Adverse Reaction Identification Driven by Semantic Information

Kaiyin Zhou[1#], Xing Zhang[2#], Pierre Zweigenbaum[3], Ruiqing Liang[2], Yingying Jiang[4], Jingbo Xia[15*]

1. Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, China
2. College of Foreign Studies, Jinan University, China
3. LIMSI, CNRS, Université Paris-Saclay, Orsay, France
4. College of Science, Huazhong Agricultural University, China
5. Institute of Mathematics, Huazhong Agricultural University, China
The authors with same contributions
* Corresponding author: xiajingbo.math@gmail.com; xjb@mail.hzau.edu.cn

Abstract

To achieve the multi-labeling task of adversary reaction task (TAC 2017), Metamap and Conditional Random Fields (CRF) are used to sequence labels of the raw text data, and one semantic dictionary is built to enhance the accuracy of the sequence-labeling classifier. Moreover, unsupervised and supervised strategies are analyzed through comparing performances of MetaMap-based and CRF-based tagger system.

keywords: Sequence labeling, CRF, MetaMap.

1 Task description

The purpose of this research is to participate in tasks assigned by the Text Analysis Conference (TAC) 2017 [1], organized by U.S. National Institute of Standards and Technology (NIST). TAC 2017 consists of six tracks in two major areas: adverse drug reaction extraction from drug labels (ADR) and knowledge base population (KBP). This research aims to solve the Task 1 of ADR track:

Task 1: Extract AdverseReactions and related mentions (Severity, Factor, DrugClass, Negation, Animal) [1].

(i) AdverseReaction: Reported ADRs that include signs and symptoms, worsening medical conditions, changes in laboratory parameters and other measures of critical body function.

(ii) Severity: Measurement of the severity of a specific AdverseReaction. For example:"major", "critical", "life-threatening") or quantitative doze, volume, etc.

(iii) DrugClass: The class of drug. For example: beta blockers.

(iv) Negation: Negation trigger word.

(v) Animal: Animal species.

(vi) Factor: This includes factors of ADRs including hedging terms (e.g., potential), references to the placebo arm, and specific sub-populations (e.g., pregnancy).

For Data retrieval, One hundred and one drug labels with fully annotation, are offered as gold standard data for classifier training, i.e., file "train_xml.tar.gz" , while 2,208 unannotated files are set as testing data. The format of the data is XML-block based, with *Mentions* annotated and *Off-set* info given [1].

[1] `http://bionlp.nlm.nih.gov/tac2017adversereactions/`

Our idea is to utilize the semantic information entailed by adverse reaction, and our system attempts to adopt MetaMap and Conditional Random Fields (CRF) to identify entities of adverse reaction from training data. The first MetaMap-based phase is an unsupervised method, where MetaMap is used for extracting metathesaurus terms from UMLS. The second CRF-based phase is a supervised method, where semantic terms and token distributions are fed into CRF to generate ADR extraction rules automatically. Both of these phases output prediction for the 2,206 unannotated testing files, and the evaluation score is obtained via the codes offered by TAC 2017.

2 Phase 1: Unsupervised learning: MetaMap-based curation

At first, an unsupervised method is used, when MetaMap is used as searching engine to preprocess training data. Codes are written to preprocess the TAC 2017 training data and convert them to a long form plain text with which MetaMap works. The batch processing module of MetaMap computation service is used to annotate the entities, including potential disorder and ADRs [2]. A big amount of results from MetaMap then are collected and filtered by our filtering rules subsequently.

Design of the filtering rules: Single words and phrases are to be matched with entries in dictionaries of MetaMap before used as scored candidates labeled with semantic types. However, there are cases when phrases provided in MetaMap dictionary can't be matched exactly with those identified in training data. What's more, labeling of adverse reactions by MetaMap needs to take further semantic information into consideration. Therefore, informative major semantic types like, Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction are selected for ADR filtering, which are believed to be informative.

Codes for Metamap-based tagging system, available in GitHub [3]:

- Step 1. Insert PMID intentionally so as to meet the file input requirement from Metamap
- Step 2. To format the file structure, e.g., table formatting, blank removal, etc.
- Step 3. Catenate files in input directory to one file, so as to make it easy for Metamap task submission.
- Step 4. Submit stitched test.txt to MetaMap, and receive its result.
- Step 5. Split text.out, result obtained from Metamap, and map each back to original files.
- Step 6. Annotation part. Collect Metamap reporting files, extract candidate ADR entries and their phrase info, locate the ADR entries with their occurrence place in original XML file, and write the ADR entries/mentions into the new XML files.

