Biomedical Information Processing (R214): Main Assignment

Chongyang Shi *(cs940)*

April 16, 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.

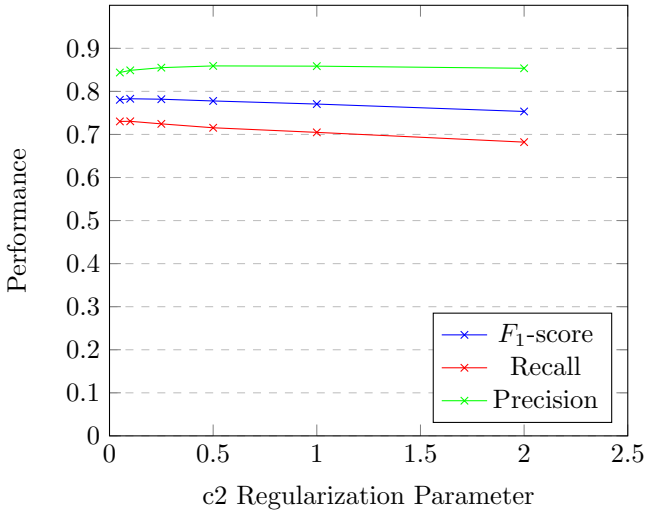

Figure 5: Resulting tagging performance on the *devel* dataset with models trained on different c2 values.

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 was 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. In the examples presented below, `**` will be used to annotate terms grounded in the sentence, with reference numbers point to the grounded entity from the dictionary after the sentence.

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 matched with separate, irrelevant 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 (which likely does not exist), with reassembly applied, "acute tubular necrosis" is now a very good biomedical description of "tubular damage" in the original text. To further resolve the lack of dictionary coverage over complex chemical constructs, efficient co-occurrence-based concept associations [5] could aid the grounding system in finding the best-matching entity based on functional groups in the term.

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). Although 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 [6]. More systematically, machine learning-based inference systems trained on biomedical databases can be used to effectively construct knowledge base from unstructured biomedical information [7]. 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 [8], "OCT" refers to Oxacalcitriol, which while nevertheless not in the dictionary, was incorrectly interpreted as methoctramine. It is possible to resolve most acronyms into canonical forms through fixed or dynamic rules based on ontology knowledge [9].

Finally, common proportions of biomedical composite words and multi-word nouns can lead to incorrect groundings of tagged terms when a direct match with the dictionary vocabulary 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 [10], as well as CRF post-processing and lexicon/dictionary-supported normalization [11] have been developed to improve accuracy when grounding compound words and multi-word nouns.

In general, a further improved grounding system may have the following features:

- Co-occurrence concept associations that can be queried efficiently [5] are used to find best-matching entities of complex chemical constructs if a direct match in vocabulary does not exist;
- For sources containing unstructured biomedical information, databases storing relational information [6] between entities are used to train inference agents [7] to better resolve terms without a direct match;
- To resolve uncommon acronyms that are not present in the dictionary, rules generated from ontology databases are applied to resolve the acronyms into canonical forms [9], which are more likely to encounter good matches during grounding;
- Unsupervised probabilistic models [10] and CRF-based normalization algorithms [11] are used to improve accuracy when grounding terms that are compositions.

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 unavailable for each surface word (defaulting to outside entities), the improved entity recognition model from Section a was used to perform entity recognition on a total of 301,084,933 surface words from 10,573,978 sentences. Tags generated through entity recognition were then cross-referenced

with the original surface words and their lemmas for grounding, which was conducted in parallelised batches to reduce memory footprint.

c.i Simple co-occurrence counts

With grounded entities grouped by the sentences they originated from, duplicate entities within the same sentence were removed, with the assumption that multiple mentions of a noun entity in the same sentence is primarily for clarity rather than emphasis [12]. Grounded entities in the sentence were then sorted in alphanumeric order to preserve consistency in composing co-mention pairs. The mention of each entity in each sentence was recorded globally as an occurrence of that entity. All possible in-order combinations of entities of length 2 were then generated with Python's `itertools.combinations`. Based on the context of the study, an additional filter was placed so that only co-mention pairs of a chemical and a disease are recorded. Implementations of these pre-processing steps can be found in Figure 14 of Appendix A.

By the number of occurrences, the ten most common grounded entities and the pairs thereof in the PubMed abstract texts are shown in Figure 8 and Figure 9.

