Biomedical Information Processing (R214): Main Assignment

Chongyang Shi *(cs940)*

April 14, 2018

For the main course assignment, I am undertaking the second practical option (**1.2**): *extracting chemical-disease associations from the biological literature.*

a Improving the Conditional Random Fields named entity recognizer

a.i Ablating features from the original feature set

Based on the default n-gram feature set in the feature extraction script, the script was modified to ablate each feature in turn. To provide a better understanding of the effects by offsets of surface form words, all surface forms of words within an offset of 1 (the trigram) were knocked out from the templates first, then just the surface forms of words before and after the current word (-1/1). Other features including the lemma, phonetic coding (*soundex*), part-of-speech, and chunk in IOB2 notation of the current word only. The resulting precisions, recall rates, and F_1-scores from ablating each feature on the *devel* dataset are presented separately in Figures 1, 2, and 3. For each named entity class as well as the overall average, with none-ablated as reference, improved performance due to ablation are presented in **bold**.

Class \ Ablated	None	*word* (all)	*word* (-1/1)	*lemma*	*soundex*	*pos*	*chunk*
B-Chemical	0.9178	**0.9345**	**0.9409**	0.9056	0.9015	**0.9495**	**0.9210**
O	0.9560	0.9471	0.9498	0.9540	0.9531	0.9499	0.9557
B-Disease	0.8403	0.8242	0.8223	**0.8418**	0.8387	**0.8412**	0.8396
I-Disease	0.7404	0.7152	0.7167	**0.7467**	**0.7506**	**0.7631**	**0.7509**
I-Chemical	0.7556	0.6488	0.6745	**0.7569**	**0.7612**	**0.7906**	**0.7682**
Macro-average	0.8420	0.8142	0.8209	0.8410	0.8410	**0.8589**	**0.8471**

Figure 1: Resulting **precisions** on different named entity classes from ablating individual features from the original feature set.

Class \ Ablated	None	*word* (all)	*word* (-1/1)	*lemma*	*soundex*	*pos*	*chunk*
B-Chemical	0.6664	0.5583	0.5955	0.6564	0.6520	0.5702	0.6652
O	0.9888	0.9888	**0.9889**	0.9888	0.9887	**0.9908**	**0.9894**
B-Disease	0.6011	0.5514	0.5672	0.5669	0.5561	0.5806	0.5992
I-Disease	0.6018	0.5530	0.5607	0.5993	0.5952	**0.6029**	0.5952
I-Chemical	0.5961	0.5114	0.5275	0.5950	0.5910	0.5938	**0.5990**
Macro-average	0.6908	0.6326	0.6479	0.6813	0.6766	0.6677	0.6896

Figure 2: Resulting **recall rates** on different named entity classes from ablating individual features from the original feature set.

Surprisingly, for chemicals at the beginning of entities (B-Chemcal), the precision (correct tags among those tagged) increased substantially when the surface form word feature is ablated. Ablating only the surface forms of before and after tokens (-1/1) produced slightly higher precision than ablating the entire surface form trigram. This is however accompanied by substantially reduced precisions on all other named entity classes, as well as reduced recall rate (correct tags among all relevant inputs that can be tagged) nearly across the board. As B-chemicals already bear a fairly high precision (91.78%), it is not advisable to ablate the surface forms and reduce recall into the 50% range.

Class \ Ablated	None	*word* (all)	*word* (-1/1)	*lemma*	*soundex*	*pos*	*chunk*
B-Chemical	0.7721	0.6992	0.7294	0.7611	0.7567	0.7125	**0.7725**
O	0.9721	0.9675	0.9690	0.9711	0.9706	0.9699	**0.9723**
B-Disease	0.7008	0.6607	0.6713	0.6776	0.6687	0.6870	0.6993
I-Disease	0.6640	0.6238	0.6292	**0.6649**	0.6639	**0.6736**	**0.6641**
I-Chemical	0.6665	0.5720	0.5920	0.6662	0.6654	**0.6782**	**0.6731**
Macro-average	0.7551	0.7046	0.7182	0.7451	0.7451	0.7443	**0.7562**

Figure 3: Resulting F_1**-scores** on different named entity classes from ablating individual features from the original feature set.

Ablation of the lemma (base word) and the phonetic coding (*soundex*) yielded minimal improvements to precisions on some named entity groups but minimal reductions on others. Recall rates all reduced by very small margins. Based on a generally negative outlook on the F_1-scores (combined metric of precision and recall), it is advisable not to alate either of he two features.

