ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201711.02026

A CRF Model Combining Multiple Features for Chemical-Disease Named Entity Recognition (Postprint)

Authors: Sui Mingshuang, Cui Lei

Date: 2017-11-08T00:00:00+00:00

Abstract

Objective: To establish a conditional random field (CRF) model integrating multiple features and to explore methods for the simultaneous automatic extraction of chemical and disease entities from large-scale biomedical texts.

Methods: We incorporated named entity recognition features—including lexical features, domain knowledge features, dictionary matching features, and unsupervised learning features—compared the effectiveness of different features on named entity recognition, and optimized the model.

Results: The CRF model incorporated lexical features, dictionary matching features, unsupervised learning features, and partial domain knowledge features. Chemical entity recognition achieved an accuracy of 97.33%, recall of 80.76%, and F-value of 8.27%; disease entity recognition achieved an accuracy of 84.20%, recall of 81.96%, and F-value of 83.07%.

Limitations: Simultaneous recognition of chemical and disease entities may result in mutual interference, and the removed partial domain features may contain useful information.

Conclusion: This study provides a reference for feature selection in biomedical named entity recognition, though further feature optimization is required to achieve improved recognition performance.

Full Text

Abstract

Objective: To develop a Conditional Random Fields (CRF) model incorporating multiple features for automatically extracting chemical and disease named entities from large-scale biomedical literature. **Methods:** We integrated various

named entity recognition features, including lexical features, domain knowledge features, dictionary matching features, and unsupervised learning features. Different feature combinations were systematically compared to optimize model performance. **Results:** The final CRF model incorporated lexical features, dictionary matching features, unsupervised learning features, and selected domain knowledge features, achieving 97.33% precision, 80.76% recall, and 88.27% F-score for chemical entity recognition, and 84.20% precision, 81.96% recall, and 83.07% F-score for disease entity recognition. **Limitations:** Simultaneous recognition of chemical and disease entities may cause mutual interference, and some eliminated domain features might contain useful information. **Conclusion:** This study provides a reference for feature selection in biomedical named entity recognition, though further optimization is needed to improve recognition performance.

Keywords: Named Entity Recognition; Conditional Random Fields; Text Mining; Unsupervised Learning

---

The mechanisms through which complex chemicals affect diseases are intricate, prompting greater vigilance regarding drug safety. Statistics show that chemical-disease relationships rank among the most frequently searched topics in the PubMed database [1]. However, the lengthy and complex nature of clinical trials, combined with delays in side-effect reporting mechanisms for approved drugs, makes early prediction of chemical-induced disease information challenging.

Concurrently, the explosive growth of biomedical literature contains chemical-disease relationships that are undoubtedly more sensitive than lengthy clinical trials. Advances in computer and text mining technologies, such as natural language processing, now enable the identification and extraction of chemical-disease associations from large, unstructured free texts. This study compares and integrates multiple features to build a Conditional Random Fields (CRF) model that simultaneously identifies chemical and disease entities from biomedical literature. The key technology involved is Named Entity Recognition (NER), which identifies chemical and disease entities from biomedical data by transforming the task into a classification labeling problem for text units. The complexity lies in the explosive growth of drug and disease entity names, diverse and irregular morphological patterns, and inconsistent naming conventions (especially for drugs) [2].

Current NER methods can be categorized into rule-based, dictionary-based, and machine learning approaches. (1) **Rule-based (template) methods** describe long-established patterns in syntax, grammar, vocabulary, morphology, and writing conventions. For chemical entities, expressions typically consist of uppercase and lowercase letters, numbers, hyphens (. and /), Greek letters, Roman numerals, quotation marks, parentheses, and other characters. Rule-based NER systems rely on domain expert-designed rules implemented through regular expressions. For instance, Xu et al. used context template methods to

construct a comprehensive drug dictionary from PubMed, identifying not only existing DrugBank entries but also drugs not present in the database [3]. Tikk et al. combined rule-based methods with CRF for drug entity recognition [4]. However, such systems lack scalability and adaptability due to their dependence on expert knowledge. (2) **Dictionary-based methods** identify named entities in free text using existing dictionaries, typically through string matching or similarity algorithms. Performance depends on the comprehensiveness of the underlying terminology resources and algorithmic effectiveness. He et al. employed a dictionary-CRF hybrid approach, constructing drug dictionaries from PubMed and applying feature coupling generalization for dictionary denoising, achieving favorable F-scores [5]. (3) **Machine learning methods** represent the current mainstream NER approach, with minimal dependence on dictionaries and rules but requiring manually annotated corpora. Model performance depends on text feature discriminability and algorithm selection [6]. Common methods include Support Vector Machines, Hidden Markov Models, Maximum Entropy Models, and the CRF model used in this study. CRF, proposed by Lafferty et al. [7], combines characteristics of Maximum Entropy and Hidden Markov Models as a typical discriminative model proven effective for biomedical NER [8]. Lee et al. used improved CRF algorithms to enhance disease NER performance [9].

