# A Knowledge-Poor Approach to BioCreative V DNER and CID Tasks

Firoj Alam[1], Anna Corazza[2], Alberto Lavelli[3], and Roberto Zanoli[3]

[1] Dept. of Information Eng. and Computer Science, University of Trento, Italy
[2] Dept. of Electrical Eng. and Information Technologies, Università di Napoli Federico II, Italy
[3] Fondazione Bruno Kessler, Trento, Italy
alam@disi.unitn.it,anna.corazza@unina.it,{lavelli,zanoli}@fbk.eu

**Abstract.** The report describes our participation in the BioCreative V track #3, both in Disease Named Entity Recognition and Normalization (DNER) and in Chemical-induced diseases relation extraction (CID). For both tasks, we have adopted a general-purpose approach based on machine learning techniques integrated with a limited number of domain-specific knowledge resources and using freely available tools for preprocessing data. Crucially, the system only uses the data sets provided by the organizers. After comparing different configurations, the one giving the best compromise between effectiveness and efficiency has been chosen. We report the results of the experiments performed during the development phase for comparing different configurations. The results of the official submission are in line with those on the development set.

**Key words:** Named Entity Recognition and Normalization, Relation Extraction, Machine Learning.

## 1 Task Description

The BioCreative V[4] task #3 consists of the automatic extraction of chemical-disease relations (CDR) from PubMed articles. It includes two subtasks: Disease Named Entity Recognition and Normalization (DNER) and Chemical-induced diseases relation extraction (CID). The dataset consists of PubMed abstracts collected from the curated articles in the Comparative Toxicogenomics Database [2]. More details on the task can be found in [5, 4].

## 2 System Architecture

In Figure 1, we present the system architecture and provide the details of each step in the following subsections.

[4] http://www.biocreative.org/

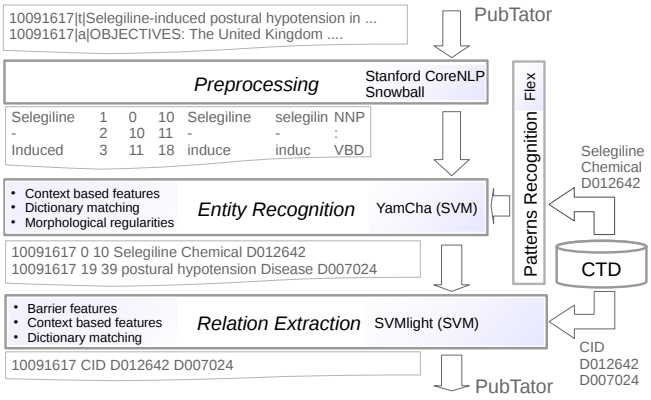


Fig. 1: System architecture.

### 2.1 Preprocessing

We use Stanford CoreNLP[5] to obtain the base form of the words, their part of speech (POS) and lemma, and to perform sentence segmentation. The Snowball tool[6] is instead used for producing the stem of the words.

### 2.2 CTD

The Comparative Toxicogenomics Database (CTD)[2] is a publicly available database that aims to advance understanding about how environmental exposures affect human health. It provides manually curated information about chemicals, and diseases that, in our approach, are used to capture the different ways the entities are mentioned in texts. Chemical and disease names are first extracted from the database, and then converted into regular expression patterns. After that, Flex[7] generates the scanners to recognize the mention patterns. We use also the chemical-disease relationships database. It includes chemical-disease pairs and it has been exploited in the Relation Extraction subtask to know the entities in texts that have a relation in the CTD.

### 2.3 Named Entity Recognition

Entity recognition is an intermediate step for automatic CDR extraction, performed in two steps: (i) detecting the mentions to the entities in texts (mention detection), and (ii) selecting the best-matching MeSH ID (normalization).

**Mention detection** is complex because of the many ways in which an entity can appear in texts. For example *acetylsalicylic acid* could be reported using the

[5] `http://nlp.stanford.edu/software/corenlp.shtml`
[6] `http://snowball.tartarus.org/`
[7] `http://flex.sourceforge.net/`

systematic nomenclature (typically multiword terms with large spelling variability), describing the compound in terms of its structure (i.e., *2-(Acetyloxy)benzoic Acid)*, rather than non-systematic nomenclature (i.e., *aspirin*) or synonyms like *acetylsalicylate*. To classify mentions we combine 3 approaches:

- *dictionary matching* consists in finding a mention in text by comparing it with a dictionary. We use the chemical and disease vocabulary to match both chemicals and diseases in texts with the CTD.
- *exploiting morphological regularities* is done by using the prefixes and suffixes of the tokenized words, and the words stem. The suffix *-emia* is for example typical of diseases (e.g., ischemia), while the prefix *meth-* is useful for chemicals discrimination (e.g., methylxanthine)
- *context based features* are implemented by considering a window of length 4 and consists of the current token, one token before and two tokens after.

