Wrappers for Feature Subset Selection in CRF-based Clinical Information Extraction

Mike Ebersbach, Robert Herms, Christina Lohr, and Maximilian Eibl

Chair Media Informatics,
Technische Universität Chemnitz, 09107 Chemnitz, Germany
{robert.herms,christina.lohr,maximilian.eibl}@cs.tu-chemnitz.de

Abstract. We present our methodology and the results for Task 1 of the CLEF eHealth Evaluation Lab 2016. This benchmark addresses clinical information extraction related to nursing shift changes, whereas the challenge is to maximize the correctness in structuring written free-text records by automatically identifying relevant text-snippets. Our approach is focused on the exploration of relevant features for conditional random fields. We use wrappers for feature subset selection in conjunction with parameter optimization to consider how the learning algorithm and the dataset interact. First, we composed a feature set based on Stanford CoreNLP, latent Dirichlet allocation, regular expressions, and the ontologies of WordNet and UMLS. Next, the heuristic methods best-first and greedy (hill-climbing) with forward and backward direction have been applied for feature evaluation and selection. Experimental results show that our system outperforms the baseline with a macro-averaged F1 of 0.311 using all features and 0.382 by performing feature selection.

Keywords: Information extraction, Natural language processing, Clinical texts, Feature subset selection, Wrapper, Conditional random fields

1 Introduction

In healthcare respectively clinical institutions treatments need to be documented very carefully in consideration of statutory guidelines. Information flow is critical in health care, because failures lead to preventable adverse events (see [1]). However, state-of-the-art technologies can assist a comprehensive workflow including verbal handover supplemented with written material. In this context, automatic speech recognition supports nursing handover by transforming verbal clinical information into written free-text records (e.g., [2]). Structured records can facilitate the information flow, e.g., by pre-filling a handover form, which requires the identification of relevant content. The CLEF eHealth Evaluation Lab 2016 [3] aims to ease patients and nurses in understanding and accessing eHealth information. Task 1 [4] of this benchmark addresses clinical information extraction related to nursing shift changes, whereas the challenge is to maximize the correctness in structuring written free-text records by automatically identifying relevant text-snippets.

Although information extraction is a challenging field, some prior works have been done concerning clinical texts. A variety of systems demonstrated the integration of Natural Language Processing (NLP) technologies for specific domains, such as radiology reports of the chest [5], mammography [6], pathology [7], dosage information [8] and discharge summaries [9, 10]. Additionally, the authors of [11] showed good results in processing clinical texts concerning principal diagnosis, co-morbidity and smoking status for asthma research. Several studies have worked on the extraction of drug names from clinical notes. In this connection, some promising methods were applied in the past, e.g., string matching [12], rule-based [13], and using lexicon sources [14]. Moreover, the medication information extraction system MedEx [15] uses a combination of lookup, regular expression and rule-based methods to tag medication information in texts.

The statistical modelling method Conditional Random Fields (CRF) [16] has been successfully applied for information extraction in clinical texts (e.g., [17] and [18]). CRF is an undirected graphical model that combines the strength of the Hidden Markov Model and the Maximum Entropy Model [19]. The work of [20] shows that CRF outperforms support vector machines in the clinical domain. Techniques of feature subset selection have been applied for CRF as described in [21] to optimize the results. In [22] the number of features could be reduced to only 3% of the original feature set with only slight loss in performance.

In this working notes paper we present our methodology and the results we obtained in Task 1 (Handover Information Extraction) of the CLEF eHealth Evaluation Lab 2016. Our approach in this work is focused on the exploration of relevant features for CRF in the context of clinical information extraction. The motivation for using feature selection is the advantage of improving the prediction performance, providing faster and more cost-effective predictors, and a better understanding of the constructed models [23]. Basically, there are two main methods: filters and wrappers. Filters use a metric to rank features and a criterion for the selection without a learning algorithm. Wrappers in contrast are considered as a black box, i.e., the feature selection algorithm exists as a wrapper around the learning algorithm [24]. In this work we use wrappers for feature subset selection in conjunction with parameter optimization to consider how the learning algorithm and the dataset interact. First, we composed a set of 41 features based on Stanford CoreNLP, latent Dirichlet allocation, regular expressions, and the ontologies of WordNet and UMLS. Next, the heuristic methods best-first and greedy (hill-climbing) with forward and backward direction have been applied for feature evaluation and selection.