Ablating the part-of-speech produced the greatest precision improvements to most groups, but mostly lowered the recall rate substantially. This is also reflected in the overall negative outlook on the combined F_1-scores. Therefore it is not advisable to ablate the part-of-speech feature.

Finally, ablating the chunk information from the feature set improved the precision without significantly affecting the recall rate in most cases, resulting in improved F_1-scores for all named entity classes barring diseases within entities (I-Disease) with a minimal decrease. Therefore, it is advisable to ablate the chunk information from the feature set used.

Vertically, precision and recall rates of terms outside entities (O) are high and only very minimally affected by ablating any of the features, which is generally expected in entity recognition operations due to the abundance of outside tokens between short named entities [1].

a.ii Improvements to the base tagger

After removing chunk information to improve performance of the base feature set (as described above), I will first experiment with expanding the n-gram feature set by expanding unigram features into trigram features. Then, I will examine the effects of adjusting several parameters of the L-BFGS training algorithm used in *crfsuite*.

a.ii.1 Expansion of unigram features

Evaluations of word representation features in entity recognition by Tang et al. [2] demonstrated the benefits of using trigram features in word stemming. With this as inspiration, I iteratively expanded the unigram features of lemma, part-of-speech, and the phonetic coding (*soundex*) in the feature set. The order of expansion was chosen due to lemma being directly related to word stemming, part-of-speech correlations between neighbouring words normally being important, and the phonetic coding being the one left. The resulting performance information are shown in Figure 4.

While the precision of B-Chemical tagging continues to follow the declining trend discussed in a.i (although not as severe in unigram expansion as in feature ablation), expanding unigram features of lemma, part-of-speech, and the phonetic coding into trigrams have resulted in improved or roughly equal precisions, recall rates – and hence F_1-scores. Precision improvements were most notable from the additions of lemma and phonetic coding on chemicals within entities (I-Chemical), showing the influence of phonetic features of nearby entities on chemical entity recognition. The strongest improvement of recall and the F_1-score originated from expanding part-of-speech to nearby entities, demonstrating the importance of expanding the semantic scope. While precisions of some named entities took a small hit when the phonetic coding (*soundex*) was added, the improved macro-average precision as well as generally improved recall rates and F_1-scores were behind my choice of retaining all three unigram expansions.

Expanded from unigram	None			*lemma*		
Entity Class	Precision	Recall	F_1-score	Precision	Recall	F_1-score
B-Chemical	0.9210	0.6652	0.7725	0.9137	0.6695	0.7728
O	0.9557	0.9894	0.9723	0.9559	0.9890	0.9722
B-Disease	0.8396	0.5992	0.6993	0.8365	0.5992	0.6982
I-Disease	0.7509	0.5952	0.6641	0.7519	0.6040	0.6699
I-Chemical	0.7682	0.5990	0.6731	0.7820	0.6013	0.6798
Macro-average	0.8471	0.6896	0.7562	0.8480	0.6926	0.7586
Expanded from unigram	*lemma + pos*			*lemma + pos + soundex*		
Entity Class	Precision	Recall	F_1-score	Precision	Recall	F_1-score
B-Chemical	0.9077	0.6864	0.7817	0.9077	0.6875	0.7824
O	0.9574	0.9894	0.9731	0.9574	0.9894	0.9731
B-Disease	0.8499	0.6162	0.7144	0.8477	0.6124	0.7111
I-Disease	0.7819	0.6103	0.6855	0.7795	0.6110	0.6850
I-Chemical	0.7884	0.6138	0.6903	0.8010	0.6241	0.7016
Macro-average	0.8570	0.7032	0.7690	0.8587	0.7049	0.7706

Figure 4: Resulting tagging performance on the *devel* dataset after expanding unigram features into trigram features in the baseline feature set. "None" represents the baseline feature set with *chunk* ablated.

a.ii.2 Adjustment of training parameters

The L-BFGS training algorithm used in entity recognition uses L2 regularization, which is normally more efficient than L1 [3], with a regularization parameter influencing the levels of bias and overfitting. Various values around the default $c2 = 1$ were tested, with broadly similar trends in variations of precision, recall, and F_1-score in individual named entities. Therefore, only the macro-average metrics from tagging the *devel* dataset with a model trained on each value of $c2$ are plotted in Figure 5.

