Semantic Annotation of UMLS using Conditional Random Fields

Shahad Kudama and Rafael Berlanga
Universitat Jaume I, Av. de Vicent Sos Baynat, s/n 12071, Castelló, Spain

Keywords: UMLS, CRF, Semantic Annotation, Biomedical Domain.

Abstract: In this work, we present a first approximation to the semantic annotation of Unified Medical Language System (UMLS®) concept descriptions based on the extraction of relevant linguistic features and its use in conditional random fields (CRF) to classify them at the different semantic groups provided by UMLS. Experiments have been carried out over the whole set of concepts of UMLS (more than 1 million). The precision scores obtained in the global system evaluation are high, between 70% and 80% approximately, depending on the percentage of semantic information provided as input. Regarding results by semantic group, the precision even reaches the 100% in those groups with highest representation in the selected descriptions of UMLS.

1 INTRODUCTION

As the biomedical literature continuously increases on the Web, a new important need is growing too: tools and algorithms to perform effective natural language processing to assist researchers in organizing, curating and retrieving all the information (Settles, 2004). To achieve this goal, the task of identifying words and phrases in free text that belong to certain classes of interest, which is known as named entity recognition (NER), is a crucial first step for many of these larger information management goals.

In recent years, much attention has been focused on the problem of recognizing different mentions in biomedical abstracts to classify them into different groups. This paper presents a framework for recognizing occurrences of different types (anatomic parts, chemical products, procedures, disorders, devices, etc.) using Conditional Random Fields with a variety of features. So, in this paper, we firstly introduce the concept of conditional random fields (CRF) and then, apply them to the set of sentences of the Unified Medical Language System (UMLS) to obtain the semantic annotation of unclassified words in one of the predefined semantic groups. As a result, different terms in the UMLS will be recognized and classified in groups with a high precision.

As in this work, we wish to predict a large number of variables that depend on each other as well as on other observed variables, we have chosen CRF as it provides good results in this kind of problems (Sutton and McCallum, 2012).

As formerly stated in (Kiryakov et al., 2004), semantic annotation (SA) can be defined as the task of processing text elements (data description fields, free texts chunks, and so on) with the purpose of assigning semantic descriptions from a knowledge resource (KR) to the mentioned entities and, in this way, to reduce the ambiguity present in most natural language expressions. In our case, SA is applied to give semantics or meaning to words and to validate and classify them into semantic categories.

Many approximations using Conditional Random Fields in the biomedical domain have been proposed so far. In (McDonald and Pereira, 2005), McDonald and Pereira use Conditional Random Fields for tagging protein and gene mentions. Proteins and genes pertain to just one out of the eleven semantic groups we handle in this work. In (He and Kayaalp, 2008) and (Friedrich et al., 2006) CRF is applied to the manually tagged GENIA corpus, which has entities that belong to one of these semantic groups: protein, DNA, RNA, cell lines and cell types (these types belong to only two UMLS semantic groups). In (Friedrich et al., 2006), different experiments show how the selected features for CRFs directly affect the precision scores. However, as far as we know there is no approach in the literature handling the high heterogeneity of semantic groups present at UMLS, and for a very large corpus such as the UMLS Metathesaurus® lexicon.

Kudama S. and Berlanga R..
Semantic Annotation of UMLS using Conditional Random Fields.
DOI: 10.5220/0005131003350341
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 335-341
ISBN: 978-989-758-048-2

Although UMLS mainly covers biomedical concepts, we can also find concepts related to procedures, devices, time, geography, people, organizations and so on. These concepts play a secondary role in UMLS but they are used in describing biomedical concepts. This fact makes UMLS a very heterogeneous source of knowledge and, of course, it also complicates the task of annotating and classifying the words contained in it. Another issue that complicates the semantic analysis is the fact that UMLS has a lot of multi-word concepts whose components are not described in the KR itself. Indeed, the results of this work can be seen as a first approximation to the semantic decomposition of the complex concepts problem.

