An Ensemble Algorithm for Sequential Labelling: A Case Study in Chemical Named Entity Recognition

Chen-Kai Wang[1], Hong-Jie Dai*[2,3], Jitendra Jonnagaddala[4,5], Emily Chia-Yu Su[1]*

[1]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan, R.O.C., [2]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan, R.O.C, [3]Interdisciplinary Program of Green and Information Technology, National Taitung University, Taitung, Taiwan, R.O.C, [3]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan, R.O.C., [4] School of Public Health and Community Medicine, UNSW Sydney, Australia
[5] Prince of Wales Clinical School, UNSW Sydney, Australia

denniscckwang@gmail.com; *hjdai@nttu.edu.tw; *emilysu@tmu.edu.tw; z3339253@unsw.edu.au

Abstract. Ensemble methods are learning algorithms that classify new data points by synthesizing the predictions of a set of classifiers. Many methods for constructing ensembles have been proposed such as weighted voting, manipulations of training samples, features, or labels. The paper proposes a novel ensemble algorithm which constructs ensembles by manipulating the label set given to the learning algorithm and then classifies a new dataset by a voting algorithm specifically designed for sequential labelling task. The dataset released in the BioCreative V.5 CEMP (Chemical Entity Mention recognition) task was used to evaluate the performance of proposed algorithm. The results revealed that the proposed algorithm can improve the precision and F-score.

Keywords. Chemical named entity recognition; ensemble method; sequential labelling problem

1 Introduction

In standard supervised sequential labelling task problems, a machine learning algorithm is given with a sequence of training examples of the form $t_i = \{(\mathbf{o}_1, y_1), \ldots (\mathbf{o}_n, y_n)\}$ and the goal of the learning algorithm is to find a function f so that $f(\mathbf{o}) = \boldsymbol{y}$. In named entity recognition (NER)

like the BioCreative V.5 CEMP (Chemical Entity Mention recognition) task [1], $\mathbf{o}_i$ is typically a vector consisted of features such as the current word, the part-of-speech of the current word, etc., and $|t_i|$ represents the length of a sentence or the number of tokens. The $\boldsymbol{y}$ values are drawn from a discrete set of labels such as B-Chemical, I-Chemical, and O if the IOB2 scheme [2] is used. Given a set of training examples **t**, a learning algorithm outputs a classifier c_i, which can predict the corresponding $\boldsymbol{y}$ values of new **x** values. An ensemble of classifiers is a set of classifiers $\mathbf{c} = \{c_1, c_2, \ldots, c_m\}$ whose individual decisions are combined in a way to classify new examples [3].

Several studies have shown that ensembles are often perform better than the individual classifiers that make them up. Ensembles can be created by several ways. For example, for one training dataset, we can apply different machine learning algorithms or even the same machine learning algorithm with different feature sets to create a set of classifiers. The labels in a dataset can also be manipulated to create different datasets for creating ensemble. For example, TG Dietterich and G Bakiri [4] randomly partitioned the classes appeared in a training set into two subsets and then relabeled the data according to the new tag set to create their ensemble. In this work we proposed an ensemble algorithm which constructs ensembles by manipulating the label set given to the learning algorithm and then classifies a new data by a voting algorithm specifically designed for sequential labelling task.

2 Method

In the BioCreative V.5 CEMP task, the goal is to recognize chemical entity mentions and classify them into seven entity categories including (1) SYSTEMATIC: the systematic names; (2) IDENTIFIERS: database IDs; (3) FORMULA: molecular formula; (4) TRIVAL: trivial, brand, common or generic names of compounds; (5) FAMILY: chemical families that can be associated to chemical structures; (6) MULTIPLE: mentions that correspond to chemicals that are not described by a continuous string of characters; (7) ABBREVIATION: abbreviations and acronyms. In our implementation, the BIESO tag set with fine-grained tokenization [5] were used. The tags, B, I, E, S and O, stand for "**B**egin", "**I**nside",

"**E**nd", "**S**ingle-word", and "**O**utside" of a particular category of chemical entities, which results in $7 \times 3 + 6 + 1 = 28$ tags[1].

