Adapting Cocoa, a multi-class entity detector, for the CHEMDNER task of BioCreative IV

S. V. Ramanan and P. Senthil Nathan
RelAgent Private Ltd., 56 Venkatratnam Nagar, Adyar, Chennai 600020, India
{ramanan,senthil}@npjoint.com
http://www.npjoint.com/

Abstract. Cocoa is an existing multi-class entity detection system for the biomedical domain. We adapted the output of this system for the CHEMDNER task, primarily by trimming entities and excluding generic entity terms, both of which are irrelevant to detection of core names as defined in the task. The system itself required only small changes, primarily to add dictionary entries or handle unusual entity mentions in abstracts. The final performance against both training and development datasets exceeded 80% for the CDI task for the key metrics, namely the F-score and the FAP. Performance in the CEM subtask was 2% lower.

Key words: chemical entity detection, CHEMDNER, multi-class tagger, rule/morphology-based system, Cocoa

1 Introduction

The annotations in the datasets provided in the CHEMDNER task of BioCreative IV provide markups for entities that in some way be correlated to chemical structures. This defines a narrow class of textual entities. Previous efforts in marking up chemical entities [2, 3] and approaches to chemical entity detection are reviewed in [1].

Cocoa [4] is a multi-class entity detection system for medical text that covers major classes such as proteins, protein parts, chemicals, anatomical parts, diseases, parameters, values and some others. While we modified the system slightly to cover increase coverage of the CHEMDNER corpus, the major problem in the output of this system with respect to the task was the large class of entities that were labeled as chemicals. These included chemical mixtures in common use, such as the term ‘solvents’, terms that described function such as ‘antibiotics’ as well as terms with imprecise mappings to chemical structure. We accordingly added a post-processing module to eliminate or trim chemical terms detected by the Cocoa system to correspond to the narrow class of entities defined by CHEMDNER task and annotations. The results were encouraging, with key metrics around 80% for the CDI subtask of CHEMDNER.
2 Systems description and methods

We describe first the chemical entity-specific processing of the input document by Cocoa. Only a few changes were made in the system with respect to the CHEMDNER task, but the general methods may be of interest. The system consists of sequential modules for:

(a) Sentence splitting and tokenization: The splitter handles the usual exceptions such as periods following abbreviations (‘Mr.’, ‘et al.’), entities that begin with a lowercase letter (‘mRNA’), and lowercase expanded greek letters (‘alpha’). Certain hyphenated words are split (‘X-activated’); expanded in-line journal references are excised to avoid complications in later processing stages.

(b) Acronym detection: We use a dynamic programming method for detection of acronyms. A simple morpheme detector is used, and first letters of words and morphemes are used for scoring the match between an acronym and its possible expansions. Substitution of element symbols for names (‘Na’ for ‘sodium’ in ‘Na pump’ for example) is also handled.

(c) POS marking and chunk detection: The Brill tagger and fnTBL are used for POS and chunk detection respectively. A post-chunker module handles common tagging and chunking errors in the biomedical domain, e.g. “increase” as verb vs adjective, and gerunds (VBG) as nouns vs verbs.

(d) Tagging tokens: For chemicals, we use a combination of dictionaries and infix/suffix detection methods. We have manually compiled an infix dictionary (≈ 400 entries) and a suffix dictionary (≈ 700 entries). A dictionary for chemical groups and names for common chemicals (‘emulsions’) has about 2000 entries. These are supplemented by a dictionary of tradenames and irregular chemical names, with ≈ 18000 entries. A dictionary of false positives is also kept for entities mislabeled as chemicals by the infix/suffix methods - examples are ‘elaborate’ (‘borate’), ‘trichlorobacter’ (‘chloro’), ‘femoracetabular’ (‘acet’), ‘ophthalmic’ (‘phthal’), ‘periodontal’ (‘iodo’), ‘pathoformic’ (‘formic’), ‘epicranial’ (‘picr’), ‘proxy’ (‘oxy’), and ‘asteroid’ (‘steroid’). The false-positive dictionary has ≈ 400 entries. We note that the tagging module also marks up proteins, body parts, diseases, and other classes; however, chemicals are marked up ahead of these other entity classes, but after detection of parameters (‘IC50’) and values (‘50 nM’). Finally orthographically distinct tokens (marked by presence of caps, numbers or other symbols in the token, or tokens with otherwise rare character bigrams) that have not yet been identified as entities are tagged as O-terms.

