Chemical Named Entity Recognition: Improving Recall Using a Comprehensive List of Lexical Features

Andre Lamurias, João Ferreira, and Francisco M. Couto

Dep. de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
alamurias@lasige.di.fc.ul.pt, joao.ferreira@lasige.di.fc.ul.pt,
fcouto@di.fc.ul.pt

Abstract. As the number of published scientific papers grows everyday, there is also an increasing necessity for automated named entity recognition (NER) systems capable of identifying relevant entities mentioned in a given text, such as chemical entities. Since high precision values are crucial to deliver useful results, we developed a NER method, Identifying Chemical Entities (ICE), which was tuned for precision. Thus, ICE achieved the second highest precision value in the BioCreative IV CHEMDNER task, but with significant low recall values. However, this paper shows how the use of simple lexical features was able to improve the recall of ICE while maintaining high levels of precision. Using a selection of the best features tested, ICE obtained a best recall of 27.2% for a precision of 92.4%.

Keywords: Text mining; Conditional Random Fields; Named Entity Recognition; Chemical Compounds; ChEBI

1 Introduction

As the number of published scientific papers grows everyday, there is also an increasing necessity for automated named entity recognition (NER) systems capable of identifying relevant entities mentioned in a given text. The BioCreative challenge is a community effort to evaluate text mining and information extraction systems applied to the biological domain. One of the tasks proposed for the fourth edition of this competition consisted in the detection of mentions of chemical compounds and drugs in MEDLINE titles and abstracts (CHEMDNER task) [1]. The chemical entities to identify were those that can be linked to a chemical structure. This task was divided in two subtasks: The first, Chemical Document Indexing (CDI), expected an ordered list of unique chemical entities referenced in given text. The second subtask, Chemical Entity Mention recognition (CEM), expected the exact position of each chemical entity mentioned in the text. The task organizers also provided a training corpus composed by 10,000 MEDLINE abstracts that were annotated manually by domain experts. The specific rules used by the annotators and the criteria used for choosing the MEDLINE entries were defined by the organization and released with the corpus.

To participate in BioCreative IV, we started by adapting our method [2] based on Conditional Random Fields (CRF) classifiers trained with the CHEMDNER corpus. Our system trained one classifier for each type of entity annotated in the corpus. The final confidence score of each token classified as a chemical entity by at least one classifier was calculated by averaging the three best classifier scores. The identified entities were validated by resolving each chemical entity recognized to the ChEBI ontology and by calculating the semantic similarity between other ChEBI entities detected on the same text. The resolution and similarity method enabled us to filter most false positives and achieve the second highest precision value on both subtasks [3]. However, the set of features used to train the classifiers was relatively small comparing with other CRF based approaches that participated in BioCreative also, which use more general and domain specific features [4, 5].

In this paper, we present a new version of our system (ICE) that achieved significantly better results using more lexical features derived from the word tokens. The rest of this paper is organized as follows: Section 2 presents an overview on the BioCreative 2013 competition which we used to evaluate our system, Section 3 (Methods) describes the approach we used for the competition and how we improved it since then, Section 4 (Results) compares the effect of different features on the precision and recall values using 3-fold cross-validation on the CHEMDNER corpus, and finally on Section 5 we express our main conclusions.

2 BioCreative 2013 - CHEMDNER Task

2.1 CHEMDNER Corpus

The CHEMDNER corpus consists in 10,000 MEDLINE titles and abstracts and was originally partitioned randomly in three sets: training, development and test. The chosen articles were sampled from a list of articles published in 2013 by the top 100 journals of a list of categories related to the chemistry field. These articles were manually annotated according to the guidelines, by a team of curators with background in chemistry. Each annotation consisted in the article identifier, type of text (title or abstract), start and end indices, the text string and the type of the CEM which could be one of the following: trivial, formula, systematic, abbreviation, family and multiple. There was no limit for the number of words that could refer to a CEM but due to the annotation format, the sequence of words had to be continuous. There was a total of 59,004 annotations on the training and development sets, which consisted in 7,000 documents.

