Improving textual medication extraction using combined conditional random fields and rule-based systems

Domonkos Tikk,^{1,2} Illés Solt¹

► An additional appendix is published online only. To view this file please visit the journal online (<http://jamia.bmj.com>).

¹Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
²Institute for Computer Science, Humboldt-University of Berlin, Berlin, Germany

Correspondence to

Dr Domonkos Tikk, Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar Tudósok krt. 2, Hungary; tikk@tmit.bme.hu

A former version of this work was presented at the Third i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, in San Francisco, California, USA, on November 13, 2009. This current paper is however quite different from the workshop paper, since in the latter no CRF-related analysis and experiments are presented. In addition, neither the workshop papers nor the presentation materials are made available to anyone, except for the workshop participants. Both i2b2's policy and the expired limited data use agreement prevent the workshop materials from being prior art.

Received 23 February 2010
Accepted 30 June 2010

ABSTRACT

Objective In the i2b2 Medication Extraction Challenge, medication names together with details of their administration were to be extracted from medical discharge summaries.

Design The task of the challenge was decomposed into three pipelined components: named entity identification, context-aware filtering and relation extraction. For named entity identification, first a rule-based (RB) method that was used in our overall fifth place-ranked solution at the challenge was investigated. Second, a conditional random fields (CRF) approach is presented for named entity identification (NEI) developed after the completion of the challenge. The CRF models are trained on the 17 ground truth documents, the output of the rule-based NEI component on all documents, a larger but potentially inaccurate training dataset. For both NEI approaches their effect on relation extraction performance was investigated. The filtering and relation extraction components are both rule-based.

Measurements In addition to the official entry level evaluation of the challenge, entity level analysis is also provided.

Results On the test data an entry level F₁-score of 80% was achieved for exact matching and 81% for inexact matching with the RB-NEI component. The CRF produces a significantly weaker result, but CRF outperforms the rule-based model with 81% exact and 82% inexact F₁-score ($p < 0.02$).

Conclusion This study shows that a simple rule-based method is on a par with more complicated machine learners; CRF models can benefit from the addition of the potentially inaccurate training data, when only very few training documents are available. Such training data could be generated using the outputs of rule-based methods.

Biomedical text mining has been a continuously growing field in the past few decades, because it has proved its efficiency in a wide range of application areas, such as the identification of biological entities (MeSH terms, proteins, genes, etc.)^{1,2} and their relationships^{3,4} in free text, assigning insurance codes to clinical records,⁵ facilitating querying in biomedical databases⁶; for a recent survey, see Cohen and Hersch.⁷

Pharmacotherapy information, including patients' responses to medications, is found in textual clinical records, such as discharge summaries. Physicians may be interested in analyzing statistically relevant data or specific cases based on clinical records. For this, such texts have to be processed extracting relevant pieces of information and arranging them into meaningful structures

automatically. These tasks are called information extraction (IE) and relation extraction (RE) in the text mining fields.

The goal of the Informatics for Integrating Biology and the Bedside (i2b2) Medication Extraction Challenge⁸ was to extract from discharge summaries information on medications experienced by the patient. This task—termed medication extraction—is a relational information extraction problem consisting of three subtasks. First, text fragments of different semantic types in free text have to be found; this is called named entity identification (NEI) or tagging. Second, filtering is performed to limit the scope of the RE. Third, in RE the entities within the scope of interest are investigated and determined whether they are in relation or not.

Here we propose a RE pipeline for medication extraction. We briefly discuss the results and lessons learned from our study. For the community, we provide an appendix to this paper (available online only at <http://jamia.bmj.com>) and the source code (available at <http://www.categorizer.tmit.bme.hu/~illes/i2b2/medication>).

BACKGROUND

The paper of the organizers⁸ gives an overview on medication extraction studies, which typically tackle the problem using drug lexicons and rule-based systems,^{9–11} and find these means invaluable.

