Biomedical entity extraction using machine-learning based approaches

Cyril Grouin
CNRS, UPR 3251, LIMSI
91403 Orsay, France
cyril.grouin@limsi.fr

Abstract

In this paper, we present the experiments we made to process entities from the biomedical domain. Depending on the task to process, we used two distinct supervised machine-learning techniques: Conditional Random Fields to perform both named entity identification and classification, and Maximum Entropy to classify given entities. Machine-learning approaches outperformed knowledge-based techniques on categories where sufficient annotated data was available. We showed that the use of external features (unsupervised clusters, information from ontology and taxonomy) improved the results significantly.

Keywords: Conditional Random Fields; Named Entity Recognition; Natural Language Processing

1. Introduction

1.1. Presentation

Scientific documents provide useful information in many domains. Because processing those documents is time-consuming for a human, NLP techniques have been designed to process a huge amount of documents quickly. In the biological domain, the availability of the GENIA corpus[1] (Kim et al., 2003), a huge corpus of 100,000 annotated terms from the biological domain out of a total of more than 400,000 terms, led many teams to design NLP-based approaches in order to extract the knowledge from the documents (Nadeau and Sekine, 2007).

Biologists need automatic methods to mine documents with scientific information in order to leverage knowledge from the scientific literature or from knowledge databases *(genome databases, pharmacology patents)*, especially to study bacteria and genome. The main issue for biologists is to discover correlations between distinct kinds of information in order to confirm or disgard a given hypothesis.

Over the past few years, an increasing number of NLP evaluation campaigns have been organized on the biomedical domain, in order to design systems that allow scientists to automatically access document content *(i2b2/VA,*[2] *BioNLP Shared-Task,*[3] *ShARe/CLEF eHealth,*[4] *Drug-Drug Interaction*[5]*)*. In each challenge, the first part of the work usually consisted in the identification of entity mentions (drug names, medical problem, test) before performing additional analysis: interactions between drug names (Segura-Bedmar et al., 2011), assertion and relationships between entities (Uzuner et al., 2011), coreference resolution (Uzuner et al., 2012), temporal links between entities (Sun et al., 2013), etc. The success of entity identification is critical for the success of the next steps.

[1] http://www.nactem.ac.uk/genia/
[2] http://www.i2b2.org/NLP/
[3] http://2013.bionlp-st.org/
[4] https://sites.google.com/site/shareclefehealth/
[5] http://labda.inf.uc3m.es/DDIExtraction2011/

1.2. Motivations

Herein, we describe the methods we used to extract knowledge from texts in biology and pharmacology, while participating in two challenges, with main objective to rapidly access the information. Our experiments rely on two machine-learning approaches: Conditional Random Fields (Lafferty et al., 2001; Sutton and McCallum, 2006) to process sequences labeling and Maximum Entropy (Guiasu and Shenitzer, 1985; Berger et al., 1996) to perform categorization on already identified mentions.

The first set of experiments consists of the identification and the categorization of bacteria and biotope mentions from scientific abstracts. The main part of this work has been done within the framework of the BioNLP Bacteria/Biotope Challenge (Grouin, 2013), with additional experiments since it occurred. The second set of experiments extends this work on a corpus of pharmacology patents with new categories. Those new experiments rely on the categorization of already identified entities *(i.e., entity frontiers were provided)* in pharmacology patents, among twelve categories of entity which are similar to semantic types from the UMLS (Lindberg et al., 1993).

2. Related Work

In 2010, the i2b2/VA NLP challenge focused on the processing of documents from the medical domain (Uzuner et al., 2011). Three tasks were proposed: (i) extraction of entities among three categories *(problem, test, treatment)*, (ii) identification of entities assertion *(present, absent, possible, hypothetical, etc.)*, and (iii) identification of relations between those entities. The participants that achieved the best results on the entity extraction task used semi-supervised or hybrid CRF-based approaches (de Bruijn et al., 2010; Jiang et al., 2011). Bacteria and biotopes identification has been addressed during the BioNLP 2011 Bacteria Biotopes shared-task (Bossy et al., 2012; Kim et al., 2011) and consisted in extracting bacteria location events from texts among eight categories *(Host, HostPart, Geographical, Environment, Food, Medical, Water and Soil)*. The 2013 edition focused on the extraction of entities among three categories *(Bacteria, Biotope, Geographical name)* (Bossy et al., 2013; Nédellec et al., 2013).

