What we have done:

- Read articles about Chinese Classifier and finalize our task.
- Our task is to predict Chinese classifiers given Chinese sentences which usually contain one classifier and one head word.
- Create a github repository.
- Download datasets including training/dev/test.

What we are going to do:

- Check the quality of the datasets, try to find some examples about why such predictions are hard.
- Build baseline model.
- Build SVM model.

- Chinese classifier assignment using SVMs

In this article, they extract data from Penn TreeBank with only classifiers and head nouns and build up noun-classifier pairs. The ontological features of nouns are acquired from HowNet, a bilingual Chinese-English lexicon and ontology. They assign each noun seen in noun-classifier pairs a word entry that contains a sememe, a categorical attribute, additional attributes, and two types of pointers (related and agent). The baseline algorithm is to assign a noun the most frequent co-occurred classifier in the test data. If the noun is unseen in the training data, they assign the classifier 个 [ge] to the noun, because 个[ge] is the most occurred classifier in the corpus. The accuracy is 50.76% to all nouns and 50.69% to the nouns that occur 2+ times in the corpus. During the experiment, they use noun features and context features. Noun features have four feature sets: (1)noun only; (2)ontological features of the noun only; (3)the feature set with both noun features and ontological features; (4)the feature set with noun features or ontological features. The article also runs two context features: (5) noun, lexical and syntactic features; (6) noun, ontological, lexical and syntactic features. The built SVM model runs all the six feature sets and all the SVMs have better performance than the baseline. There is no significant difference between the performance with the 1st, 2nd, 3rd and 4th feature sets. But the performance of the SVMs using lexical and syntactic features (experiments 5 and 6) is significantly worse than those without ($p < 0.05$).

- ClassifierGuesser: A Context-based Classifier Prediction System for Chinese Language Learners

As an improvement of the first paper, this paper adapt the prediction task for a sentence-based scenario rather than word-based classifier prediction.

Corpus:
There are three corpus used in this article : The Lancaster Corpus of Mandarin Chinese, the UCLA Corpus of Written Chinese and the Leiden Weibo Corpus.

Data processing:
In order to improve the data quality, several filters are applied first, such as removing duplicate sentence, filter sentence with less than 4 tokens or more than 60 tokens. Then parsed the remaining sentences with the Stanford constituent parser and extracted the head of the

classifier in each sentence based on the parse tree. Finally, the sentences with the head word and corresponding classifier are used as final data set.

Baseline:
The baseline comes from the results in previous papers, the first paper is one of them.

Context-based models:

1. Training word embeddings with word2vec on sentences from the original three corpora and also obtain pre-trained word embeddings.
2. Training two widely used machine learning models (SVM, Logistic Regression) on the embedding vector of the head word
3. Gradually add more contextual features to the models: With the motivation of reducing head word ambiguity, they include embedding vectors of words within window size n=2 of the head word.
4. Using a bidirectional LSTM to encode the entire sentence excluding any head word annotation and predict classifiers based on the last hidden state

Results:
The Micro F1 on test set is 71.51 and the Macro F1 on test set is 30.56.

References

Hui Guo and Huayan Zhong. 2005. Chinese classifier assignment using SVMs.
Peinelt, Nicole, Maria Liakata, and Shu-Kai Hsieh. 2017. "ClassifierGuesser: A Context-based Classifier Prediction System for Chinese Language Learners." *Proceedings of the IJCNLP 2017, System Demonstrations*: 41-44.