This Paper is organized as follows: In the next section we introduce our designed feature set for clinical information extraction and the implemented system including the methods for feature evaluation and selection. In Section 3 we describe the applied dataset, the experimental setup, and the evaluation results. Finally, we conclude this paper in Section 4 and give some future directions.

2 Method

Our method is based on the exploration of relevant features for CRF in the context of clinical information extraction. First, we extract a set of features using a variety of technologies from the field of NLP. In order to obtain relevant features we perform feature selection using wrappers in conjunction with hyperparameter optimization of CRF.

2.1 System Overview

In order to verify the potential of features for CRF modeling we introduce an analysis system (see Fig. 1) which includes two stages: feature extraction and feature evaluation. The former comprises a number of toolkits and frameworks that are used to extract information on word, sentence, and document-level. The result of this stage is the enrichment of text data, i.e., a feature vector is assigned to each word.

In the second stage, these features will be evaluated by a wrapper-algorithm. A CRF-based classification is performed using a feature subset. In each iteration the resulting evaluation score will be compared and is crucial for further processing. Depending on the applied heuristic the subset can be modified or selected as the final set of features. Thus, by experimental evaluation we obtain the best performing features concerning the constructed CRF model and a given dataset.

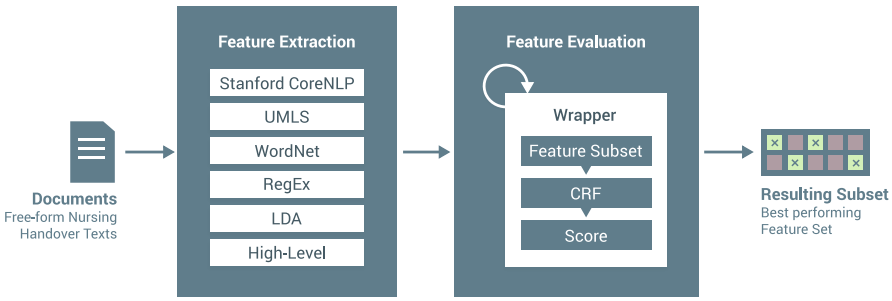

Fig. 1. *System overview*

2.2 Feature Extraction

Based on different toolkits and frameworks from the field of NLP we composed a set of features which are aimed to perform well in combination with the classifier. In order to conduct a linguistic analysis we used the Stanford CoreNLP 3.6.0 [25] with the factory extension for English language [26]. This tool has been

applied for the extraction of named entities and Part-of-Speech (POS) tags as well as structured POS-trees. The lexical database WordNet [27] has been used to obtain meaningfully related words and concepts. Regarding the medical domain, UMLS [28] - the Unified Medical Language System - was integrated including its Metathesaurus. The latent Dirichlet allocation (LDA) implementation [29] of the data mining toolkit KNIME 3.1.2 [30] was used to observe similarities between data. Our preliminary experiments showed appropriate results by using only three LDA topics.

We implemented the feature extraction component and obtained the following 41 features:

- **2 basic features:** the word itself and the corresponding lemma
- **6 Named Entity features:** We extracted the named entity of a word (e.g., PERSON, LOCATION, DURATION). Additionally, we computed the occurrence of the five entities PERSON, NUMBER, DATE, TIME, and DURATION per sentence as separate features.
- **8 POS features:** A POS-tag is used as a feature, e.g., ADJ or NN. In this connection, we used the three features POS-tree, depth of POS-tree, and depth of word-path as nominal values. Based on these structured features we grouped words of a syntactic unit (phrase), calculated the depth of word-path to depth of POS-tree ratio, and determined the tense (past, present, future) as a feature.
- **3 WordNet features:** first hypernym, synonym, and hyponym of a word ranked by WordNet
- **14 UMLS features:** First, we determined a generic category as a feature. For this purpose, we performed a mapping of 22 UMLS thesauri by six self-defined as well as prioritized categories: 1) anatomy – FMA; 2) medical devices – UMD; 3) nursing – NIC, NOC, LNC, ICNP, PCDS, RCD, CCPCC, COSTAR; 4) drugs – AOD, ATC, GC, VANDF, NDDF; 5) diagnosis and diseases – ICD10, ICPC2ICD10ENG, DXP, ICD10AM, ICD10CM, MTHICD9; 6) vaccines – MVX. Next, we used UMLS MetaMap to obtain semantic types (133 categories) and a reduced form comprising 15 semantic groups as described in [31]. Both, semantic types and semantic groups, were applied on word-level and phrase-level resulting in four features. We computed the occurrence of the four semantic groups ANAT, CHEM, DISO, and PROC per sentence and as relative frequency per sentence, i.e., the occurrence divided by the total number of tokens in a sentence (eight features). Finally, the sum of the occurrences of the four semantic groups per sentence was added as a feature.
- **1 LDA feature:** one out of three LDA groups assigned to each word of a sentence
- **1 Regular expression feature:** We conducted rule-based matching on the original free text-form (e.g., based on "came in with" or "under Dr.") to determine the name of patient, name of physician, room number, bed number, age in years, gender, patient admission reason and diagnosis.
- **6 High-Level features:** For each word we extracted high-level features derived from the original free text-form. Three position-based features were