3. Material and Methods

3.1. Corpora

3.1.1. Bacteria/Biotope corpus

The corpus used during the challenge comprises a total number of 131 web pages about bacterial species written for non-experts *(description of individual bacterium and groups of bacteria, first observation, characteristics, evolution and biotopes)*. This corpus includes both raw textual documents—without any tokenization performed over texts—and external reference annotations. Table 1 shows a few statistics on the annotations obtained over training and development corpora[6] for each type of entity to be annotated *(bacteria, habitat,* and *geographical)*.

Corpus	Training	Development
# Documents	52	26
# Words	16,294	9,534
Average # words/doc	313.3	366.7
# Bacteria	832	515
# Habitat	934	611
# Geographical	91	77

Table 1: Statistics on the bacteria/biotope corpus

Besides challenge data, we used an additional corpus of 1,884 textual documents[7] from the same sources as those used in the BioNLP 2013 shared task corpus. No external reference annotations have been made on those documents. We used this new corpus of 2,015 documents to build unsupervised clusters of words (see section 3.2.).

3.1.2. Pharmacology patents corpus

The corpus is composed of web pages of pharmacology patents valid for Europe. Each patent includes a detailed description of the invention in English, and a shorter description written in German and in French. The pharmacology entities are annotated within the English part of each document with embedded tags, i.e., entity frontiers are provided and an entity categorization must be performed.

Table 2 shows a few statistics on both training and development corpora, with the number of annotations per category from the gold standard annotations in decreasing order. According to those statistics, the number of annotations per category is clearly unbalanced: 50% of entities are, either of the category "Treatment_Composition_Active substance" or the category "Treatment_Final product_Pharmacology form" and four categories of entities include each one less than 2% of all entities. Figure 1 shows an extract from the development corpus with annotations of entity and relations.

[6]The reference annotations for the test corpus have not been released by the organizers to the participants.

[7]We obtained these documents as part of Quaero program, a research project in which both the Bacteria/Biotope shared task organizers and us are involved.

Corpus	Train	Dev
# Documents	52	15
# Words	22,259	4,615
Average # words/doc	428.1	307.6
# Entity	16,918	4,136
Average # entity/doc	325.3	275.7
Treatment_Composition_Active substance	6,538 (38.6%)	1,611 (39.0%)
Treatment_Final product_Pharmacology form	1,947 (11.5%)	459 (11.1%)
Treatment_Target_Organ	1,659 (9.8%)	413 (10.0%)
Pathology_Illness	1,562 (9.2%)	361 (8.7%)
Treatment_Target_Population	1,295 (7.6%)	341 (8.2%)
Treatment_Final product _Medical device	1,247 (7.4%)	315 (7.6%)
Treatment_Pharmacology action	1,242 (7.3%)	294 (7.1%)
Treatment_Administration_Mode	751 (4.4%)	168 (4.1%)
Pathology_Sign or Symptom	261 (1.5%)	61 (1.5%)
Treatment_Administration _Posology	251 (1.5%)	85 (2.1%)
Treatment_Final product_Device name	104 (0.6%)	19 (0.5%)
Treatment_Final product_Drug name	61 (0.4%)	9 (0.2%)

Table 2: Statistics on the corpora of pharmacology patents

3.2. Methods

3.2.1. Entity identification and categorization

Resources. We used external resources to build our models. First, we used OntoBiotope,[8] an ontology tailored for the biotopes domain that includes 1,756 concepts.
Second, we built a list of 357,387 bacteria taxa based on the NCBI taxonomy database[9] (Federhen, 2012) so as to help our system to identify the bacteria names. This taxonomy includes twelve categories of entities from the biological domain.[10] From this taxonomy, we extracted all names belonging to the *Bacteria* category (24.3% of the content).
Third, we used lexical annotations produced by the Cocoa[11] tool (Ramanan and Nathan, 2013). These annotations emphasize 37 pre-defined categories, mainly from the molecular biology domain.

[8]`http://bibliome.jouy.inra.fr/MEM-OntoBiotope/OntoBiotope_BioNLP-ST13.obo`

[9]`http://www.ncbi.nlm.nih.gov/taxonomy/`

[10]Bacteria, invertebrates, mammals, phages, plants, primates, rodents, synthetics, unassigned, viruses, vertebrates and environmental samples.