Named entity identification systems typically apply linguistic grammar-based methods or statistical (probabilistic) models. Grammar-based systems use hand-crafted rules and are therefore costly, while statistical approaches create classification models in a supervised manner requiring preferably large manually annotated training corpora. State-of-the-art statistical NEI methods can achieve approximately 85–90% F₁-score on their application domain.¹² The portability of NEI systems is limited; their performance usually drops significantly when applied to another domain.¹³

Conditional random fields (CRF)^{14,15} are a probabilistic framework applied successfully for NEI, also on the biomedical domains (for a survey, see He and Kayaalp).¹⁶ The CRF-based solutions were among the best systems^{17,18} at the i2b2 de-identification challenge¹⁹ producing 80–99% F₁-measure depending on the entity type. Despite of its effectiveness at entity recognition we found no references using CRF to extract medication information.

METHODS

Problem description

In the medication extraction challenge, a relation, called a medication entry, consists of a medication

name, dosage information, mode of administration, frequency and duration of taking the medication and reason for administration. Except the medication name, all other five constituents are optional. The content of medication entry constituents (termed fields) are specified by a token-level standoff annotation using the simple whitespace tokenization provided. The very same text fragment could belong to several entries. Restrictions apply on the context of medication: allergic reactions and negative statements should be excluded (details in Uzuner and Cadag).⁸

Evaluation of system outputs was done in two different settings: exact and inexact, the former works on phrase, the latter on token level. The training set contained 17 fully annotated discharge summaries (termed 17gt). The development set had 679 unlabelled documents. The test set contained 553 unlabelled documents, among which 251 were selected as the basis of evaluation (termed 251gt). More details are provided in the online version.

Approaches

We decomposed the medication extraction problem into three subtasks. First, named entity identification is applied for the six entity types. Second, a filtering is performed to discard contexts that are out of the challenge scope. Third, the RE is executed, when recognized entities are arranged into medication entries with a medication name head. The components are connected to form a complete medication extraction pipeline (see figure 1).

As the training set is very small, we developed a rule-based NEI (figure 1A) for which we made use of dictionaries and manually created regular expressions. We participated at the challenge using this NEI component in our solution, achieving fifth place among 20 systems.

After the completion of the challenge, we also developed a CRF-based NEI approach and applied it with various settings. As a baseline, we used a CRF model exclusively trained on 17gt

Figure 1 Data flow in the pipeline and in the different named entity identification (NEI) components. (A) Rule-based (RB)—NEI uses manually compiled rules, while conditional random fields (CRF)—NEI is either trained on (B) the small training set with gold standard entity annotations or (C) the entire document set with the potentially inaccurate rule-based NEI annotations.

(figure 1B). Second, we tagged all documents with rule-based NEI, and trained a CRF model on this corpus (termed 1249rb, figure 1C). We performed experiments with two tokenization approaches and three feature sets in the CRF modules.

Filtering is performed by a context-aware rule-based approach. We partitioned discharge summaries into zones, filtered out allergy and family history-related parts. Furthermore, negative statements were discarded.

Relation extraction is again performed by a rule-based model. The text zones are further segmented using identified medication names. With custom rules we determined their scopes, and finally arranged the extracted entities into medication entries.

Rule-based NEI

We made use of two information sources at the creation of rule-based NEI: vocabularies (mainly for medication and partly for reason) and rules defined as regular expressions. The former is self-explaining, while the latter brings the ability to deal with linguistic variations and add context sensitivity through the so-called look-arounds. See more details in the online version.

To support NEI we created a custom grammar that combines the benefits of using vocabularies and regular expression rules (see the online appendix at <http://jamia.bmj.com> and an excerpt in figure 2). The grammar defines entities as named regular expressions that may contain references to other entities. A terminal entity is one without references to other entities.

Vocabularies and regular expressions are encoded as terminal entities, while non-terminal entities are used to combine them with the rich syntax of regular expressions. For example, based on the dosage unit vocabulary and the number regular expression one can define the dosage entity to be ‘{number} {dosage unit}s?’ that is expanded to an ordinary regular expression. Note that defining a regular expression this way is intuitive and facilitates re-use.

The medication name vocabulary was populated from the following sources:

- Public (eg, DrugBank drugcards (<http://www.drugbank.ca/downloads>), Freebase topics of type Drug (<http://www.freebase.com/view/medicine/drug>));

Figure 2 Example of regular expression grammar to match dosage information.