computed: relative position of the word in a sentence as well as in a document and relative position of the sentence in a document. Moreover, we extracted the number of tokens as well as commas per sentence and added a feature as boolean value which describes the occurrence of a word that ends with a comma.

The utilization of all features without its verification can be considered as a brute-force approach. In order to obtain an appropriate feature subset we applied the wrapper approach for feature selection as described in the next section.

2.3 Feature Evaluation

Regarding the wrapper based approach for feature selection, the optimal subset of features can typically be found by using exhaustive search. To accomplish this with n features, 2^n combinations have to be tested which is not feasible for a large number of features. In this work we assess four search algorithms in connection with CRF: best-first and greedy (hill-climbing) with forward and backward direction.

The forward best-first algorithm starts with an empty set of features. In each iteration, a feature will only be added if the performance of the resulting set leads to the maximum performance. It can be lower than the one of the previous iteration. Therefore, this algorithm stops when all features are added. This approach is also called Sequential Forward Selection (SFS) and because of its good tradeoff between accuracy and number of iterations this method is one of the most used wrapper-based algorithms [32]. The forward hill-climbing algorithm also starts with an empty feature set and adds a feature in each iteration. However, a feature will only be added if the performance of the new set is better than the performance of the present best set. If the performance is worse, the feature will be skipped and the next one will be tested. This process is repeated until no better feature set can be found. At this point the algorithm stops. The advantage of this approach is the low number of iterations needed and, subsequently, its shorter execution time. Generally, the determined maximum is a local maximum, i.e., the algorithm stops if no better solution was found for the same level.

One big disadvantage of both algorithms is the so called "nesting" effect: once a feature has been added to the final subset, it can not be removed anymore [33]. Both algorithms also work backwards starting with a full set and removing a feature from the set in each iteration.

3 Experiments and Results

The main goals of the experiments are to verify the performance using our introduced feature set as well as the proposed feature selection methods and to improve the baseline results of Task 1 (Handover Information Extraction) of the CLEF eHealth Evaluation Lab 2016. Before describing the details of the experimental setup, the used dataset is introduced. Afterwards, we discuss results

obtained by the utilization of the two wrapper methods best-first and greedy (hill-climbing) for feature selection and the final results of the evaluation.

3.1 Dataset

We worked with the NICTA Synthetic Nursing Handover Data [1, 34] which was created in 2012 for clinical speech recognition and information extraction related to nursing shift-change handover. Basically, each handover document can be summarized using the five main categories: "Patient Introduction", "My Shift", "Appointments", "Medication", and "Future Care".

In total, there are 35 sub-categories (e.g., last name, age, and current bed of a patient) which represent the slots of a typical handover form and have to be assigned to each word across the dataset. The data consists of 301 synthetic patient cases (handover documents) as text data and is partitioned into 101 training, 100 validation, and 100 testing cases. In this work we used the prepared partitions provided by the organizers, i.e., each partition includes all corresponding handover documents whereas tokens are separated line by line. The training and validation set including the ground truth labels are aimed for method development. The ground truth labels of the test set were not released for evaluation purposes.

3.2 Experimental Setup

The official evaluation measure of Task 1 is the macro-averaged F1 which calculates the average F1 across all categories. As classification algorithm we used the CRF implementation of CRF++ [35]. In order to apply the proposed methods for feature subset selection as described in section 2.3, we implemented a wrapper that handles the datasets in connection with CRF and conducts the search strategies best-first and greedy (hill-climbing), both in forward and backward direction.

In the development phase we optimized the hyperparameter C of the CRF classifier in order to investigate the balance between overfitting and underfitting. In more detail, we computed the macro-average F1 for each feature set using five different values of C (0.01, 0.1, 1, 10, and 100).

In order to obtain the results on the test set we concatenated the training and validation sets to a new training set. The idea was to construct a robust classifier by using more training data. Moreover, we applied the best performing feature set and the corresponding value of C obtained in the development phase. For comparison, we also applied the full feature set.