[11]Compact cover annotator for biological noun phrases, `http://npjoint.com/annotate.php`

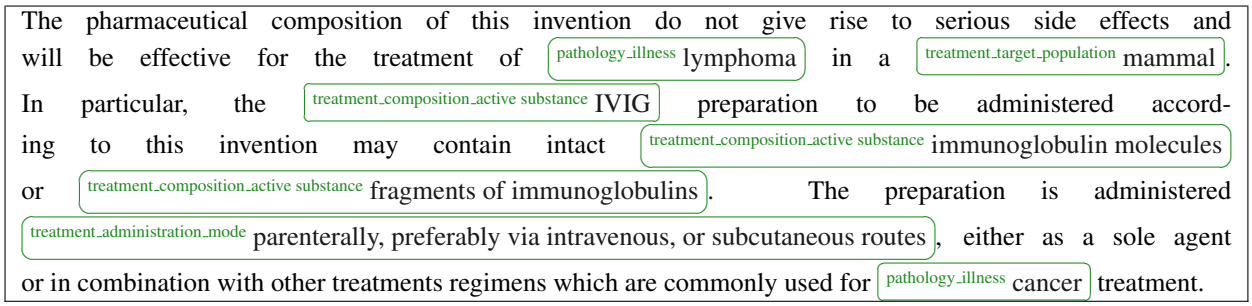

Figure 1: Extract from the corpus of pharmacology patents. Expected answers are inside green boxes

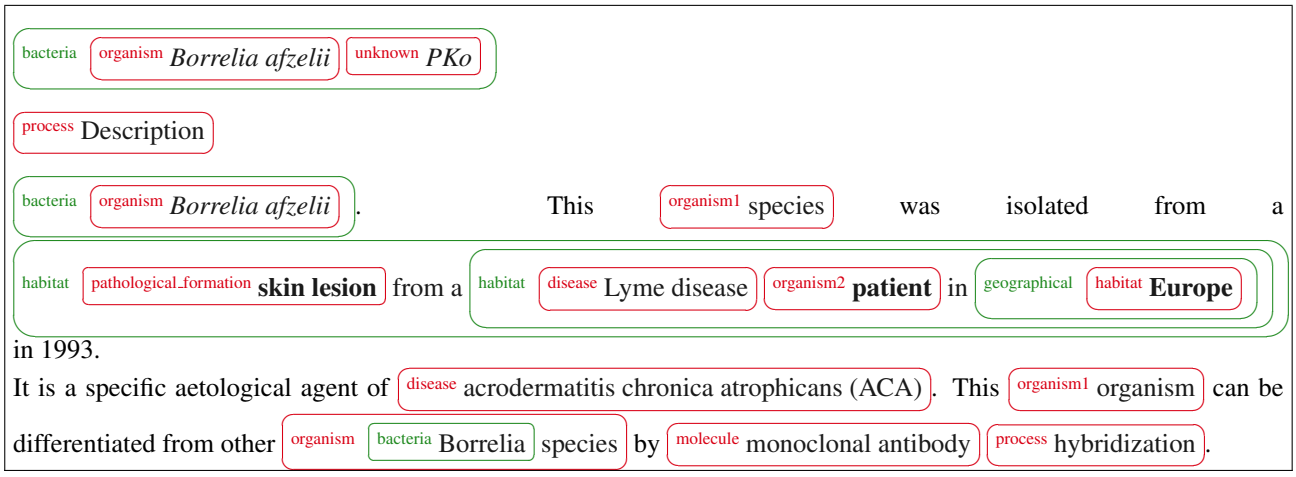

Figure 2: Annotated extract from the Bacteria Biotope corpus. Reference annotations are inside green boxes (bacteria, habitat, geographical), Cocoa annotations are inside red boxes. Entities found in the NCBI taxonomy are in *italic*, entities found in the MBTO ontology are in **bold font**

Last, we also used part-of-speech sequence annotations provided by the BioYaTeA tool (Golik et al., 2013), a term extractor designed to process data from the biology domain based on YaTeA tool (Aubin and Hamon, 2006). We did not used BioYaTeA annotations when participating to the challenge.

We completed those resources producing unsupervised clusters using the Brown's algorithm (Brown et al., 1992) as implemented in Liang's tool[12] (Liang, 2005). We performed clustering on the whole corpus of 2,015 documents, producing a total amount of 120 classes of tokens, based on the tokens occurring at least two times in the corpus.