In the next section, we present the CRF algorithm to understand their mathematical basis. In Section 3, we explain what UMLS is and the modifications done over it to our experiments. In Section 4, we present the proposed method to obtain the seed tagged words that feed the CRF algorithm. Then, in Section 5, a complete evaluation of the process is made and, finally, the conclusion is presented in the last section.

2 CONDITIONAL RANDOM FIELDS

As biomedical NER can be thought of as a sequence segmentation problem where each word is a token in a sequence to be assigned a label, CRF method was chosen as a good option to annotate the UMLS concept descriptions. CRF is a structured prediction method, which is essentially a combination of classification and graphical modeling, combining the ability of graphical models to compactly model multivariate data with the ability of classification methods to perform prediction using large sets of input features (Sutton and McCallum, 2012).

CRFs are undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine. As explained in (Settles, 2004), such models are well suited to sequence analysis, and CRFs in particular have been shown to be useful in part-of-speech tagging (Lafferty, McCallum and Pereira, 2001), shallow parsing (Sha and Pereira, 2003) and named entity recognition for newswire data (McCallum and Li, 2003). They have also just recently been applied to the more limited task of finding gene and protein mentions (McDonald and Pereira, 2005), with promising early results.

As explained in (Settles, 2004), CRFs are probabilistic tagging models that give the conditional probability of a possible tag sequence given the input token sequence. Let o = *{o1, o2..., on}* be a sequence of observed words of length n, this is the input token sequence. Let *S* be a set of states in a finite state machine, each corresponding to a label $l \in L$. Let *s = {s1, s2, ... sn}* be the sequence of states in S that correspond to the labels assigned to words in the input sequence *o*. Linear-chain CRFs define the conditional probability of a state sequence given an input sequence to be:

$$P(s|o) = \frac{1}{Z_0} \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j (s_{i-1}, s_i, o, i) \right) \quad (1)$$

Where *Zo* is a normalization factor of all state sequences and it is constant for the given input, $f_j(s_{i-1}, s_i, o, i)$ is one of *m* functions that describes a feature and specifies an association between the predicates that hold at a position and the state for that position and λ_j is a learnt feature weight for each feature function, that specifies whether that association should be favored or disfavored. We assume that the *ith* input token is represented by a set *oi* of predicates that hold of the token or its neighborhood in the input sequence (McDonald and Pereira, 2005).

The learnt feature weight λ_j for each feature f_j should be highly positive for features that are correlated with the target label, highly negative for features that are anti-correlated with the label and around zero for relatively uninformative features. These weights are set to maximize the conditional log likelihood of labeled sequences in a training set *D = {<o, l>(1)... <o, l>(n)}:*

$$LL(D) = \sum_{i=1}^{n} \log\left(P \right) l_{(i)}\left(\right) - \sum_{j=1}^{m} \frac{\lambda_j^2}{2\sigma^2} \quad (2)$$

When the training state sequences are fully labeled and unambiguous, the objective function is convex, thus the model is guaranteed to find the optimal weight settings in terms of *LL(D)* (Settles, 2004). Once these settings are found, the most probable tag sequence for a given input unlabeled sequence *o* can be obtained applying a Viterbi-style algorithm to the maximization (Lafferty, McCallum and Pereira, 2001).

Typical features considered in the approaches of the literature are mainly divided in two groups: the orthographic features (capitalization, affixes, alphanumerical text, etc.) and semantic features (using, for example, external lexicons) (Settles, 2004).

It is important to point that our goal is different from those of the literature using CRFs: we aim at annotating at word-level complex entries of a biomedical KR, whereas these approaches aim at predicting the semantic group of a text chunk. Indeed the existing Gold Standard used in these approaches cannot be used for evaluating the word-level annotation problem, as they do not provide the semantic groups of the words belonging to each Gold Standard sample.

3 UMLS

In this work, we use the whole lexicon of UMLS MetaThesaurus® for building the word sequences dataset. UMLS Metathesaurus (UMLS from now on) is a compendium of more than 100 controlled vocabularies in the biomedical sciences. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems. UMLS further provides facilities for natural language processing, so it is intended to be mainly used by developers of systems in biomedical informatics. As UMLS also provides relationships between concepts, it can be also regarded as a big "ontology" that enables physicians to classify signs, symptoms, and diseases using medical concepts.