Ensemble Generation

In this work, an ensemble of nine classifiers were generated. For all the created classifiers, they were based on the same feature sets as described in our previous work [5]. The first two classifiers were trained by using the conditional random fields (CRFs) [6] and maximum entropy (ME) [7] with the original training dataset released by the CEMP task. The other seven classifiers were created by using CRFs with seven relabeled datasets. The relabeled datasets were compiled by duplicating the original dataset into seven copies. For a copy, we kept one entity type and merged the other six categories into one label. The above process repeated seven times to create seven individual dataset; each of which contains only two entity types: one is the original entity type and another represents the other six categories.

Ensemble Algorithm

The above ensemble generation step created nine classifiers. We denote the created classifiers as **c**. Assume that the given **x** is a sequence of observations $\{\mathbf{o}_1, \ldots, \mathbf{o}_n\}$. For each of the observation $\mathbf{o}_i$, a classifier c_k can output the scores for all possible categories. For instance, each of the first two models, c_1 and c_2, trained with the original dataset by CRF and ME can output 28 scores for each **o**. Given a sequence **x** of $|t_i|$ tokens, both output a $|t_i| \times 28$ matrix m^{k} and each entry of the matrix (denoted as $m^{k}_{(i,j)}$) represents the score given by c_k for the *i*th token in case that it labeled with the *j*th class. However, for the classifier trained with the relabeled dataset, only the columns corresponding to the unmerged label and O tags in the dataset can be filled in. In our implementation, the uniform score was assigned for those columns; the score was evenly distributed among all merged categories.

Now we can average all entries of the $|\mathbf{c}|$ score matrixes to get an averaged matrix *m* and apply the maximum function of each row to determine the best outputs of our ensemble, which is similar to the idea of voting-based ensemble learning. Unfortunately, we cannot do this in se-

[1] After fine-grained tokenization, all MULTIPLE entities consist more than one token.

quential labelling tasks, because we need considering the labelling sequence to avoid illegal label combinations, such as {B- SYSTEMATIC, I-FORMULA}. Instead of applying an additional machine-learning model to fuse the variety of label combination [8, 9], we propose the algorithm shown in Figure 1 to deal with the problem.

The algorithm starts by pruning the possible starting points in the first row. We ignore all classes that cannot be the initial seed for the following calculation. The step removes elements such as E-IDENTIFIER, I-FAMILY, from the first row of the given *m* to generate the seed vector *s*. For each seed label in *s*, we apply depth-first search (DFS) algorithm to traverse *m* to generate all possible sequence combinations from the initial seed label *s* and their aggregated scores. The results are returned by the DFS function and the maximum score of all sequences is compared with the best score to determine whether or not to update the current best ensemble result. After finishing the calculation of all seeds, the mapped label sequence of the best ensemble result is returned.

INPUT:
A score matrix *m*: Each entry $m_{i,j}$ represents the score for the *i*th token labeled with the *j*th class.
An array (*a*) of *c* classes: Each element of the array *a* represents the exact class label for the *j*th column of *m*.
An array (*b*) consists of all possible begin labels: The array *b* contains all possible begin labels, such as B-ABBREVIATION, S-FORMULA.
OUTPUT: An array represents the final predicted labels.

SequentialLabellingEnsemble(*m*, *a*, *b*)
1. *s* ← Pre-pruning(*m*, *a*, *b*)
2. *best* ← **nil**
3. **for** *k* ← 1 **to** len(*s*)
4. *dfs_p* ← DFS(s_k, *m*, *a*, *b*)
5. *max_dfs* ← max(score(dfs_p_1), score(dfs_p_2), …, score(dfs_p_n))
6. **if** *max_dfs* > score(*best*)
7. *best* ← *max_dfs*
8. return map(*best*, *a*)

Figure 1. The proposed ensemble algorithm.

The computation complexity of the algorithm shown in Figure 1 is high, because of the possible label combinations are huge. Therefore, in our implementation, we set two parameters, *w* and *nb*, to control the complexity. *w* controls the depth of the DFS algorithm to traverse and *nb* controls the number of best sequence hold for calculation.