3.3 Feature Selection

For the NICTA Synthetic Nursing Handover Data some features can be irrelevant and redundant. We performed feature evaluation in conjunction with CRF and the proposed wrapper approaches. Applying all 41 introduced features on the

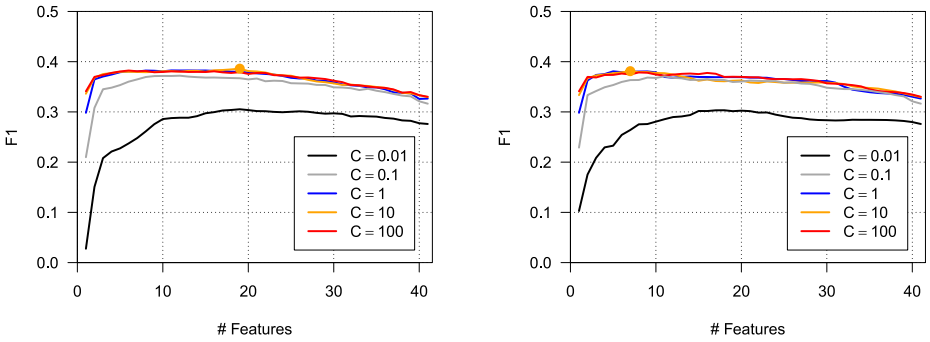

Fig. 2. *Macro-averaged F1 on the validation set according to feature selection using best-first search in forward (left) and backward direction (right) with different values of the hyperparameter C.*

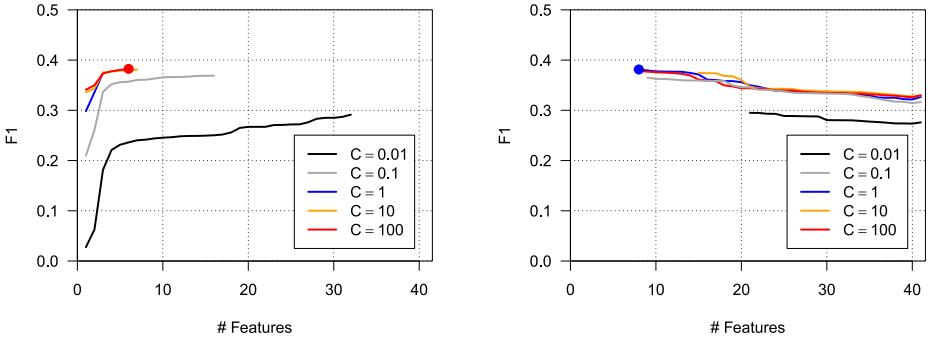

Fig. 3. *Macro-averaged F1 on the validation set according to feature selection using greedy search (hill-climbing) in forward (left) and backward direction (right) with different values of the hyperparameter C.*

validation set the best macro-averaged F1 measure was obtained with 0.330 using hyperparameter optimization C=100 for CRF.

Fig. 2 illustrates the macro-averaged F1 using the best-first search algorithm in forward and backward direction. It can be seen that using only one feature with C=10 and C=100 outperforms the comprehensive feature set. However, using forward direction and C=10 we obtained the best result at the number of 19 features with 0.386 whereas using backward and C=10 the highest value is 0.381 at the number of 7 features.

Fig. 3 shows the results of the macro-averaged F1 according to feature selection using hill-climbing in forward and backward direction. Concerning the forward direction, the local maximum could be reached at an early stage with 6 features and C=100 which results in 0.382. In contrast, the backward direction conducts more evaluations. The search algorithm achieved the best result with 0.381 using 8 selected features and C=1.

Table 1. Summary of the best performing feature set.

Feature group	Feature
Basic	Word
	Lemma
POS	POS tag
	Phrase
Named Entity	Named Entity tag
	PERSON occurrence per sentence
	TIME occurrence per sentence
	DURATION occurrence per sentence
WordNet	Hypernym
	Synonym
	Hyponym
UMLS	Generic category
	Word-based semantic
	Word-based group
	Phrase-based semantic
	Phrase-based group
	ANAT occurrence per sentence
	PROC occurrence per sentence
LDA	LDA group

Our criterion for the final configuration of the system on the test set is the highest macro-averaged F1 measure in the development phase. Thus, we select the 19 features obtained with best-first in forward direction and the hyperparameter C=10 for CRF. Table 1 gives an overview of the final feature set. Most of these features have semantic properties derived by UMLS, WordNet, and Named Entity recognition. The two feature groups Regular Expressions as well as High-Level were left out.