Such approaches are combined by means of YamCha[8], an open source customizable text chunker based on Support Vector Machines (SVMs). With YamCha it is possible to redefine the feature sets (window-size) and we considered the stem and the POS of the current token, its prefixes/suffixes, and whether or not the token matches with the vocabulary. The system also considers the POS of the token before the current token, the prefixes/suffixes of the 2 following tokens, and the entity labels assigned during the tagging to the 2 tokens before.

**Normalization** selects the best-matching MeSH ID by means of *dictionary matching* based on CTD (see the pseudocode in Algorithm 1).

**Algorithm 1** Pseudocode for Mention Normalization. pred_mentions are the mentions recognized by the NE system. gold_mentions, ctd_chemical, ctd_disease are dictionaries in which mentions are associated with MeSH IDs

```
Input: gold_mentions, ctd_chemical, ctd_disease, pred_mentions
Output: normalized_mentions
  procedure normalization(gold_mentions, ctd_chemical, ctd_disease, pred_mentions)
    for all mention_i ∈ pred_mentions do
      if mention_i ∈ gold_mentions then
        mention_i_id ← gold_mentions.get(mention_i)
      else if mention_i = chemical and mention_i ∈ ctd_chemical then
        mention_i_id ← ctd_chemical.get(mention_i)
      else if mention_i = disease and mention_i ∈ ctd_disease then
        mention_i_id ← ctd_disease.get(mention_i)
      else
        mention_i_id ← −1
      end if
    end for
  end procedure
```

### 2.4 Relation Extraction

Relation extraction is formulated as a binary classification problem in the vector space model. A Feature Vector (FV) is therefore defined for each instance of the

[8] `http://chasen.org/~taku/software/yamcha/`

problem, corresponding to a pair of chemical and disease entities, constructed by the juxtaposition of the FVs corresponding to the two entities and a set of *relation features*, which take into account both entities. The classifier will decide whether a relation exists between the two entities.

As each entity is associated with one or more mentions, we define a FV for each mention, and then combine them to obtain the FV of the entity. All the features we consider are Boolean, and mention FVs are combined by means of an OR operation. Again, each mention FV is built by considering the OR of each token FV, which are based on token characteristics and on word and POS unigrams, bigrams and trigrams from a window of length 5 centered in the token. In addition to these more standard features, in some configurations of our system we also considered Barrier Features (BFs) [1].

As mentioned above, in addition to entity features, we also consider 4 binary relation features, depending on both entities. They signal whether the entity pair is listed as a positive chemical-disease relation in the Comparative Toxicogenomics Database [2], whether all mentions occur in the same sentence and if such sentence is in the title or in the abstract. Note that the first of these relation features is the only feature based on an external knowledge source. As relation features are more likely to predict the existence of an actual relation, we overweigh them with respect to entity features by introducing a Relation Features Weight (RFW) larger than 1.

Feature selection is needed because of the potentially very large number of $n$-grams and BFs. We then only keep features occurring a large enough number of times in the corpus used for the BioCreative IV CHEMDNER task [3]. The threshold to decide which features to prune is set to the mean of all counters.

Classification is performed using SVMlight[9], while we applied a post-processing phase to recognize those relations that, even though annotated in the training set, have not been identified by the system in the test set.

## 3 Experiments

To find the best system configuration for the official evaluation, we trained both the systems for DNER and CID on the provided training data set and tested it on the development set. The following sections describe the experiments done.

### 3.1 DNER

Given that the distributed data set contains annotations for both chemical and disease entities, we have implemented a single system for recognizing both the entity types in the DNER and CID subtask even though DNER does not require it. Table 1 reports the results of chemical-disease mention detection and normalization with the default configuration of the system described in Section 2.3. These results were compared with 2 baselines: baseline1 is calculated matching

[9] `http://svmlight.joachims.org/`

the chemical and disease mentions in the texts with the CTD and normalizing them with the MeSH ID associated to those mentions in the CTD; baseline2 is calculated by training the system on the tokenized articles in the training set without any additional source of information. Finally, we retrained the system on the training set plus the development set and evaluated it on the test set.