- ▶ Labeled training documents (17gt);
- ▶ Unlabeled development set documents—matches of regular expression patterns that exhibit high precision on labeled documents.

Using multiple sources allowed for better coverage of spelling and abbreviation variations, but all misspellings and shortened medication names could not be covered.

For other fields, vocabularies were first bootstrapped from the training annotations, and then rules were created manually by also observing the development set documents. Altogether, we defined 87 entities in the custom grammar (see source code), including the six top level entities directly used in annotation, 19 domain-specific entities (eg, dosage units, verb phrases used with medications, adjectives of insulin).

Conditional random fields-based NEI

At post-challenge analysis we developed an alternative NEI component using CRF. A CRF is trained by a token sequence when tokens are annotated by single labels (here the six entity types or other). Tokens are represented by their feature vectors. Typical features are binary properties, for example, whether the token matches a pattern. Based on the sequence of labels in training documents and the observed corresponding feature vectors, CRF learns the relations between labels and features, and builds a discriminative model that is applied to predict the most likely label sequence of unlabeled token sequences.

The predictive performance of a CRF model depends heavily on the tokenization and the feature set applied,²⁰ thus we experimented with two types of tokenizations and three feature sets. We used whitespace tokenization (token boundaries are specified by whitespace) and a custom tokenization that preserves abbreviations and unusually spelled floating point numbers as a single token by defining context-dependent token boundaries (details in the online version).

As for feature sets (FSs), first we considered identity FS in which features correspond to unique words (for FS sizes on 17gt see table 1). Standard FS also includes surface, offset conjunction (feature vectors are augmented by features in a two-window with marked relative position) and character n-gram features. The domain dependent FS domain includes vocabularies and typical entity patterns taken from rule-based NEI (see examples in the online version), in addition to the identity FS and surface features. In FS full, all features are included.

Table 1 Evaluation of the performance of NEI components on the 17gt and 251gt sets measured at entity level

NEI approach	No of features	Micro F ₁ on 17gt		Micro F ₁ on 251gt	
		Med name	Overall	Med name	Overall
Rule-based NEI	—	0.935	0.972	0.877	0.934
CRF (whitespace, identity)	4543	0.708	0.944	0.755	0.951
CRF (whitespace, standard)	834 892	0.744	0.944	0.764	0.949
CRF (whitespace, domain)	8511	0.897	0.963	0.865	0.962
CRF (whitespace, full)	943 641	0.816	0.950	0.818	0.955
CRF (custom, identity)	4111	0.732	0.950	0.743	0.954
CRF (custom, standard)	914 936	0.727	0.947	0.735	0.952
CRF (custom, domain)	7399	0.899	0.963	0.843	0.962
CRF (custom, full)	918 224	0.814	0.953	0.802	0.958

At the former, the rule-based named entity identification (NEI) are evaluated on the entire 17gt, while for conditional random fields (CRF) we used 17-fold cross-validation. CRF models evaluated on 251gt were trained on the entire 17gt. The third and fifth rows indicate performance at medication name identification. Note that the overall F₁-micro values are independent from the final performance measures of the medication extraction challenge, as 87.4% of the tokens are not part of an entity, so the constant other-prediction would achieve a 0.874 micro F₁-score. Results of best CRF feature sets emboldened.

We evaluated the various tokenizations and feature sets on the 17 ground truth documents (17gt) using a 17-fold cross-validation.

We integrated the CRF model into the pipeline using two different scenarios (see figure 1).

Filtering

Discharge summaries were first partitioned into larger text segments called zones. Zones consist of a heading and a body. As the documents were made available as unstructured text, we used hand-crafted regular expressions to identify zone headings. Zones that typically contain medication listings (eg, 'discharge medications', 'meds') were marked as list context, while the others as narrative.