3.4 Results

A series of experiments was carried out for the prediction of categories in clinical information extraction. Regarding the test set, we submitted two runs: method A including all 41 proposed features in combination with the hyperparameter C=100 for CRF and method B with 19 selected features and C=10 which achieved the best result in the development phase as described in section 3.3.

Table 2 gives an overview of the final results. In addition to the macro-averaged F1 across all categories as the official measure, we show the results of Precision and Recall for comparison purposes. It can be seen that our methods outperformed the baseline (0.324) on the validation set. Method A achieved a F1 score of 0.330 with a corresponding improvement of 0.006. It is noticeable that the Precision of the baseline is slightly better. However, the best performance was achieved by method B with a F1 of 0.386 which corresponds to an improvement of 0.062 over the baseline and 0.056 over method A.

Table 2. Summary of the final results on the validation set and the independent test set. The best validation and test results are highlighted in bold.

Method	Validation			Test		
	Precision	Recall	F1	Precision	Recall	F1
NICTA (baseline)	0.485	0.297	0.324	0.435	0.233	0.246
Method A	0.461	0.322	0.330	0.423	0.300	0.311
Method B	**0.511**	**0.382**	**0.386**	**0.493**	**0.369**	**0.382**

Table 3. Top 10 classified categories by method B on the test set.

Category	F1
PatientIntroduction_CurrentRoom	1.000
PatientIntroduction_GivenNames/Initials	0.985
PatientIntroduction_Lastname	0.985
PatientIntroduction_CurrentBed	0.957
PatientIntroduction_Ageinyears	0.934
PatientIntroduction_Gender	0.832
PatientIntroduction_UnderDr_Lastname	0.795
MyShift_Input/Diet	0.759
MyShift_Status	0.694
MyShift_ActivitiesOfDailyLiving	0.556

Concerning method A, there is only a slight drop to 0.311 on the test set. Moreover, the Precision of the baseline system is slightly better with a difference of 0.012. Method B performed best on the test set with a F1 of 0.382 and a corresponding improvement of 0.136 over the baseline and 0.071 over method A. Moreover, method B outperformed the baseline in terms of Precision, Recall, and F1 on the validation as well as on the test set.

The best classified categories by method B on the test set are assigned to "PatientIntroduction" as shown in Table 3. The top 10 categories are ranked concerning the F1 score which is at least 0.5. Furthermore, we observed that each of the top 7 categories reached a F1 of at least 0.75 in both runs, A and B. Nevertheless, our system had some difficulties in classifying words at categories with only few training data. For instance, in case of the label "PatientIntroduction_CarePlan" we obtained many false positives in both runs but no true positives.

4 Conclusions

We presented a methodology concerning feature subset selection in clinical information extraction for Task 1 of the CLEF eHealth Evaluation Lab 2016. Our approach is focused on the exploration of relevant features for conditional random fields. We use wrappers for feature subset selection in conjunction with

parameter optimization to consider how the learning algorithm and the dataset interact. First, we composed a feature set based on Stanford CoreNLP, latent Dirichlet allocation, regular expressions, and the ontologies of WordNet and UMLS. Next, the heuristic methods best-first and greedy (hill-climbing) with forward and backward direction have been applied for feature evaluation and selection. In the development phase we observed that 19 out of 41 features performed best in combination with the hyperparameter C=10 of the CRF method on the validation set. Experimental results show that our system outperforms the baseline on the validation set with a macro-averaged F1 of 0.330 using all features as a brute-force approach and 0.386 by performing feature selection. Moreover, we could achieve better results than the baseline (0.246) on the test set with all features resulting in 0.311 and a higher performance using feature selection with 0.382 which corresponds to an improvement of 0.136.

Further improvements could be achieved by more appropriate features from the field of Natural Language Processing. Moreover, the investigation of relevant features using other heuristic methods for wrapper-based feature selection, such as genetic search, would be interesting in order to enhance the system performance.

