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

Nicole Peinelt[1,2] and **Maria Liakata**[1,2] and **Shu-Kai Hsieh**[3]
[1]The Alan Turing Institute, London, UK
[2]Department of Computer Science, University of Warwick, Coventry, UK
[3]Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan
{n.peinelt, m.liakata}@warwick.ac.uk, shukaihsieh@ntu.edu.tw

Abstract

Classifiers are function words that are used to express quantities in Chinese and are especially difficult for language learners. In contrast to previous studies, we argue that the choice of classifiers is highly contextual and train context-aware machine learning models based on a novel publicly available dataset, outperforming previous baselines. We further present use cases for our database and models in an interactive demo system.

1 Introduction

Languages such as Chinese are characterized by the existence of a class of words commonly referred to as ‘classifiers’ or ‘measure words’. Based on syntactic criteria, classifiers are the obligatory component of a quantifier phrase which is contained in a noun phrase or verb phrase.[1] Semantically, a classifier modifies the quantity or frequency of its head word and requires a certain degree of shared properties between classifier and head. Although native speakers select classifiers intuitively, language learners often struggle with the correct usage of classifiers due to the lack of a similar word class in their native language. Moreover, no dictionary or finite set of rules covers all possible classifier-head combinations exhaustively.

Previous research has focused on associations between classifiers and nominal head words in isolation and included approaches based on ontologies (Mok et al., 2012; Morgado da Costa et al., 2016), databases with semantic features of Chinese classifiers (Gao, 2011), as well as an SVM with syntactic and ontological features (Guo and Zhong, 2005). However, without any context classifier assignment can be ambiguous. For instance, the noun 球 ‘ball’ can be modified by *ke* - a classifier for round objects - when referring to the object itself as in (1), but requires the event classifier *chang* in the context of a ball match as in (2). We argue that context is an important factor for classifier selection, since a head word may have multiple associated classifiers, but the final classifier selection is restricted by the context.

(1) 一 颗 红色 的 球
one *ke* red DE ball
‘a red ball’

(2) 一 场 精彩 的 球
one *chang* exciting DE ball
‘an exciting match’

This study introduces a large-scale dataset of everyday Chinese classifier usage for machine learning experiments. We present a model that outperforms previous frequency and ontology baselines for classifier prediction without the need for extensive linguistic preprocessing and head word identification. We further demonstrate the usefulness of the database and our models in use cases.

2 System Design

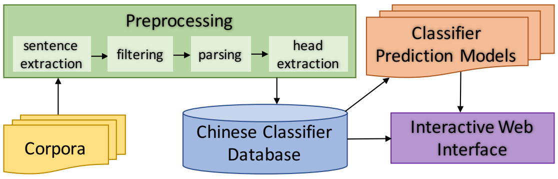

Preprocessing
sentence extraction
filtering
parsing
head extraction
Classifier Prediction Models
Corpora
Chinese Classifier Database
Interactive Web Interface

Figure 1: Overview of proposed system

[1] Following Huang (1998) and 何杰 (2008) we include verbal as well as nominal classifiers.

The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations, pages 41–44,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

Figure 1 gives an overview of our system. It comprises data collection, pre-processing and the compilation of the Chinese Classifier Database (section 2.1), the training of classifier prediction models (section 2.2), and the interactive online interface (section 3).

2.1 The Chinese Classifier Database

The database is based on three openly available POS tagged Chinese language corpora: The Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004), the UCLA Corpus of Written Chinese (Tao and Xiao, 2012) and the Leiden Weibo Corpus (van Esch, 2012). Sentences from the corpora were assigned unique ids, filtered for the occurrence of classifier POS tags and cleaned in a number of filtering steps in order to improve the data quality (Table 1). We further parsed the remaining sentences with the Stanford constituent parser (Levy and Manning, 2003) and extracted the head of the classifier in each sentence based on the parse tree.[2] By manually evaluating 100 randomly sampled sentences from the database, we estimate a classifier identification accuracy of 91% and head identification accuracy of 78%. Based on our observations, most errors are due to accumulating tokenisation, tagging and parsing errors, as well as elliptic classifier usage. In addition to the example sentences, we also included lexical information from CC-Cedict[3] for the 176 unique classifier types.

Applied filters	Sentences	%
None (initial corpus)	2,258,003	100
1. duplicate sentence	1,553,430	69
2. <4 or >60 tokens in sentence	1,470,946	65
3. classifiers consisting of letters/numbers; or <70% of Chinese material in sentence	1,437,491	64
4. tagged classifiers are in fact measure units (e.g. 毫米)	1,150,749	51
5. classifiers with <10 examples	1,109,871	49
6. classifier fails manual check	1,103,338	49
7. frequent error patterns	1,083,135	48
8. multiple classifiers in a single sentence	858,472	38

Table 1: Number of remaining sentences in database. Matching sentences are excluded.