Until the deadline of TAC2017, we mainly focused on the labels of ADRs, and the partial results are listed in the following table. On all 99 labels, the F1 score is as low as 0.2433; meanwhile, for these 34 labels we are working on, the F1 score is 0.5044. The performance is shown in Table 1. This result makes it a baseline system for analysis of a performance comparison between an unsupervised tagging system and a supervised one.

Table 1: Task 1 Partial results of the TAC2017

	TP	FP	FN	Precision	Recall(%)	F1(%)
On all 99 labels	2200	1307	12377	62.73	15.09	24.33
On submitted 34 labels	2200	1307	3023	62.73	42.12	50.40

[2] `https://ii.nlm.nih.gov/Batch/UTS_Required/metamap.shtml`
[3] `https://github.com/kyzhouhzau/Identification-of-drug-side-effects`

3 Phase 2: Supervised learning: CRF-based prediction

Generally, the result curation of MetaMap is a typical baseline for entity extraction. To enhance the merely searching strategy, a supervised machine learning based sequence labeling tool, Conditional Random Field (CRF), is used.

CRF is a widely used sequence labeling tool for NLP tasks, which defines the probability of a label sequence $\mathbf{L} = (l_1, l_2, ..., l_l)$, given the observation sequence $\mathbf{O}$. For flexibility, $\mathbf{O}$ is treated as tokens, part of speech (POS), or semantic types according to different feature functions as defined below [2]:

$$exp(\sum_j \lambda_j t_j(l_{i-1}, l_i, \mathbf{O}, i)) + \sum_k \mu_k s_k(l_i, \mathbf{O}, i)),$$

where $t_j(l_{i-1}, l_i, \mathbf{O}, i))$ stands for transition feature function that represents the transition distribution of label pair $\{l_{i-1}, l_i\}$ based on observation sequence $\mathbf{O}$, while $s_k(l_i, \mathbf{O}, i)$ stands for state feature function that quantify the state distribution of the label y_i given the observation sequence $\mathbf{O}$. The mechanism of CRF is the optimization of the parameters λ_j and μ_k, so as to maximize the probability of $P(\mathbf{L}|\mathbf{O})$,

$$P(\mathbf{L}|\mathbf{O}, \lambda, \mu) = \frac{1}{Z(\mathbf{O})} exp(\sum_j \lambda_j t_j(l_{i-1}, l_i, \mathbf{O}, i)) + \sum_k \mu_k s_k(l_i, \mathbf{O}, i)),$$

where $Z(\mathbf{O})$ is for normalization [2].

For efficient implementation of CRF algorithm, the efficient computation package Wapiti ([1]) is used. As potential feature functions, five various patter files were used in Wapiti, see Table 2. 'Tok' refers to the tokens info, 'Pre' for prefix, 'Suf' for suffix, 'Pos' for part of speech (POS), and 'Dis' for disorder controlled vocabulary.

Table 2: Features used for CRF patterns

Name	Description	Generation method	Scale
Tok	Tokenized features: word	NLTK tool kit [4]	Uni-, Bi-, Trigram
Pre	Tokenized features: prefix	Regular expression	Unigram
Suf	Tokenized features: suffix	Regular expression	Unigram
Pos	Lexicon features: Part of Speech	NLTK took kit [4]	Uni-, Bi-, Trigram
Dis	Semantic features: Disorder	Controlled vocabulary	Uni-, Bi-, Tri-, 4-gram

Among the pattern features, Unigram, Bigram and Trigram are all considered for each feature. The rule of N-gram is shown in Table 3. Taking 'Tok' as an example, which is the basic feature that matches the word appearance distribution, Unigram, Bigram and Trigram are all considered. For Unigram, %x[-1,0]/%x[0,0]/%x[1,0] is list in pattern file and the precede word, current word, and the subsequent word are all captured; and for Bigram, both pair of %x[-1,i]/%x[0,i] and %x[0,i]/%x[1,i] are captured; similarly for Trigram, as shown in Table 3.