According to the annotation guidelines, mentions of medications, which are not taken by or prescribed for the patient, should not be extracted. Such medication mentions can be found in negated context (eg, 'patient did not take X'), in allergy listings (eg, 'allergies: penicillin', 'allergic to aspirin') or even in the family history. We identified context semantics at two levels. At the zone level, we filtered out entire zones with headings referring to allergy or family history (typical headings are 'Allergies:', 'All:;', 'Fam HX:', 'Family history:'). Below the sentence level, we identified allergies, family history, and negation trigger words and their scopes using our i2b2 obesity challenge semantic classifier.²¹ Entities in such scopes were also disregarded in the RE step.

Filtering also handles the rare case of overlapping entities generated by rule-based NEI—for example 'every two days' (frequency) versus 'two days' (duration).

Relation extraction: assembling medication entries

The task of extracting relations between medication names and entities of other types was simplified to determining the scopes of each medication entity. Given a medication entity and its scope, a medication entry was constructed by populating its fields with all constituents found in its scope.

To determine the scopes, zone bodies were further split into sentences (narrative context) or list items (list context). If a sentence contained exactly one medication name, then the context was assumed to be the whole sentence. If a sentence contained several medication names, then it was initially partitioned using the medication names as the start of scope boundaries. The partitioning was then fine tuned also taking into account that medication entries may have constituents preceding the medication name (like in 'IV Lasix'). Fine tuning extended the scopes of medication names to the left until a scope terminating pattern is matched or another medication name is found (an example is depicted in figure 3). For details on scope terminators, see Solt *et al.*²¹

Implementation

The pipeline was implemented within the generic UIMA (<http://www.uima.apache.org/>) framework that proved to be well-suited

She was treated for CHF with digoxin, Ace inhibitor, and IV Lasix and she continued to get IV Lasix b.i.d. 40 mg	She was treated for CHF with digoxin, Ace inhibitor and IV Lasix and she
---	--

Figure 3 Fine tuning of the scope detection. Medication entities are marked with solid boxes, other entity types with dashed boxes, and new lines indicate partition boundaries. On the left a sentence is shown in which medication entries are used as scope terminators. Using fine additional rules, the partition boundaries are moved as shown on the right.

Table 2 Entity and entry level precision, recall and F₁-micro results achieved on 251gt using the rule-based NEI component (official evaluation) and the best performing CRF–NEI

NEI	Fields	Ratio of tokens	Inexact			Exact		
			P	R	F ₁	P	R	F ₁
Rule-based NEI	Medication	4.25%	0.9199	0.8225	0.8685	0.8729	0.8250	0.8482
	Dosage	2.59%	0.8998	0.8372	0.8674	0.8956	0.7687	0.8274
	Mode	1.01%	0.9235	0.853	0.8869	0.9233	0.8398	0.8796
	Frequency	1.99%	0.8414	0.8322	0.8368	0.8434	0.7853	0.8133
	Duration	0.63%	0.5489	0.3954	0.4597	0.4107	0.3764	0.3928
	Reason	0.88%	0.4590	0.2899	0.3554	0.3716	0.2555	0.3028
	Overall	11.35%	0.8580	0.7623	0.8073	0.8408	0.7580	0.7972
CRF (custom, domain) trained on 1249rb	Medication		0.8793	0.8144	0.8456	0.9336	0.8156	0.8706
	Dosage		0.9216	0.7757	0.8424	0.9263	0.8273	0.8740
	Mode		0.9397	0.8289	0.8808	0.9404	0.8369	0.8856
	Frequency		0.8866	0.8004	0.8413	0.8969	0.8143	0.8536
	Duration		0.5597	0.3236	0.4101	0.6981	0.3056	0.4251
	Reason		0.6186	0.1705	0.2674	0.6817	0.1435	0.2371
	Overall		0.8868	0.749	0.8121	0.9137	0.7359	0.8152

The ratio of tokens is measured on the 251gt (whitespace tokenization). Note that performance on the medication field is not directly comparable with the corresponding numbers in table 1, as there is no one-to-one correspondence between entities and entries (the same occurrence of medication name can be in multiple entries), and precision is increased here, because filtering disregards entities occurring in improper contexts. CRF, conditional random field; NEI, named entity identification.

for this information extraction task. First, the annotation schema was created. Second, conversion between i2b2 standoff format and the UIMA representation was developed. Third, the workflow of our approach was broken down into hierarchically structured components (so-called annotators). For CRF, we integrated MALLET,²² a natural language processing-focused Java machine learning library into the workflow. For further details, see the online version and source code.