Formalism. As we have to identify first the entity in the text, i.e., to determine whether a word or a phrase is an entity or not, and second which kind of entity the identified mention is, we used the CRF formalism (Lafferty et al., 2001; Sutton and McCallum, 2006) as implemented in the Wapiti toolkit (Lavergne et al., 2010).

Features. We used both surface, lexical and external features to build our model:

- Token from the text;
- Surface feature: capitalization of the token, presence of digit or punctuation mark in the token;
- Lexical features:
  - presence of the token in the NCBI taxonomy;
  - presence of the token in the OntoBiotope ontology;
  - category of the token based on the Cocoa annotations;
  - the cluster ID of each token from the Brown clusterization;
  - the longest sequence of part-of-speech tags in which the token occurs, based on the BioYaTeA annotations.

Figure 2 shows an annotated extract from the bacteria biotope corpus. We represented inside green boxes the reference annotations, and inside red boxes the Cocoa annotations. We observed a strong correlation between some Cocoa annotations and the expected answer. Cocoa annotations were used to train the CRF model.

3.2.2. Term categorization

Formalism. As entity frontiers were provided, we used the Maximum Entropy (Guiasu and Shenitzer, 1985; Berger et al., 1996) formalism from the Wapiti toolkit to build our models. We used the following features to build our model:

- Terms: both the whole entity and each token from this entity;

[12] http://www.cs.berkeley.edu/~pliang/software/

#	Features	Expected answer
9	entity=Theophylline token=Theophylline cap=Mm local=Mm sty=Traitement_Composition_SubstanceActive	Treatment_Composition _ActiveSubstance
21	entity=17β-estradiol token=17β-estradiol cap=O local=O digit=DIG brown=11111001111	Treatment_Composition _ActiveSubstance
18	entity=powder_form token=powder token=form cap=mm local=mm local=mm brown=111101111001 brown=111100111101 key=T_FormPharma	Treatment_FinalProduct _PharmaForm
168	entity=target_cell token=target token=cell cap=mm local=mm local=mm brown=111110000111 brown=11111001010 sty=Treatment_Target_Organ	Treatment_Target_Organ

Figure 3: Extract from the train file

- Surface feature:
  - capitalization of the whole entity and of each token among four schemas *(all in upper case, all in lower case, combination of upper and lower case, not relevant)*;
  - presence of digit within the entity;
  - presence of key-concept in the entity name: this would be beneficial for the categorization *("form" in "powder form" is a clue for the "Treatment_FinalProduct_PharmaForm" category)*;
  - presence of special symbols within the name which are useful to detect drug name or device name: copyright "©", registered mark "®", and trade mark "™".
- External feature:
  - the cluster ID of each token based on a Brown clusterization performed on the whole corpus;
  - the semantic type of the whole entity and of each token from the UMLS (Lindberg et al., 1993), based on the UMLS file provided by the organizers (i.e., only the term and its semantic type, no CUI).

Figure 3 presents an extract from the file used to build the model. As an example, the entity "power_form" (#18) is composed of two tokens "powder" and "form"; both the whole entity and each token are capitalized in lower case *(cap=mm, local=mm, local=mm)*; the Brown cluster ID for the first token is "111101111001" and the ID for the second token is "111100111101"; and the token "form" is a key-concept for the expected answer "Treatment_FinalProduct_PharmaForm".

Experiments. We conducted four experiments on this task:

1. The first experiment is the result of the MaxEnt model;
2. The second experiment relies on post-processing of the output from the MaxEnt model: for each unpredicted category, if a key-concept suggested the category should be modified, we changed it accordingly, else we gave the mostly used category: "Treatment_Composition_ActiveSubstance";
3. The third experiment is similar to the first experiment, except we did not use information from the UMLS thesaurus;
4. The last experiment is only based on the thesaurus; it consists in applying the UMLS list provided by the organizers, either on the full corpus or on the test corpus.

4. Results

4.1. Entity identification and categorization

Table 3 shows the detailed results we achieved on the development set (in terms of recall, precision and F-measure) from the bacteria biotopes entity identification and categorization task. The results we achieved during the challenge are on the upper line while the results we obtained with the additional experiments are on the lower line.