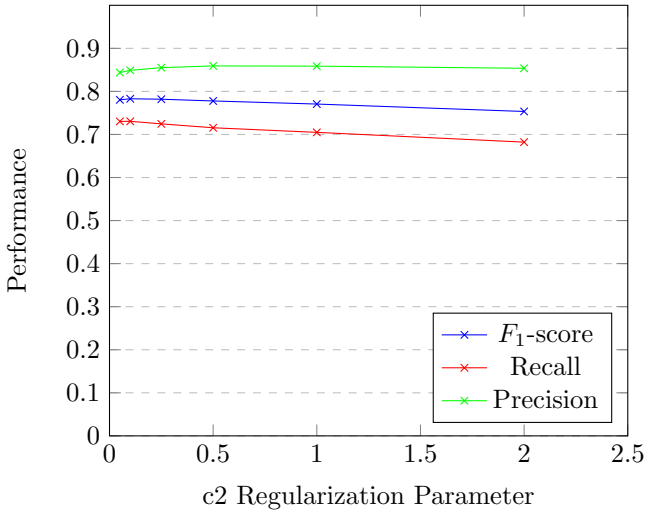

Figure 5: Resulting tagging performance on the *devel* dataset with models trained on different c2 values.

A reduced $c2$ parameter results in longer training, but also generally improved recall and the overall F_1-score. With the adverse effects of extended training time and reduced precision (due to increased overfitting), as well as diminished gains in recall, $c2 = 0.25$ was chosen as the $c2$ value of choice.

With $c2 = 0.25$, L-BFGS also allows different line search algorithms, which yield roughly the same training speed, as well as almost the same performance with all named entity classes, as shown in Figure 6. Therefore the default More and Thuente's method is kept unchanged.

Line Search	*MoreThuente*			*Backtracking*			*StrongBacktracking*		
Entity Class	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
B-Chemical	0.9223	0.6972	0.7941	0.9223	0.6972	0.7941	0.9223	0.6972	0.7941
O	0.9609	0.9884	0.9745	0.9609	0.9884	0.9744	0.9609	0.9884	0.9744
B-Disease	0.8363	0.6671	0.7422	0.8353	0.6656	0.7409	0.8352	0.6661	0.7411
I-Disease	0.7515	0.6338	0.6876	0.7520	0.6330	0.6874	0.7520	0.6330	0.6874
I-Chemical	0.8048	0.6367	0.7110	0.8048	0.6367	0.7110	0.8048	0.6367	0.7110
Macro-average	0.8552	0.7246	0.7819	0.8550	0.7242	0.7816	0.8550	0.7243	0.7816

Figure 6: Resulting tagging performance on the *devel* dataset with different line search algorithms: More and Thuente's method, backtracking method with regular Wolfe condition, and backtracking method with strong Wolfe condition.

a.iii Evaluate the improved model on the test set

To summarise, the changes made on the baseline entity recognizer include the ablation of chunk information from the feature set during feature extraction; expanding lemma, part-of-speech, and phonetic coding (*soundex*) features from unigrams into trigrams (covering feature information of entities before and after the current word); as well as adjusting L-BFGS' $c2$ regularization parameter to 0.25. A comparison between the improved model and the original model is shown in Figure 7.

Model	**Improved**			**Original**			*Change*
Entity Class	Precision	Recall	F_1-score	Precision	Recall	F_1-score	F_1-score
B-Chemical	0.9168	0.6771	0.7789	0.9085	0.6457	0.7549	+3.18%
O	0.9619	0.9890	0.9753	0.9576	0.9882	0.9726	+0.28%
B-Disease	0.8218	0.6514	0.7268	0.8217	0.5895	0.6865	+5.87%
I-Disease	0.7522	0.6434	0.6936	0.7311	0.6178	0.6697	+3.57%
I-Chemical	0.7998	0.6087	0.6913	0.7438	0.6081	0.6691	+3.32%
Macro-average	0.8505	0.7139	0.7732	0.8325	0.6899	0.7506	+3.01%

Figure 7: A comparison between the improved model and the original entity recognition model on tagging performance on the *test* dataset.