2.2 CEM and CDI Subtasks

There were two types of predictions the participants could submit for the CHEMDNER task: a ranked list of unique chemical entities described on each document (CDI task) and the start and end indices of each chemical entity mentioned on each document (CEM task). Each list should be ordered by how confident the

system is that each prediction is a chemical entity. Using the CEM predictions, it was possible to generate results for the CDI subtask, by excluding multiple mentions of the same entity in a text.

A gold standard for both subtasks was included with the corpus, which could be used to calculate precision and recall of the results, with the evaluation script released by the organization. Each team was allowed to submit up to five different runs for each subtask.

3 Methods

3.1 Submission to BioCreative 2013

Our method uses Conditional Random Fields (CRFs) for building probabilistic models based on training datasets. We used the MALLET [6] implementation of CRFs, adapted to also output the probability of the most probable sequence. This probability was used as a confidence score for each prediction, making it possible to filter predictions with low confidence.

To train models and classify new text, it is necessary to tokenize the text and generate features from word tokens. Then, the corresponding label is added to the feature list. This label could be "Not Chemical", "Single", "Start", "Middle" or "End", to include chemical entities composed by more than one token. We have used a specifically adapted word tokenizer for chemical text adapted from an open source project [7]. Four features were being extracted from each word token by our system: Stem, Prefix and suffix (size 3) and a boolean which indicates if the token contains a number (Has number). We merged the training and development sets of the CHEMDNER corpus into one training set and generated one dataset for each type of CEM. With this method we expected to identify more correct chemical entities since we were including the results of classifiers focused on just one type of CEM. The confidence score used when more than one of the classifiers identified the same CEM was the average of the three best confidence scores. This system was then evaluated with 3-fold cross-validation.

With the terms identified as chemical entities, we employed an adaptation of FiGO, a lexical similarity method [8], to perform the search for the most likely ChEBI terms. Then, we were able to calculate the Gentleman's simUI [9] semantic similarity measure for each pair of entities identified in the text and successfully mapped to the ChEBI ontology. We used the maximum semantic similarity value for each entity as a feature for filtering and ranking. This value has shown to be crucial to achieve high precision results [10].

Since each team could submit up to five runs for each subtask, we generated three runs to achieve our best F-measure, precision and recall, based on the cross-validation results we obtained on the training set. For the other two runs, we filtered the predictions by semantic similarity only. The best results we obtained were with the run we submitted for best precision (run 2), achieving the second highest precision value in the competition. For this run, we excluded results with the classifier confidence score and the semantic similarity measure lower

than 0.8. We now focused on keeping the precision of our system at high values, while improving the recall and F-measure.

3.2 New Features

After implementing thirteen new features, we studied the effect of adding one new feature at a time, while always keeping the four original features constant. These new features are based on orthographic and morphological properties of the words used to represent the entity, inspired by other CRF-based chemical NER systems [4, 5, 11–13]. We integrated the following features:

Prefix and Suffix sizes 1, 2 and 4: The first and last n characters of a word token.

Greek symbol: Boolean that indicates if the token contains greek symbols.

Non-alphanumeric character: Boolean that indicates if the token contains non-alphanumeric symbols.

Case pattern: "Lower" if all characters are lower case, "Upper" if all characters are upper case, "Title" if only the first character is upper case and "Mixed" if none of the others apply.

Word shape: Normalized form of the token by replacing every number with '0', every letter with 'A' or 'a' and every other character with 'x'.

Simple word shape: Simplified version of the word shape feature where consecutive symbols of the same kind are merged.

Periodic Table element: Boolean that indicates if the token matches a periodic table symbols or name.

Amino acid: Boolean that indicates if the token matches a 3 letter code amino acids.

For example, for the sentence fragment "Cells exposed to α-MeDA showed an increase in intracellular glutathione (GSH) levels", the list of tokens obtained by the tokenizer and some possible features are shown on Table 1.

After applying the same methods described on Section 3.1 for each new feature, we were able to compare the effect of each one on the results. Then, we selected the features that achieved a higher precision, recall and F-measure, creating three sets of features for each metric and a fourth set with all the features tested.