Category	Recall	Precision	F-measure
Bacteria	0.8794	**0.9397**	**0.9085**
	0.8872	0.9249	0.9057
Geographical	0.6533	0.7903	0.7153
	0.6933	**0.8387**	**0.7591**
Habitat	0.6951	0.8102	0.7482
	0.7197	**0.8444**	**0.7771**
Overall	0.7771	0.8715	0.8216
	0.7950	**0.8836**	**0.8369**

Table 3: Results from the challenge (upper line) and from the additional experiments (lower line) on the bacteria biotopes identification (development corpus). The best results appear in bold font

4.2. Term categorization

Table 4 shows the results our system achieved on the development set for term categorization. For the primary experiment (#1), results are shown in terms of recall, precision and F-measure for each category and overall. This experiment corresponds to the output of the MaxEnt model out of the box. For the other two experiments, we only provide F-measure.

The second experiment (#2) consists in post-processing the previous output to deal with unpredicted categories while the third experiment (#3) did not make use of information from the UMLS thesaurus. The evaluation was performed using the Wapiti toolkit on the predicted categories.

Experiment	#1 (primary)			#2	#3
Category	R	P	F	F	F
Pathology_Illness	0.745	0.891	0.811	**0.830**	0.734
Pathology_Sign or Symptom	0.738	0.763	**0.750**	0.510	0.374
Treatment_Pharmacology action	0.684	0.827	0.749	0.374	**0.808**
Treatment_Administration_Posology	0.847	0.632	0.724	**0.873**	0.832
Treatment_Administration_Mode	0.958	0.856	**0.904**	0.900	**0.904**
Treatment_Target_Organ	0.927	0.916	0.921	**0.922**	0.858
Treatment_Target_Population	0.988	0.921	0.953	**0.959**	0.862
Treatment_Composition_Active substance	0.939	0.873	**0.905**	0.880	0.885
Treatment_Final product_Pharmacology form	0.874	0.907	0.890	0.886	**0.914**
Treatment_Final product_Medical device	0.765	0.941	0.844	**0.854**	0.568
Treatment_Final product_Device name	0.263	0.333	**0.294**	**0.294**	0.250
Treatment_Final product_Drug name	0.111	0.500	0.182	0.182	0.182
Overall	0.804	0.851	**0.827**	0.798	0.784

Table 4: Term categorization evaluation on the development corpus depending on the experiment. Bold font stands for the best result

Table 5 shows the results on the test set.

#	Experiment	F-measure
1	MaxEnt (primary)	0.7694
2	MaxEnt + post-processing	0.7545
3	MaxEnt w/o UMLS	0.7368
4a	Thesaurus only (on the full corpus)	0.6621
4b	Thesaurus only (on the test set)	0.6366

Table 5: Results on the test set

5. Discussion

5.1. Entity identification and categorization

The results on the bacteria biotope corpus show that our CRF-based system succeed to correctly identify bacteria mentions (F=0.906). Nevertheless, the biotope entities are more difficult to process than the bacteria entities (*geographical* F=0.759, *habitat* F=0.777). A similar observation has been made on the 2011 BioNLP Bacteria Biotope shared-task for all participants.

The use of clusters of tokens produced on a huge corpus and annotations of part-of-speech sequence provided by BioYaTeA slightly improved the results. In comparison with the results we achieved during the BioNLP challenge (Grouin, 2013), our overall F-measure increased from 0.8216 to 0.8369. In details, the use of BioYaTeA and the new clustering increased the results for both *geographical* (+4.38 points) and *habitat* (+2.89 points) categories for which the gain is well balanced between precision and recall. Nevertheless, global results are slightly lower for *bacteria* (-0.28 point); the recall increased while the precision decreased. While precision is good on the whole, there is still room for improvement, especially in order to improve the recall. For *geographical* and *habitat* categories, post-processing rules would be beneficial.

5.2. Term categorization

On both development and test set, our best results were achieved on the primary experiment *(i.e., the MaxEnt model applied as it is without any post-processing task).* Our MaxEnt model obtained higher precision than recall on 8 categories out of 12. Higher recall was obtained for the three most frequent categories in the corpus (see Table 2).

While using post-processing improved our results on a few categories *(especially on "Pathology illness" from F=0.7816 on the first experiment to F=0.7955 on the second one)*, the overall results are lower in the second experiment. Performance is lower in the third experiment. Applying the thesaurus without any training method on the corpus did not improve the results (fourth experiment).

The pharmacology patents challenge was restricted to partners from a research project. As we were the sole participant, we can not compare our results with others. We assume our results would constitute a baseline.