Table 3: Feature pattern for each N-gram setting

N-gram	Feature pattern for the i-th feature
Unigram	%x[-1,i]/%x[0,i]/%x[1,i]
Bigram	%x[-1,i]/%x[0,i]
	%x[0,i]/%x[1,i]
Trigram	%x[-2,i]/%x[-1,i]/%x[0,i]
	%x[-1,i]/%x[0,i]/%x[1,i]
	%x[0,i]/%x[1,i]/%x[2,i]
4-gram	%x[-3,i]/%x[-2,i]/%x[-1,i]/%x[0,i]
	%x[-2,i]/%x[-1,i]/%x[0,i]/%x[1,i]
	%x[-1,i]/%x[0,i]/%x[1,i]/%x[2,i]
	%x[0,i]/%x[1,i]/%x[2,i]/%x[3,i]

Note: %x[0,i] refers to the current position of the i-th feature.

[4] `http://www.nltk.org/`

Besides token info, POS are extracted automatically by using NLTK, while suffix and prefix are also considered by introducing direct regular expression coding.

In addition, semantic group info from Metamap is integrated into 'Dis' feature, with a form of controlled vocabulary which is composed of 12 subtypes, i.e., acquired abnormality, anatomical abnormality, cell or molecular dysfunction, congenital abnormality, disease or syndrome, experimental model of disease, finding, injury or poisoning, mental or behavioral dysfunction, neoplastic process, pathologic function, sign or symptom, as shown in Table 4. Thus, a controlled dictionary with 522,852 terms is built up.

Table 4: Semantic features used for CRF features [3]

Abbreviation	Semantic Name	SubType ID	SubType
DISO	Disorders	T020	Acquired Abnormality
DISO	Disorders	T190	Anatomical Abnormality
DISO	Disorders	T049	Cell or Molecular Dysfunction
DISO	Disorders	T019	Congenital Abnormality
DISO	Disorders	T047	Disease or Syndrome
DISO	Disorders	T050	Experimental Model of Disease
DISO	Disorders	T033	Finding
DISO	Disorders	T037	Injury or Poisoning
DISO	Disorders	T048	Mental or Behavioral Dysfunction
DISO	Disorders	T191	Neoplastic Process
DISO	Disorders	T046	Pathologic Function
DISO	Disorders	T184	Sign or Symptom

4 Comparison and discussion upon MetaMap- and CRF-based tagging systems upon the test and train dataset

4.1 Performance comparison of supervised and unsupervised tagging system upon the test set

The comparison between MetaMap- and CRF- based system showed that the performances varied between the unsupervised tagger system and the supervised one. As an unsupervised system, MetaMap-based system relied heavily on the built-in UMLS concepts thesaurus to label the potential ADR terms. Though regular expression was used to generalize its matching rule, Metamap-based system does not perform well towards unknown phrases, nor did phrase boundary detection. It's noted that the MetaMap-based system failed to target all labels but only "ADR" labels, and that made this comparison not complete equal. Comparatively, CRF-based system was capable of learning both the tokenization distribution rule and the phrase boundary info from the training corpus, and it had higher recalling rate than the former. As shown in Table 5, CRF achieves higher score in F-score.

Table 5: Comparison of MetaMap and CRF upon test set

	TP	FP	FN	Precision	Recall(%)	F1(%)
MetaMap-based system	2200	1307	12377	62.73	15.09	24.33
CRF-based system	6299	15367	8276	29.07	43.21	34.76

4.2 Features contribution in listed methods, from the analysis upon the train set

For all possible feature combinations, thorough experiments have been conducted and those sampled feature combinations are shown in Figure 1. This result shows an inconsistency in the level of features. However, the POS features does appear more frequently in the classifier with higher F-score.

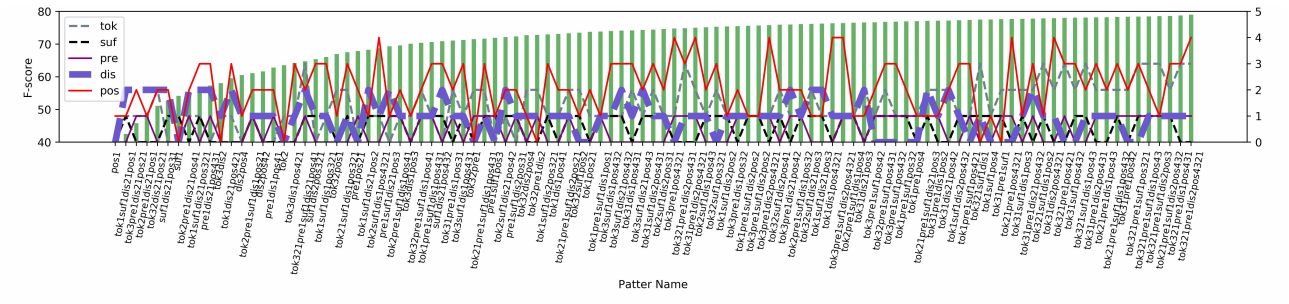

Figure 1: Feature contribution for performance of sequence labeling model – Sampling from the whole experiments