Chemicals				**Diseases**			
Count	Probability	MESH ID	Name	Count	Probability	MESH ID	Name
248227	0.045%	D006859	H	166561	0.030%	D009369	tumour
135662	0.024%	D009569	NO	155455	0.028%	D004714	endometrial hyperplasia or cancer
102140	0.018%	D008694	METH	128998	0.023%	D064420	Toxicity
79642	0.014%	C034818	methyl 6,7-dimethoxy-4-ethyl-B-carboline-3-carboxylate	60568	0.011%	D020511	disorder of neuro-muscular transmission
78439	0.014%	C066430	3-aminopropyl-diethoxy-methyl-phosphinic acid	58780	0.011%	D007239	infections
70656	0.013%	C025136	phenylacetic acid	54426	0.010%	D047508	massive hepatocellular necrosis
68107	0.012%	D004298	Dopamine	51521	0.009%	D003643	deaths
65758	0.012%	D018698	glutamine	44556	0.008%	D012140	respiratory and cardiovascular depression
63797	0.011%	D005947	glucose	40723	0.007%	D031901	gestational trophoblastic disease
62878	0.011%	D002118	Calcium	38994	0.007%	D008103	cirrhosis of the liver

Figure 8: The top ten most commonly mentioned chemical and disease entities by appearance in number of sentences.

It was expected that by the sole criterion of number of occurrences, chemicals and syndromes of conditions commonly involved in the studies of environments and diseases will occupy a significant proportions of the top tens. This was observed in Figure 8 with the presence of hydrogen, nitrogen oxide, calcium, as well as "tumour", "infections" and "deaths". These concepts alone do not confer a significant amount of useful information about chemical-induced diseases. Entities related to more specific concepts do exist, however, such as two chemicals (C034818, C066430) associated with modulation of GABAA receptors, which affect anxiety-related mental states, suggesting the prominence of studies in chemically induced disorders among the abstract texts. Diseases related to different organs within the top ten are also good indicators of lesions involved in the studies.

Association		Chemical		Disease	
Count	Probability	MESH ID	Name	MESH ID	Name
6425	0.00116%	D003404	creatinine	D009369	tumour
5860	0.00105%	D003404	creatinine	D047508	massive hepatocellular necrosis
5180	0.00093%	D006859	H	D020511	disorder of neuromuscular transmission
3995	0.00072%	D004967	estrogen	D004714	endometrial hyperplasia or cancer
3850	0.00069%	D005472	5-FU	D004714	endometrial hyperplasia or cancer
3259	0.00059%	D006859	H	D012140	respiratory and cardiovascular depression
3162	0.00057%	D002945	cisplatinum	D004714	endometrial hyperplasia or cancer
3112	0.00056%	D004317	Doxorubicin	D004714	endometrial hyperplasia or cancer
2986	0.00054%	D006859	H	D004714	endometrial hyperplasia or cancer
2902	0.00052%	D002945	cisplatinum	D009369	tumour

Figure 9: The top ten most common associations of chemical and disease entities by appearance in number of sentences.

Apart from the potential links between creatinine, cisplatinum (Cisplatin), and tumours, most of the top co-mentions based on simple co-occurrence counts are associated with a single class of diseases – endometrial hyperplasia or cancer. This signals a key limitation of ranking associations based on simple co-occurrence counts: source texts with imbalanced compositions of study can make broad observations of associations between chemicals and diseases difficult. Once again the three associations between hydrogen and diseases provide very little useful information, as hydrogen is an almost-universal component of organic molecules.

To improve the effectiveness of co-occurrence count-based rankings, it may be possible to take ideas from object categorisation techniques by introducing relative locations (or in text mining, locations of grounded entities in sentences) of terms into consideration [13], or to borrow from machine translation techniques by applying arithmetic and geometric mean functions to co-occurrence counts from an imbalanced corpus [14]. Ultimately however, the aforementioned issues prompt more sophisticated statistical measures to be employed for better identifying the links between environmental chemicals and diseases.

c.ii Statistical measures

For the purpose of calculating occurrence probabilities, all sentences with any mention of a chemical or disease are counted into the total, even if a pair of chemical and disease cannot be established within the sentence. This is to maintain the consistency between chemical occurrences and disease occurrences. In addition to occurrences of individual and pairs of entities, four statistical measure are considered: Pointwise Mutual Information (PMI), Normalized Pointwise Mutual Information (NPMI), the Jaccard coefficient/index, and Symmetric Conditional Probability (SCP). With occurrences of pairs of chemical and disease entities recorded in dictionaries indexed by 2-tuples of IDs, calculations of these statistical co-occurrence metrics for pair of entities are straightforward through dictionary comprehensions, as shown in Figure 15 of Appendix A. Top ten associations as measured by these metrics are shown in Figures 10, 11, 12, and 13 respectively. Logarithms with base 2 were used for all relevant calculations.