References

1. Suominen, H., Zhou, L., Hanlen, L., Ferraro, G.: Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics*, 3(2), 2015.
2. Herms, R., Richter, D., Eibl, M., Ritter, M.: Unsupervised language model adaptation using utterance-based web search for clinical speech recognition. *CLEF 2015 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, 2015.
3. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. *CLEF 2016 - 7th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS)*, Springer, September, 2016.
4. Suominen, H., Zhou, L., Goeuriot, L., Kelly, L.: Task 1 of the CLEF eHealth evaluation lab 2016: Handover information extraction. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*, CEUR-WS, September 2016.
5. Hripcsak, G., Austin, J., Alderson, P., Friedman, C.: Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. *Radiology*, 224(1):157–163, 2002.
6. Friedman, C.: Towards a comprehensive medical language processing system: methods and issues. In *Proceedings of the AMIA annual fall symposium*, page 595. American Medical Informatics Association, 1997.
7. Hahn, U., Romacker, M., Schulz, S.: Medsyndikate—a natural language system for the extraction of medical information from findings reports. *International journal of medical informatics*, 67(1):63–74, 2002.
8. Evans, D., Brownlow, N., Hersh, W., Campbell, E. M.: Automating concept identification in the electronic medical record: an experiment in extracting dosage information. In *Proceedings of the AMIA Annual Fall Symposium*, page 388. American Medical Informatics Association, 1996.

9. Friedman, C., Knirsch, C., Shagina, L., Hripcsak, G.: Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. In *Proceedings of the AMIA Symposium*, page 256. American Medical Informatics Association, 1999.
10. Patrick, J., Li, M.: High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527, 2010.
11. Zeng, Q., Goryachev, S., Weiss, S., Sordo, M., Murphy, S., Lazarus, R.: Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1):1, 2006.
12. Chhieng, D., Day, T., Gordon, G., Hicks, J.: Use of natural language programming to extract medication from unstructured electronic medical records. In *AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium*, pages 908–908, 2006.
13. Levin, M., Krol, M., Doshi, A., Reich, D.: Extraction and mapping of drug names from free text to a standardized nomenclature. In *AMIA*, pages 438–42, 2007.
14. Sirohi, E., Peissig,P.: Study of effect of drug lexicons on medication extraction from electronic medical records. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 308–318, 2004.
15. Xu, H., Stenner, S., Doan, S., Johnson, K., Waitman, L., Joshua C Denny. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.
16. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
17. Uzuner, Ö., Solti, I., Cadag, E.: Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.
18. Bundschus, M., Dejori, M., Stetter, M., Tresp,V., Kriegel, H.: Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9(1):1, 2008.
19. Lin, S., Ng, J., Pradhan, S., Shah, J., Pietrobon, R., Kan, M.: Extracting formulaic and free text clinical research articles metadata using conditional random fields. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 90–95. Association for Computational Linguistics, 2010.
20. Li, D., Kipper-Schuler, K., Savova, G.: Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 94–95. Association for Computational Linguistics, 2008.
21. Skeppstedt, M., Kvist, M., Nilsson, G., Dalianis, H.: Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158, 2014.
22. Klinger, R., Friedrich, C.: Feature subset selection in conditional random fields for named entity recognition. In *RANLP*, pages 185–191, 2009.
23. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
24. Kohavi, R., John, G.: Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

25. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. *ACL (System Demonstrations)*, pages 55–60, 2014.
26. Klein, D., Manning, C. D.: Fast Exact Inference with a Factored Model for Natural Language Parsing. *Neural Information Processing Systems 15 (NIPS 2002)*, Cambridge, MA: MIT Press, pages 3–10, 2003.
27. George A. Miller. *WordNet: A Lexical Database for English.* Communications of the ACM Vol. 38, No. 11, pages 39–41, 1995.
28. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.
29. Newman, D., Asuncion, A., Smyth, P. Welling, M.: Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10:1801–1828, 2009.
30. Berthold, M., Cebron, N., Dill, F., Gabriel, T. Kötter,T., Meinl,T., Ohl,P., Thiel,K., Wiswedel,B.: Knime-the konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1):26–31, 2009.
31. McCray, A. T., Burgun, A., Bodenreider, O.: Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(1), page 216, 2001.
32. Bermejo, P., Gámez, J., Puerta, J.: Incremental wrapper-based subset selection with replacement: An advantageous alternative to sequential forward selection. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 367–374. IEEE, 2009.
33. Schenk, J., Kaiser, M., Rigoll, G.: Selecting features in on-line handwritten whiteboard note recognition: Sfs or sffs? In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 1251–1254. IEEE, 2009.
34. Suominen, H., Hanlen, L., Goeuriot, L., Kelly, L., Jones, G. J.: Task 1a of the CLEF eHealth Evaluation Lab 2015. *6th Conference and Labs of the Evaluation Forum*, page 1391, 2015.
35. Kudo, T.: Crf++: Yet another crf toolkit (2005). *Software available at https://taku910.github.io/crfpp/*, 2013.