In this work, we deal with the 2012AB version of UMLS. The latest set of UMLS semantic groups (SGRs) and UMLS semantic types (STYs) have been retrieved as well. From the analysis of STYs, SGRs and vocabularies included in Biotea (Garcia, McLaughlin and Garcia, 2013), we defined a set of customized semantic groups.

UMLS concepts are attributed to the STYs; these are then categorized into SGRs. Currently UMLS makes use of 15 SGRs that are assigned to 99.5% of the UMLS concepts (Castro, Berlanga and Garcia, 2014). The SGRs have been defined for organizational reasons in order to better manage the conceptual complexity of STYs (McCray, Burgun and Bodenreider, 2001). The proposed semantic groups deliver a finer grained grouping in comparison to those from UMLS. For instance, proteins or drugs have been separated as specific groups that are different from more generic chemicals. These specific groups have been defined by interpreting the descriptions for the SGRs and the STYs according to the UMLS Semantic Network.

The SGRs have been modified as follows. We split the UMLS CHEMicals group into GeNe & ProTeins (GNPT) for types closely related to either genes or proteins, DRUG for drugs, and CHEM for the rest of the chemicals. Our GNPT group also includes types from the UMLS GENEs group. From LIVB group (living organisms) we extracted PEOP group (people) and from CONC group (concepts) we extracted SPAT group (spatial concepts).

A first step before selecting sequence examples for CRF consists of identifying those concepts that are described with a single word. These words will serve as seed input as they provide semantic information to the learning process. However, there are words that are assigned to more than one semantic group, that is, they are ambiguous. These words cannot be used as input in CRF as it cannot handle ambiguous examples. To identify the single-word concepts, we process the MRCONSO file from UMLS.

4 METHODOLOGY

In order to illustrate the goal of this work, let us consider the UMLS Metathesaurus concept entry "*abdominal computed tomography adrenal gland calcification"* for an unambiguous concept tagged as DISO (disorder). Our goal is to annotate each word of the entry with one of our predefined semantic groups. In this case, we would have as a result: *abdominal* ANAT, *computed* INDC, *tomography* PROC, *adrenal* INDC, *gland* ANAT and *calcification* DISO.

In order to achieve this goal, firstly, tagging part of the UMLS words was necessary in order to generate a training file for the CRF algorithm.

The first approximation in tagging UMLS single words consisted in taking the words that only had one associated semantic group in the UMLS; words that are unambiguous according to the KR. For example, if the concept of one word *stomach* is tagged as ANAT, anatomical part, then we can assume that the word *stomach* belongs to the semantic group ANAT.

In the second approximation, a simple process of statistical inference had been added: for each word that acts as "head" of a multi-word concept, the semantic group of the concept is associated to it. After observing all the UMLS, each word has been associated to its most probable semantic group, as long as it goes beyond a certain threshold (>0.7 in the experiments). This threshold has been set manually observing the results to choose a point where all words are properly tagged.

Once the words explained before were annotated, the next step was to create the two necessary files

Table 1: Semantic groups, percentage of appearance in training file and description of each one.

Semantic group	% train	Description
LIVB	19.49%	Alga, virus, human, animal, organism, etc.
ANAT	17.86%	Body location, cell component, tissue, embryonic structure, etc.
DISO	7.90%	Anatomical abnormality, disease or syndrome, mental or behavioral dysfunction, etc.
INDC	7.77%	Qualitative or quantitative concept, classification, idea, etc.
SPAT	7.45%	Spatial concept
OBSV	7.45%	Finding, sign or symptom, clinical attribute, etc.
PROC	4.58%	Laboratory or test result, health care activity, research activity, etc.
GNPT	3.16%	Enzyme, receptor, molecular or nucleotide sequence, gene or genome, etc.
CHEM	2.52%	Immunologic factor, vitamin, biologically active substance, etc.
DEVI	1.46%	Medical, research, or drug delivery device
PHYS	1.33%	Physiological, cell, genetic or organism function

for CRF: the training and testing files. The training file should include as much information as possible in order to get a good learning model. All the UMLS entries with all words unambiguously annotated were used to create the training file. The rest of UMLS entries constituted the test file. As a result, the training file contains 183.275 different words, and the test file has 529.117 different words.