PMI	Chemical		Disease	
	MESH ID	Name	MESH ID	Name
8.95518	C476217	cinacalcet HCl	D006961	hyperparathyroidism
8.72841	D005702	Galanthamine hydrobromide	D014826	vocal fold palsy
8.35901	D013390	Suxamethonium chloride	D005207	Fasciculations
8.35375	D011441	Propylthiouracil	D006980	hyperthyroidism
8.31867	D013390	Suxamethonium chloride	D012019	reflex sympathetic dystrophy
8.31393	C031942	argatroban	D013684	telangiectasis
8.16299	D005013	ethosuximide	D004832	absence seizures
8.07865	D008972	molindone	D002819	Choreoathetoid movements
7.99780	D007464	clioquinol	C538178	acrodermatitis enteropathica
7.83060	D004025	dicyclomine	D004211	intravascular coagulation

Figure 10: The top ten associations of chemical and disease entities as measured by Pointwise Mutual Information (PMI).

With mutual dependence between pairs of chemicals and diseases identified by PMI in Figure 10, a broad range of associations between medicinal chemicals and the diseases or conditions they treat or induce were established. For suxamethonium chloride, which induces fasciculations, top ten associations also covered its primary contra-indication (muscular dystrophy) [15]. To properly distinguish between indications and contra-indications among diseases or conditions, an enhanced dictionary or a set of defined relations [6] is required.

NPMI	**Chemical**		**Disease**	
	MESH ID	Name	MESH ID	Name
0.594802	D002248	carbon monoxide	D011041	poisoning
0.550872	C476217	cinacalcet HCl	D006961	hyperparathyroidism
0.535089	D014673	vecuronium bromide	D020879	neuromuscular blockade
0.530970	D010622	phencyclidine	D006996	hypocalcemia
0.521261	D004025	dicyclomine	D004211	intravascular coagulation
0.515581	D018170	sumatriptan	D008881	Migraine
0.507942	D007980	Levodopa	D055154	dysphonia
0.501047	C010012	adriamycinone	D000160	adverse effect on the proximal eighth nerve
0.491024	D002996	clomiphene citrate	D011085	polycystic ovary syndrome
0.490998	D011441	Propylthiouracil	D006980	hyperthyroidism

Figure 11: The top ten associations of chemical and disease entities as measured by Normalized Pointwise Mutual Information (NPMI).

With Normalized PMI (NPMI) to constrain the level of co-occurrence indicated by PMI, PMI's occasional over-sensitivity to low frequency data is mitigated [16]. Some pairs of medicinal chemicals and diseases remain in the top ten, while others have been replaced. Replacements include the fairly obvious poisoning effect of carbon monoxide, and an adverse nerve effect of adriamycinone (Doxorubicinone) [17, ch. 55] if improperly administered via intrathecal injection.

Jaccard	**Chemical**		**Disease**	
	MESH ID	Name	MESH ID	Name
0.102216	D002248	carbon monoxide	D011041	poisoning
0.0690245	D007980	Levodopa	D055154	dysphonia
0.0632168	D010622	phencyclidine	D006996	hypocalcemia
0.0556352	D003404	creatinine	D047508	massive hepatocellular necrosis
0.0497766	D014673	vecuronium bromide	D020879	neuromuscular blockade
0.0476688	C047426	venlafaxine	D001281	Atrial Fibrillation
0.0467789	D007538	Isoniazid	D014376	tuberculosis
0.0466248	D000244	ADP	D001791	platelet aggregations
0.0442772	D002245	CO2	D011020	Pneumocystis pneumonia
0.0428755	D004025	dicyclomine	D004211	intravascular coagulation

Figure 12: The top ten associations of chemical and disease entities as measured by the Jaccard index.

The Jaccard index disregards the shapes and distributions of inputs [18] to mitigate the undesirable influences thereof when identifying concept associations. A majority of the top ten associations determined by the Jaccard Index are also present in NMPI's top ten. Among the differences, it is notable that the arterial carbon dioxide strengthen effects of pneumocystis pneumonia [16] has been identified. The direction of this association is opposite to most others with chemicals as sources of effects.

Finally, being the most focused on conditional cause and effects, Symmetric Conditional Probability (SCP) returned its top ten associations broadly identical to those already identified by NPMI and the Jaccard index. The only new addition appears to be from the use of streptozotocin to induce diabetes in experimental animals [19].

c.iii Limitations

In addition to limitations of ranking associations through simple co-occurrence counts discussed in c.i, two further areas of limitations exist in identifying chemically induced diseases through the method studied

SCP	Chemical		Disease	
	MESH ID	Name	MESH ID	Name
0.0396361	D002248	carbon monoxide	D011041	poisoning
0.0177305	D007980	Levodopa	D055154	dysphonia
0.0144888	D010622	phencyclidine	D006996	hypocalcemia
0.0111154	D003404	creatinine	D047508	massive hepatocellular necrosis
0.00975593	D013311	streptozotocin	D003920	Diabetic
0.00924036	D014673	vecuronium bromide	D020879	neuromuscular blockade
0.00903808	D000244	ADP	D001791	platelet aggregations
0.00865604	C047426	venlafaxine	D001281	Atrial Fibrillation
0.00822348	D018170	sumatriptan	D008881	Migraine
0.00799549	D007538	Isoniazid	D014376	tuberculosis

Figure 13: The top ten associations of chemical and disease entities as measured by Symmetric Conditional Probability (SCP).

in this report.