3 Results and Discussion

We submitted five runs in the CEMP task. Table 1 shows the results. The first run is an ensemble with seven classifiers trained with CRFs and the dataset modified from the training set of the BioCreative V.5 CEMP task. The CRF toolkit we used is CRF++[2]. Unfortunately, we failed to use the toolkit with the entire training set of BioCreative V.5 CEMP task to train the model with the seven tag set. Therefore, we used the dataset of BioCreative V CHEMDNER-patents track [10], which is smaller than the dataset released in the CEMP task, to build our classifier for the second run. The third run combined the first run with the classifier trained with ME on the BioCreative V.5 CEMP task. The fourth run combined the second and third run with the classifier trained with CRF on another relabeled dataset of the CEMP corpus. In the dataset, we used one label to represent all seven categories. The final run was an ensemble of the second and third runs.

Table 1. The official evaluation results on the CEMP test test.

Run	F-score	Precision	Recall
1. Ensemble (7 classifiers)	0.8377	0.8567	0.8194
2. BC V	0.8387	0.8479	**0.8296**
3. 1+ME	0.8387	0.8575	0.8207
4. 2+3+All Merged	0.8389	**0.8583**	0.8203
5. 2+3	**0.8395**	0.8568	0.8228

Although it cannot be directly comparable, we can observe that the proposed algorithm can improve the precision and F-score comparing with the baseline configuration (Run-2). In addition, we observed some bugs in our algorithm implementation after the evaluation phase. We believe

[2] https://taku910.github.io/crfpp/

that the performance of the ensemble results can be much better than the official results as shown in Table 1.

4 Conclusion

In the paper, we give a briefly introduction of our ensemble algorithm specifically designed for the sequential labelling task. Unlike other previous works which employs additional machine-learning models or post-processing rules to combine the results of all classifiers for the sequential label problem, our algorithm considers the confidence scores of each individual classifiers and the possibility of label transition. The results on the CEMP task demonstrates the proposed ensemble algorithm can improve the performance of the individual classifier. In the future, we will compare the performance of the proposed algorithm with that of other traditional ensemble methods. We will also apply different relabeling technique to generate the ensembles and study their performance with the proposed algorithm.

5 Acknowledgment

The study was supported by Ministry of Science and Technology (MOST) 105-2221-E-143-003 and MOST-106-2922-I-038-014.

REFERENCES

1. Perez M, Rabal O, Rodriguez GP, Vazquez M, Fdez-Riverola F, Oyarzabal J, Valencia A, Lourenco A, Krallinger M: **Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks**. In: *Proceedings of the BioCreative V5 Challenge Evaluation Workshop; Barcelona, Spain*. 2017: 3-11.
2. Sang EF, Veenstra J: **Representing text chunks**. In: *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics: 1999*. Association for Computational Linguistics: 173-179.
3. Dietterich TG: **Ensemble Methods in Machine Learning**. In: *Proceedings of the First International Workshop on Multiple*

Classifier Systems: June 21-23; Cagliari, Italy. Springer-Verlag 2000.

4. Dietterich TG, Bakiri G: **Solving multiclass learning problems via error-correcting output codes**. *Journal of artificial intelligence research* 1995, **2**:263-286.
5. Dai HJ, Lai PT, Chang YC, Tsai RT: **Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization**. *Journal of Cheminformatics* 2015, **7**(Suppl 1 Text mining for chemistry and the CHEMDNER track):S14.
6. Lafferty J, McCallum A, Pereira F: **Conditional random fields: Probabilistic models for segmenting and labeling sequence data**. In: *Proceedings of the 18th International Conference on Machine Learning (ICML): June 28*. 2001: 282–289.
7. Berger AL, Pietra SAD, Pietra VJD: **A maximum entropy approach to natural language processing**. *Computational Linguistics* 1996, **22**(1):39-71.
8. Dieb TM, Yoshioka M: **Extraction of Chemical and Drug Named Entities by Ensemble Learning Using Chemical NER Tools Based on Different Extraction Guidelines**. *Transactions on Machine Learning and Data Mining* 2015, **8**(2):61-76.
9. Wei Q, Chen T, Xu R, He Y, Gui L: **Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks**. *Database* 2016, **2016**:baw140-baw140.
10. Krallinger M, Rabal O, Lourenço A, Perez MP, Rodriguez GP, Vazquez M, Leitner F, Oyarzabal J, Valencia A: **Overview of the CHEMDNER patents task**. In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop: 2015*. 63-75.