Biomedical Information Processing (R214): main assignment report

Chongyang Shi *(cs940)*

April 12, 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 in exchange for lowering recall into the 50%s.

Ablated / Class	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. [**?**] 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 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

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.