4 Results

4.1 BioCreative 2013

Using 3-fold cross-validation on the training and development sets, we obtained the results presented in Table 2. The first three runs were aimed at achieving a high F-measure, precision and recall, respectively. On runs 4 and 5 we filtered only by semantic similarity. We used as reference the results of run 2 since the precision value obtained with the test set was the second highest in the CHEMDNER task. Our objective was to improve recall and F-measure values with minimal effect on the precision.

Table 1. Example of a sequence of some the new features, and the corresponding label, derived from a sentence fragment (PMID 23194825).

Token	Prefix 4	Suffix 4	Case pattern	Word shape	Label
Cells	Cell	ells	titlecase	Aaaaa	Not Chemical
exposed	expo	osed	lowercase	aaaaaaa	Not Chemical
to	to	to	lowercase	aa	Not Chemical
α-MeDA	α-Me	MeDA	mixed	xxAaAA	Chemical
showed	show	owed	lowercase	aaaaaa	Not Chemical
an	an	an	lowercase	aa	Not Chemical
increase	incr	ease	lowercase	aaaaaaaa	Not Chemical
in	in	in	lowercase	aa	Not Chemical
intracellular	intr	ular	lowercase	aaaaaaaaaaaaa	Not Chemical
glutathione	glut	ione	lowercase	aaaaaaaaaaa	Chemical
(	(	(	-	x	Not Chemical
GSH	GSH	GSH	uppercase	AAA	Chemical
)	)	)	-	x	Not Chemical
levels	leve	vels	lowercase	aaaaaa	Not Chemical

Table 2. Precision, Recall and F-measure estimates for each run submitted to BioCreative 2013, obtained with cross-validation on the training and development dataset for the CDI and CEM subtasks.

	CDI			CEM		
	P	R	F_1	P	R	F_1
Run 1	84.8%	71.2%	77.4%	87.3%	70.2%	77.8%
Run 2	95.0%	6.5%	12.2%	95.0%	6.0%	11.1%
Run 3	52.1%	80.4%	63.3%	57.1%	76.6%	65.4%
Run 4	87.9%	22.7%	36.1%	89.7%	21.2%	34.3%
Run 5	87.9%	22.7%	36.1%	79.9%	22.6%	35.3%

4.2 New Features

The precision, recall and F-measure values obtained using our four original features plus one new one are presented in Table 4.2 For each metric, we added a shaded column which compares that value with the corresponding one on Table 2, for the run with best precision.

The features that returned the best recall and F-measure were the simple word shape and prefix and suffix with size=2. Using prefix and suffix with size=1 and the alphanumeric boolean decreased our precision the most, without improving the other metrics as much as other features. The periodic table feature, which was one of our two domain-specific features, achieved a recall value of 16.4%, while maintaining the precision at 94%. Our other domain-specific feature, amino acid, achieved our highest precision in this work. The general effect of using five features instead of the original four was a decrease in precision by 0.8%-4.5% and increase in recall and F-measure by 0.4%-19.5%.

For each subtask, we performed another cross-validation run with the original four features to use as baseline values. We created three feature sets composed by the original features we used for BioCreative and the features that improved precision, recall or F-measure on any subtask, compared to the baseline. The three feature sets created were:

Best precision: Stem, Prefix/suffix 3, Has number, Prefix/suffix 4, Has greek symbol, Has periodic table element, Has amino acid.

Best recall: Stem, Prefix/suffix 3, Has number, Prefix/suffix 1, Prefix/suffix 2, Has greek symbol, Has periodic table element, Case pattern, Word shape, Simple word shape.

Best F-measure: Stem, Prefix/suffix 3, Has number, Prefix/suffix 1, Prefix/suffix 2, Has greek symbol, Has periodic table element, Has amino acid, Case pattern, Word shape, Simple word shape.

The results obtained with these sets are presented in Table 4.2 Although there was a decrease in precision in every case, the difference in recall and F-measure values was always much higher. The feature set with best F-measure was able to improve the recall by 21.0% while taking only 3.2% of the precision.