6. Conclusions

In this paper, we presented the two methods we used to process entities from the biomedical domain. We used CRF to both identify and categorize entities among three categories *(bacteria, biotopes, geographical names)*, while we used a Maximum Entropy approach to only categorize entities on corpora of pharmacology patents.

To detect bacteria and biotopes names, we used a machine-learning approach based on CRFs. We used several resources to build the model, among them the NCBI taxonomy, the OntoBiotope ontology, the Cocoa annotations, and unsupervised clusters created through Brown's algorithm. This formalism and those external resources proved to be relevant to process both identification and categorization of entities from the biomedical domain (F=0.8369).

We achieved our best results on the term categorization experiment, using a Maximum Entropy-based approach based on knowledge-based resources (F=0.7694); post-processing, dit not help (F=0.7545); we also noticed that using external semantic resources—here, the UMLS—is beneficial (F=0.7368 when not using those resources). Last, only using an existing thesaurus is not sufficient to handle term categorization (F=0.6366).

7. Acknowledgments

This work was supported by Quaero Programme, funded by Oseo, French State agency for innovation.

8. References

Aubin, S. and Hamon, T. (2006). Improving term extraction with terminological resources. In Saloski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Proc of FinTAL*, pages 380–7, Turku, Finland.

Berger, A. L., Della Pietra, S., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bossy, R., Jourde, J., Manine, A.-P., Veber, P., Alphonse, E., van de Guchte, M., Bessières, P., and Nédellec, C. (2012). BioNLP shared task – the bacteria track. *BMC Bioinformatics*, 13(Suppl 11):S3.

Bossy, R., Golik, W., Ratkovic, Z., Bessières, P., and Nédellec, C. (2013). Bionlp shared task 2013 – an overview of the bacteria biotope task. In *Proc of BioNLP Shared Task Workshop*, pages 161–169, Sofia, Bulgaria. Association for Computational Linguistics.

Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79.

de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J. D., and Zhu, X. (2010). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc*, 18(5):557–62.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Res*, 40(Database issue):D136–43.

Golik, W., Bossy, R., Ratkovic, Z., and Nédellec, C. (2013). Improving term extraction with linguistic analysis in the biomedical domain. In *Proc of CICLing*, Samos, Greece. Special Issue of the journal Research in Computing Science.

Grouin, C. (2013). Building a contrasting taxa extractor for relation identification from assertions: Biological taxonomy & ontology phrase extraction system. In *Proc of BioNLP Shared Task Workshop*, pages 144–152, Sofia, Bulgaria. Association for Computational Linguistics.

Guiasu, S. and Shenitzer, A. (1985). The principle of maximum entropy. *The Mathematical Intelligence*, 7(1).

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. *J Am Med Inform Assoc*, 18(5):601–6.

Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):180–2.

Kim, J.-D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N., and Tsujii, J. (2011). Overview of BioNLP shared task 2011. In *BioNLP Shared Task 2011 Workshop Proc*, pages 1–6, Portland, OR. ACL.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. *Proc of ACL*, pages 504–13, July.

Liang, P. (2005). Semi-supervised learning for natural language. Master's thesis, MIT.

Lindberg, D. A., Humphreys, B. L., and McRay, A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, 32(4):281–91.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes*, 30(1):3–26.

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-J., Ohta, T., Pyysalo, S., and Zweigenbaum, P. (2013). Overview of bionlp shared task 2013. In *Proc of BioNLP Shared Task Workshop*, pages 1–7, Sofia, Bulgaria. Association for Computational Linguistics.

Ramanan, S. and Nathan, P. S. (2013). Performance and limitations of the linguistically motivated Cocoa/Peaberry system in a broad biomedical domain. In *Proc of BioNLP Shared Task Workshop*, pages 86–93, Sofia, Bulgaria. Association for Computational Linguistics.

Segura-Bedmar, I., Martinez, P., and Sánchez-Cisneros, D. (2011). The 1st DDIExtraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. In *Proc of Drug-Drug Interaction Extraction Challenge*, pages 1–9.

Sun, W., Rumshisky, A., and Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc*, 20(5).

Sutton, C. and McCallum, A. (2006). An introduction to Conditional Random Fields for relational learning. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*. MIT Press.

Uzuner, O., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–6.

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J. P., and South, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*, 19(5):786–91.