Extraction of Drug-Drug Interactions from Biomedical texts

Lavanya Mandadapu

Albert Espín

Table of Contents

Introduction

- This project solves the problems proposed in the SemEval 2013 Task 9 competition:
  - Task 9.1: Recognition and Classification of drug names in four categories:
    - Drug (Generic drug name).
    - Brand (Branded drug names).
    - Group (Drug group names).
    - Drug-n (Active substances not approved for human use).
  - Task 9.2: Extraction and Classification of Drug-Drug Interactions in five categories:
    - No interaction
    - Advice
    - Effect
    - Mechanism
    - Generic interaction
- For both Tasks, two different types of models are created:
  - Baseline rule-based model, to see if a simple approach can solve the problem.
  - Machine Learning models:
    - Naive Bayes and CRF for Task 9.1.
    - Ensemble of Deep Neural Networks for Task 9.2.
- The results are evaluated using the F1 measure on the full test corpora (DrugBank and MedLine).

Task 9.1

Baseline Rule-based System

- Initially, tokenize the sentences with NLTK word tokenizer.
- Rule 1: If word is capitalized ⇒ Brand name.
- Rule 2: If word contains the suffix “azole”, “idine”, “amine” or “mycin” ⇒ Generic drug name.
- Default: Not a drug name.
- Result: overall F1-score is **0.14**, with precision of **0.42** and recall of **0.1**. Very poor, only decent for brands and generic drugs.

Naive Bayes Model

- Statistical model where 8 categorical features are used for learning the class of tokens (not drug or drug+type).
- The overall F1 score is **0.36** with a precision of **0.37** and a recall of **0.38**.
- Drawback: Sequential relations between words cannot be learned.

Categorical Features:

- POS tag.
- Syntactic chunk.
- Name entity type.
- Whether words end with the suffixes of Rule-based model.
- Orthography of word.
- Contains hyphen.
- Word present in DrugBank list of words.
- Word present in NIH list of words.

Note: words and lemmas are discarded because of having large number of possible values for binary vector representation.

CRF Model

- Conditional Random Fields are sequential models: they take into account the local context of the word to classify each word as drug or not (and its category if it is a drug).
- The validation set (10% of training partition) is used as to optimize the model parameters:
  - Minimum frequency of features: 1 (to preserve all the information, e.g. uncommon words).
  - Regularization coefficient (for Stochastic Gradient Descent): 0.15 (lower values caused overfitting, higher ones underfitting).
  - Maximum iteration number: 1000, usually not reached.
  - Early stopping after 10 iterations without improvement.
  - Initial learning rate is calculated performing 20 trials, and decreased through iterations.

CRF Features

The following features were considered, 66 in total (including variants for current word, previous and next word):

- Base word (original and lower case).
- Affixes: 3/4/5-character prefix and suffix of the words.
- Orthography (all-capitalized, starts with upper-case, all-numeric, combines letters and numbers, etc.).
- Contains hyphen or not.
- Whether word is in DrugBank (external, not test set), FDA or NIH drug names databases.
- Lemma, POS tag, syntactic chunk and named entity, all obtained with the biomedical-specialized Genia tagger (although results were similar to NLTK).
- Word shape feature, summarizing the word structure (X: upper-case letter, x: lower-case letter, 0: digit, O: other character). Also a version that removes consecutive equal characters.
- Semantic cluster to which the word belongs, obtained using batch-based K-Means to group 200-dimensional word embeddings (pre-trained in medical texts) in 450 clusters (random initialization outperformed K-Means++).
- Start/end of sentence.

CRF Feature selection

- The validation set is used to select the best subset of features.
  - Phase 1 with a small number of runs:
    - Those features that when present increased the validation F1 significantly (more than 0.5%) were added to a **WhiteList** (31 features).
    - Those features whose contribution was not significant or slightly negative were added to the **CandidateList**.
    - The features that significantly decreased the accuracy were removed.
  - Phase 2 finds the best combination of features: WhiteList + Subset(CandidateList), among 50 subsets.
- There are 12 combinations in tie with the highest validation F1, and they are used to calculate the test F1, all giving very similar results (less than 3% difference).
- The BIO and BILOU sequential term labeling schemes were tested, with BIO being 4% better despite of its lower specialization.

CRF Results

- F1-score of **0.74** for DrugBank+Medline is obtained, better than all the original participants (the highest macro average was **0.64**).
- Two models produce the best result:
  - WhiteList + Subset of CandidateList (**49** features).
  - WhiteList alone (**31** features); preferred due to its greater simplicity (Occam's razor).