To determine the statistical significance of the improvement between the expanded feature set and the original, we ran a bootstrap resampling simulation similar to the BioCreative II gene mention task [14] and BioCreative CHEMD-NER task evaluations. We picked 1000 PMIDs from the train and development sets and computed the recall and F-measure for this subset of documents. Then we repeated this process 10,000 times, and estimated the average recall and F-measure, and respective standard deviation for each feature set. With the original features, the average recall was 8.00% (SD=0.53%) and the average F-measure was 14.74% (SD=0.90%) while using the expanded feature set, the average recall was 27.20% (SD=0.92%) and the average F-measure was 42.02% (SD=1.13%).

Table 3. Precision, Recall and F-measure estimates for each new features used with the original set, obtained with cross-validation on the training and development dataset for the CDI subtask.

	CDI						CEM					
Feature set	P	ΔP	R	ΔR	F_1	ΔF_1	P	ΔP	R	ΔR	F_1	ΔF_1
Prefix/suffix 1	91.0%	-4.0%	14.0%	+7.5%	24.3%	+12.1%	92.4%	-2.6%	13.4%	+7.4%	23.4%	+12.3%
Prefix/suffix 2	92.4%	-2.6%	19.1%	+12.6%	31.6%	+19.4%	93.5%	-1.5%	18.3%	+12.3%	30.6%	+19.5%
Prefix/suffix 4	93.3%	-1.7%	6.9%	+0.4%	12.9%	+0.7%	94.2%	-0.8%	6.6%	+0.6%	12.2%	+1.1%
Greek letter	93.4%	-1.6%	12.0%	+5.5%	21.2%	+9.0%	94.2%	-0.8%	11.8%	+5.8%	20.9%	+9.8%
Periodic table	94.0%	-1.0%	16.3%	+9.8%	27.8%	+15.6%	94.7%	-0.3%	16.4%	+10.4%	28.0%	+16.9%
Amino acid	95.0%	0.0%	9.0%	+2.5%	16.4%	+4.2%	95.1%	+0.1%	8.7%	+2.7%	16.0%	+4.9%
Alphanumeric	90.4%	-4.6%	5.3%	-1.2%	10.0%	-2.2%	92.0%	-3.0%	4.4%	-1.6%	8.4%	-2.7%
Case pattern	93.0%	-2.0%	15.7%	+9.2%	26.9%	+14.7%	93.5%	-1.5%	14.9%	+8.9%	25.6%	+14.5%
Word shape	93.9%	-1.1%	11.8%	+5.3%	20.9%	+8.7%	93.3%	-1.7%	12.7%	+6.7%	22.4%	+11.3%
Simple word shape	92.2%	-2.8%	17.1%	+10.6%	28.9%	+16.7%	92.4%	-2.6%	16.9%	+10.9%	28.7%	+17.6%

Table 4. Precision, Recall and F-measure estimates for each feature set used with the original set, obtained with cross-validation on the training and development dataset for the CDI and CEM subtasks

	CDI						CEM					
Feature set	P	ΔP	R	ΔR	F_1	ΔF_1	P	ΔP	R	ΔR	F_1	ΔF_1
Precision	93.7%	-1.3%	15.4%	+8.9%	26.5%	+14.3%	94.1%	-0.9%	15.0%	+9.0%	25.9%	+14.8%
Recall	91.5%	-3.5%	24.7%	+18.2%	38.9%	+26.7%	92.0%	-3.0%	23.9%	+17.9%	37.9%	+26.8%
F-measure	91.7%	-3.3%	28.3%	+21.8%	43.2%	+31.0%	92.3%	-2.7%	28.0%	+22.0%	43.0%	+31.9%
All features	91.5%	-3.5%	24.5%	+18.0%	38.7%	+26.5%	93.0%	-2.0%	24.2%	+18.2%	38.4%	+27.3%

5 Conclusion

Our participation in the CHEMDNER task of BioCreative 2013 achieved high precision values for both subtasks, but at the expense of a low recall. This manuscript shows how ICE improved its recall and F-measure maintaining the same levels of precision, by using a more comprehensive feature set. The effect of adding each new feature to ICE was evaluated by cross-validation on the CHEMDNER corpus. We then evaluated feature sets composed by the features that achieved the best precision, recall and F-measure, using the same method.