Table 1: Results of entity normalization and mention detection (in brackets) on the development set. The last row shows the DNER values on the test set.

| | P | R | $F_1$ |
|---|---|---|---|
| Chemical | 88.11(92.24) | 88.05(86.95) | 88.08(89.51) |
| Disease | 84.31(83.50) | 77.57(80.75) | 80.80(82.10) |
| Chemical+Disease | 86.09(88.32) | 82.26(84.20) | 84.13(86.21) |
| baseline1 | 76.03(81.07) | 64.01(69.47) | 69.51(74.82) |
| baseline2 | 88.14(78.40) | 64.13(64.21) | 74.24(70.60) |
| DNER | 86.82 | 81.84 | 84.26 |

To measure the impact of the different source of information to the final system performance on the development set, one type of information is removed at a time from the system default configuration. Table 2 reports these results.

Table 2: Variation in results of entity normalization and mention detection (in brackets) when one type of information is removed at a time.

| | P | R | $F_1$ |
|---|---|---|---|
| Chemical+Disease Entities | | | |
| -dictionary matching | +1.84(-1.32) | -17.62(-5.95) | -9.62(-3.81) |
| -context based features | -3.46(-9.41) | -0.3(-2.26) | -1.84(-5.82) |
| -morphological regularities | -1.63(-0.43) | +0.59(-0.23) | -0.43(-0.48) |
| Disease Entities | | | |
| -dictionary matching | +0.89(-2.66) | -13.94(-4.57) | -7.95(-3.66) |
| -context based features | -2.53(-10.53) | -0.75(-4.03) | -1.58(-7.30) |
| -morphological regularities | -0.89(-0.39) | +0.70(-0.61) | -0.04(-0.50) |

### 3.2 CID

We compared the performance of different configurations of the approach described in Section 2.4 on the development set. Given that in the training data there is a strong unbalance between negative and positive examples, we set the cost-factor parameter in SVMlight to their ratio, i.e. 4.3. On the basis of preliminary experiments (not reported here for space reasons), we set RFW to 5, apply lemmatization to deal with data sparseness and consider the linear kernel for SVM. The most difficult choice has been to decide whether to use BFs and a list

of stopwords. In Table 3 we report the experimental results obtained with gold standard mentions. As all four settings perform very similarly, for the sake of both efficiency and robustness, we tried to minimize the number of features and introduced the stopword filtering, but not the BFs. When such configuration is applied to the mentions automatically extracted, we obtain: P=36.94, R=56.03, $F_1$=44.52. This performance is in line with the one obtained on the test set in the official evaluation (P=35.39, R=56.47, $F_1$ = 43.51).

Table 3: Experimental results for relation extraction with gold standard mentions.

| | With stopword filtering | | | Without stopword filtering | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| With BFs | 37.32 | 83.20 | 51.53 | 39.12 | 81.03 | 52.77 |
| Without BFs | 39.20 | 76.98 | 51.95 | 40.81 | 74.60 | 52.76 |

## 4 Conclusions

We proposed an approach to the problem which is characterized by a minimal requirement of specialized knowledge source and by a more than acceptable speed (less than 4 secs for DNER, and about 9 secs for CID on a PC with Intel Xeon E5-2600 with 32GB of RAM during the official evaluation). Further experimentation is required to optimize the choice of the most effective features by means of a composition of feature design and feature selection. Last, but not least, a more sophisticated choice of potential relation candidates among all possible entity pairs could help improving performance.

## References

1. Alicante, A., Corazza, A.: Barrier features for classification of semantic relations. In: Proc. of the 8th Recent Advances in Natural Language Processing (2011)
2. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., et al.: The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Research 43(D1), D914–D920 (2015), http://nar.oxfordjournals.org/content/43/D1/D914.abstract
3. Krallinger, M., Rabal, O., Leitner, F., et al.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. J. Cheminformatics 7(S-1), S2 (2015), http://dx.doi.org/10.1186/1758-2946-7-S1-S2
4. Li, J., Sun, Y., Johnson, R., et al.: Annotating chemicals, diseases, and their interactions in biomedical literature. In: Proceedings of the fifth BioCreative challenge evaluation workshop (2015)
5. Wei, C., Peng, Y., Leaman, R., et al.: Overview of the BioCreative V Chemical Disease Relation (CDR) task. In: Proceedings of the fifth BioCreative challenge evaluation workshop (2015)