In practice, method combinations often yield better results. Lowe et al. proposed a grammar-dictionary hybrid approach for chemical NER [10], while tools such as tmChem [11] and DNorm [12] and the NCBI disease corpus [13] have facilitated NER development.

Research Framework

Following the BioCreative V corpus [14], this study constructs a CRF model integrating lexical, dictionary, domain knowledge, and unsupervised learning features to simultaneously identify chemical and disease entities from biomedical literature. Through iterative testing and comparison, the optimal CRF model was determined, as illustrated in [Figure 1: see original paper]. Implementation primarily relies on various natural language processing tools and Perl.

GENIA Preprocessing

GENIA Tagger is a specialized analyzer for biomedical text that serves as a preprocessing tool for NER [15]. Running on Linux systems, the command `./geniatagger <input> output` processes sentence-segmented input files (one sentence per line), outputting each word' s lemma, part-of-speech, chunk information, and recognition results for protein, DNA, RNA, and cell line entities.

Feature Set Construction

Lexical Features: (1) Word features: using the word itself as an NER feature. (2) Stem, POS, and chunk features: obtained from GENIA output. (3) Stopword

features: matching each word against a stopword list; features are marked Y for stopwords and N otherwise, using the stopword list from tmChem [11].

Domain Knowledge Features: (1) Morphological features: using regular expression matching to identify patterns with uppercase/lowercase letters, numbers, hyphens (. and /), Greek letters, Roman numerals, quotation marks, parentheses, etc. Words matching these patterns receive feature value Y, others N, generating a morphological feature matrix. (2) High-frequency word features: high-frequency words appear frequently in chemical and disease entity names. The high-frequency word list was constructed as follows [16]: First, count words labeled as chemical or disease entities in the training corpus, recording occurrence count within entities (CF) and total occurrence count in the corpus (TF). Second, calculate weight: Weight = CF/TF $\times$ 100%. Third, extract words meeting these criteria: TF $\geq$ 100 and Weight $\geq$ 0.5; or 10 $\leq$ TF $<$ 100 and Weight $\geq$ 0.6; or 5 $\leq$ TF $<$ 10 and Weight $\geq$ 0.7; or 2 $\leq$ TF $<$ 5 and Weight $\geq$ 0.8. This yields separate high-frequency word lists for chemicals and diseases. Feature value is Y if the current word appears in the list, N otherwise. (3) Affix features: extract 3-character prefixes and suffixes for each word. Using the same method as high-frequency words, generate prefix/suffix lists for chemicals and diseases from words longer than 5 characters in the training corpus. Feature value is Y if the current word's affix appears in the list. (4) Word shape features: chemical entities often share specific shapes. Transform words using "AaX0" pattern (uppercase$\rightarrow$A, lowercase$\rightarrow$a, digits$\rightarrow$0, *other*$\rightarrow$X). Count TF for each shape and CF for entity shapes, generating a chemical word shape list. Feature value is Y if the current word's shape appears in the list. (5) Boundary word features: boundary words are the first and last words of multi-word entities. Using boundary information improves boundary detection and reduces errors in compound entity recognition. Construct left/right boundary word lists for chemicals and diseases using the same method. (6) Context features: context information refers to words immediately preceding and following entities, which improves boundary detection. Construct chemical/disease context word lists using the same method. (7) Unigram and nested word features: unigrams are single-word entities; nested words can function as independent entities or combine with other words to form compound entities. Build unigram and nested word lists from the training corpus. (8) tmChem and DNorm features: incorporate results from tmChem [11] and DNorm [12] as CRF features, marking Y if these tools annotate the current word as an entity. These established NER tools were included to compare performance and assess whether their integration improves results.

Dictionary Matching Features: Construct chemical and disease dictionaries. Match corpus words against dictionary entries, marking Y if the word appears in the dictionary.

Unsupervised Learning Features: (1) Word vector features: Use Word2Vec [17-18] to generate low-dimensional, continuous, real-valued vector representations, with better performance from larger training corpora. This study used

PubMed abstracts retrieved with "Chemicals and Drugs Category" [Mesh] AND "Diseases Category" [Mesh] AND hasabstract[text] (4,417,929 abstracts) plus BioCreative V training and test sets. Using 50-dimensional vectors, we simplified the word vector matrix to (+, -, 0) form following Wu et al. [19]:

```
MEAN( j) MEAN( j) , if M MEAN( j) , if M MEAN( j) 0, otherwise
```

where MEAN(j)+ and MEAN(j)-represent positive and negative means of column j. (2) Brown clustering features: Brown et al. [20] proposed a hierarchical clustering algorithm that builds a tree structure from bottom to top. Using leaf node paths as features, each word is encoded as a long binary string similar to Huffman coding. (3) K-means clustering on word vectors: Using Word2vec' s built-in K-means algorithm on the generated word vectors, grouping similar words into 256 clusters.