Changes made to features and L-BFGS parameters resulted in a model which yields improved precision and recall across the board. Combining the metrics of precision and recall, changes in F_1-scores are shown on the right of Figure 7. The already high-precision and high-recall terms outside entities received the least improvement, while disease terms at the start of entities benefited the most from the improved model. More than 3% of improvements were made to all other named entity classes and the macro-average. Based on these results from testing on the unseen *test* input dataset, I believe it is reasonably conclusive that the improved model is superior to the original in entity recognition of environment and disease terms.

b Grounding named entities through approximate string matching

The MESH concept dictionary provides terms for two classes of entities: chemicals and diseases. Based on the class of entity determined by the improved entity recognition model from the previous section, all possible choices of the specified class in the MESH dictionary were supplied to the approximate string matching process facilitated by the `fuzzywuzzy` [4] library. With the initial approach, best matching dictionary terms were grouped by their original sentences, which were annotated with their respective matched terms in output.

Unfortunately, this initial approach clearly appeared to be inadequate due to the lack of connecting recognised terms belonging to a continuous entity:

```
**Urine**{1} **N**{2} **-**{3} **acetyl**{4} **-**{5} **beta**{6} **-**{7} **D**{8}
    **-**{9} **glucosaminidase**{10} - - a marker of **tubular**{11} **damage**{12} ?
{1} purine (Score: 91)
{2} alanine (Score: 90)
{3} 11-deoxycortisol (Score: 0)
```

```
{4} 2-acetylaminofluorene (Score: 90)
{5} 11-deoxycortisol (Score: 0)
{6} 17beta-estradiol (Score: 90)
{7} 11-deoxycortisol (Score: 0)
{8} 1,2-DMH (Score: 90)
{9} 11-deoxycortisol (Score: 0)
{10} AMI (Score: 90)
{11} acute tubular necrosis (Score: 90)
{12} axonal damage (Score: 90)
```

The first ten terms in the above sentence included in the *devel* dataset are designated as being outside a named entity class ("O"), but mis-recognised by the entity recognition model as a chemical spanning across multiple terms (in fact, the assembled term is an enzyme, which can be counted as a chemical but not considered as such by the dataset's context). The result of the fragmented individual approximate string matching process is a long list of poor matches annotated. Similarly, the fragmentation of "tabular" and "damage" caused them to be individually (and inaccurately) matched with individual terms in the dictionary.

These two errors highlight the need of reassembling neighbouring terms belonging to the same entity before attempting approximate string matching. Therefore, the grounding process was modified to first reassemble these neighbouring terms together (e.g. from "... $+ O + B - Chemical + I - Chemical + I - Chemical + O +$ ..." to "... $+ O + Chemical + O +$ ...") before matching the assembled term with the best approximation in the dictionary. The above example now becomes:

```
**Urine N - acetyl - beta - D - glucosaminidase ** {1} - - a marker of **tubular
    damage ** {2} ?
{1} 9-[[2-methoxy-4-[(methylsulphonyl)amino]phenyl]amino] -N,5-dimethyl- 4-
    acridinecarboxamide (Score: 86)
{2} acute tubular necrosis (Score: 86)
```

While the enzyme is still not matched by an appropriate term in the dictionary (likely simply does not exist), "acute tubular necrosis" is now a very good biomedical description of "tubular damage" in the original text. This demonstrates the improvement in approximate string matching quality by applying reassembly.

Overall, the grounding method with entity reassembly works fairly well, such as on the followed sentence:

```
BACKGROUND : **Calcitriol ** {1} therapy suppresses serum levels of parathyroid
    hormone ( PTH ) in patients with **renal failure ** {2} but has several drawbacks
    , including **hypercalcemia ** {3} and / or marked suppression of bone turnover ,
    which may lead to adynamic bone disease .
{1} Ca (Score: 90)
{2} renal failure (Score: 100)
{3} hypercalcemia (Score: 100)
```

While Calcitriol does not exist in the MESH dictionary, it does increase the body's intake of its closest match in the dictionary – calcium (Ca). While this is mostly a lucky match due to the lack of a less relevant term with a shorter edit distance, similar processes of inferring terms through biomedical relations have already been applied elsewhere to improve grounding, such as utilising contrastive information between proteins [5]. More systematically, database and machine learning-aided inference can used to effectively construct knowledge base from unstructured biomedical information [6]. In the above example, "renal failure" was also correctly matched with the corresponding term in the dictionary after entity reassembly before grounding.

