R. (2014). Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association*, 21, pp. e11--e19.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, 2, pp. 768--774.

Lita, L.V., Ittycheriah, A., Roukos, S., and Kambhatla, N. (2003). tRuEcasIng. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1, pp. 152--159.

Murphy, E.C., Ferris, F.L., and O'Donnell, W.R. (2007). An electronic medical records system for clinical research and the EMR–EDC interface. *Investigative Ophthalmology & Visual Science*, 48(10), pp. 4383--4389.

Okazaki, N. (2007). CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite.

Okazaki, N., Ananiadou, S., and Tsujii, J. (2010). Building a high quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9), pp. 1246--1253.

Patrick, J., and Li, M. (2010). High accuracy information

extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5), pp. 524--527.

Penrose, L.S. (1946). The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1), pp. 53--57.

Pyysalo, S., and Ananiadou, S. (2013). Anatomical entity mention recognition at literature scale. *Bioinformatics*, btt580.

Pedersen, T. (2006). Determining smoker status using supervised and unsupervised learning with lexical features. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Rak, R., Kolluru, B., and Ananiadou, S. (2012a). Building trainable taggers in a web-based, UIMA-supported NLP workbench. In *Proceedings of the ACL 2012 System Demonstration*, pp. 121--126.

Rak, R., Rowley, A., Black, W., and Ananiadou, S. (2012b). Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database: the journal of biological databases and curation*.

Scott, D.J., Lee, J., Silva, I., Park, S., Moody, G.B., Celi, L.A., and Mark, R.G. (2013). Accessing the public MIMIC-II intensive care relational database for clinical research. *BMC medical informatics and decision making*, 13(1), pp. 9.

Saeed, M., Lieu, C., Raber, G., and Mark, R.G. (2002). MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in Cardiology*, pp. 641--644.

Uzuner, Ö., South, B.R., Shen, S., and DuVall, S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5), pp. 552--556.

Xu, Y., Tsujii, J., and Chang, E. (2012a). Named entity recognition of follow-up and time information in 20 000 radiology reports. *Journal of the American Medical Informatics Association*, 19(5), pp. 792--799.

Xu, Y., Hong, K., Tsujii, J., and Chang, C. (2012b). Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5), pp. 824--832.

Yang, H. (2010). Automatic extraction of medication information from medical discharge summaries. *Journal of the American Medical Informatics Association*, 17(5), pp. 545--548.