2.2 Classifier Prediction

2.2.1 Task

Following the only previous machine learning approach (Guo and Zhong, 2005), we frame classifier prediction as a multi-class classification problem. However, in contrast to previous work that focused on word-based classifier prediction, we adapt the prediction task for a sentence-based scenario, which is a more natural and less ambiguous task than predicting classifiers without context. Not all sentences in the Chinese classifier database contain head words, due to co-referential and anaphoric usage. Hence, we query the database for sentences in which both the head word and corresponding classifier were identified, resulting in 681,102 sentences. This subset is randomly split into training (50%), development (25%) and test set (25%). In each sentence with an identified classifier and corresponding head word, we substitute the classifier with the gap token <CL> and use the classifier as its class label. For example, the tagged sentence

我们是一 <c> 家 </c> <h> 人 </h>。

is transformed into the training example

我们是一 <CL> <h> 人 </h>。

with the label '家'. Labels are simplified from tokens to types by reducing duplicate classifiers (e.g. 个个 → 个) and mapping traditional characters to simplified characters (e.g. 個 → 个), resulting in a dataset[4] with 172 distinct classes.[5] Given a training set of observed sentences and classifiers, the task is to fill the gap in a sentence with the most appropriate classifier.

2.2.2 Baseline approaches

As previous studies have evaluated algorithms on individually collected unpublished data, we implement the following baselines to compare our models with previous results:

- *ge*: always assign the universal and most common noun classifier 个 (Guo and Zhong, 2005; Morgado da Costa et al., 2016).

[2] Starting from the position of the classifier, we move one node up in the tree at a time until reaching a noun or verb phrase and extract its head word.

[3] https://cc-cedict.org/

[4] We make our dataset publicly available at https://github.com/wuningxi/ChineseClassifierDataset.

[5] The number of unique classifiers differs from the full database because only example sentences with identified head words are taken into account.

- *pairs*: assign the classifier most frequently observed in combination with this head word during training; assign 个 for unseen words (Guo and Zhong, 2005).
- *concepts*: assign classifiers based on classifier-concept pair counts using the Chinese Open Wordnet and 个 for unseen words (Morgado da Costa et al., 2016).

2.2.3 Context-based models

Previous approaches predominantly rely on ontological resources, which require a lot of human effort to build and maintain, resulting in limited coverage for new words and domains. We use distributed representations to capture word similarity based on syntactic behaviour, as they can be trained unsupervised on a large scale and are easily adapted on new language material. We train word embeddings with word2vec (Mikolov et al., 2013) on sentences from the original three corpora and also obtain pre-trained word embeddings from Bojanowski et al. (2017). The pre-trained embeddings consistently achieve better results and are hence used in all subsequent experiments.

Since the head word is linguistically the most important factor for classifier selection, we first train two widely used machine learning models (SVM, Logistic Regression) on the embedding vector of the head word (*head*). In order to investigate to which extend context may help with classifier prediction, we then gradually add more contextual features to the models: With the motivation of reducing head word ambiguity, we include embedding vectors of words within window size n=2 of the head word ($cont_h$). Furthermore, we add embedding vectors of words surrounding the classifier gap ($cont_{cl}$) to capture the typical immediate environment of different classifiers. As preliminary experiments indicate that increasing the window size to n>2 increases computation costs without significant performance gains, a better approach to include more context is needed. We hence use a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to encode the entire sentence excluding any head word annotation ($cont_{cl}$) and predict classifiers based on the last hidden state (Figure 2).

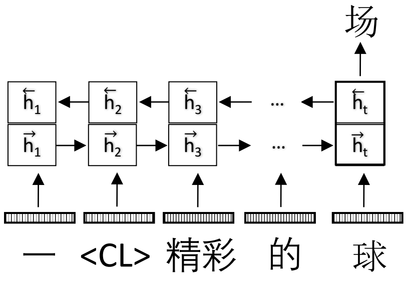


Figure 2: LSTM architecture for context-based classifier prediction.

2.2.4 Results

We report micro F1 (accuracy) and macro F1 scores for each model after hyper-parameter tuning in Table 3. The head-classifier combination baseline gives a strong result, which the SVM and Logistic Regression models trained on only headword embedding vectors cannot surpass. Global corpus statistics on classifiers outperform the local information captured by the word embeddings in this case. Adding head word context features successfully reduces the ambiguity of head words and results in a significant improvement over the baseline. Including contextual features of the classifier gap slightly decreases the performance, but still outperforms the context-unaware models. The best model is the LSTM which achieves micro F1 71.51 and macro F1 30.56 on the test set based on the full sentence context without the need for headword identification (hyper-parameters as reported in Table 2, optimiser: Adam, learning rate: 0.001).