Field	Precision	Recall	F1
Strict Matching	0.82	0.72	0.77
Exact Matching	0.9	0.79	0.84
Partial Matching	0.9	0.81	0.86
Type Matching	0.86	0.75	0.8
Macro Average	0.95	0.68	**0.74**

Task 9.2

Baseline Rule-based System

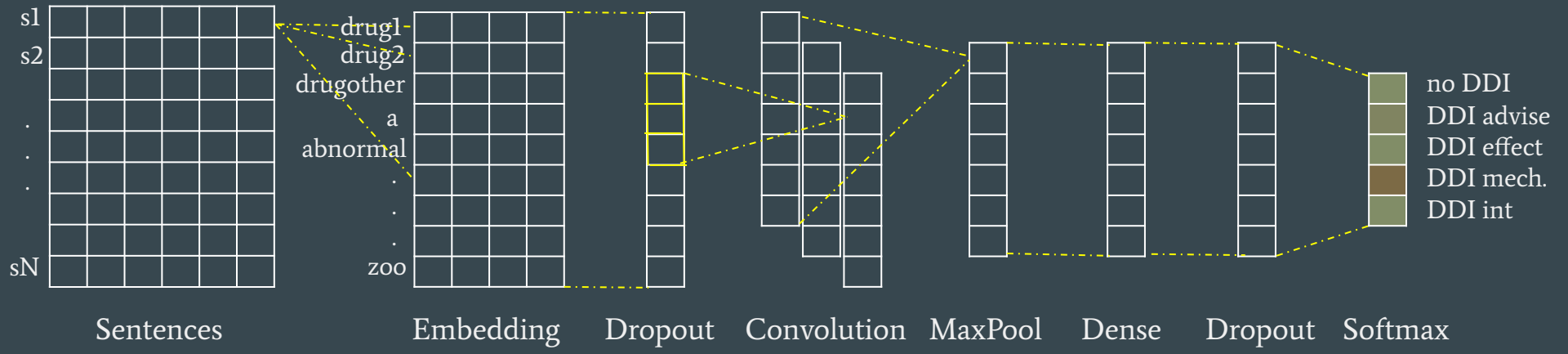
- For each sentence:
  - Find tokens with NLTK word tokenizer, and lemmatize them.
  - Locate the tokens between the two drugs. If any of them is included in a dictionary of keywords, return the associated DDI type, otherwise return that there is no interaction.
  - The dictionary is based on the official “Annotation guide” examples, for instance:
    - “should” ⇒ advise
    - “elevation” ⇒ effect
    - “inhibit” ⇒ mechanism
    - “interact” ⇒ generic interaction
- Result: precision of 0.25, recall of 0.31 and 0.28 F1-score. It is low, but better than the worst participants’ run (0.21).

Ensemble of Deep Neural Networks

- Preprocessing:
  - Drug anonymization to learn interactions, not particular drugs: generic strings “drug1” and “drug2” are used for the main pair, “drugother” for the rest.
  - Only alphanumeric tokens are accepted (no other characters, except hyphen). Lower-case transform.
- Ensemble of multiple Deep Neural Network models:
  - 90% of training data for learning the models, 10% for validation.
  - All models predict the DDI class (or no DDI) for each drug pair per sentence.
  - Voting is used to decide the final class (majority vote, i.e. mode class).
- Architecture uses Nadam optimizer, categorical cross-entropy as loss and the following components:
  - Input sentences, represented as lists of unique integer indices for each word in the vocabulary.
  - Embedding layer that fine-tunes pre-trained word embeddings of corpora of medical articles.
  - Alternative layers (some models use all of them, others part of them):
    - 1D-Convolution (filters: 100, 300, 500...; mask size: 3, 4, 5, 10...), to find patterns of relations between words in sentences; Max Pooling used to rule out irrelevant ones.
    - Recurrent layers (GRU/LSTM; units: 32, 64, 128...) to exploit sequential relations of words.
    - Fully-Connected (units: 500, 1000...) for detail and Dropout (ratio: 0.1, 0.3...) to generalize.
  - Fully-connected 5-unit output layer with softmax activation, to determine the DDI class (or no DDI).

Best individual model of the Ensemble

- CNN with 100 convolution filters of mask size 4, with Max Pooling and two Dropout layers of 0.3 ratio, and one Fully-Connected layer with 500 units. This model obtains a test F1 score of 0.60, individually.
- Other models obtain lower individual F1 scores (e.g. 0.3 for GRU-only model, 0.51 for GRU+CNN).



Ensemble Results

- A combination of 8 Neural Networks (CNNs, RNNs, CNNs+RNNs) achieves 0.63 test F1, almost as high as the winner of the original competition (0.648), who used hybrid SVM classifiers. All the other participant models, including one of the winner's runs, are outperformed by this Ensemble.

Field	Precision	Recall	F1
Partial Evaluation: only detection of DDI	0.74	0.67	0.69
Detection and Classification of DDI	0.67	0.59	0.63
DDI with type mechanism	0.69	0.59	0.64
DDI with type effect	0.63	0.61	0.62
DDI with type advise	0.69	0.68	0.68
DDI with type int	0.97	0.30	0.46
Macro-average	0.74	0.55	**0.63**

Conclusions and Future work

- The machine learning solutions developed for the tasks (CRF for Task 9.1 and Ensemble of Neural Networks for Task 9.2) have significantly outperformed the baselines, since the advanced systems extract features to learn complex models that infer better than simple rules.
- The results are significantly better than the winner (CRF in Task 9.1) or very close (Ensemble in Task 9.2), but some additions may further improve F1:
  - Sub-optimal feature selection strategies for CRF, e.g. based on Information Gain or Chi-squared. In this work all-subsets search was used, with a limited number of combinations (exponential total, optimal only if all are explored).
  - Attention mechanisms for the CNNs.
  - Genetic algorithms to evolve different populations of ensembles (with validation F1 as the fitness function) to select the most suitable combination.

Do you have any questions?

Thank you!