The semantic groups that are being managed are the most frequent in UMLS and are presented in Table 1, where the percentage of the occurrences in the training file and a brief description of each one are provided.

Almost half of the semantic groups are not directly related to biomedicine field. Words pertaining to one of the three special groups have been removed, namely: stop words (a list of English stop words like "of", "the", "in", "and", etc.), numbers (a word formed completely by digits, e.g. 21, 000, 9.2, 2'45, etc.) and letters (a text with one letter, as "a", "Z", etc.). Non-alphanumerical characters have been also automatically removed.

As the CRF algorithm needs features of the words in order to predict the category of each one of them, we discussed and decided which features are the most interesting according to the set of words present in the UMLS. The features should be able to discriminate the entities correctly, even on new, unseen examples (Friedrich et al., 2006):

- **POS (part-of-speech) tagging**. It is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition, as well as its context, for example, relationship with adjacent and related words in a sentence or a paragraph. POS tagging is used in the identification of words as nouns, verbs, adjectives, adverbs, etc. OpenNLP has been used to obtain POS tag annotations (OpenNLP, 2010).
- **Chunking label**. A common tagging format, IOB format, for tagging tokens in a chunking task in computational linguistics, indicating the beginning and ending of a chunk. OpenNLP has been used to obtain the chunking labels (OpenNLP, 2010).
- **Prefix**. A prefix is an affix placed before the stem of a word. A list of English prefixes related to the biomedicine field has been used to detect prefixes in words, for example *carcino-*, *gastro-*, *immuno-*, etc. We assume that words sharing the same prefix are likely to bring similar semantics.
- **Suffix**. A suffix is an affix placed after the stem of a word. A list of English suffixes related to the biomedicine field has been used to detect suffixes in words, for example *-kinesis*, *-lepsia*, *-malacia*, *-derma*, etc. Again, words sharing the same suffix are likely to bring similar semantics.
- **Number.** This feature can take two values: HASNUMBER or NONUMBER, the first one is set when the word contains a digit and the second one, if not. Having or not digits in words gives clues about the category of the word, which are likely to be CHEM or GNPT, for example.

- **Semantic type**. One of the subgroups of the groups that we have presented, which correspond to the STYs of UMLS. If the first method had not been able to select one, the value of this feature will be NOTYPE.
- **Semantic group**. One of the groups that we have presented. If the first method had not been able to select one, the value of this feature will be NOGROUP.

5 EVALUATION

We have used CRFsuite (Okazaki, 2011), an implementation of Conditional Random Fields (Lafferty, McCallum and Pereira, 2001) for labeling sequential data. The software provides fast training and tagging, simple data format for training and tagging, performance evaluation on training, etc.

Once configured both training and test files with the necessary format for CRFsuite and all the features explained before, the steps for the evaluation were the following:

1. Creating manually a Gold Standard with the most frequent words of the test file with unknown semantic group. Words like "protein", "branch", "oral", "lower", "cervical", "receptor", and so on, are used in the GS for the evaluation of the system. The GS includes 300 terms manually annotated.
2. Training the system and generating the probabilistic model. The system spends almost 10 hours training.
3. Tagging the test file using the model created before. It is very important to remark that the test file and the train file are disjoint, as they do not share UMLS entries, so the test is being done over contexts that the system had not seen before.
4. Comparing the CRF annotations with the GS annotations.

In Table 2, the percentages and number of examples of the most representative semantic groups in the GS are presented. The representation of each semantic group follows a similar distribution of the semantic groups in UMLS.

Once the file test is created with all the concepts that do not have all the terms annotated, in Table 3, the percentages of each big group in the file test are presented: special groups (stop words, numbers and letters), tagged concepts and no tagged concepts.

Table 2: Percentages of each group in the G.S.