Individually, the features that were specific to chemical compounds achieved the best balance between precision and recall. Adding only the prefixes and suffixes with size 2, we were able to increase the recall and F-measure by 12.3% and 19.5%, while decreasing the precision by 1.5%. Using a combination of the features that achieved the best results individually, we were able to increase the recall and F-measure by 21.2% and 31.0% respectively while decreasing the precision by 2.6% (Table 4.2).

Considering the run that achieved the highest precision in the official BioCreative results for the CDI task, our precision is 6.9% lower, but the recall and F-measure are 10.9% and 13.9% higher, respectively. Considering the run with best precision in the CEM task, our precision is 5.7% lower, but the recall and F-measure are 9.3% and 11.8% higher. Our precision values would be the third and sixth highest in the CDI and CEM subtasks, respectively. However, notice that the results presented here were not obtained with CHEMDNER test set, and for the competition, using the official test set, our results were higher than the cross-validation estimates we obtained.

In the future we intend to use more domain-specific features, and filter predictions with a more powerful semantic similarity measure [15].

Acknowledgments. The authors want to thank the Portuguese Fundação para a Ciência e Tecnologia through the financial support of the SPNet project (PTDC/EBB-EBI/113824/2009), the SOMER project (PTDC/EIA-EIA/119119/2010) and through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

References

1. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: Overview of the chemical compound and drug name recognition (CHEMDNER) task. In: BioCreative Challenge Evaluation Workshop vol. 2. (2013) 2
2. Grego, T., Pęzik, P., Couto, F.M., Rebholz-Schuhmann, D.: Identification of chemical entities in patent documents. In: Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living. Springer (2009) 942–949
3. Lamurias, A., Grego, T., Couto, F.M.: Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. In: BioCreative Challenge Evaluation Workshop vol. 2. Volume 489. (2013) 75
4. Huber, T., Rocktäschel, T., Weidlich, M., Thomas, P., Leser, U.: Extended feature set for chemical named entity recognition and indexing. In: BioCreative Challenge Evaluation Workshop vol. 2. (2013) 88
5. Leaman, R., Wei, C.H., Lu, Z.: NCBI at the biocreative IV CHEMDNER task: Recognizing chemical names in PubMed articles with tmChem. In: BioCreative Challenge Evaluation Workshop vol. 2. (2013) 34
6. McCallum, A.K.: Mallet: A machine learning for language toolkit. (2002)
7. Corbett, P., Batchelor, C., Teufel, S.: Annotation of chemical named entities. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics (2007) 57–64
8. Couto, F.M., Silva, M.J., Coutinho, P.M.: Finding genomic ontology terms in text using evidence content. BMC bioinformatics **6**(Suppl 1) (2005) S21
9. Gentleman, R.: Visualizing and distances using GO. URL http://www. bioconductor. org/docs/vignettes. html (2005)
10. Grego, T., Couto, F.M.: Enhancement of chemical entity identification in text using semantic similarity validation. PloS one **8**(5) (2013) e62984
11. Batista-Navarro, R.T., Rak, R., Ananiadou, S.: Chemistry-specific features and heuristics for developing a CRF-based chemical named entity recogniser. In: BioCreative Challenge Evaluation Workshop vol. 2. (2013) 55
12. Usié, A., Cruz, J., Comas, J., Solsona, F., Alves, R.: A tool for the identification of chemical entities (CheNER-BioC). In: BioCreative Challenge Evaluation Workshop vol. 2. (2013) 66
13. Campos, D., Matos, S., Oliveira, J.L.: Chemical name recognition with harmonized feature-rich conditional random fields. In: BioCreative Challenge Evaluation Workshop vol. 2. (2013) 82
14. Smith, L., Tanabe, L.K., Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., et al.: Overview of BioCreative II gene mention recognition. Genome biology **9**(Suppl 2) (2008) S2
15. Couto, F., Pinto, H.: The next generation of similarity measures that fully explore the semantics in biomedical ontologies. Journal of Bioinformatics and Computational Biology **11**(5 (1371001)) (2013) 1–12