Data Sources and Implementation

Data Sources

BioCreative (Critical Assessment of Information Extraction in Biology) is an international competition evaluating text mining systems for biology. BioCreative V (2015) included Disease Named Entity Recognition and Normalization (DNER) and Chemical-Disease Relations (CDR) tasks, providing our corpus [13]. The project treats drugs and chemicals as interchangeable. We downloaded the training and test sets [1], extracted PMID, TI, and AB fields, performed sentence segmentation, built indexes, and obtained chemical and disease vocabularies.

Dictionary Construction

The Medical Subject Headings (MeSH) is an authoritative, expandable thesaurus from the U.S. National Library of Medicine for NER and normalization. Downloaded from NLM [2], we: (1) extracted all terms from categories C (Chemicals and Drugs) and D (Diseases), mapping each chemical to its MeSH ID; (2) extracted additional terms from Supplementary Concept Records (SCRs) where the HM field pointed to category C or D entries, creating a comprehensive dictionary of 578,475 chemical terms and 195,151 disease terms.

Feature Matrix Construction

Following the experimental design, we matched feature lists to generate feature matrices using the "IBO" annotation scheme: I (Inside) for words within entities, O (Outside) for non-entity words, and B (Beginning) for entity first words .

CRF Model Execution

We used CRF++ 0.58 [1] to build the chemical-disease NER model. CRF++ extracts features through templates where each line defines a feature extraction pattern using `%x[row, col]` to reference tokens with relative row and column

offsets . All templates were Unigram type, adjusted for optimal NER performance.

Results

Data Processing

The training and test sets each contained 500 PubMed abstracts with 111,990 and 116,840 words, respectively, including 5,203 and 5,385 chemical entities and 4,182 and 4,424 disease entities.

Feature Template and NER Performance

The final template included 21 features: 5 chemical-specific, 5 disease-specific, and 11 shared features. Word, lemma, and chunk features used a context window of 5. Performance results show that dictionary-only matching achieved low precision and F-score but high recall. Adding lexical features improved chemical and disease F-scores by approximately 10% each, significantly boosting precision while recall decreased. Incorporating domain knowledge features increased chemical precision but substantially decreased recall, reducing F-score. After extensive testing, we removed most domain knowledge features except chemical context and tmChem results. For diseases, however, retaining context and DNorm features improved both precision and recall. Adding unsupervised learning features enhanced both NER tasks, achieving final F-scores of 88.27% for chemicals and 83.07% for diseases, with overall precision, recall, and F-score of 90.64%, 81.32%, and 85.73%.

Before incorporating tmChem and DNorm, our model achieved F-scores of 86.45% for chemicals and 80.13% for diseases, surpassing DNorm (78.2%) and tmChem (83.6%). Adding these tool features improved overall F-score by approximately 3%.

However, simultaneous chemical and disease recognition may have caused interference, as disease NER performance ranked 7th, 4th, and 7th among BioCreative V teams for precision, recall, and F-score (top scores: 90.53%, 86.17%, 86.46%; averages: 78.99%, 74.81%, 76.03%).

Error analysis revealed three main categories: (1) False positives (non-entities marked as entities, e.g., “high Adherence group”); (2) Missed entities (e.g., complex chemical entities like “3,4-methylenedioxymethamphetamine,” uncommon hyphenated entities like “piperacillin/tazobactam,” and abbreviations like “iAs” for inorganic arsenic); (3) Boundary errors (incomplete or over-extended entities, e.g., including degree words in “acute myocardial ischemia”). The performance degradation after adding domain knowledge features led us to remove potentially important morphological, word shape, and boundary features that might help recognize these entity types.

Conclusion

Building upon previous research, this study explores a multi-feature CRF model for simultaneous chemical and disease entity recognition from biomedical literature. Our model incorporates popular features including lexical, domain knowledge, dictionary matching, and unsupervised learning features. Results indicate that high-frequency word, morphological, word shape, and boundary features performed poorly for chemical NER and were eliminated, retaining only context and tmChem features—a trade-off that sacrificed some entity recognition capability.

Future work will address error types by defining special features for parenthetical abbreviations, hyphenated connectors (and/or/-), and degree word restrictions. For complex chemical entities, improved dictionaries and enhanced morphological/word shape feature weights may help. This study provides a reference for feature selection in chemical-disease NER models, with plans to develop relation extraction algorithms that combine syntactic analysis and machine learning to generate complete chemical-semantics-disease triples, ultimately creating a platform for automatic identification and extraction of chemical-disease entities and their relationships.