The first area of limitation lays on the limited effectiveness of automated CRF entity recognition and grounding through approximate string matching. Imperfect entity recognition will cause some entities that can otherwise be appropriately grounded to be mislabelled as outside named entities, which will not be grounded. Limitations of grounding as discussed in Section b will supply improperly-grounded entities into the calculations of statistical co-occurrence metrics, which in turn affect the accuracies of top associations identified. Entity recognition and grounding algorithms with improved performance are required to mitigate this area of limitation.

The second area of limitation was caused by the simplifications made when calculating probabilities and statistical metrics. Duplicates of the same grounded entity within the same sentence were deducted from consideration. Orders of occurrences between different grounded chemical and disease entities within the same sentence, or between different sentences from the same abstract are not considered, which may hold useful information related to their associations [13].

As each entity or pair of entities is counted at most once in each sentence, the total number of sentences was used as the denominator in probability calculations, so that probabilities of all entities or pairs of entities will have a common denominator. This additional simplification resulted in the omission of sentences with only chemical entities or only disease entities grounded. While at sentence-level they cannot be said to have any associations, additional associations may have been possible if the same sentences within the same abstract could be considered together.

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] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Facta: a text search engine for finding associated biomedical concepts," *Bioinformatics*, vol. 24, no. 21, pp. 2559–2560, 2008.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] H. H. Clark and C. Sengul, "In search of referents for nouns and pronouns," *Memory & Cognition*, vol. 7, no. 1, pp. 35–41, 1979.

[13] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[14] X. Zhu, Z. He, H. Wu, C. Zhu, H. Wang, and T. Zhao, "Improving pivot-based statistical machine translation by pivoting the co-occurrence count of phrase pairs," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1665–1675.

[15] National Institute for Health and Care Excellence, "Suxamethonium chloride." [Online]. Available: https://bnfc.nice.org.uk/drug/suxamethonium-chloride.html

[16] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, pp. 31–40, 2009.

[17] K. Kompoliti and S. S. Horn, "Drug-induced and iatrogenic neurological disorders," in *Textbook of Clinical Neurology (Third Edition)*. Elsevier, 2007, pp. 1285–1318.

[18] L. Leydesdorff, "On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index," *Journal of the Association for Information Science and Technology*, vol. 59, no. 1, pp. 77–85, 2008.

[19] A. A. Rossini, A. A. Like, W. L. Chick, M. C. Appel, and G. F. Cahill, "Studies of streptozotocin-induced insulitis and diabetes," *Proceedings of the National Academy of Sciences*, vol. 74, no. 6, pp. 2485–2489, 1977.

Appendices

A Excerpts of implementation from statistical processing of grounded entities.

```
mentions = defaultdict(int)
associations = defaultdict(int)
total_lines = 0 # Excluding lines without any grounded entities.
while True:

    assoc = assoc_file.readline()
    if not assoc:
        break

    assoc = assoc.strip()
    if len(assoc) > 0:

        # Multiples of the same in a sentence are only counted once.
        elements = sorted(list(set([i.strip() for i in assoc.split(",")])))

        # Generate pairs.
        pairs = combinations(elements, 2)

        # Tally total.
        total_lines += 1

        for pair in pairs:

            # Individual occurrences.
            for item in pair:
                mentions[item] += 1

            # If set, only consider pairs of chemical and disease.
            # Dictionary composition: {id: (name, type), ...}
            if MIXED_ONLY:
                first_t = dictionary[pair[0]][1]
                second_t = dictionary[pair[1]][1]
                if first_t.lower() == second_t.lower():
                    continue

            # Co-occurrences.
            associations[pair] += 1
```

Figure 14: Pre-processing of grounded entities in sentences separated by new lines.

```
# This keeps probabilities consistent between chemical and disease entities.
element_prob = {i: mentions[i] / total_lines for i in mentions}
joint_prob = {i: associations[i] / total_lines for i in associations}
pmi = {i: log(joint_prob[i] / (element_prob[i[0]] * element_prob[i[1]]), 2) for i in joint_prob}
npmi = {i: pmi[i] / (-1 * log(joint_prob[i], 2)) for i in pmi}
scp = {i: joint_prob[i] ** 2 / (element_prob[i[0]] * element_prob[i[1]]) for i in joint_prob}
jaccard = {i: associations[i] / (mentions[i[0]] + mentions[i[1]] - associations[i]) for i in associations}
```

Figure 15: Calculations of statistical occurrence metrics for pairs of chemical and disease entities.