Semantic group	% in G.S.	Number of examples
GNPT	26.66%	80
CHEM	26.33%	79
ANAT	15.66%	47
INDC	14.00%	42
LIVB	12.33%	37
SPAT	8.33%	25
DEVI	6.33%	19
DISO	3.66%	11
OBSV	3.33%	10
PHYS	3.33%	10
PROC	3.00%	9

Table 3: Statistics in test file.

Group	Number of words	% in test file
Special groups (STOP, NUMBER and LETTER)	838.435	13.23%
Tagged	2.816.149	44.45%
No tagged	2.680.631	42.31%

Table 4: Precision for each coverage percentage.

% coverage (over the 44.45% of the test file)	% precision
0%	68.98%
10%	69.16%
20%	70.64%
30%	70.90%
40%	72.64%
50%	74.04%
60%	75.08%
70%	76.13%
80%	76.48%
90%	77.52%
100%	80%

Having the tags of the 44.45% of the test file, we decided to study the behavior of the CRF depending on the quantity of annotated words that it is being provided to it. So finally, 11 executions were made including different percentages of the tagged words (44.45% of the test file). As shown in Table 4, precision increases as the algorithm receives more information of tagged words, starting in 68.98%, when no semantic information is provided, to 80%, when the 44.45% of the test file is tagged.

As words tagged in the test file are randomly selected, we performed 5 different executions for each coverage point and calculate the average precision of these executions (see Tables 4 and 5).

Really interesting results have been found in the annotated terms that can show us the capability of the learnt model:

- Words as "second", "cervical", "middle", "forth", "central, "external" and "nasal" are tagged as SPAT (anatomic spatial concept).
- "Body", "fasciculus", radius", "thyroid" are tagged as ANAT (anatomical concept).
- "eiphosoma", "petiolaris" and "reptans" are associated to LIVB (living organisms).
- Words like "methylenetetrahydrofolate" and "adenosyltransferase" are in CHEM (chemical product).
- "neurolysis", "peristalsis" and "venereal" are tagged as DISO (disorder).

All plural words were wrongly tagged. We noticed that a word presented in singular was tagged with the expected semantic group, but in plural it was not. Probably, this is happening because plural words do not appear frequently in the training file, so the CRF method cannot make a good model to predict their semantic group. We are working to handle this problem with different strategies in order to raise the precision of the system: adding a feature with the singular form of the word, preprocess the test file to change all plurals into singulars, etc. If these approaches do not solve the problem, we will add a new feature with the root of the word using more advanced stemming approaches.

Table 5 presents the precision scores by semantic group and coverage points. For each group in the rows and for each percentage of coverage inthe execution, the precision obtained is presented.

As we can see in Tables 4 and 5, the system is able to recognize different terms and assign them to a correct group among the semantic groups we are dealing with high precision.

As expected, the worst results are obtained in groups with little representation in the training file. As they almost do not appear, the system cannot make a good model to predict them. Conversely, well represented groups like DEVI, OBSV, LIVB and SPAT, obtain very good results: 100%, 100%, 97.29% and 92%.

6 CONCLUSIONS

We have presented a framework to use CRFs for tagging concepts in UMLS and then classifying them into very different and heterogeneous predefined categories. The obtained results in the evaluation are encouraging: the global precision goes from 68.98% to 80% depending on the percentage of information included in the test file (from 0% to 100% of tagged words, which are the 44.45% of the test file). The precision of each semantic group depends on its representation in the training file. As many times the group appears in the train file, its results improve, achieving the 100% of precision in groups like OBSV and DEVI.

We think that results are promising but no complete comparison with other methods has been made, so in future work, this kind of study will be made using, for example, Inductive Logic Programming or Statistical Relational Learning.

Furthermore, there are many options to try to improve the results. Another KR, like Babelnet, could be used to the semantic groups poorly defined in UMLS. Semantically decomposing and studying the coherence of the semantic annotations in concepts semantically related in the KR would be another interesting task to do.

Regarding the features, our future work will be concentrated on making a bigger set of features and automatically studying the results depending on the features selected. It would give us information about which features better discriminate between the semantic groups in the UMLS.