Some other issues do persist after entity reassembly, such as the lack of ability to match complex acronyms with their full base words:

```
(Simple acronyms)
CBA / **Ca ** {1} male mice started on **AZT ** {2} 0 . 75 mg / ml **H2O ** {3} at 84
    days of age and kept on it for 687 days when dosage reduced to 0 . 5 mg / ml **H2O
```

```
     ** {4} for a group , another group removed from **AZT ** {5} to see recovery ,
    and third group remained on 0 . 75 mg .
{1} Ca (Score: 100)
{2} AZT (Score: 100)
{3} H2O (Score: 100)
{4} H2O (Score: 100)
{5} AZT (Score: 100)

(Complex, lesser-known acronyms)
RESULTS : In Nx dogs , **OCT ** {1} significantly decreased serum PTH levels soon
    after the induction of **renal insufficiency ** {2} .
{1} methoctramine (Score: 90)
{2} renal insufficiency (Score: 100)
```

From the original literature [7], "OCT" refers to Oxacalcitriol, which while is nevertheless not in the dictionary, was incorrectly interpreted as methoctramine. It is possible to resolve most acronyms into canonical forms through fixed or machine learning-established rules based on ontology knowledge [8].

Finally, common proportions of biomedical compound words and multi-word nouns can lead to incorrect groundings of tagged terms when a direct match a dictionary compound word does not occur. For example:

```
Histological examination on 9 of 10 mice with such **thrombocytopenia ** {1} showed
    changes compatible with **myelodysplastic syndrome ** {2} ( **MDS ** {3} ) .
{1} thrombocytopenia (Score: 100)
{2} Fanconi syndrome (Score: 86)
{3} DES (Score: 67)
```

Myelodysplastic syndromes concern bone marrows, while Fanconi syndrome describes kidney conditions. With the nature of the conditions entirely changed, this result from grounding is entirely erroneous. Methods to better distinguish semantic compositions of compound words ("compound bracketing") through unsupervised probabilistic models [9], as well as CRF post-processing and lexicon/dictionary-supported normalization [10] have been developed to improve accuracies of grounding on compound words and multi-word nouns.

c Identifying associations between disease and chemical mentions

The full abstract collection of PubMed texts on chemically induced disorders numbered 301,084,933 lines, with words already processed into a format identical to those used in Section a. With reference named entity classes available for each surface word, I applied the same feature template and L-BFGS training configurations of the improved entity recognition model from Section a, and split the abstract collection at breaks (empty lines) between sentences, as described in Figure 8 into three datasets: *training*, *validation*, *testing*. The first dataset will be used to train the CRF tagger, whose performance will be validated on the second dataset. Tags applied by the trained model on the last dataset will be used for the rest of the experiments.

Dataset	Number of Lines	Approx. Number of Sentences
training	50,180,846	1,757,828
validation	50,180,840	1,726,773
testing	200,723,247	7,089,377
Total	301,084,933	10,573,978

Figure 8: The approximate of words (and empty lines between sentences) in the three datasets after splitting.

References

[1] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2009, pp. 147–155.

[2] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed research international*, vol. 2014, 2014.

[3] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 109–116.

[4] seatgeek, "fuzzywuzzy: Fuzzy string matching in python," 2018 April. [Online]. Available: https://github.com/seatgeek/fuzzywuzzy

[5] J.-J. Kim, Z. Zhang, J. C. Park, and S.-K. Ng, "Biocontrasts: extracting and exploiting protein–protein contrastive relations from biomedical literature," *Bioinformatics*, vol. 22, no. 5, pp. 597–605, 2005.

[6] J. Shin, S. Wu, F. Wang, C. De Sa, C. Zhang, and C. Ré, "Incremental knowledge base construction using deepdive," *Proceedings of the VLDB Endowment*, vol. 8, no. 11, pp. 1310–1321, 2015.

[7] M.-C. Monier-Faugere, Z. Geng, R. M. Friedler, Q. Qi, N. Kubodera, E. Slatopolsky, and H. H. Malluche, "22-oxacalcitriol suppresses secondary hyperparathyroidism without inducing low bone turnover in dogs with renal failure," *Kidney international*, vol. 55, no. 3, pp. 821–832, 1999.

[8] N. Naderi, T. Kappler, C. J. Baker, and R. Witte, "Organismtagger: detection, normalization and grounding of organism entities in biomedical documents," *Bioinformatics*, vol. 27, no. 19, pp. 2721–2729, 2011.

[9] P. Pecina, "Lexical association measures and collocation extraction," *Language resources and evaluation*, vol. 44, no. 1-2, pp. 137–158, 2010.

[10] H.-C. Lee, Y.-Y. Hsu, and H.-Y. Kao, "Audis: an automatic crf-enhanced disease normalization in biomedical text," *Database*, vol. 2016, 2016.