RESULTS

Named entity identification components

Table 1 presents entity level evaluation results of NEI components. Recall that we used solely the rule-based NEI component at the challenge. Here we report on the medication name and overall (six fields plus other) micro F₁-score. The former is an indicative upper bound for the performance of the extraction task, but potentially includes false positives (eg, in negative context). The latter numbers indicate the relative performance of different CRF models and help in identifying the best CRF settings, but are inadequate to estimate the RE performance.

Relation extraction

Table 2 includes the inexact and exact F₁-scores on 251gt of our shared task submission and post-challenge CRF components. In table 3 we report on the performance of CRF models trained on 17gt and 1249rb, measured on the 251gt. Observe that CRF trained on the 17gt brings somewhat inferior performance

compared with the rule-based solution. As expected, the domain FS achieves significantly better results compared with identity and standard FSs (8–10% F₁-score gain). Such differences cannot be observed between types of tokenization.

Models trained on the 1249rb outperform 17gt trained CRF. Interestingly, here the tokenization makes a difference: with whitespace only the performance of the rule-based approach is reached, while custom CRF outperforms that with 0.7 (inexact) and 1.9 (exact) at F₁-score.

DISCUSSION

Table 1 shows that the rule-based approach is more inclined for overfitting than CRF at the NEI components. As expected, the RE achieved better overall performance using the rule-based NEI component when compared with the CRF–NEI component trained only on 17gt, which can be explained by an insufficient number of training documents. When, however, CRF–NEI was trained on 1249rb, that is additional, although potentially inaccurate training data were available, it already outperformed the rule-based solution ($p < 0.02$).

We achieved best results with custom tokenization and domain feature set. Offset conjunction and n-gram features added only noise to training, because CRF could not retrieve useful information from them, due to insufficient training data.

The F₁-score on medication names is an upper bound of the overall RE F₁-score. Therefore, better performance in medication identification would also increase the performance measured on

Table 3 Entry level precision, recall and F₁-micro results achieved on 251gt with rule based and CRF-based solutions

Method	Training set	Inexact			Exact		
		P	R	F ₁	P	R	F ₁
Rule based (challenge submission)	—	0.8580	0.7623	0.8073	0.8408	0.7580	0.7972
CRF (whitespace, identity)	17gt	0.8951	0.5606	0.6894	0.8550	0.5461	0.6665
CRF (whitespace, standard)	17gt	0.9522	0.5418	0.6906	0.925	0.5539	0.6929
CRF (whitespace, domain)	17gt	0.8999	0.6765	0.7724	0.8732	0.6802	0.7647
CRF (custom, identity)	17gt	0.8794	0.5699	0.6916	0.8496	0.5607	0.6755
CRF (custom, standard)	17gt	0.9443	0.5493	0.6946	0.9203	0.5628	0.6985
CRF (custom, domain)	17gt	0.8822	0.6832	0.7701	0.8619	0.6753	0.7573
CRF (whitespace, domain)	1249rb	0.8616	0.7594	0.8073	0.8339	0.7567	0.7934
CRF (custom, domain)	1249rb	0.9137	0.7359	0.8152	0.8868	0.749	0.8121

CRF, conditional random fields.

other fields as only entities being constituents of entries were evaluated.

We remark that duration and reason were the most difficult fields to identify properly; these were also the least frequent ones (see table 2) with the larger average token/phrase ratio (table 1 in).⁸ In addition, duration exhibited a large variety of surface forms, while reason was often located outside the sentence boundary, and was thus disregarded by our sentence scope-based algorithm.

CONCLUSION

In this paper we reported on our approach for the i2b2 medication extraction challenge. We developed a context-aware rule-based pipeline that identified medication entries in three steps: named entity identification, filtering, and RE using scope detection. For the NEI component we presented results for the rule-based approach used in the contest and various CRF-based approaches developed after completion of the challenge.