Once we have all the words annotated with their semantic groups in the UMLS concepts, we will be able to present the next step: try to infer the semantic group of the whole concept. We are working on a Bayesian model that uses the co-occurrence probabilities of each pair of semantic groups and we are expecting to obtain interesting results soon.

ACKNOWLEDGEMENTS

This work has been partially funded by the

Table 5: Precision by semantic group at different coverage points.

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
LIVB	94.44%	94.44%	94.59%	94.59%	95.13%	97.29%	97.29%	97.29%	97.29%	97.29%	97.29%
ANAT	63.82%	65.95%	70.21%	71.27%	72.34%	72.34%	72.34%	74.46%	74.46%	74.46%	74.46%
DISO	36.36%	38.63%	45.45%	45.45%	54.54%	63.63%	63.63%	63.63%	63.63%	72.72%	72.72%
INDC	64.28%	64.28%	64.28%	66.66%	66.66%	69.04%	69.04%	69.04%	69.04%	69.04%	69.04%
SPAT	84%	84%	88%	88%	92%	92%	92%	92%	92%	92%	92%
OBSV	20%	20%	20%	20%	40%	40%	50%	60%	80%	80%	100%
PROC	22.22%	22.22%	22.22%	44.44%	33.33%	33.33%	33.33%	44.44%	44.44%	55.55%	66.66%
GNPT	83.75%	82.5%	83.75%	85%	85%	86.25%	86.25%	86.25%	87.5%	90%	88.75%
CHEM	85.89%	84.21%	86.07%	87.34%	87.34%	88.60%	88.60%	88.60%	89.87%	92.40%	91.13%
DEVI	50%	50%	50%	50%	50%	50%	50%	100%	100%	100%	100%
PHYS	66.66%	66.66%	66.66%	66.66%	66.66%	66.66%	66.66%	66.66%	66.66%	66.66%	66.66%

"Ministerio de Economía y Competitividad" with contract number TIN2011-24147, and the Fundació Caixa Castelló project P1-1B2010-49. Shahad Kudama has been supported by Universitat Jaume I predoctoral grant PREDOC/2011/61.

REFERENCES

Burr Settles. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. *Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*. Geneva, Switzerland. 2004. Pages 104-107.

Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web 2* (2004). Pages 49-79.

Ryan McDonald and Fernando Pereira. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* 2005,6:S6.

Ying He and Mehmet Kayaalp. Biological Entity Recognition with Conditional Random Fields. *AMIA Annual Symposium Proceedings*. 2008: 293–297.

Christoph M. Friedrich, Thomas Revillion, Martin Hofmann and Juliane Fluck. Biomedical and Chemical Named Entity Recognition with Conditional Random Fields: The Advantage of Dictionary Features. *Proceedings of the International Symposium of Semantic Mining in Biomedicine (SMBM)*. 2006. Pages 85-89.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the 18th International Conference on Machine Learning*. 2001, pages 282-289.

Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. *In Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference.* 2003. Pages 134-141.

Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *In Proceedings of the Conference on Natural Language Learning*. 2003. Pages 188–191.

Leyla Jael Garcia Castro, Rafael Berlanga and Alexander Garcia. Biolinks, a semantic similarity score based on UMLS groups for identifying relevant related full-text articles in PubMed Central. *To appear.*

Garcia Castro, L.J., McLaughlin, C. and Garcia, A. Biotea: RDFizing PubMed Central in Support for the Paper as an Interface to the Web of Data. *Biomedical semantic*s, 2003. 15;4 Suppl 1:S5.

McCray, A.T., Burgun, A. and Bodenreider, O. Aggregating UMLS semantic types for reducing conceptual complexity, *Proceedings of Medinfo*, 2001. 10, 216-220.

Charles Sutton and Andrew McCallum. An introduction to Conditional Random Fields, *Foundations and Trends in Machine Learning,* 4 2012.

NaoakiOkazaki. CRFSsuite Software 2011. http://www.chokkan.org/software/crfsuite/ (18 April 2014)

Apache OpenNLP 2010. http://opennlp.apache.org/ (18 June 2014)