The suffix dictionary was constructed primarily by manual examination of entries from dictionaries (such as the CTD chemical dictionary) after sorting the entries by 4- or 5-character suffixes. Official suffix recommendations such as the USAN stem list were also consulted. Infix dictionaries were initially constructed by a domain expert, vetted by a computational linguist, and then checked for false-positives against online dictionaries such as OneLook.

(e) Detecting multi-word expressions: First, we merge adjoining chemical tokens. Certain word-specific adjectives are then merged (‘fuming H2SO4’, ‘dry ether/ice’, ‘Lewis acid’). Common chemical adjectives are also handled (‘aliphatic’, ‘aromatic’, ‘enolic’, ‘saturated’, ‘disubstituted’). O-terms following certain chem-
BioCreative IV - CHEMDNER track

Chemical classes are recognized ('vitamin B1', 'Triton X-100'). Terms in the 'molecule' category, such as 'hormone', are recognized as chemicals if they are preceded by relevant qualifiers ('sex hormones', 'prolactin-inhibiting hormone'). This also holds for some geometrical parts ('ring' - 'aromatic ring'). A limited amount of word-sense disambiguation is also taken care of in this module, e.g., 'radical' ('free radical' vs. 'radical prostatectomy'). Chemical premodifiers in parenthesis ['(3R,3S)'] are merged with succeeding chemical entities.

(f) Coordination and apposition: Conventional techniques are used for identifying noun phrases (NPs) in coordination (e.g., 'NP, NP, and NP'). We use some care in not marking possible appositional phrases as coordinated (e.g., 'each of the ERK cascade components, Raf, MEK, and ERK'). We then mark O-terms which can be identified as entities through an appositional context (NP, such as O-term; O-term, which is a NP), where the context provides clues on the tag of the O-term ('enzyme', 'drug'). Frequently occurring, but non-standard, definitions of acronyms are also recognized ['noradrenaline (NA; 50 microM)'].

(g) Acronym extender: O-terms that are variants of acronyms are recognized by this module. These include plurals ('ACs' from 'AC'), suffix variants ('EGR2' from 'EGR1') and greek-variants ('PKM-zeta' from 'PKM').

(h) Formula detection and other: This module detects chemical formula. These include aminoacid chains, sugars (MurNAcGlcNAc) as well as simple terms ('CHCl3'). Simple extensions such as the use of 'Me' for 'methyl', use of the generic 'R' group ('R-N(CH2COOH)2') and charged ionic forms ['Ca(2-)'] are also recognized. Finally, chemicals defined by established labels for pharmaceutical companies (SCH-1234) are also handled.

The output from these modules (the 'Cocoa' system) contains a large number of chemical entities that are outside the scope of the CHEMDNER task. These include function-based descriptors ('antibiotics', 'fungicides', 'reductants', 'electron donors'), chemical mixtures ('paints', 'lotions'), broad categories ('lipids', 'cations') and state descriptions ('vapor', 'gas'). There are also a large number of entities which are too broad for structural definition as defined by the absence of annotations for such entities in the CHEMDNER corpus. Examples are 'heterocyclic compounds', 'nitroheterocycles', 'peroxynotrates' and 'homogalacturonans'. Many of these excluded entities are defined in the annotation notes for the CHEMDNER task, but the majority were discovered by manual comparison with the gold annotations. Further, many adjectives need to be eliminated from the core entity for this task e.g. 'oxidized' 'hydrated' and 'tridentate', as do many headwords ('polysaccharides', 'adjuvants', 'chelants', 'substituents'). Altogether, we had about ≈ 450 excluded adjectives, a similar number of excluded headwords, and ≈ 1200 excluded generic entity names. It must be noted that these excluded words came primarily from their definition or inclusion in the Cocoa dictionaries or multi-word expression modules as parts of chemical entities. Finally, acronyms were retained in the chemical entity list based on the retention of the headword of the expansion after pruning.

