Biomedical Information Processing (R214): main assignment report

Chongyang Shi *(cs940)*

April 3, 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

Ablated Class	None	*word*	*lemma*	*soundex*	*pos*	*chunk*
B-Chemical	0.9178	0.9345	0.9056	0.9015	0.9495	0.9210
O	0.9560	0.9471	0.9540	0.9531	0.9499	0.9557
B-Disease	0.8403	0.8242	0.8418	0.8387	0.8412	0.8396
I-Disease	0.7404	0.7152	0.7467	0.7506	0.7631	0.7509
I-Chemical	0.7556	0.6488	0.7569	0.7612	0.7906	0.7682
Macro-average	0.8420	0.8142	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. All results are from the *devel* dataset.

Ablated Class	None	*word*	*lemma*	*soundex*	*pos*	*chunk*
B-Chemical	0.6664	0.5583	0.6564	0.6520	0.5702	0.6652
O	0.9888	0.9888	0.9888	0.9887	0.9908	0.9894
B-Disease	0.6011	0.5514	0.5669	0.5561	0.5806	0.5992
I-Disease	0.6018	0.5530	0.5993	0.5952	0.6029	0.5952
I-Chemical	0.5961	0.5114	0.5950	0.5910	0.5938	0.5990
Macro-average	0.6908	0.6326	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. All results are from the *devel* dataset.

Ablated Class	None	*word*	*lemma*	*soundex*	*pos*	*chunk*
B-Chemical	0.7721	0.6992	0.7611	0.7567	0.7125	0.7725
O	0.9721	0.9675	0.9711	0.9706	0.9699	0.9723
B-Disease	0.7008	0.6607	0.6776	0.6687	0.6870	0.6993
I-Disease	0.6640	0.6238	0.6649	0.6639	0.6736	0.6641
I-Chemical	0.6665	0.5720	0.6662	0.6654	0.6782	0.6731
Macro-average	0.7551	0.7046	0.7451	0.071	0.7443	0.7562

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

References