Parameter	Values
Hidden units	160, 224, 320, 384, **480**
Dropout rate	**0.0,** 0.25, 0.5
Batch size	32, **64,** 96, 128

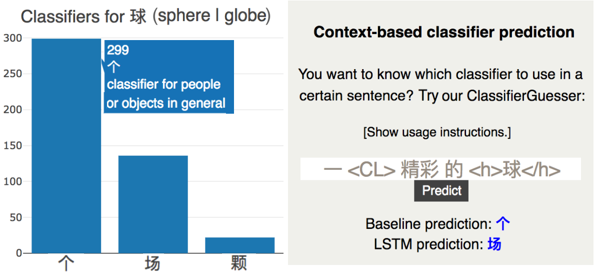
Table 2: Tuned hyper-parameters for LSTM. Terms in bold represent final settings.

		Micro F1		Macro F1	
	Features	dev	test	dev	test
base line	ge	45.12	45.21	0.36	0.37
	pairs	61.82	61.72	24.40	23.80
	concepts	49.08	49.11	8.40	7.94
svm	head	53.67	53.72	13.33	13.56
	+cont$_h$	66.02	66.02	24.86	24.39
	+cont$_{cl}$	58.97	58.83	22.23	21.75
log reg	head	57.61	57.72	15.99	15.66
	+cont$_h$	67.81	67.67	28.95	27.37
	+cont$_{cl}$	67.43	67.29	27.51	26.70
lstm	cont$_s$	**71.69**	**71.51**	**31.56**	**30.56**

Table 3: Model performance on the classifier prediction task (logreg = Logistic Regression).

3 Use Cases

When learning new classifiers, Chinese language learners can obtain frequency statistics from the online interface of the Chinese Classifier Database [6] to focus on the most commonly used and most important classifiers. Learners can explore a visualisation of frequently used classifier-head word combinations in an interactive bar plot (Figure 3, left) which displays example sentences from the database when clicking on the bars. Furthermore, the ClassifierGuesser (Figure 3, right) can be used when learners want to compose a sentence but don't know the appropriate classifier. After inputing a sentence with a gap, the system predicts the best classifier candidate based on the *pairs* baseline and the best LSTM model.


Classifiers for 球 (sphere | globe)
300
250
200
150
100
50
0
299
个
classifier for people or objects in general
个
场
颗

Context-based classifier prediction

You want to know which classifier to use in a certain sentence? Try our ClassifierGuesser:

[Show usage instructions.]

一 <CL> 精彩 的 <h>球</h>

Predict

Baseline prediction: 个
LSTM prediction: 场

Figure 3: Screenshot of classifier-head pair visualisation (left) and ClassifierGuesser (right).

4 Conclusion

This paper introduced a system for predicting Chinese classifiers in a sentence. Based on a novel dataset of example sentences for authentic usage of Chinese classifiers, we conducted multiple machine learning experiments and found that incorporating context improves Chinese classifier prediction over word-based models. Our best model clearly outperforms the baselines and does not require manual feature engineering or extensive preprocessing. We argue that including contextual features can help resolve ambiguities and context-based classifier prediction is a more realistic task than isolated head word-based prediction. We further presented an interactive web system to access our database and pretrained models and demonstrated possible use cases for language learners.

Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1, Microsoft Azure and the German Academic Exchange Service (DAAD).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Helena Hong Gao. 2011. E-learning design for Chinese classifiers: Reclassification. *Communications in Computer and Information Science*, 177:186–199.

Hui Guo and Huayan Zhong. 2005. Chinese classifier assignment using SVMs.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

C.-T. James Huang. 1998. *Logical relations in Chinese and the theory of grammar*. Taylor & Francis, New York & London.

Roger Levy and Christopher Manning. 2003. Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting on ACL-Volume 1*, pages 439–446.

Anthony McEnery and Zhonghua Xiao. 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Hazel Mok, Shu Wen, Gao Huini Eshley, and Francis Bond. 2012. Using WordNet to predict numeral classifiers in Chinese and Japanese. In *GWC 2012 6th International Global Wordnet Conference*, pages 264–271.

Luis Morgado da Costa, Francis Bond, and Helena Gao. 2016. Mapping and Generating Classifiers using an Open Chinese Ontology. In *Proceedings of the 8th Global WordNet Conference*, pages 247–254.

Hongyin Tao and Richard Xiao. 2012. *The UCLA Chinese Corpus (2nd edition)*. UCREL.

Daan van Esch. 2012. Leiden Weibo Corpus. http://lwc.daanvanesch.nl/.

何杰. 2008. 现代汉语量词研究: 增编版. 北京语言大学出版社, 北京市.

[6] chinese-classifier-database.azurewebsites.net