As Cocoa is a multi-class tagger, certain entities tagged as 'protein parts' were also redefined as chemical entities. These included single or multiple amino
acids or residues, motifs, ‘N’ or ‘C’ in ‘N-terminal’ in the protein-part class. Some entities defined by the system as proteins are marked up as chemicals in CHEMDNER, such as angiotensin, bradykinin, gentisin, and oxytocin. The ‘N’ in proteins such as ‘N-glycosyltransferase’ is also marked up for the task as a chemical, as are both ‘O’ and ‘ethoxyresorufine’ in ‘ethoxyresorufine-O-deethylase’. Similarly ‘N’ and ‘O’ in process terms, such as ‘N- and O-demethylation’, are also tagged. Finally, entities marked as an ‘anatomical entity’ by Cocoa such as ‘bile acid’ are also tagged as chemicals for the task. All these latter special cases were found and marked up by manual examination of precision and recall errors with respect to the gold annotations.

After excluding or pruning words as described above, we added a second stage of post-processing for the CHEMDNER task. This consists primarily of splitting certain combination terms in the Cocoa output such as ‘dopamine precursor L-3,4-dihydroxyphenylalanine’, ‘citrate reduced gold’ or ‘(14)C-labeled LO’. These terms arise in Cocoa as single terms due to a system design decision to markup the longest possible sequence of words as a single entity (to facilitate event extraction). Another need for a split is from expressions such as ”the isoflavone genistein”, where a class descriptor precedes a particular entity. Sometimes the descriptors are themselves too broad, such as in ‘monoterpene hydrocarbons’ and are discarded. The need and the rules for such splitting were again derived from manual examination of the gold annotations.

Ranking was done for these entities by modeling precision and recall errors. Character counts were calculated for errors, and a cutoff was set based on these counts. Most errors came from O-terms, i.e. terms all in CAPS, and we assigned an initial score to an entity as equal to the numbers of capitalized letters, hyphens and spaces in an entity. We also plotted the frequency of errors based on the length of the entities. There was a small dependence of error on the length of the entity, with errors peaking at about 8-9 character length. Unmatched parentheses also were a minor source of error. Both the latter cases were given a lesser weight in the score (score increased by 2-5). Certain terms such as ‘phenolic’ and ‘triterpene’ were determined to have some inconsistency in markup in the gold annotations, and the score was increased (by 5) if these terms were detected. The final rank was determined by sorting the entities according to their score as determined above.

3 Discussion

We added a post-processing module to mark up entities according to the conventions or rules followed in the CHEMDNER task. The Cocoa system itself was modified only slightly to account for missed entities by either dictionary augmentation or modification in the multi-word entity detection module. We initially used the training dataset for refining the system based on its earlier release. The results on the development set (when released) was about 2% lower for all measures. Subsequently, we alternated between training and development sets for improving the system while testing against the other dataset. The final
micro-evaluation figures against the development dataset for the CDI and CEM subtasks are:

**CDI subtask:**

- Micro prec.: 0.84658
- Micro recall: 0.86089
- Micro F-scr.: 0.85368
- Micro Avrg P: 0.75972
- Micro FAP-s.: 0.80396

**CEM subtask:**

- Micro prec.: 0.85349
- Micro recall: 0.82927
- Micro F-scr.: 0.84121
- Micro Avrg P: 0.72834
- Micro FAP-s.: 0.78072

The results for the training dataset were marginally (1%) higher.

The CHEMDNER task had a reasonably large dataset (7000 abstracts in training+development) that allowed for rules to be formed by inspection of the annotations. Our manual inspection showed that inter-annotator consistency was probably quite high (different annotations for the same entity were about 2-3%). Altogether, the annotations enabled us to achieve ≥ 80% for key measures (F-score and FAP) on the CDI task; the scores for the CEM task were lower primarily due to poor recall, implying that all instances of an entity in a document were not detected. Inspection of precision and recall errors also showed that the majority of errors arose from incorrect or missed detection of O-terms, i.e. orthographically defined entities. We are still exploring the reasons for these errors.

Extracting chemical entities from full-length articles is an interesting next step. Tagging reactions and extracting quantitative parameter-value-reaction triplets would be of relevance to both annotators and SAR-queries.

### 3.1 Figures

### References

Fig. 1. Flow chart of the Cocoa system. The main modification for the CHEMDNER task is the addition of post-processing module (in gray) to filter out chemical entities that are too generic for correlation with structures, an aspect of the task.