We showed that the standard CRF-based approach did not improve upon the rule-based approach due to the limited amount of training data. However, when additional, although potentially inaccurate, training data were made available, applying CRF resulted in a generalized model that showed considerably better performance.

In this exercise, model creation using the CRF was easier, because feature definition was more straightforward and less error-sensitive than rule creation. On the other hand, rule-based methods are easier to comprehend and lack the lengthy training phase, which supports iterative trial-and-error development. Performance improvement can be expected with less effort with CRF by adding more features, compared with the cumbersome rule fine-tuning.

Acknowledgments The authors would like to thank the i2b2 organizers and affiliates for providing them with this research experience; and other shared task participants for developing the valuable ground truth dataset. They also wish to thank Roman Klinger for his invaluable advice on training CRF models.

Funding DT was supported by the Humboldt Foundation. The project described was supported in part by the i2b2 initiative, Award Number U54LM008748 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Cooper GF**, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. *J Am Med Inform Assoc* 1998;**5**:62–75.

2. **Krallinger M**, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. *Drug Discov Today* 2005;**10**:439–45.
3. **Hakenberg J**, Plake C, Royer L, et al. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol* 2008;**9**:S14.
4. **Zhou D**, He Y. Extracting interactions between proteins from the literature. *J Biomed Inform* 2008;**41**:393–407.
5. **Farkas R**, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics* 2008;**9**(S3):S10.
6. **Plake C**, Schiemann T, Pankalla M, et al. AliBaba: PubMed as a graph. *Bioinformatics* 2006;**22**:2444–5.
7. **Cohen AM**, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;**6**:57–71.
8. **Uzuner Ö**, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514–8.
9. **Evans DA**, Brownlow ND, Hersh WR, et al. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proc AMIA Annu Fall Symp* 1996:388–92.
10. **Gold S**, Elhadad N, Zhu X, et al. Extracting structured medication event information from discharge summaries. *Proc AMIA Annu Fall Symp* 2008:237–41.
11. **Jagannathan V**, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *Int J Med Inform* 2009;**78**:284–91.
12. **Finkel JR**, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. *ACL 2005: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 25–30 June 2005; Ann Arbor, MI, USA; The Association for Computer Linguistics. 2005:363–70.
13. **Poibeau T**, Kosem L. *Proper name extraction from non-journalistic texts. CLIN 2001: Selected papers from the Eleventh Computational Linguistics in the Netherlands Meeting*. 3 November 2000; Tilburg, The Netherlands: Rodopi, 2000:144–57.
14. **Lafferty JD**, McCallum A, Pereira FCN. *Conditional random fields: probabilistic models for segmenting and labeling sequence data. ICML 2001: Proceedings of the 18th International Conference on Machine Learning*; 28 June–1 July 2001. Williamstown, MA, USA. San Francisco: Morgan Kaufmann, 2001:282–9.
15. **McDonald R**, Pereira FCN. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* 2005;**6**(Suppl 1):S6.
16. **He Y**, Kayaalp M. Biological entity recognition with conditional random fields. *AMIA Annu Symp Proc* 2008:293–7.
17. **Aramaki E**, Miyo K. Automatic deidentification by using sentence features and label consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006. http://www.m.u-tokyo.ac.jp/medinfo/ont/paper/2006_i2b2-deid.pdf (accessed 24 Jul 2010).
18. **Wellner B**, Huyck M, Mardis S, et al. Rapidly retargetable approaches to de-identification in medical records. *J Am Med Inform Assoc* 2007;**14**:564–73.
19. **Uzuner Ö**, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic deidentification. *J Am Med Inform Assoc* 2007;**14**:550–63.
20. **McCallum AK**. Efficiently inducing features of conditional random fields. *UAI 2003: Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*; 7–10 August 2003. Acapulco, Mexico; San Francisco: Morgan Kaufmann, 2003:403–10.
21. **Solti I**, Tikk D, Gál V, et al. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *J Am Med Inform Assoc* 2009;**16**:580–4. doi:10.1197/jamia.M3087.
22. **McCallum AK**. MALLLET: A machine learning for language toolkit. 2002. <http://www.mallet.cs.umass.edu> (accessed 24 Jul 2010).