References

[1] Wei C H, Peng Y, Leaman R. et al. Overview of the BioCreative V Chemical Disease Relation (CDR) Task[C]. In: Proceedings of the 5th BioCreative Challenge Evaluation Workshop. 2015.

[2] Sui Mingshuang, Cui Lei. Detection of Drug Adverse Effects by Text-mining[J]. Chinese Journal of Medical Library and Information Science, 2015, 24(11): 67-72.

[3] Xu Bo, Lin Hongfei, Yang Zhihao. Generating a Drug Name Dictionary Based on Pattern Extraction and Rich Feature Sets[C]. In: Proceedings of the 5th China Conference on Information Retrieval. 2009.

[4] Tikk D, Solt L. Improving Textual Medication Extraction Using Combined Conditional Random Fields and Rule-based Systems[J]. Journal of the American Medical Informatics Association, 2010, 17(5): 540-544.

[5] He Linna, Yang Zhihao, Lin Hongfei, et al. Drug Name Entity Recognition Based on Feature Coupling Generalization[J]. Journal of Chinese Information Processing, 2014, 28(2): 72-77.

[6] Krauthammer M, Nenadic G. Term Identification in the Biomedical Literature[J]. Journal of Biomedical Informatics, 2004, 37(6): 512-526.

[7] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In: Proceedings of the 2002 International Conference on Machine Learning. 2002.

[8] Chowdhury Md F M, Lavelli A. Disease Mention Recognition with Specific Features[C]. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010.

[9] Lee H C, Hsu Y Y, Kao H Y. An Enhanced CRF-based System for Disease Name Entity Recognition and Normalization on BioCreative V DNER Task[C]. In: Proceedings of the 5th BioCreative Challenge Evaluation Workshop. 2015.

[10] Lowe D M, Sayle R A. LeadMine: A Grammar and Dictionary Driven Approach to Entity Recognition[J]. Journal of Cheminformatics, 2015, 7(S1): 1-9.

[11] Leaman R, Wei C H, Lu Z. tmChem: A High Performance Approach for Chemical Named Entity Recognition and Normalization[J]. Journal of Cheminformatics, 2015, 7(S1).

[12] Leaman R, Islamaj Dogan R, Lu Z. DNorm: Disease Name Normalization with Pairwise Learning to Rank[J]. Bioinformatics, 2013, 29(22): 2909-2917.

[13] Doğan R I, Leaman R, Lu Z. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization[J]. Journal of Biomedical Informatics, 2014, 47(2): 1-10.

[14] Li J, Sun Y, Johnson R J, et al. Annotating Chemicals, Diseases and Their Interactions in Biomedical Literature[C]. In: Proceedings of the 5th BioCreative Challenge Evaluation Workshop. 2015.

[15] Kim J D, Ohta T, Tateisi Y, et al. GENIA Corpus–Semantically Annotated Corpus for Bio-textmining[J]. Bioinformatics, 2003, 19(S1): 180-182.

[16] Xia Guanghui. The Research of Gene Name Entity Recognition Mechanism by Combining Dictionary Method and Machine Learning Method[D]. Beijing: Peking Union Medical College, 2013.

[17] Zhang Y, Xu J, Chen H, et al. Chemical Named Entity Recognition in Patents by Domain Knowledge and Unsupervised Feature Learning[J/OL]. The Journal of Biological Databases and Curation[2016-06-10]. http://database.oxfordjournals.org/content/2016/baw049.

[18] He Honglei. Research of Word Representations on Biomedical Named Entity Recognition[D]. Dalian: Dalian University of Technology, 2015.

[19] Wu Y, Xu J, Jiang M, et al. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text[C]. In: Proceedings of the 2015 AMIA Annual Symposium.

[20] Brown P F, Desouza P V, Mercer R L, et al. Class-based N-gram Models of Natural Language[J]. Computational Linguistics, 1992, 18(4): 467-479.

Author Contributions

Sui Mingshuang: Designed the research, conducted experiments, collected and analyzed data, drafted the manuscript. Cui Lei: Conceived the research idea, revised the final manuscript.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors and available upon request at suims1107@163.com: [1] Sui Mingshuang. Allchemical.txt; alldisease.txt. Chemical and disease dictionaries. [2] Sui Mingshuang. corpus.pl. Data processing and dictionary matching algorithms. [3] Sui Mingshuang. feature.zip. Feature construction algorithms. [4] Sui Mingshuang. CRF model.zip. Final training/test sets and CRF model.

Received: 2016-06-24 Revised: 2016-07-19

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.