As previously designed, 'dis' was treated as a semantic feature, and it appeared in most classifiers among the Top 50. Compared with 'dis' and 'suf', the features 'pre', 'pos' and 'tok' are necessary parts of features among the Top 50 classifiers, and these results indicate that tokenized information and Part of Speech information played vital role in the classification.

Among the top 50, the top three all consist of 'tok' trigram, 'tok' unigram, 'pre', 'suf', 'dis' bigram or unigram, 'pos' 4-gram or trigram, thus it can prove that 'dis' and 'suf' are proper addition to token and POS features.

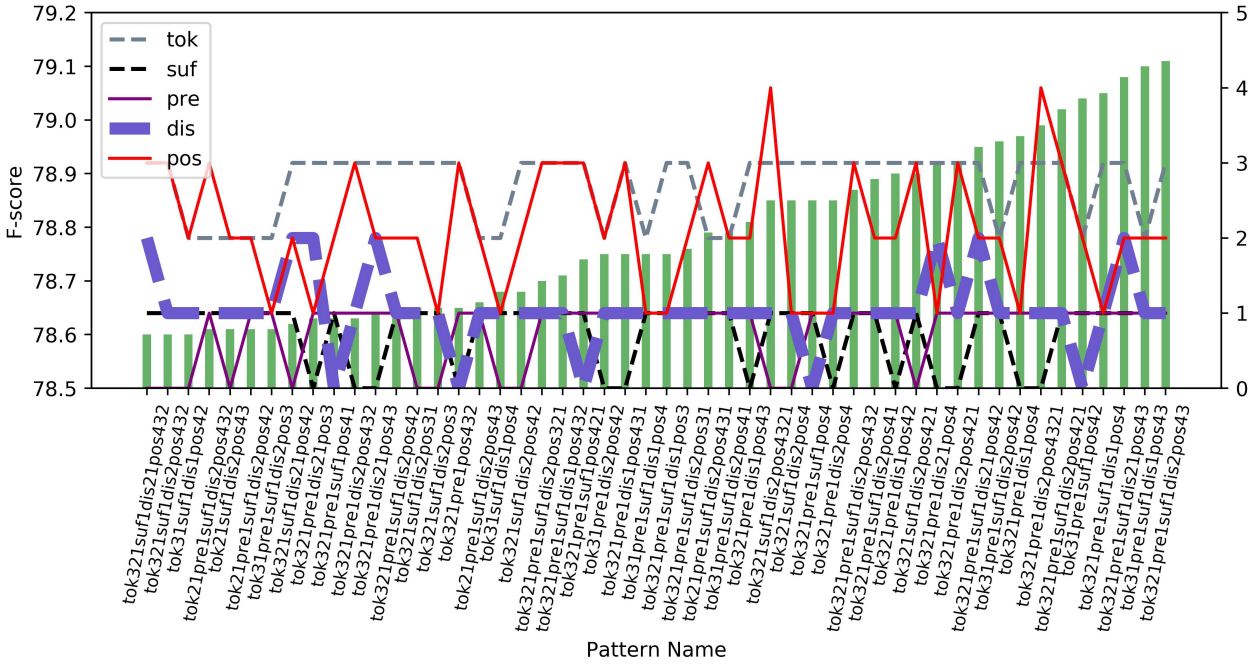

Figure 2: Feature contribution for performance of sequence labeling model – The top 50

Results obtained from the experiments fully prove that careful selection of features can contribute to performance of the sequence tagger system.

Acknowledgements

This paper is supported by the Fundamental Research Funds for the Central Universities (grant no. 15JNYH007) – Constructing a Semantic Knowledge Network of English Words: A Multimodal Perspective. We also thank Kirt Roberts for offering additional code to evaluate our partial results in Phase 1.

References

[1] Lavergne, T., Cappé, O., & Yvon, F. (2010, July). Practical very large scale CRFs. InProceedings of the 48th Annual Meeting of the Association for Computational Linguistics(pp. 504-513). Association for Computational Linguistics.

[2] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[3] Metamap semantic group. https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt.