UTRECHT UNIVERSITY

Computational Models of Classifier Choice in Mandarin:

from Rule-based to BERT

Author: **Jani J. Järnfors**

First Supervisor: **Prof. dr. C.J. van Deemter**
Second Supervisor: **Dr. M.P. Schraagen**

A thesis submitted in fulfillment of the requirements for the degree of

Master of Science
in
Business Informatics

Department of Information and Computing Sciences
Faculty of Science
Utrecht University
The Netherlands
16 July 2021

Computational Models of Classifier Choice in Mandarin:
from Rule-based to BERT

Jani J. Järnfors

Abstract

As in many Asian languages, Mandarin grammar often requires that a noun is preceded by a classifier word in certain syntactic positions. Many nouns can appear with multiple different classifiers, and the choice of classifier can have a significant impact on the meaning of the whole sentence. There is no dictionary-based approach or a finite set of rules covering all possible classifier-head noun combinations exhaustively, as sometimes the larger context surrounding the noun is required to make the correct choice of classifier. This makes the problem of predicting the correct classifier in a given context challenging.

Earlier studies have suggested different kinds of computational methods for this task. However, as the studies have examined very broad selections of classifiers, which have not been explicitly listed, or defined in a computational way, their results and subsequent analyses have left room for further study. Hence, this study aims to produce an extensive and transparent analysis on the task of classifier prediction, including providing explicit lists for different categorizations of classifiers. Namely, it considers the fact that linguists generally agree that Mandarin classifiers should be categorized into two clearly different categories: true classifiers and measure words. In order to also evaluate the impact of context on classifier choice, a subset of classifiers, that we consider to *add information*, is examined.

We used four different models, a simple rule-based model, a bidirectional LSTM model, a BERT masked language model and a BERT classification model, to predict classifiers in sentences. We were able to produce a new state-of-the-art result for generating Mandarin classifiers by using a BERT classification model. We also showed that all the models perform better with true classifier compared to measure words or other types of classifiers. As a consequence, the results indicated that a simple rule-based model can be used to generate true classifiers reasonably well. In addition, the context-aware BERT classification model clearly outperforms the other models in predicting both measure words and classifiers that *add information*. However, we theorize that certain classifiers may still not be possible to accurately predict in all situations using current solutions, as the classifiers themselves carry meaning, which is not obvious from the rest of the sentence context.

Acknowledgements

First of all, I want to thank Kees van Deemter and Guanyi Chen for all their help during the writing of this thesis. This thesis would not exist without their numerous hours of guidance. At times I felt lost related to the theory and approach of this topic, but all our discussions during the past eight months were extremely helpful in finding the right direction in completing this study. In addition, Guanyi's help with creating the different models used in this study has been invaluable. Marijn Schraagen's feedback has also been very beneficial as it has provided a different point of view to the topic of this thesis.

Lastly, I also want to express a huge sincere thank you to professor Rint Sybesma for all the feedback, comments and insights he has provided during the past months. He has brought up extremely interesting issues related to the theory of classifiers in Mandarin with regards to this work, and I hope I have done justice to at least some of them in this thesis.

This thesis project has really shown me that good research is really a team effort — so much more can be achieved working together compared to working alone. I will definitely take this mindset with me going forward.

Contents

1 Introduction and Research Questions

Classifiers (CL) are present in multiple languages of the world, and Mandarin Chinese[1] is a prime example of a classifier language. Classifiers are words that normally appear with numerals. They are marginal in European languages, but common in East Asian languages (Zhang, 2013). For example, if one were to translate the phrase "One dog is walking on the road" to Mandarin, it would be phrased like "One CL dog is walking on the road": Yī zhī gǒu zài lùshàng zǒu, 一只 狗在路上走.

Let us compare two other sentences between English and Mandarin. The examples have been adapted from Zhang (2013, p. 5):

1a.
瑶瑶看见了三支 笔。
Yáoyáo kànjiànle sān zhī bǐ.
Yaoyao saw three CL pen.
Yaoyao saw three pens.

1b.
瑶瑶看见了三滴 油。
Yáoyáo kànjiànle sān dī yóu.
Yaoyao saw three CL oil.
Yaoyao saw three drops of oil.

Some languages have the counterpart of example *1b*, while not the counterpart of example *1a*. These include English, where the Mandarin classifier *dī* 滴 corresponds with the English word *drop*. However, in example *1a*, English does not have a corresponding word for *zhī* 支. Languages that have both of these types of classifiers are called classifier languages (Zhang, 2013; Bisang, 2011; Cheng and Sybesma, 2005, 2012). Classifiers in the category of example *1a* often give an indication of what kind of entity the noun denotes. For example, the classifier *zhī* 支 in example *1a*, indicates that the object bǐ 笔, "pen", is a long, thin and inflexible object. Other similar long, thin and inflexible objects also take the same classifier. In Mandarin, classifiers are mostly obligatory when counting a head noun and in demonstrative expressions (Zhang, 2007).

The present Master's thesis project investigates how, and to what extent, it may be possible to predict the choice of classifier in Mandarin given a noun and the wider context in which it appears. *Prima facie* evidence suggests that this prediction may not always be straightforward. Mandarin contains a large

[1]Henceforth Mandarin Chinese will be referred to as "Mandarin".

number of classifiers[2], and although the choice of classifier is limited by the (head) noun with which the classifier is associated, this may still leave several options, which may sometimes produce a different meaning. Let us consider the following examples:

(a) Yí gè diànnǎo / Yì tái diànnǎo (a computer)

(b) Yí gè lǎoshī / Yí wèi lǎoshī (a teacher)

(c) Yí gè rén / Yì qún rén (a person/a group of people)

(d) Yì bēi kāfēi / Yì tīng kāfēi (a cup/can of coffee)

Although each of these cases involves classifier choice, the problem of choosing a classifier is likely to be more challenging in those cases, such as (b)-(d), where the classifier adds information, for example, in terms of politeness ((b), neutral vs. polite), number ((c), singular vs. plural), or quantity ((d), a cup vs. a can of coffee). This is perhaps clearest in the case of 1(d), where bēi and tīng indicate different containers, and consequently different quantities, of coffee. Controversies relate to whether some of the above examples should be considered measure words instead of *true* classifiers. This is discussed extensively in section 2.

Researchers have asked what determines the choice of classifier, constructing algorithms that predict what classifier suits a given discourse context. The most sophisticated model we are aware of is by Peinelt et al. (2017). The authors construct a large-scale classifier dataset by extracting and filtering data from publicly available Chinese corpora. They conduct experiments on their dataset using several different models as baselines, including a rule-based system, two machine learning based systems, and a bidirectional LSTM system (Hochreiter and Schmidhuber, 1997). An initial evaluation study indicated that the LSTM achieved the best performance.

The present thesis aims to follow the research by Peinelt et al. (2017) by training a BERT (Devlin et al., 2018) model to perform the same task. The study by Peinelt et al. (2017) is used as a reference as it is the most recent study, and the most state-of-the-art model, on using machine learning to predict Mandarin classifiers. In addition, other studies have only considered classifiers and nominal headwords in isolation, while the study by Peinelt et al. (2017) considers whole sentences. As Peinelt et al. (2017) found that incorporating contextual features of sentences increased the LSTM model's performance, the question comes up whether BERT, with its superior ability to take context into account, might do better.

A limitation of the Peinelt et al. (2017) study is that it does not analyse the performance of the algorithm in terms of different types of classifiers. This should be an important aspect to investigate further as there is evidence that

[2]Some linguists do not consider certain measure words to be classifiers and thus the exact number of classifiers varies widely. Estimates range from several dozen (Li and Thompson, 1989), to about 50 (Chao, 1968) to over 900 (Zhang, 2007).

some types of classifiers could be more difficult to predict than others. Peinelt et al. (2017) also does not provide an explicit list of the classifiers they have analyzed, which makes later analysis of the results difficult.

As linguists generally agree that not all classifiers have the same purpose or behave in the same way, we hypothesize that an algorithm might produce different kinds of results (e.g., be more or less successful) for different types of classifiers. However, there is no agreement on the exact set of classifiers that should be consider *true classifiers* in literature. Some linguist do not consider certain measure expressions to be *true classifiers*, so the inventories of classifiers vary widely between studies. Due to these factors, we approach the problem by first looking at a very broad set of classifiers and then focusing on certain subsets of classifiers that we define based on the literature. These subsets include a categorization into *true classifiers, dual classifies and measure words* and a subset of what we define as classifiers which *add information*.

We then train four different models, a simple rule-based model, a bidirectional LSTM, a BERT masked language model and a BERT classification model, to perform the task of generating classifiers. Afterwards we compare the performance of these models by looking at their general performance and their performance in terms of our different subsets of classifiers.

This leads us to consider the main research questions and hypotheses of this study. The main research goal is to:

- Find out how well different kinds of models are able to predict classifiers in Mandarin sentences.

We consider the following research questions:

1. How well is it possible to do on this task?
2. How do the different models compare to each other in performance?
3. Are some kinds of classifiers harder to predict than others? In order to answer this question, we look at certain subsets of classifiers.
4. How do the models perform on a subset of classifiers which *add information*?

We also formalize the following hypotheses based on our knowledge that context plays a part in predicting certain classifiers:

1. The BERT classification model's advantage over its competitors is greater for *measure words* than for *true classifiers*.
2. The RULE model's disadvantage over the BERT classification model is smaller for *true classifiers* than for other cases.
3. The BERT classification model's advantage over its competitors is greater for classifiers which *add information* than for more general classifiers, such as ge 个。

Apart from adding to existing theoretical insights, we hope that natural language generation systems could benefit from the insights of this study. We want to understand what it means to choose classifiers, how hard the problem is and whether there are differences between different kinds of classifiers. In addition, we want to find out if there are particularly difficult cases. By knowing more about the theory behind classifiers, we hope to provide benefit for those working with language technologies where Mandarin is generated. For instance, insights from this study could be beneficial in the field of automatic machine translation and automatic text generation. The problem, as it is framed in this thesis (i.e. generate a missing classifier in a sentence), may occur in practice in machine translation where a sentence with an English noun, that requires a classifier in Mandarin, needs to be translated into Mandarin. In this case, a suitable classifier (possibly from multiple options) needs to be generated to make the sentence both grammatically and semantically correct.

Section 2 of this thesis will provide more information on Mandarin classifiers and provide more details on why some cases of classifiers promise to be particularly difficult. Section 3 will then outline the dataset used in this study and the models we build, and section 4 will present the results of the experiments. Lastly, section 5 will consider our results and answer the research questions.

This thesis uses simplified Chinese characters and Hanyu pinyin for transcription.

2 Background

The following subsections focus on the theories related to the usage of classifiers in Mandarin with especially highlighting what makes classifier choice a non-trivial problem. The section starts with looking at the semantic features of classifiers and describing how ontologies of Chinese classifiers are problematic. After this, the contextual nature of some classifiers and how context is an important aspect of choosing classifiers is discussed. Subsequently, a categorization of Mandarin classifiers into true classifiers and measure words is described.

2.1 Classifier Choice in Mandarin

2.1.1 Semantic Features of Nouns and General Classifiers

The choice of classifier is often thought to depend on semantic features of the head noun, but determining what these features are is not always straightforward. We introduce the problem using the classifier *tiáo* 条, which has a semantic indication for *long and rope-like* object. This classifier is applied to many kinds of objects, including animate and inanimate, which feature a long and rope-like shape. For example, this classifier must be used when counting fish (Zhang, 2007):

一条 鱼
Yī tiáo yú。
One CL fish.

两 条鱼
Liǎng tiáo yú.
Two CL fish.

However, *tiáo* 条 is also used for multiple other nouns that have a long shape and are flexible like a rope, including ones which are not animals. Tai and Wang (1990) state that the longness of the shape is identified as the cognitive basis of the classifier *tiáo* 条. However, there are many objects that are long, that do not take the classifier *tiáo* 条 (Tai and Wang, 1990). For example, many animals, which have long shapes, take *zhī* 只, which is the default classifier for *animacy* (excluding humans), instead of *tiáo* 条 (Tai, 1994). In addition, some long objects take *gēn* 根 or *zhī* 枝 instead of *tiáo* 条. Tai (1994) provides the following example:

一条 蛇 - yī tiáo shé - one snake
一条 鱼 - yī tiáo yú - one fish
一条 黄瓜 - Yī tiáo huángguā - one cucumber
一条 凳子 - Yī tiáo dèngzǐ - one bench

一根 棍子 - Yī gēn gùnzi - one rod
一根 火柴 - Yī gēn huǒchái - one match
一根 针 - Yī gēn zhēn - one needle
一根 甘蔗 - Yī gēn gānzhè - one sugarcane

一支 笔 - Yī zhī bǐ - one writing instrument
一支 蜡烛 - Yī zhī làzhú - one candle
一支 枪 - Yī zhī qiāng - one gun
一支 香烟 - Yī zhī xiāngyān - one cigarette

Consequently, the choice of classifier in many of these cases is not always obvious from the object itself: a long animal might not take the classifier for long objects, but instead the general classifier for animals, *zhī* 只. Mandarin includes many examples of similar cases (Tai, 1994)[3].

This brings us to consider the aspect of general classifiers. The feature of the classifier denoting some kinds of inherent aspects of the noun does not apply to so-called general classifiers. Examples include the aforementioned *zhī* 只, which is a general classifier for animals and also the classifier *ge* 个. *Ge* 个 can be used very generally for humans, abstract entities, different kinds of containers,

[3]Tai (1994) does not provide explicit numbers. This is presumably because there are thousands upon thousands of nouns, so evaluating all possible cases is very difficult. In addition, according to Tai (1994) the usage of classifiers for certain semantic categories varies between dialects and regions. Nevertheless, the author implies a large variety of similar cases can be found.

rings and frames, objects with a large enclosed hollow interior and a multitude of other 3-dimensional objects (Loke, 1994, p. 40-41). Conclusively, *ge* 个 is a general classifier which acts as an option for nouns that lack more specific classifiers, and is one of the most common classifiers in Mandarin (Sybesma et al., 2017). Hence, it does not denote any kind of inherent semantic feature of the noun.

2.1.2 Effect of Context on Classifier Choice

An interesting aspect to consider is also that inanimate objects tend to not have a general classifier (Tai, 1994). In fact, there are a large number of specific classifiers for inanimate objects (Tai, 1994). Zhang (2007) referring to multiple studies (Loke, 1996; Polio, 1994; Sun, 1988; Tai, 1992; Tai and Wang, 1990) mentions that in these cases, a different classifier might be used for stylistic effects or to imply the speaker's intentions. The author adds that the differences relate to semantic qualities such as formal vs. informal, written vs. colloquial, educated vs. uneducated, positive vs. negative, and common usage vs. local dialects.[4] Hence, context can play an important part in choosing a classifier.

For instance, according to Zhang (2007), the noun *huà* 画, "painting", can appear with at least three different classifiers: *zhāng* 张, *fú* 幅, and *zhēn* 帧. Each of these classifiers associates the noun differently. Firstly, *zhāng* 张 associates the painting with the class of objects with a flat surface. *Fú* 幅, on the other hand, suggest a more sophisticated use of classifiers: the person might want to appear as more educated or the situation might be more formal. Lastly, *zhēn* 帧 implies that the painting is particularly valuable or exclusive. In these cases, there are different semantic connotations depending on the usage of the classifier (Zhang, 2007).

Hence the choice of classifier in the above mentioned situations can be completely dependent on how the speaker or author wants to portray the object — or how they want to present their personal status. In addition, this shows that a word in a corpus could go together with multiple different classifiers. In the case of the noun 画 huà, "painting", a look-up-table would list at least three different classifiers for the word, and it is not clear which one should be chosen for each context. Mandarin has a significant amount of similar cases, for example, the word 树 shù, "tree", can also take at least 8 different classifiers, all of which provide the word with different semantic meanings (Tai, 1994)[5].

Sometimes, the choice of classifier can also change the meaning of the whole sentence. The following example is presented in Peinelt et al. (2017):

1.
一颗 红色的球
Yī kē hóngsè de qiú
A red ball

[4]It has been noted that there are some differences in classifier usage between different dialects of Mandarin (Tai, 1992, 1994), but these differences will not be studied here.

[5]Again, Tai (1994) does not provide explicit numbers, but rather provides a few examples, while implying that a large amount of similar cases exist.

2.
一场 精彩得球
Yī chǎng jīngcǎi dé qiú
A wonderful game

This shows that depending on the classifier assigned to the noun *qiú* 球, the meaning of the whole sentence changes. In this case we could say the classifier *adds information* to the sentence. It could also be interpreted that the noun *qiú* 球 has two different meanings and and each correlates with its own classifier, thus helping us interpret the whole sentence correctly. No matter how the example is interpreted, the choice of classifier is still not obvious only from the head noun qiú 球. To pick the right classifier we need to understand whether we are talking about a ball or a game.

Finally, Zhang (2013) demonstrates that in many cases, the choice between two or more classifiers can be entirely *arbitrary* — in the sense that it is difficult to exactly define what the differences in meaning between certain classifiers are.[6] The author provides the following example (Zhang, 2013, p. 59):

> "The same noun may arbitrarily occur with different individual classifiers. For example, mousha-an "murder case" can be counted by the classifiers zong, qi, jian, chu, zhuang, and chang, but qiangdao-an "robbery case" can be counted by the same set of classifiers except the last two (zhuang and chang)."

2.1.3 Conclusions on Classifier Choice

We have demonstrated that even though many times the choice of classifier is based on inherent semantic features of the noun, sometimes the choice of classifier depends on context and other times the choice is arbitrary to an extent.[7] Furthermore, the choice of classifier can change the meaning of a whole sentence.

It follows that using ontological resources or categorizations to assign classifiers semantically to certain objects is an incomplete solution for choosing classifiers. In addition, using dictionary based approaches or look-up-tables is problematic as one noun can take multiple classifiers depending on the context. Consequently, understanding the larger context is important in making the appropriate choice of classifier. Due to this, we suspect a context-aware model such as BERT could perform better than dictionary based approaches or look-up-tables in the task of generating classifiers.

[6] Specifically, it should be noted that even in these cases where the choice seems arbitrary, in many cases the different classifiers do affect the meaning to an extent, but it might be difficult to define exactly in what way. The speakers themselves might not be aware of the slight differences in meaning.

[7] It should be noted, that in empirical studies there is a high tendency to use the general classifier 个 ge in conversations and speech acts (Erbaugh, 1986; Guo, 2002; He, 2001). According to Zhang (2007), both children and adults use the general classifier in situations where they are uncertain about what specific classifier should be used for a particular noun. In addition, they use the general classifier also in cases where there are known specific classifiers (Zhang, 2007). Thus, in many situations 个 ge can be an acceptable choice, even if more specific classifiers are available.

2.2 Categorization of Chinese Classifiers

It is generally understood that not all classifiers play similar roles or have similar characteristics (Cheng and Sybesma, 1998, 1999, 2005; Sybesma et al., 2017). Consequently, the problem of predicting classifiers may not always be equally difficult, and may even require different solutions. To compound the problem, there seems to be no universally-agreed formal way to categorize classifiers, and terminology differs between different studies. This situation creates a challenge to the present project, because our problem could be demarcated in different ways (i.e., by focussing on different types of classifiers). The way we propose to tackle this challenge is as follows:

- Initially, we will investigate a very broad set of classifiers disregarding the distinctions that have been made in the literature.
- Later, when evaluating the results of our prediction algorithms, we will zoom in on several subsets of classifiers proposed by researchers. We expect to find that our algorithms will perform better on some classifiers than on others.

In what follows, let us briefly sketch some of the issues that have been raised in the literature surrounding the classification of classifiers.

2.2.1 True Classifiers Versus Measure Words

Cheng and Sybesma (1998, 1999, 2005) follow Tai and Wang (1990) and Croft (1994) in describing a distinction between two types of classifiers, the first of which creates a unit of measurement and the second of which names the unit in which the entities denoted by the noun come naturally. As mentioned by Cheng and Sybesma (1998), in past research, for instance in (Chao, 1968, p. 509), these categories have been referred to as individual and non-individual classifiers. Individual means that the classifier singles out one countable discreet unit, whilst non-individual means that the classifier makes the noun countable, but does not pick out a discrete unit.

As an example, words such as *píng* 瓶, "bottle", *bǎ* 把, "handful", *wǎn* 碗, "bowl", create units by which the amount of liquor, rice and soup is measured — but liquor and rice and soup do not come naturally in bottles, handfuls or bowls (Cheng and Sybesma, 1998). These classifiers function as non-individual classifiers. However, concepts like people, pens, cows, tails and tables have natural units they can be counted with. These are classifiers such as *ge* 个, *zhī* 只 and *běn* 本, and they do not create a unit of measurement; they just name it (Cheng and Sybesma, 1998). These classifiers are individual classifiers.

In their paper, Cheng and Sybesma (1998) refer to individual classifiers as "count-classifiers" and non-individual classifiers as "massifiers." These terms are used widely in later studies, but other terms such as "count-noun classifiers and mass-noun classifiers" and "sortal classifiers and mensural classifiers" are also common. It should not be assumed that different studies refer to the exact same set of classifiers even while using the exact same terminology. This paper

will generally use the terms *true classifier* to refer to individual classifiers and the term *measure words* to refer to non-individual classifiers.[8]

This concept is further discussed by Zhang (2007), who also points out that scholars generally agree that there should be a distinction between count-noun classifiers and mass-noun classifiers. The author refers to the work of Allen (1977), Tai and Wang (1990), and Tai (1992), who agree that count-noun classifiers denote an inherent and permanent property of an object while mass-noun classifiers indicate temporary states of the object and give a quantifying description of the object.

For instance, *gēn* 根 is a count-noun classifier that indicates a long and stick-like property. *jié* 节, *bāo* 包, and *tiáo* 条, on the other hand, are mass-noun classifiers that denote a temporary state of being. All these classifiers can be applied to the object *cigarette*, which has the property of being long and stick-like (Zhang, 2007, p. 48):

a. 一根 香烟 - Yī gēn xiāngyān (one CL cigarette; 'a cigarette')
b. 一节 香烟 - Yī jié xiāngyān (one CL cigarette; 'a section of cigarette')
c. 一包 香烟 - Yī bāo xiāngyān (one CL cigarette; 'a pack of cigarette')
d. 一条 香烟 - Yī tiáo xiāngyān (one CL cigarette; 'a carton of cigarette')

In the above example, *a.* uses a count-noun classifier that indicates a permanent property of the object, namely that the object is long and stick-like. *B*, *c* and *d* on the other hand are mass-noun classifiers and represent temporary states of the object, i.e. a pack of cigarettes can be separated into individual cigarettes, meaning it is not a permanent representation of the state of the object.

Cheng and Sybesma (1998) also discuss another way of categorizing classifiers. They mention that although all classifiers are nominal, one group consists of classifiers that can be completely grammaticalized as classifiers, meaning they cannot occur as independent nouns. The examples provided include *ge* 个 (classifier for a person and other general objects) and *zhī* 只 (classifier for animals). The authors add that this group of classifiers constitutes a closed class. Conversely, the other group of classifiers consists of nouns, which create a unit for measuring mass nouns and which can also appear as independent nouns. For example, *bēi* 杯, "glass", can be used to measure mass nouns, such as water: yī bēi shuǐ, 一杯水; a glass of water. However, bēi 杯 can also used independently as a noun: zhuōzi shàng yǒu bēi, 桌子上有杯; there is a glass on the table. This group of classifiers includes most types of containers, including *jié* 节, *bāo* 包, and *tiáo* 条 from the earlier example, and constitutes an open class.

Cheng and Sybesma (1998) note that the division between the two ways of categorizing classifiers roughly coincides: count-classifiers form a closed class and classify nouns that are cognitively singularizable, meaning that the classifier

[8]When referring to earlier research, the terms used in the original papers are also used in this thesis. However, in all other cases the terms *true classifiers* and *measure words* will be used to generally refer to this categorization. It should not be assumed the terms refer to an explicit set of classifiers unless stated so.

singles out one discreet countable unit. Massifiers on the other hand form an open class and classify nouns that are cognitively masses.

However, Zhang (2013) divides classifiers into seven groups as illustrated by the following examples (Zhang, 2013, p. 27):

1.
瑶瑶看见了三支 笔。[Individual CL]
Yáoyáo kànjiànle sān zhī bǐ.
Yaoyao saw three CL pen.
Yaoyao saw three pens.

2.
瑶瑶看见了三滴 油。[Individuating CL]
Yáoyáo kànjiànle sān dī yóu.
Yaoyao saw three CL oil.
Yaoyao saw three drops of oil.

3. Kim bought three liters of milk. [Standard measure]
4. Kim bought three bottles of milk. [Container measure]
5. three kinds of chocolate [Kind CL]
6. three sections of orange [Partitive CL]
7. three groups of students [Collective CL]

(Zhang, 2013, p. 57) does not agree with the notion that classifiers can be straightforwardly sorted into sortal classifiers and mensural classifiers. The author mentions that, in earlier studies, individual classifiers are generally categorized as sortal classifiers. However, they point out that the categorization for the rest of the classes of classifiers varies widely between different studies. Figure 1 presents a summary of how earlier studies have sorted different types of classifiers into sortal classifiers and mensural classifiers.

Figure 1: Summary of classifier categorization in earlier studies by Zhang (2013, p. 57) (S = sortal CL; M = mensural CL)

(150)	Individual CL	Individuating CL	Partitive CL	Collective CL	Kind CL
Grinvald (2002: 261), Li et al. (2008)	S	M	M	M	
Rijkhoff (2002: 48)	S	M		S	
Gerner & Bisang (2010)	S	S	M	M	
Velupillai (2012)	S	S			
Li et al. (2010: 209, 217)	S	S	S	M	

Zhang (2013) adds that the subtypes are usually defined by listing the subtypes and then described by examples: no formal criterion is provided. They conclude that there being no formal definitions explains why the groupings vary

so much. The author also mentions that it is unclear from these works if *kind classifiers* are sortal or mensural. Her and Lai (2012) agree that previous studies on classifiers and measure words provide drastically different inventories. They add that even those who support a formal classifier-measure word distinction have not made the distinctions explicit, and many works assume that the categories are distinguishable, and then distinguish the categories "rather subjectively" (Her and Lai, 2012, p. 2).

Zhang (2007) also observes that, in general, it seems like distinguishing mass-noun classifiers from count-classifiers is easy. They mention that typical mass-noun classifiers express scalar concepts such as length, weight, aggregation, and any kind of container of objects that can see as a unit such as a box, pack, bottle, bowl or plate. However, the author adds that the distinction is not always clear. Zhang (2007) cites Becker (1975, p. 114), who mentions that the distinction should be considered a continuum and that no clear boundary exists. The author mentions that one classifier, such as *tiáo* 条 can belong to both categories. For example, it can be a count-classifier for long and rope-like objects as shown in the examples in section 2.1.1. However, *tiáo* 条 can also function as a mass-noun classifier, which indicates a unit or temporary state of an object[9] (Zhang, 2007). From Zhang (2007, p. 7):

一条 香烟
yī tiáo xiāngyān
a carton of cigarettes

一条 面包
yī tiáo miànbāo
a loaf of bread

一条 肥皂
yī tiáo féizào
a bar of soap

As demonstrated, *tiáo* 条 could be considered to belong in both the categories of true classifiers and measure words, depending on where it is used. Hence, the classifier could be considered a type of *dual classifier*, which means that depending on the noun it appears with, it can act either as a true classifier or as a measure word. This shows that categorizing each classifier into just true classifiers or measure words is not possible in all cases, as some classifiers, depending on the context, may belong in both categories. Hence, *dual classifiers* may not necessarily be considered a separate category *linguistically*, but rather just for the sake of clarity present a explicit category for classifiers that may function as both true classifiers and measure words depending on the situation.

[9]It should be noted, that even in this function, the classifier, *tiáo* 条, still retains its semantic connotation of long and rope-like (Zhang, 2007). So in these examples, the reader would assume that the carton of cigarettes, loaf of bread and bar of soap have a long shape.

2.2.2 Using Semantics to Categorize Classifiers

This leads us to consider another proposed method of categorizing classifiers, which is to consider whether they provide semantic information to the noun. Zhang (2013) states that some earlier studies have implied that sortal classifiers provide semantic information about the noun, whilst mensural classifiers do not provide semantic information. The examples above and further examples in Zhang (2013, p. 58) show that both sortal classifiers and mensural classifiers can provide semantic meaning to the noun. Zhang (2013, p. 60) adds that even though *ge* 个 has often in earlier studies been categorized as a sortal classifier, it actually does not classify the semantic types of nouns, like for instance, *tiáo* 条, (long and rope-like), does. Thus some classifiers that are traditionally considered mensural classifiers, such as *tiáo* 条, carry semantic meaning, whilst some that are traditionally considered count-noun classifiers, such as *ge* 个, do not carry semantic meaning. This means that the feature of whether a classifier provides semantic meaning to a noun is also not accurate for categorizing classifiers into count-noun classifiers and mass-noun classifiers.

2.2.3 The *DE* Test

Another well-known method for sorting classifiers into sortal classifiers and mensural classifiers is described by Cheng and Sybesma (1998). They claim that a *de* 的 test can be used to check which category a classifier belongs in. The authors present evidence that a count-classifier cannot be followed by *de* 的, although massifiers can be. Also, they mention that it appears that an adjective cannot occur in front of a count-classifier, while an adjective can appear in front of a massifier. However, Zhang (2013, p. 62)) provides evidence against both of these claims. Firstly, the author demonstrates that all types of classifiers can be followed by *de* 的 in an appropriate context. The author also provides extensive evidence of cases where an adjective can occur in front of a count-classifier. Zhang claims the context for *de* 的 does not concern the count-mass contrast, but rather the syntactic position of *de* 的: in one position, it is a comparative ellipsis construction while in another position, it introduces a constituent directly. Thus, the author describes Cheng and Sybesma's claims as "descriptively inadequate." (Zhang, 2013, p. 74)

2.2.4 Conclusions on Categorizing Classifiers

We conclude that there is some agreement in the literature that Mandarin classifiers could be categorized into two different categories: true classifiers and measure words. However, Zhang (2013) calls this categorization into question, especially as few formal definitions have been formed for the categories: each researcher sorts classifiers into the two categories in different ways. Her (2012) also states that terminology is part of the confusion: the terms used for different types of classifiers are not systematic. In addition, some classifiers may have multiple functions (*dual classifiers*), further increasing the difficulty of categorizing them.

According to Wu and Her (2021, p. 5), only the studies by Lai (2011) and Her and Lai (2012) provide a comprehensive list of classifiers in Mandarin that is based on explicit and testable criteria. Her and Lai (2012) take the list of general measure words *(yiban liangci 一般量词)* from the *Mandarin Daily Dictionary of Chinese Classifiers (MDDCC)*, which has been compiled with data from the *Academia Sinica Balanced Corpus of Modern Mandarin Chinese*. The authors have re-examined the list of 173 classifiers in the *MDDCC* and applied the tests they formalize in their paper. Using these tests, the authors have come up with the following three lists of classifiers:[10]

1. True Classifiers — 76 classifiers
2. Dual Classifiers (Classifiers that can function as both true classifiers and measure words) — 21 classifiers
3. Measure Words — 76 classifiers

The work by Her and Lai (2012) presents the most complete formal categorization of classifiers into true classifiers, dual classifiers and measure words — even though even it still is an incomplete presentation (Wu and Her, 2021). As there is currently no agreement on a perfect method for categorizing classifiers, we believe applying the list from Her and Lai (2012) to our study can provide interesting insights into the task of predicting classifiers. We use the categorizations by Her and Lai (2012) as a tool to analyze and compare the performance of our algorithms in terms of the subsets of *true classifiers*, *dual classifiers* and *measure words*.

2.3 Earlier Studies in Computational Linguistics

There have been a few studies that have attempted to predict classifiers in Mandarin texts. However, there has been much more work done analyzing classifiers than generating them using natural language processing (Wen et al., 2012). The correctness of this assessment is evident from the limited amount of studies available. During this literature review, we found six studies that focus on generating classifiers in Mandarin. The following approaches have been implemented in these earlier studies:

- Ontology-based approaches (Wen et al., 2012; Da Costa et al., 2016)
- Databases with semantic features of Chinese classifiers (Gao, 2011)
- SVM with syntactic ontological features (Guo and Zhong, 2005)
- A context-aware machine learning model (Peinelt et al., 2017)

Besides, there has been a study by Zhang et al. (2008) related to statistical machine translation of Mandarin with a focus on classifiers:

[10] Please refer to appendix A for these lists.

- Measure word generation for statistical machine translation (Zhang et al., 2008)

Other similar machine translation studies do not focus on classifiers, so they have not been evaluated here.

2.3.1 Ontology, Database and SVM Approaches

Earlier studies present multiple different kinds of approaches to the generation of classifiers. Wen et al. (2012) create a method for generating classifiers using both Chinese and Japanese WordNets, which are lexical databases of semantic relations between words. They assign synsets from WordNet to classifiers by hand and use a modified algorithm to generate sortal classifiers based on semantic hierarchies. Wen et al. (2012) mention that one of the goals of their study is to improve the accuracy and efficiency of machine translation. The authors point out two pieces of research, one in the Thai language by Sornlertlamvanich et al. (1994) and a derivative of their work for Korean and Japanese by Bond and Paik (2000); Paik and Bond (2001). These studies involve using noun-classifier pairs, where the default classifier is assigned to the most common noun. The improved algorithm by Bond and Paik compared to Sornlertlamvanich can handle nouns, which belong in multiple semantic classes better.

In Wen et al. (2012), classifiers are categorized into seven categories: sortal, mensural, event, group, anaphoric, non-classifier, and other. Only sortal-classifier are considered in their study, and these classifiers are annotated by hand. For the Chinese language, they were able to correctly analyze 79.37% of semantic classes of the noun phrases, and by using the default classifier for each semantic class, they were able to generate 78.8% of the classifiers correctly. In their case, each classifier had to match the original classifier in the annotated corpus exactly. They note that the study's WordNet does not have full coverage of all the nouns in the world, so the result might be slightly skewed.

Da Costa et al. (2016) do a similar study by creating an algorithm that automatically builds a frequency-based dictionary of noun-classifier pairs, which are mapped to concepts in the Chinese Open WordNet as described in Wang and Bond (2013). The set of sortal classifiers they use is not the same as Wen et al. (2012). They achieve a human-validated performance of 87% for their algorithm. In their study, a native Chinese speaker evaluated whether the generated classifiers were acceptable, so it is not precisely comparable to the study by Wen et al. (2012), where the classifiers were compared directly to the original corpus, meaning there was only one correct option.

The study by Gao (2011), on the other hand, attempts to create an e-learning tool, which is an automatic system for associating classifiers with nouns to help learners of Chinese. The system they create associates the semantic meanings of nouns from a database to classifiers. They mention that their system works well for associating the most commonly used classifiers and their associated nouns. However, they have not tested all pairs, such as those with "fuzzy boundaries", and they do not provide quantitative results. In relation to this study, Da Costa

et al. (2016) reference the work by Huang et al. (1998) by saying that the many-to-one selective associations between nouns and classifiers in the work by Gao (2011) are challenging to keep track of because they depend much on context.

According to Peinelt et al. (2017), the study by Guo and Zhong (2005) is the only previous machine learning approach to classifier prediction. Guo and Zhong (2005) use a Support Vector Machine multiclass implementation called LIBSVM (Chang and Lin, 2011) for their study. The best accuracy they can achieve is 58.71% using two SVMS, of which one is trained on the noun and one on the ontological features. However, they had a native Chinese-speaking human participant evaluate the acceptability of 241 randomly chosen generated classifiers from the experiment mentioned above. The judge rated 94.2% of the generated classifiers as either acceptable or good. With a rating of 1 meaning acceptable and a rating of 2 meaning good, the average score, in this case, is 1.80 (Guo and Zhong, 2005). This shows how much acceptable variability there is to choosing a classifier: sometimes, two or more different classifiers are equally acceptable choices without changing the meaning of the sentence at all. Other times, a different classifier can subtly affect the sentence's meaning but still be an acceptable choice.

2.3.2 Context-Aware Models

The most recent study by Peinelt et al. (2017) focuses on using a context-aware machine learning model. According to the authors, there is no dictionary-based approach or a finite set of rules covering all possible classifier-head combinations exhaustively. They also mention that previous approaches have predominantly relied on ontological resources, which require much human effort to build and maintain. According to them, this results in limited coverage for new words and domains. They further add that without context, classifier assignment can be ambiguous. Thus, Peinelt et al. (2017) argue that context is an essential factor for classifier selection, as context defines which classifier should be assigned to the noun. These observations also reflect those made in section 2.1.2 of this present thesis.

In the study, Peinelt et al. (2017) compare their models to the following baselines:

1. *Ge* 个: always assign the universal classifier in every case (Guo and Zhong, 2005; Da Costa et al., 2016)

2. Pairs: assign the most frequently observed classifier in combination with the headword. Assign *ge* 个 for all unseen words (Guo and Zhong, 2005).

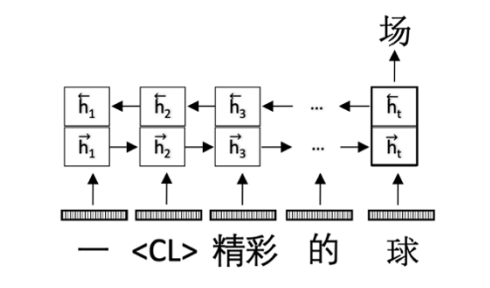
3. Concepts: assign classifiers based on classifier-concept pair counts using the Chinese Open Wordnet and *ge* 个 for unseen words (Da Costa et al., 2016).

They train word embeddings with Word2vec (Mikolov et al., 2013) on sentences from the original three corpora and obtain pre-trained word embeddings

from Bojanowski et al. (2017). According to them, the pre-trained word embeddings produce better results and were thus used in subsequent experiments.

They also tried to find out how much context adds to the performance of the models, so they first trained two models using SVM and Logistic regression on the headword's embedding vector. They also used a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to encode the entire sentence, excluding any headword annotation and predicting classifiers based on the last hidden state. Figure 2 describes how the LSTM system encodes sentences.

Figure 2: Description of how the LSTM system by Peinelt et al. (2017, p. 43) encodes sentences.



The results from Peinelt et al. (2017) provide the following essential points:

1. Always assigning *ge* 个 to all cases produces an accuracy score of 45.21.
2. Using pairs produces a better result (61.72) than just using SVM (53.72) and Logistic regression models (57.72) trained on only headword embeddings.
3. Adding headword context features results in improvements for SVM (66.02) and for Logistic Regression (67.67).
4. The best model, LSTM, achieves an accuracy score of 71.51 on the test set based on the full sentence context.

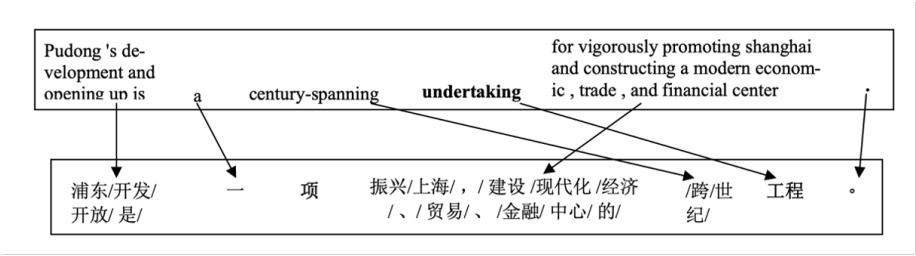
From the results in Peinelt et al. (2017), it is clear that adding contextual features to the models improves the prediction accuracy. This again strengthens the observation that, in some cases of choosing a classifier, sentence context plays a large role, and that assigning classifiers just based on the head noun is not accurate.

2.4 Other Observations

Zhang et al. (2008) mention that a significant problem in statistical machine translation is the long-distance dependency between the classifier and the headword, meaning the headword and classifier are separated by other words in the

sentences. According to them, most classifiers are close to the headword they modify, but more than 16% of classifiers are separated from their corresponding headword by at least five words. Figure 3 presents a case where the classifier and headword are separated by 15 words. Many of the earlier studies differ in whether they consider classifiers separated by such a great extent, and this information is not always available.

Figure 3: Example from Zhang et al. (2008, p. 90) where the classifier and headword are separated by 15 words.



We want to also find out whether a large distance between the classifier and the head noun could have an effect on generating certain classifiers. We intend to analyze this aspect post hoc.

2.5 Conclusions from Earlier Studies

The studies related to the prediction of Chinese classifiers have a considerable variety. Thus, it is difficult to compare the different studies in classifier generation to each other due to the significant differences in their characteristics. When evaluating the results of these studies, one should ask the following questions:

- Is the focus on true classifiers only or also on measure words? How is this distinction made?
- Is the study considering classifier-noun pairs or complete sentences?
- Are the generated classifiers required to match the original corpus exactly or are they evaluated by a human participant, which means there could be more than one acceptable choice?

Moreover, all of this information is not available in all of the studies, so any comparisons made between the results of different studies should be taken with a grain of salt. Most of the studies provide multiple different kinds of results: accuracy results, F1 scores, other accuracy measures. Besides, the

studies provide multiple different scores for different kinds of pre-processing, small changes in the implementations of the methods, and different datasets. Thus, listing the results and comparing the results directly is difficult. However, we can make some general conclusions about the nature of the task of classifier generation.

Firstly, we should consider that *ge* 个 represents a large number of all classifiers in a given dataset. In Peinelt et al. (2017), guessing *ge* 个 for all cases results in an accuracy of 45.21%. Similarly, Da Costa et al. (2016) receive a score of "roughly 40%" while assigning *ge* 个 to all entries in their dictionary. Thus, we should assume that the general classifier *ge* 个 constitutes a huge part of all classifier usage in Mandarin.

Furthermore, we should consider that language production is not always deterministic. For instance, in Guo and Zhong (2005), the best result achieved by an SVM implementation is 58.71% when comparing the generated classifiers to the original corpus. However, when a human judge rates the generated classifiers' acceptability, 94.2% of classifiers are rated as being acceptable. As multiple different choices of classifiers might be acceptable, we should not assume that the original classifier is the only *right* choice. However, further human experiments would need to be completed to produce a corpus that would help to analyze where multiple choices are equally acceptable. However, completing human experiments is outside the scope of this study, so we frame the classifier prediction problem as attempting to generate the exact classifier in the original sentence. Suggestions for future human experiments will be further discussed in section 5.3.

There is also more evidence pointing to the fact that certain types of classifiers are harder to predict than others. For example, the study by Wen et al. (2012), which only considered true classifiers, achieved an accuracy score 78.8%. The results of studies that consider all classifiers (except certain measure units), such as Peinelt et al. (2017), achieve a worse accuracy (71.51%), which may indicate that true classifiers are more straightforward to generate than other kinds of classifiers. This is precisely why we want to analyse accuracy of the algorithms on different categories of classifiers, so that we can demonstrate if this is indeed the case.

3 The Models

In this section we outline how we approach the problem of classifier prediction, the dataset used in this study and the models we build for the task of predicting classifiers in Mandarin.

3.1 The Approach to Predicting Classifiers

As the study by Peinelt et al. (2017) features the most sophisticated model for classifier prediction, we follow their approach. This also allows us to make

comparisons to their study, and to provide new insights into their results by analyzing certain subsets of classifiers.

They approach the problem as follows: given a sentence in which a classifier is yet to be realised, and the head noun is flagged, predict the missing classifier. For example, in the input:

Yī <CL> jīngcǎi de <h>qiúsài</h>.

<CL> indicates where the missing classifier is, and the <h> tag pair flags the head noun.

The job of the algorithms is to predict the missing classifier. This is achieved differently for each model we build. In the case of the RULE model, the model assigns each head noun a classifier, which is then compared to the original classifier. For the MLM model, the word generated by the model is compared to the original classifier. For the classifier models, LSTM and BERT, each sentence is classified in one of 172 classes. The original class label is then compared to the assigned class. Through this process, we are able to find out how well the algorithms are able to predict the original classifiers.

3.2 The Dataset

This study uses a large-scale dataset of everyday Chinese classifier usage constructed by Peinelt et al. (2017), namely ChineseClassifierDataset[11] (henceforth, CCD). The dataset is based on three POS tagged Chinese language corpora:

1. The Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004)
2. The UCLA Corpus of Written Chinese (Tao and Xiao, 2012)
3. Leiden Weibo Corpus (Van Esch, 2012)

To construct the dataset, Peinelt et al. (2017) perform the following actions on the original corpus data:

1. Sentences are assigned unique IDs.
2. Sentences are filtered for the occurrence of classifier POS tags.
3. The data is filtered. Figure 4 presents the filtering steps that have been taken to improve the data quality.

[11] `github.com/wuningxi/ChineseClassifierDataset`

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

Figure 4: Filtering steps taken to improve data quality. From Peinelt et al. (2017, p. 42).

4. Sentences are parsed with the Stanford constituent parses (Levy and Manning, 2003).

5. The head of the classifier is extracted in each sentence based on the parse tree.

Through randomly sampling 100 sentences, the authors estimate that the system can identify classifiers with an accuracy of 91% and headwords with an accuracy of 78%. They observe that most errors are due to accumulating tokenization, tagging and parsing errors, and elliptic classifier usage. As is clear, the dataset is lacking in this aspect and this limitation should be considered when evaluating the results.

Peinelt et al. (2017) also mention that not all sentences in their database contain headwords due to "co-referential and anaphoric usage." They query the database for only sentences in which both the headword and corresponding classifier were identified. In sentences where both a classifier and a headword are recognized, the classifiers are substituted with gap tokens <CL>, and the classifier is used as its class label:

我们是一 <cl> 家 </cl><h> 人 </h>。

→ 我们是一 <CL><h> 人 </h>。Class label = 家

After cleaning and applying the aforementioned filters, there are a total of 681,102 sentences in the CCD dataset. The dataset is further split into 60% training set, 20% development set, and 20% test set.

Through personal communication with the authors of Peinelt et al. (2017), we have also learned the details of filtering step 4, which removes all "measure units". The measure units were filtered out using two lists. The first list[12] was created using Baidu Baike articles[13] for common, international and scientific measure units. The second list[14] is based on a list of measure units from the appendix of a Chinese classifier dictionary.[15] The lists focus on exactly defined units for weights, volumes, distances, time, temperature, money and more. However, container units are not filtered out from the dataset.

It is important to note that some linguists might also consider some of the measure units, that have been filtered out, as classifiers. For this reason, the dataset is not necessarily a complete overview of Mandarin classifiers.

The final dataset contains the following 172 classifiers: 个, 种, 次, 张, 件, 句, 条, 位, 家, 场, 只, 点, 部, 首, 段, 篇, 滴, 份, 号, 块, 颗, 名, 群, 款, 片, 堆, 本, 些, 级, 分, 杯, 起, 步, 顿, 套, 把, 集, 对, 辆, 碗, 回, 代, 声, 座, 届, 阵, 道, 类, 班, 层, 项, 番, 口, 双, 支, 台, 朵, 瓶, 等, 股, 丝, 根, 趟, 封, 包, 轮, 头, 幅, 遍, 副, 门, 粒, 枚, 盘, 组, 批, 间, 笔, 身, 棵, 波, 样, 桌, 楼, 季, 盒, 下, 盆, 面, 箱, 处, 页, 节, 串, 排, 栋, 系列, 袋, 锅, 盏, 束, 通, 团, 圈, 所, 发, 世, 扇, 桶, 壶, 餐, 堂, 则, 艘, 架, 曲, 线, 匹, 户, 肚子, 笼, 手, 伙, 枝, 卷, 罐, 幕, 码, 行, 株, 刀, 任, 脸, 环, 幢, 辈, 般, 桩, 顶, 尾, 尊, 册, 列, 章, 路, 宗, 版, 杆, 袭, 拨, 记, 剂, 具, 帖, 队, 例, 局, 味, 席, 管, 档子, 人次, 缸, 缕, 遭, 拳, 员, 堵, 棒, 眼, 帮, 族.

We further use the categorization by Her and Lai (2012) to split these 172 classifiers into four categories for our analysis in section 4. The categories are: *true classifiers*, *dual classifiers*, *measure words*. In addition, those classifiers which are not present in the listing by Her and Lai (2012), we categorize as *not in list*[16]. The 172 classifiers are cateforized in the following way:

True Classifiers: 个, 张, 件, 句, 条, 位, 只, 首, 篇, 颗, 名, 本, 辆, 声, 座, 道, 朵, 根, 封, 头, 幅, 粒, 枚, 间, 笔, 棵, 面, 处, 栋, 盏, 所, 发, 扇, 则, 艘, 架, 曲, 匹, 枝, 卷, 株, 幢, 顶, 尾, 尊, 册, 杆, 袭, 记, 剂, 具, 席, 管, 员, 堵.

Dual Classifiers: 家, 点, 部, 份, 块, 片, 分, 起, 把, 班, 口, 支, 台, 轮, 门, 页, 节, 线, 宗, 缕.

Measure Words: 段, 滴, 号, 群, 款, 堆, 级, 套, 集, 对, 回, 代, 层, 项, 双, 股,

[12]The first list of measure units: 米, 公斤, 秒, 安培, 安, 开, 开尔文, 摩, 微克, 摩尔, 坎, 坎德拉, 弧度, 球面度, 赫兹, 赫, 牛, 牛顿, 帕斯卡, 帕, 焦, 焦耳, 瓦特, 毫克, 瓦, 库仑, 库, 伏特, 伏, 法拉, 法, 欧姆, 欧, 西门子, 西, 韦伯, 韦, 特斯拉, 特, 亨利, 亨, 摄氏度, 流明, 勒克斯, 勒, 贝可, 贝克勒尔, 戈瑞, 戈, 希瓦特, 希, 分钟, 时, 小时, 度, 秒, 角分, 角秒, 升, 吨, 海里, 电子伏, 分贝, 特克斯, 公顷, 顷, 十, 百, 千, 兆, 吉咖, 吉, 太, 太拉, 拍它, 拍, 艾可萨, 艾, 泽它, 尧它, 厘, 毫, 微, 纳, 纳诺, 皮克, 皮, 飞母托, 飞, 阿, 阿托, 分米, 厘米, 毫米, 微米, 纳米, 千米, 公里, 公尺, 公分, 公里, 公丝, 丝米, 忽米, 公微, 毫微米, 市尺, 费密, 埃, 英里, 英尺, 英寸, 英寻, 平米, 平方米, 公亩, 亩, 英亩, 平方码, 平方英尺, 平方英里, 平方米, 公升, 立升, 立米, 立方, 立方码, 立方英尺, 美加仑, 英加仑, 加仑, 立方米, 毫升, 升, 年, 日, 天, 时, 秒, 周, 兆周, 千周, 赫, 兆赫, 兆赫兹, 千赫兹, 千赫, 吨, 千克, 公斤, 斤, 公两, 两, 克, 英吨, 美吨, 盎司, 格令, 道尔顿, 开氏度, 度, 华氏度, 列氏度, 磅, 巴, 托, 泊, 厘泊.

[13]https://baike.baidu.com/

[14]The second list of measure units: 秒, 刻, 天, 日, 美元, 周岁, 月, 星期, 年, 载, 克, 两, 加仑, 斤, 吨, 公分, 米, 厘米, 毫米, 寸, 尺, 里, 哩, 度, 单位, 秒差距, 圆, 美金, 元, 蚊, 角, 毛, 毫分, 岁, 丈, 亩, 顷, 升, 千瓦, 合, 欧元, 千瓦时, 小时, 时辰, 秒, 时, 钟, 摄氏度, 歲, 英镑, ℃, 毫, 平方, 厘.

[15]刘子平 (2015). 汉语量词大词典. 上海: 上海辞书出版社.

[16]Please see appendix A for the original lists by Her and Lai (2012).

丝, 副, 组, 批, 身, 桌, 串, 排, 束, 团, 圈, 堂, 户, 伙, 行, 列, 章, 路, 版, 帖, 队, 味.

Not in list: 种, 次, 场, 些, 杯, 步, 顿, 碗, 届, 阵, 类, 番, 瓶, 等, 趟, 包, 遍, 盘, 波, 样, 楼, 季, 盒, 下, 盆, 箱, 系列, 袋, 锅, 通, 世, 桶, 壶, 餐, 肚子, 笼, 手, 罐, 幕, 码, 刀, 任, 脸, 环, 辈, 般, 桩, 拨, 例, 局, 档子, 人次, 缸, 遭, 拳, 棒.

3.3 Setup

In this section, we describe the setups of our four algorithms, that are used in this study. We build four models for the task of predicting classifiers: a rule-based model, an LSTM implementation, a BERT masked language model and a BERT classification model. Each model is described in detail in the following subsections. These models will allow us to make overall comparisons and to also analyse the performances of the different models with respect to our subsets of classifiers.

3.3.1 Rule-based Model

We build a simple rule-based model (RULE) following the work by Peinelt et al. (2017). The model uses the following rules to assign classifiers:

1. given a head noun, assign the most frequent classifier associated with it in the training data.
2. if two or more classifiers are equally frequent, one of the classifiers is randomly assigned.
3. If the head noun does not appear in the training data, then the classifier *ge* 个, is assigned.

The RULE model is only given the head noun of each sentence in the dataset. For example, for the sentence:

我们是一 <CL><h> 人 </h>。Class label = 家.

The model is provided with just the head noun:

人

The assigned classifier is then compared to the the original class label = 家.

3.3.2 LSTM

We also build a bidirectional LSTM model similarly to Peinelt et al. (2017), and frame the task as a 172 class multi-class classification task. For the training, we set the batch size to 256, the hidden size to 300, and the learning rate to 2e-5. As Peinelt et al. (2017) used novel pre-trained word embeddings, we also use

pre-trained Chinese word embeddings from Li et al. (2018)[17] in order to make our implementation as comparable to that of Peinelt et al. (2017) as possible.

Furthermore, we observe that the average length of the sentences in the training set is 49.99, with the longest sentence being 243. We decide to pad all the sentences to a length of 100 for computational reasons.

We train for 8 epochs and save the best model. Our hyperparameters have been tuned manually (i.e. using sensible defaults and testing a couple of options) and we have not completed comprehensive hyperparameter tuning on the model due to computational and time limits.

We follow the LSTM implementation by Peinelt et al. (2017) and drop the head flag for each sentence. Thus, each sentence in the dataset is provided to the model in the following format:

我们是一 <CL> 人。

The model then assigns each sentence to one of the 172 classes, and the assigned class is compared to the real class label.

3.3.3 The BERT Models

We use a pre-trained language representation model called BERT (Bidirectional Encoder Representations from Transformers) as presented in Devlin et al. (2018). Since its release, BERT has produced state-of-the-art results on multiple natural language processing tasks (Devlin et al., 2018). According to the authors, BERT is able to perform multiple different kinds tasks including question answering and language inference. In addition, as the model is publicly available, multiple modifications have been created for other kinds of tasks, too (Wolf et al., 2020).

It should be noted that the performance of BERT can depend on how the language model has been trained. There are dozens of different implementations publicly available (Wolf et al., 2020). Because the models are trained on different data, namely the pre-training is done on differently sized and differently structured data, the training data influences the model in different ways. In addition, some models employ different kinds of training tasks. This means that the exact results of this experiment might change if we used a different implementation (Liu et al., 2019). For both of our BERT implementations, we use the "bert-base-chinese" version[18] of BERT.

BERT Masked Language Model BERT is a suitable choice for the task presented here, as BERT is based on a *masked language model (MLM)* pre-training objective (Devlin et al., 2018). The authors mention that this involves masking a part of the tokens of the input and then predicting the original vocabulary based only on its context.

[17] These are word embeddings trained by skip-gram on 9 large Chinese corpora with 300 dimensions. It is available at: `github.com/Embedding/Chinese-Word-Vectors`

[18] `huggingface.co/bert-base-chinese`

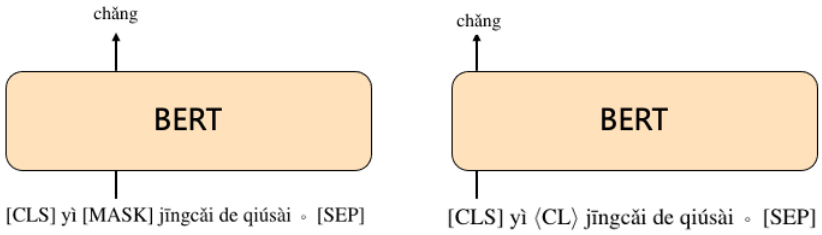


Figure 5: Sketch of our BERT-based Classifier selection models: predicting the classifier by unmasking the [MASK] (left); predicting the classifier as classification (right).

In order to assess how well BERT, as a masked language model, can model classifiers, we tried to use BERT without any fine-tuning on the task of classifier selection. Specifically, as shown in Figure 5 (left), we replace the classifier indicator <CL> with the [MASK] symbol of BERT and ask BERT to unmask it.[19] The unmasked token serves as the predicted classifier. Note that addressing the classifier selection task in this way will sometimes produce words that are not classifiers.

We refer to this model as MLM.

BERT Classification Model Additionally, we test BERT in its classic use. We use the Simple Transformers package[20], which has been built over the Huggingface infrastructure (Wolf et al., 2020) to construct a multi-class classification model. We fine-tune BERT on the CCD as a multi-class classification task, where there are 172 classes (i.e., 172 classifier words) in total, and make a prediction with the help of the [CLS] symbol (see Figure 5 (right)). We refer to this model as BERT.

When fine-tuning, we set the learning rate to 2e-5 and batch size to 150. Our hyperparameters have been tuned manually (i.e. using sensible defaults and testing a couple of options) and we have not completed comprehensive hyperparameter tuning on the model due to computational and time limits.

4 Comparison of the Models

In this section, in line with the plan that was formulated in sections 1 and 3.1, we want to dive deeper into how well each of these models performs, and how they differ in terms of what they do well and what they do not do well. Hence, we present the results of the algorithms for the whole dataset, for different categories of classifiers and for a subset of special classifiers, that *add information*.

[19]Since our experiments suggested that the head flag (i.e., ⟨h⟩ and ⟨/h⟩) makes no contribution to classifier selection, we drop it to speed up the prediction.

[20]https://simpletransformers.ai/

4.1 Overall Results

		Macro-averaged			Weighted-averaged		
Model	Accuracy	Precision	Recall	F1	Precision	Recall	F1
`RULE`	61.89	34.87	20.50	23.39	58.23	61.90	58.24
`LSTM`	<u>70.44</u>	33.11	20.12	22.48	67.90	<u>70.44</u>	68.12
`MLM`	62.22	<u>51.91</u>	<u>33.40</u>	<u>37.68</u>	<u>77.28</u>	62.23	<u>68.21</u>
`BERT`	**81.71**	**52.86**	**38.10**	**40.77**	**80.70**	**81.71**	**80.77**

Table 1: Evaluation Results of each model on CCD. The best results are **boldfaced**, whereas the second best are <u>underlined</u>. `MLM` is the model that uses BERT as a masked language model, while `BERT` is the fine-tuned BERT. The macro-averaged scores are the arithmetic means of each metric. The weighted-averaged scores are weighted by the number of instances in each class.

Table 1 charts the performance of each model.[21] BERT clearly performs the best on all metrics, achieving an accuracy score of 81.71%, while the second best accuracy score, 70.44%, is achieved by the LSTM model. In addition, the LSTM produced the second best weighted-average recall. The MLM performs the second best on all other metrics, which include macro-averaged precision, recall and F1, and also on weighted-averaged Precision and F1 score. The rule-based model achieves the lowest accuracy with 61.89%.[22]

4.2 True Classifiers, Dual Classifiers, Measure Words

Model	RULE	LSTM	MLM	BERT	Frequency
`Whole Dataset`	61.90	<u>70.44</u>	62.23	**81.71**	136221
`True Classifiers`	78.30	<u>80.57</u>	68.70	**87.81**	85917
`Dual Classifiers`	29.91	40.12	<u>47.29</u>	**65.19**	10817
`Measure Words`	22.47	<u>37.69</u>	36.99	**61.51**	11317
`Not in List`	39.98	<u>64.35</u>	58.35	**77.56**	28170

Table 2: Evaluation Results of each model on different categories of classifiers. The best results are **boldfaced**, whereas the second best are <u>underlined</u>. `MLM` is the model that uses BERT as a masked language model, while `BERT` is the fine-tuned BERT. `Frequency` represents how many cases are present in the test set.

Table 2 presents the results for the whole dataset and breaks down the performance of the models by subsets of classifiers based on the categorization by

[21] The raw results for each model are available in appendix B.

[22] The results of both our RULE model and bidirectional LSTM model are comparable to those presented in Peinelt et al. (2017).

Her and Lai (2012). These subtypes include *True Classifiers, Dual Classifiers, Measure Words* and those classifiers that are not listed by Her and Lai (2012), which are labelled *Not in List*.

The BERT model performs the best on all metrics, while the LSTM performs the second best on everything except *Dual Classifiers*, as the MLM achieves the second best result in this category. All the models perform worse on *Dual Classifiers, Measure Words* and those classifiers *Not in List* compared to *True Classifiers.*

We can also see that the differences in accuracies for the *True Classifiers* between RULE, LSTM and BERT are significantly smaller, 78.30 vs 80.57 vs 87.81, than compared to the differences between *Dual Classifiers*, 29.91 vs 40.12 vs 65.19 and *Measure Words* 22.47 vs 37.69 vs 61.51. The same applies to those classifiers *Not in List*. It is clear that the BERT model makes up a big part of its overall accuracy improvement compared to the other models in *Dual Classifiers, Measure Words* and *Not in List* instead of *True Classifiers*.

4.3 Classifiers that *Add Information*

Classifier	RULE	LSTM	MLM	BERT	Frequency
个 `Ge`	86.98	88.65	71.63	**92,79**	61581
`Range 800-2200 avg.`	41.48	54.44	52.00	**73.20**	20156
位 `wèi`	18.30	35.08	39.99	**59.87**	2158
名 `míng`	29.63	46.54	59.51	**70.99**	810
`Range 280-800 avg.`	26.79	36.28	46.31	**64.24**	14976
群 `qún`	0.88	16.67	16.92	**52.51**	798
些 `xiē`	2.86	21.75	37.30	**56.51**	630
堆 `duī`	4.38	30.36	23.50	**52.12**	685
套 `tào`	8.10	16,63	21,88	**34,57**	457
对 `dùi`	35,55	49,31	44,72	**62,39**	436
双 `shuāng`	42,46	41,40	63,16	**76,49**	285

Table 3: Evaluation Results of each model on the general classifier ge 个, politeness and plurality. As a comparison, average results are also provided for all classifiers in their respective range. The best results are **boldfaced**, whereas the second best are underlined. `MLM` is the model that uses BERT as a masked language model, while `BERT` is the fine-tuned BERT. `Frequency` represents how many cases are present in the test set.

To analyse a subset of classifiers that we consider to *add information*, we choose classifiers that appear frequently enough in the dataset. We consider classifiers that appear more than 200 times in the test set to be frequent enough. This inclusion criteria is applied as we believe data sparsity could have an effect on the prediction results, i.e. less common classifiers are more difficult to predict

because the model has not seen them in training often enough.

We pick classifiers that imply politeness and plurality, as we believe context plays a large part in predicting these classifiers. The subsets were chosen by using insights from section 2 to find classifiers that we observe as having a large effect on the meaning of the whole sentence, i.e. choosing another classifier such as *ge* 个 instead of the classifiers in the plurality subset changes the meaning of the sentence drastically (plurality vs. singularity). It should be noted that this subset of classifiers is based on rather subjective opinions, and that *adding information* is not generally agreed to be a well-established concept in the literature. This subset should neither be seen as suggesting a formal definition nor being exhaustive. We choose to call this subset of classifiers as the classifiers that *add information* for clarity and practical purposes.

To offer comparisons of this subset to a general classifier, which we believe does not require as much context to predict, we show the results of the models for ge 个 in table 3. BERT outperformed the other models with a score of 92.79. The MLM was the second best with a result of 88.65, closely followed by RULE with 86.98. LSTM clearly had the worst performance with 71.63. We also provide the average scores for the frequency ranges of 800-2200 and 280-800 as comparisons for the respective subsets of classifiers.

For a politeness classifier, wèi 位 and míng 名 were the only classifiers that were frequent enough in the dataset. Both of the classifiers are considered *true classifiers* by Her and Lai (2012). From table 3 we can see that BERT clearly performed the best in predicting both wèi 位 and míng 名 with accuracy scores of 59.87 and 70.99, respectively. This is considerably higher than the scores for the RULE model, 18.30 and 29.68 respectively. The results of the MLM are the second best for both classifiers with scores of 39.99 and 59.51, respectively. However, the results of all models are below their respective average for the frequency range of 800-2200.

To represent plurality, we chose six classifiers of which four refer to "multiple" (i.e., qún 群, duī 堆, xiē 些, and 套 tào) and two refer to "pair" (i.e., duì 对 and shuāng 双). Five out of the six classifiers are considered measure words by Her and Lai (2012), while one of them, xiē 些, is not in their lists. BERT also performed the best for all the plurality classifiers. LSTM and MLM both had some second best places, while the RULE model performed the worst in all cases.

The differences between the models are generally larger for the politeness and plurality classifiers compared to the general classifier ge 个. BERT clearly performs the best in all these cases, and RULE performs the worst. It is also interesting to note that the MLM model performs the second best for 6 out of the 8 cases. The results of all the models for those classifiers referring to multiple, qún 群, duī 堆, xiē 些, and 套 tào, are below their respective averages for the frequency range of 280-800. For the classifiers referring to pair, duì 对 and shuāng 双, the results are generally better than for the average of the frequency range of 280-800.

5 Discussion

5.1 Answering the Research Questions

This section will focus on answering the research question as introduced in section 1 of this thesis by analysing the results presented in section 4, and discussing what kinds of inferences can be made based on the results. In addition, we will provide some post-hoc analyses and suggestions for further study.

5.1.1 Baselines vs. BERT

First, we consider research question 1:

- How well is it possible to do on this task?

In terms of the accuracy score, the BERT classification model has clearly beat the bidirectional LSTM model by Peinelt et al. (2017), 81.71 vs. 71.51. Thus, we have produced a new state-of-the-art result for predicting Chinese classifiers in the CCD dataset. However, to completely answer this question, we have to consider our other research questions first.

Hence, next we consider research question 2:

- How do the different models compare to each other in performance?

The BERT model clearly outperforms all other models in every metric. Looking at only the accuracy score, the LSTM performs the second best. It should be noted that the MLM model produces both considerably high macro-averaged and also weighted-averaged precision, recall and F1 scores.

It should be examined why the MLM achieves such high scores in these metrics. The reason for this is that the MLM model is able to produce a larger variety of classifiers, many of which have very low support in the dataset. To demonstrate this, we calculate the macro-averaged recall for all classifiers which have less than 50 support in the dataset. This includes 53 classifiers.[23] As a comparison, we also calculate the macro-averaged recall for all classifiers which have more than 1000 support in the dataset. This includes 20 classifiers.[24]

Table 4 clearly shows that the MLM model is able to predict rare cases of classifiers the best. It even beats the overall best model BERT. The LSTM model is especially bad at generating these rare classifiers. This is confirmed by looking at only those classifiers with over 1000 support which reflect the overall results — with the BERT model performing the best and the LSTM the second best.

The low performance of both the supervised LSTM and BERT classification models on rare classifiers could be explained by the highly imbalanced dataset:

[23]The classifier are the following: 曲, 匹, 户, 肚子, 笼, 手, 伙, 枝, 罐, 卷, 幕, 行, 码, 刀, 任, 株, 脸, 环, 辈, 幢, 般, 桩, 顶, 尾, 尊, 列, 册, 章, 路, 宗, 版, 杆, 拨, 袭, 记, 剂, 具, 帖, 队, 味, 例, 局, 席, 管, 档子, 人次, 缸, 缕, 遭, 拳, 棒, 员, 堵.

[24]The classifiers are the following: 个, 种, 次, 张, 件, 句, 条, 位, 家, 场, 只, 点, 部, 首, 段, 篇, 滴, 份, 号, 块.

Support	Rule	LSTM	MLM	BERT	Frequency
<50	8.70	2.37	**19.39**	<u>11.58</u>	1185
>1000	49.50	<u>62.35</u>	56.30	**77.83**	110827

Table 4: Macro-Averaged Recall of each model on those classifiers with <50 support and >1000 support. The best results are **boldfaced**, whereas the second best are <u>underlined</u>. `MLM` is the model that uses BERT as a masked language model, while `BERT` is the fine-tuned BERT. `Frequency` represents how many cases are in the range.

the classification models have not seen the classifiers often enough to be able to predict them correctly. Instead, they produce more commonly appearing classifiers. Imbalanced datasets present a challenge for most classification models, and thus this observation is not suprising. However, the unsupervised MLM model is able to generate even very rare classifiers surprisingly well. This could be due to the model being pretrained on a very large scale dataset. Although BERT is also pretrained on the same dataset, as it is a supervised model, it is also fine tuned on a classification task using the CCD dataset. Thus, it likely learns to predict more common classifiers more confidently from the imbalanced CCD dataset instead of the rare cases. It should be noted that the imbalanced distribution of classifiers in the dataset is a natural feature of classifiers in Mandarin: some classifiers appear dozens of times more frequently than others. The results for the BERT and LSTM classification models could be different if the imbalanced dataset issue was mitigated by balacing the dataset.

The high weighted-average scores of the MLM can also be explained by its ability to predict rare classifiers. Out of all the 169 classifiers in the test dataset, the RULE model never correctly predicted 34 classifiers, while the LSTM never correctly predicted 66, the MLM never correctly predicted 14 and the BERT model never correctly predicted 34 classifiers. It is clear from this that the MLM can generate a much larger variety of classifiers than the other models, which have been trained using an imbalanced dataset.

In summary, the BERT model clearly performs the best on the whole dataset, and the performance of the MLM model and RULE model are similar. However, the MLM model is undoubtedly able to produce a larger variety of classifiers and generate rare classifiers better than any other model. This is likely due to highly imbalanced distribution of classifiers in the CCD dataset, which affects the performance of the supervised models on the rare cases.

5.1.2 True Classifiers, Dual Classifiers, Measure Words

Now we look at research question 3:

- Are some kinds of classifiers harder to predict than others? In order to answer this question, we look at certain subsets of classifiers.

The results of the models in terms of the categorization into *true classifiers*, *dual classifiers, measure words* and those *not in list* shows some interesting aspects. Firstly, BERT clearly performs the best in all categories, while the LSTM performs the second best for everything except dual classifiers. A common feature of all the models is that true classifiers are much easier to predict than the other categories. Compared to true classifiers, the scores for measure words are much lower. This makes sense as many measure words could be considered lexical elements, i.e. predicting some measure words is comparable to trying to predict any random noun in a given sentence. For example, generating measure words would not be needed in automatic machine translation as most measure words, such as containers or quantities, would be present in the original text and thus would only need to be translated into their equivalent words in Mandarin. This is different from generating true classifiers as those are not present, for example, in an English text.

The reason BERT performs the best on measure words is also not surprising given that BERT is able to understand the context of sentences better than the RULE or LSTM models. Thus, it is able to produce the correct lexical element more often. This also confirms our first hypothesis that BERT's advantage over its competitors is larger for measure words. The reason the MLM model, which has the same pretraining as BERT, does not perform so well is likely because the model can produce any word in its vocabulary. Thus, there is a much larger inventory of possible lexical elements for it to choose from than for the BERT model which can only choose 172 classes. In fact, the MLM model generated 1566 words that are not classifiers. If the results of the MLM model were evaluated by a human participant, it might be possible that they may rate some of these generated words as acceptable choices, especially if they are measure words such as containers, i.e. cup vs. glass vs. mug, etc.

The RULE model does a very bad job in predicting both dual classifiers and measure words. It is not surprising as these classifiers likely have a high variety of head nouns. For instance, a measure word meaning *pair* can likely occur with a large amount of different head nouns.

If we were to only consider true classifiers, the RULE model performs remarkably well — almost in line with the LSTM and only around 10 percentage points below the performance of BERT. This observation confirms our second hypothesis that the RULE model's disadvantage over BERT is smaller for true classifiers than other cases. Thus, it could be said that using a rule-based approach for assigning true classifiers is very reasonable, especially considering its simplicity. This aspect was not clear from the results of Peinelt et al. (2017), and thus we have shown that a simple rule-based model should not be underrated.

As a summary, dual classifiers and measure words are clearly more difficult to predict than true classifiers. This relates to the literature which states that true classifiers should be considered different from measure words. It is clear that many measure words behave differently than most true classifiers (i.e. they are lexical elements), and this explains why the models are unable to predict them well. If we were to only consider true classifiers, the BERT model predicts almost 9 in 10 classifiers correctly. The RULE model, considering how simple

it is compared to the BERT classification model, performs remarkably well, predicting almost 8 in 10 true classifiers correctly.

5.1.3 Classifiers that *Add Information*

Lastly, we consider research question 4:

- How do the models perform on a subset of classifiers which *add information*?

For the general classifier, ge 个, we can observe that all the models, except the MLM perform remarkably well. This might be explained again by the fact that ge is extremely prevalent in the dataset, representing 61581 of the 136221 cases (45.2%). Due to this, the classification models likely learn to predict it very well. In addition, as the RULE model performs so well, it is clear that ge 个 appears with certain headwords very often, i.e. this subset of headwords does not have much variability.

The polite classifiers, wèi 位 and míng 名, are predicted much less well compared to the general classifier, ge 个. In addition, all the prediction results for wèi 位 and míng 名 are below the average accuracy for the range of 800-2200 frequency. The RULE model performs especially badly for both classifiers. This is likely because many headwords that take wèi 位 and míng 名 can also take ge 个 in cases where the speaker does not want to emphasize politeness.

BERT performs considerably well in picking míng 名, but less well picking wèi 位, even though it still clearly performs the best. The accuracy scores for the polite classifiers are 59.87 and 70.99, compared to 73.2 for all classifiers in the frequency range of 800-2200. Thus, the score of wèi 位 is significantly lower than for other classifiers in the same frequency range. However, the score for míng 名 is very similar to the average. It is likely that the rest of the sentence context is able to help BERT make the decision between ge 个 and a politeness classifier to some extent, but not nearly perfectly. It is possible that sometimes the politeness aspect is not obvious from the rest of sentence context, which means the classifier itself is the only thing marking politeness.

We also look at the confusion matrix for BERT and observe the following findings. It is clear from table 5 that BERT most commonly misclassifies wèi 位 as ge 个 or míng 名. Futhermore, we can see from table 6 that míng 名 is most commonly misclassified as ge 个 or wèi 位. This confirms that ge 个, wèi 位 and míng 名 get mixed up often.

The fact that the MLM performs the second best for both of these classifiers implies that the pretrained BERT is superior in understanding classifier context compared to the LSTM with word embeddings. The importance of the sentence context also relates to the fact that the RULE model cannot be used reliably to generate politeness classifiers. This is likely explained by the fact that many head nouns can take either a non-polite classifier or a polite classifier depending on the context, and the non-polite classifiers are much more common. There are likely few nouns which always take the polite classifier.

Classifier	Number of Predictions	Percentage of Total
位 wèi	1292	59.87
个 ge	628	29.10
名 míng	72	3.34
次 cì	19	0.88
种 zhǒng	19	0.88
群 qún	17	0.79
张 zhāng	12	0.56
家 jiā	10	0.46
Others	89	4.12
Total	2158	100

Table 5: All predictions for wèi 位 cases for BERT. Showing only classifiers with >10 predictions.

Classifier	Number of Predictions	Percentage of Total
名 míng	575	70.99
个 ge	105	12.96
位 wèi	64	7.90
Others	66	8.15
Total	810	100

Table 6: All predictions for ming 位 cases for BERT. Showing only classifiers with >10 predictions.

The plurality classifiers, qún 群, duī 堆, xiē 些, and 套 tào, are always predicted significantly worse than the average score for classifiers in the frequency range of 200-800. On the other hand, duì 对 and shuāng 双 are easier to predict than the other plurality classifiers. The results of the RULE model imply that these classifiers, which mean *pair*, appear quite frequently with certain head nouns. These likely appear in sentences like "pair of shoes" or "pair of chopsticks". This may explain why the other models also predict these with much higher accuracy.

Looking again at the confusion matrix for BERT, we can make some observations of plurality classifiers. Table 7 shows that qún 群 is also most often misclassified as ge 个. This shows that in many cases the plurality vs. singularity issue causes difficulty for the model. It is likely not always clear from the rest of the sentence context whether the noun should be considered plural or singular. Other true classifiers such as jiā 家, zhī 只 and wèi 位, which also imply singularity, appear as well. Duī 堆 is also observed, as it can sometimes be a synonym for qún 群.

We can see from table 8, that xiē 些 is most commonly misclassified as ge 个, too. Diǎn 点 is also often mixed up with xiē 些, which is understandable as both of them can mean "a little bit". Table 9 shows that duī 堆 is as well most commonly misclassified as ge 个. Qún 群, which also refers to multiple of something, also appears.

Table 10 shows that tào 套 is most often misclassified as ge 个. In addition, jiàn 件, which can occur as a classifier for items of clothing, appears. As tào 套 could mean a set of clothing (such as a suit), it is not surprising that it gets mixed up with jiàn 件, which is a singular classifier for pieces of clothing.

For the classifiers meaning pair, duì 对 and shuāng 双, we observe from table 11, that for duì 对, ge 个 is the most common misclassification. Shuāng 双 is the second most common, which is not surprising as it can be a synonym for duì 对. Table 12 shows that for shuāng 双, zhī 只 is the most common misclassification.

BERT, in general, does a comparatively good job predicting the plurality classifiers, but the results are still below average for most of them. The context somehow likely helps BERT understand which sentences should be considered plural and which singular.

Classifier	Number of Predictions	Percentage of Total
群 qún	419	52.51
个 ge	220	27.57
家 jiā	33	4.14
只 zhī	26	3.26
位 wèi	16	2.01
堆 duī	16	2.01
种 zhǒng	14	1.75
Others	54	6.77
Total	798	1

Table 7: All predictions for qún 群 cases for BERT. Showing only classifiers with >10 predictions.

Classifier	Number of Predictions	Percentage of Total
些 xiē	356	56.51
个 ge	110	17.46
点 diǎn	108	17.14
种 zhǒng	11	1.75
Others	45	7.14
Total	630	100

Table 8: All predictions for xiē 些 cases for BERT. Showing only classifiers with >10 predictions.

Classifier	Number of Predictions	Percentage of Total
堆 duī	357	52.12
个 ge	89	12.99
群 qún	68	9.93
件 jiàn	16	2.34
张 zhāng	15	2.19
本 běn	11	1.61
Others	129	18.83
Total 685 100,00		

Table 9: All predictions for duī 堆 cases for BERT. Showing only classifiers with >10 predictions.

Classifier	Number of Predictions	Percentage of Total
套 tào	158	34.57
个 ge	60	13.13
件 jiàn	57	12.47
张 zhāng	33	7.22
种 zhǒng	18	3.94
款 kuǎn	16	3.50
部 bù	11	2.41
Others	104	22.76
Total	457	100.00

Table 10: All predictions for tào 套 cases for BERT. Showing only classifiers with >10 predictions.

Classifier	Number of Predictions	Percentage of Total
对 duì	272	62.39
个 ge	60	13.76
双 shuāng	18	4.13
只 zhī	13	2.98
群 qún	13	2.98
Others	60	13.76
Total	436	100.00

Table 11: All predictions for duì 对 cases for BERT. Showing only classifiers with >10 predictions.

Classifier	Number of Predictions	Percentage of Total
双 shuāng	218	76.49
只 zhī	20	7.02
个 ge	13	4.56
Others	34	11.93
Total	285	100

Table 12: All predictions for shuāng 双 cases for BERT. Showing only classifiers with >10 predictions.

All in all, it is clear that the politeness classifiers, wèi 位 and míng 名, get mixed up with each other and with the general classifier ge 个. Whether to use a polite classifier or a non-polite classifier is likely not evident from the sentence context always. For instance, the noun rén 人, "person", could take either ge 个, a non-polite singular classifier, or wèi 位, a polite singular classifier, depending on the situation — and the *politeness* aspect might not be obvious from the

rest of sentence. Thus, the model cannot know just based on the rest of the sentence which one to choose as the classifier itself is the only element implying politeness.

The plurality classifiers also get most often mixed up with each other — and the singular ge 个 and other singular classifiers such as jiàn 件 or zhī 只. Hence, in these cases also it is likely not always evident from the sentence context alone whether a sentence should be plural or singular. For example, rén 人, "person", could take either the singular classifier, ge 个, or a plurality classifier, qún 群, "group". In these cases it might again not be obvious from the rest of the sentence whether we are talking about one person or a group of people: the classifier itself is the only element which provides this information. In addition, as many of the plurality classifiers can be considered synonyms to an extent, it is not surprising that one plurality classifier is generated instead of another plurality classifier.

However, the above observations confirm our third hypothesis that the BERT classification model's advantage over its competitors is greater for classifiers which *add information* than for other more general classifiers, such as ge 个. Still, even BERT does not do a great job with these classifiers.

5.1.4 Conclusions on the Research Questions

Now, let us come back to research question 1:

- How well is it possible to do on this task?

As we have seen, there are certain examples of classifiers which are clearly more difficult to predict than others. These include measure words and dual classifiers as well as those classifiers which *add information*. Even though the context-aware BERT classification model does a better job at predicting these types of classifiers, even it does not perform well. As we have discussed above, it is likely that in certain cases, even if the model understands the rest of sentence context, it is not enough to generate the correct classifier. In these cases, the classifier could be considered to add information, which is not obvious from the rest of the sentence context: for instance, whether a polite classifier should be chosen instead of non-polite one. Due to this, it is likely that current solutions will not be able to predict the correct classifier in these cases with 100% accuracy.

Consequently, it is clear that a BERT classification model can do remarkably well with most classifiers, but in order for it to achieve 100% accuracy, it would have to be extremely lucky in the choices it makes in the difficult cases mentioned above. This is not realistic, so it is likely not possible to do perfectly in this task with current solutions.

5.2 Post Hoc Analysis

5.2.1 Effect of Distance between Classifier and Headword

We also investigate other factors which might influence the decisions of BERT. As mentioned in section 2.4, we want consider the effect of the *distance* between

the classifier and the head noun. For instance, let us take the following example:

Yī <cl>chǎng</cl> jīngcǎi de <h>qiúsài</h>.

In the example, there is a pre-modifier consisting of two words between the classifier *chǎng* and the head noun *qiúsài*. Thus, the distance for this example is 2. We expect that the larger the distance is, the worse BERT performs. Because some sentences in the dataset are a combination of Chinese characters and Latin characters, we define two ways to calculate the distance.

Distance 1 represents the number of words, excluding any punctuation. A word is a unit with spaces around it.

Distance 1 example:

流行于 iphone 和 Android 平台上的经典小游戏也推出了 html5 版本来是为 IE9 订制的不过所有支持 html5 的浏览器都可以很好的运行 = 33 words

Distance 2 represents the number of characters, including all punctuation. Each Latin character is counted, punctuation marks are counted, spaces are eliminated.

Distance 2 example:

流行于 iphone 和 Android 平台上的经典小游戏也推出了 html5 版, 本来是为 IE9 订制的, 不过所有支持 html5 的浏览器都可以很好的运行。= 72 characters

Dataset	Distance 1	Distance 2
`Whole Dataset`	1.057	1.681
`Correct Predictions`	1.036	1.638
`Incorrect Predictions`	1.151	1.872

Table 13: Distance 1 represents the number of words, excluding any punctuation. A word is a unit with spaces around it. Distance 2 represents the number of characters, including all punctuation. Each Latin character is counted, punctuation marks are counted, spaces are eliminated.

Looking at the results in table 13, we can see that the distances of the correct predictions are shorter than for the incorrect predictions, for both distance 1 and distance 2. An un-paired t-test for both distance 1 and distance 2 confirms that distance has a negative effect on the model's performance ($p < .001$).

5.3 Limitations and Further Studies

Having looked at a broad set of classifiers, we are aware that many linguists could have framed this problem differently. As many linguists typically regard true

classifiers and measure words as completely different phenomena, we acknowledge that many would have likely only focused on true classifiers. Predicting certain measure words such as containers, for example if someone drank a cup of coffee or a bottle of coffee, is very much a different problem compared to predicting a true classifier in a practical application such as automatic machine translation.

However, as we have presented in this thesis, making the distinction between true classifiers and measure words is not always straightforward. Thus, our approach with its subset analyses offers insight into different classes of classifiers, including a categorization by Her and Lai (2012) into true classifiers, dual classifiers and measure words. In addition, our approach allows to analyse any subset of classifiers as long as we have a way of telling which classifiers belong in which class. Furthermore, next to predicting true classifiers, the problem of predicting measure words, such as containers, might be of interest to some researchers in natural language processing and artificial intelligence. For instance, how well does a model understand the context surrounding drinking coffee, i.e. does it understand that coffee is normally drunk from cups instead of bottles. (Generating the measure word cup instead of bottle.)

It should also be noted, that the dataset used in this study, as outlined in section 3.2, is not perfect. It contains a certain amount of sentences where the classifier or the headword (9% and 22% respectively) has not been identified correctly. Even though Peinelt et al. (2017) take steps to mitigate this problem, it clearly is still present in the dataset. Thus, the predictions that are made for these incorrectly labelled sentences do not necessarily represent the task as was intended, which in turn influences the results to some extent.

Furthermore, since the choice of classifier is not deterministic, the type of corpus evaluation that was performed in this paper arguably does not "tell the whole story" regarding the quality of the different models. To remedy this issue, we suggest two further experiments as further studies, each of which involves human participants. One is a *speaker* experiment, in which several participants would be asked to choose classifiers given a linguistic context. By comparing the outcomes of this experiment with the CCD corpus, we would obtain a better understanding of the difficulty of the task that we have set our algorithms. By thus asking multiple participants to accomplish the same task as our algorithms, we would obtain a new corpus, in which each linguistic context is associated with a bag of (1 or more) possible classifiers. This new dataset would enable to conduct a new, non-deterministic evaluation of the models.

The other experiment would have human *readers* judge the acceptability of each classifier choice that is made by a given model. Reader experiments of this kind are a standard tool in judging the quality of decisions taken by an NLG algorithm (cf. Van Der Lee et al. (2019)) and will give rise to a new set of analyses analogous to the ones in the present paper, which will complete our understanding of the quality of the decisions that are taken by each model. A similar small-scale experiment was conducted in Guo and Zhong (2005), which showed that the acceptability of generated classifiers is rated much higher by human participants than the results of just comparing the generated classifier

to the one in the original corpus suggest.

5.4 Conclusions

In this study, we have provided new insights into multiple aspects of the task of classifier prediction in Mandarin. Firstly, we have added to the work by Peinelt et al. (2017) by showing that a context-aware model such as BERT has superior performance compared to other models. This means we have produced a new state-of-the-art result for the CCD dataset.

In addition, we have shown that while the performance of the RULE model seems lacking on the surface — when it is only applied to true classifiers, the performance of the model is quite remarkable. The analysis by Peinelt et al. (2017) does not consider different categorizations of classifiers and instead looks into a very large range of classifiers. Due to this the performance of the rule-based model seems very poor. By diving deeper into the categorisation of classifiers, we have shown that a simple rule-based model can perform particularly well in most true classifier cases. Thus the rule-based model should not be underrated for the task.

However, there are a subset of true classifiers, such as the polite classifiers, wèi 位 and míng 名, that are very difficult for all models, especially the RULE model. There clearly needs to be more of an understanding of the context of the sentence in order to correctly pick these classifiers, and even in those cases, BERT, which is the most context-aware model available, does not perform particularly well. This implies that for this subset of classifiers the classifier is *adding information*, that is not possible to infer from the rest of sentence context. Thus the task of predicting these kind of cases might not be possible with current solutions.

Furthermore, many of our findings relate to aspects of the literature related to the categorization of Mandarin classifiers. For example, for most measure words, it clearly seems that the task is more comparable to predicting lexical elements. Thus, it is not surprising that the RULE model cannot perform well. As BERT is good at predicting lexical elements in general, it is much more suitable for generating measure words, too. Nevertheless, true classifiers and measure words should clearly be considered different kinds of elements even in the context of this task. It is not sensible to compare the prediction of a classifier such as wèi 位 to the prediction of a container such as 杯 bēi. These are clearly different kinds of tasks and ask different kinds of questions.

Finally, we have shown that an unsupervised model such as the MLM is able to perform remarkably well with predicting especially rare classifiers. None of the classification models are able to perform as well with less frequently appearing classifiers. This is likely because classification models are known to struggle with imbalanced datasets. The classification models could perform better on these cases too, if steps were taken during the training to mitigate the issue of the imbalanced dataset, for instance, by balancing the training dataset.

The topic of Mandarin classifiers is complicated and there continues to be some controversy among linguistics on what should be considered *true classi-*

fiers. As earlier studies have not explicitly listed, or defined in a computationally explicit way, the sets of classifiers they have analyzed, we hope this extensive and transparent study can shed more light on this challenging aspect, and inspire further machine learning and computational linguistics approaches in studying and generating Mandarin classifiers.

6 Bibliography

Allen, K. (1977). Classifier. *Language*, 53(2):285–311.

Becker, A. L. (1975). A linguistic image of nature: the Burmese numerative classifier system. Walter de Gruyter, Berlin/New York Berlin, New York.

Bisang, W. (2011). Classifiers in East and Southeast Asian languages: counting and beyond. In *Numeral types and changes worldwide*, pages 113–186. De Gruyter Mouton.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bond, F. and Paik, K. (2000). Reusing an ontology to generate numeral classifiers. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Univ of California Press.

Cheng, L. L.-S. and Sybesma, R. (1998). Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua journal of Chinese studies*, 28(3):385–412.

Cheng, L. L.-S. and Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of NP. *Linguistic inquiry*, 30(4):509–542.

Cheng, L. L.-S. and Sybesma, R. (2005). *Classifiers in four varieties of Chinese*, pages 259–292. Oxford University Press, Oxford.

Cheng, L. L.-S. and Sybesma, R. (2012). Classifiers and DP. *Linguistic inquiry*, 43(4):634–650.

Croft, W. (1994). Semantic universals in classifier systems. *Word*, 45(2):145–171.

Da Costa, L. M., Bond, F., and Gao, H. H. (2016). Mapping and generating classifiers using an open Chinese ontology. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 249–256.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Erbaugh, M. S. (1986). *Taking stock: The development of Chinese noun classifiers historically and in young children*, pages 399–436. John Benjamins Amsterdam.

Gao, H. H. (2011). E-learning design for Chinese classifiers: Reclassification of nouns for a novel approach. In *International Conference on ICT in Teaching and Learning*, pages 186–199. Springer.

Guo, H. and Zhong, H. (2005). Chinese classifier assignment using SVMs. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Guo, X. (2002). *Xiandai Hanyu Liangci Yongfa Cidian [A Dictionary of Usage of Contemporary Chinese Classifiers]*. Beijing: Yuwen Chubanshe.

He, J. (2001). *Xiandai Hanyu Liangci Yanjiu [Studies on classifiers in Modern Chinese]*. Beijing: Minzu Publishing House.

Her, O.-S. (2012). Distinguishing classifiers and measure words: A mathematical perspective and implications. *Lingua*, 122(14):1668–1691.

Her, O.-S. and Lai, W.-J. (2012). Classifiers: The many ways to profile'one'—a case study of Taiwan Mandarin. *International Journal of Computer Processing Of Languages*, 24(01):79–94.

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

Huang, C.-R., Chen, K.-j., and Gao, Z.-m. (1998). Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. *Quantitative and Computational Studies of Chinese Linguistics*, pages 339–352.

Lai, W.-C. (2011). *Identifying true classifiers in Mandarin Chinese*. PhD thesis, Taipei: National Chengchi University.

Levy, R. and Manning, C. D. (2003). Is it harder to parse Chinese, or the Chinese treebank? In *proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 439–446.

Li, C. N. and Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*, volume 3. Univ of California Press.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., and Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized Bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loke, K.-k. (1994). Is ge merely a "general classifier". *Journal of the Chinese Language Teachers Association*, 29(3):35–50.

Loke, K.-K. (1996). Norms and realities of Mandarin shape classifiers. *Journal of the Chinese Language Teachers Association*, 31:1–22.

McEnery, A. and Xiao, Z. (2004). The lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *Religion*, 17:3–4.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Paik, K. and Bond, F. (2001). Multilingual generation of numeral classifiers using a common ontology. In *Proceedings of the 19th International Conference on the Computer Processing of Oriental Languages*, pages 141–147.

Peinelt, N., Liakata, M., and Hsieh, S.-K. (2017). Classifierguesser: A context-based classifier prediction system for Chinese language learners. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 41–44.

Polio, C. (1994). Non-native speakers' use of nominal classifiers in Mandarin Chinese. *Journal of the Chinese Language Teachers Association*, 29(3):51–66.

Sornlertlamvanich, V., Pantachat, W., and Meknavin, S. (1994). Classifier assignment by corpus-based approach. *arXiv preprint cmp-lg/9411027*.

Sun, C. (1988). The discourse function of numeral classifiers in Mandarin Chinese. *Journal of Chinese Linguistics*, 16(2):298–322+.

Sybesma, R. P. E., Behr, W., Gu, Y., Handel, Z. J., Huang, C.-T. J., and Myers, J. (2017). *Encyclopedia of Chinese language and linguistics*. Brill.

Tai, J. and Wang, L. (1990). A semantic study of the classifier tiao. *Journal of the Chinese Language Teachers Association*, 25(1):35–56.

Tai, J. H. (1992). Variation in classifier systems across Chinese dialects: towards a cognition-based semantic approach. *Chinese Language and Linguistics*, 1:587–608.

Tai, J. H. (1994). Chinese classifier systems and human categorization. *In honor of William S.-Y. Wang: Interdisciplinary studies on language and language change*, pages 479–494.

Tao, H. and Xiao, R. (2012). The UCLA Chinese corpus. lancaster: Ucrel.

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., and Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Van Esch, D. (2012). Leiden Weibo corpus.

Wang, S. and Bond, F. (2013). Building the Chinese open WordNet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Wen, H. M. S., Eshley, G. H., and Bond, F. (2012). Using WordNet to predict numeral classifiers in Chinese and Japanese. In *Proceedings of the 6th Global WordNet Conference*.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Wu, J.-S. and Her, O.-S. (2021). Taxonomy of numeral classifiers: A formal semantic proposal. In *Numeral Classifiers and Classifier Languages*, pages 40–71. Routledge.

Zhang, D., Li, M., Duan, N., Li, C.-H., and Zhou, M. (2008). Measure word generation for English-Chinese SMT systems. In *Proceedings of ACL-08: HLT*, pages 89–96.

Zhang, H. (2007). Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics*, 16(1):43–59.

Zhang, N. N. (2013). *Classifier structures in Mandarin Chinese*, volume 263. Walter de Gruyter.

Appendices

A Lists of True Classifiers, Dual Classifiers and Measure Words

The lists on the following pages are excerpts from Her and Lai (2012, p. 10-14). *Table 2* lists true classifiers, *table 3* lists dual classifiers and *table 4* lists measure words. The numbering of the tables are from Her and Lai (2012).

Table 2: 76 Classifiers out of the MDDCC's 173

ben3 本	yi1 ben3 shu1 一本書	jian4 件	yi1 jian4 da4yi1 一件大衣	ting3 挺	yi1 ting3 ji1qiang1 一挺機槍
bi3 筆	yi1 bi3 shou1ru4 一筆收入	jie4 介	yi1 jie4 shu1sheng1 一介書生	tou2 頭	yi1 tou2 da4xiang4 一頭大象
bing3 柄	yi1 bing3 fu3tou2 一柄斧頭	jing1 莖	yi1 jing1 bai2fa3 一莖白髮	wan1 彎	yi1 wan1 ming2yue4 一彎明月
ce4 冊	yi1 ce4 shu1 一冊書	ju4 句	yi1 ju4 kou3hao4 一句口號	wan1 灣	yi1 wan1 liu2shui3 一灣流水
chu4 處	yi1 chu4 shang1kou3 一處傷口	ju4 具	yi1 ju4 shi1ti3 一具屍體	wan2 丸	yi1 wan2 yao4wan2 一丸藥丸
chuang2 床	yi1 chuang2 mian2bei4 一床棉被	juan3 卷	yi1 juan3 lu4yin1dai4 一卷錄音帶	wei3 尾	yi1 wei3 yu2 一尾魚
chuang2 幢	yi1 chuang2 lou2fang2 一幢樓房	ke1 棵	yi1 ke1 song1shu4 一棵松樹	wei4 位	yi1 wei4 lao3shi1 一位老師
dang3 檔	yi1 dang3 gu3piao4 一檔股票	ke1 顆	yi1 ke1 xi1gua1 一顆西瓜	xi2 席	yi1 xi2 dong3shi4 一席董事
dao4 道	yi1 dao4 zhuan1qiang2 一道磚牆	li4 粒	yi1 li4 hong2dou4 一粒紅豆	xi2 襲	yi1 xi2 bo2sha1 一襲薄紗
ding3 頂	yi1 ding3 mao4zi 一頂帽子	liang4 輛	yi1 liang4 jing3che1 一輛警車	yuan2 員	yi1 yuan2 da4jiang4 一員大將
ding4 錠	yi1 ding4 yuan2bao3 一錠元寶	mei2 枚	yi1 mei2 jiang3zhang1 一枚獎章	ze2 則	yi1 ze2 xiao4hua4 一則笑話
dong4 棟	yi1 dong4 da4 lou2 一棟大樓	mian4 面	yi1 mian4 jing4zi 一面鏡子	zhan3 盞	yi1 zhan3 deng1 一盞燈
du3 堵	yi1 du3 qiang2 一堵牆	ming2 名	yi1 ming2 xue2sheng1 一名學生	zhang1 張	yi1 zhang1 chunag2 一張床
duo3 朵	yi1 duo3 mei2gui1 一朵玫瑰	pi1 匹	yi1 pi1 ma3 一匹馬	zhao1 招	yi1 zhao1 ce4lüe4 一招策略
fa1 發	yi1 fa1 zi3dan4	pian1 篇	yi1 pian1	zheng4	yi1zheng4

	一發子彈		wen2zhang1 一篇文章	幀	jie2hun1zhao4 一幀結婚照
fang1 方	yi1 fang1 yin4zhang1 一方印章	qi2 畦	yi1 qi2 dao4tian2 一畦稻田	zhi1 只	yi1 zhi1 jiu3tan2 一只酒罈
feng1 封	yi4 feng1xin4 一封信	qu3 曲	yi1 qu3 liu2xing2ge1 一曲流行歌	zhi1 枝	yi1 zhi1 shu4zhi1 一枝樹枝
fu2 幅	yi1 fu2 hua4 一幅畫	que4 闋	yi1 que4 gu3ci2 一闋古詞	zhi1 隻	yi1 zhi1 mao1 一隻貓
gan3 桿	yi1 gan3 qiang1 一桿槍	shan4 扇	yi1 shan4 men2 一扇門	zhi3 紙	yi1 zhi3 qie4jie2shu1 一紙切結書
gen1 根	yi1 gen1 tou2fa3 一根頭髮	sheng1 聲	yi1 sheng1 jian1jiao4 一聲尖叫	zhou2 軸	yi1 zhou2 hua4 一軸畫
ge 個	yi1 ge ren2 一個人	shou3 首	yi1 shou3 er2ge1 一首兒歌	zhu1 株	yi1 zhu1ying1hua1 一株櫻花
guan3 管	yi1 guan3 mao2bi3 一管毛筆	sao1 艘	yi1 sao1 chuang2 一艘船	zhu4 柱	yi1 shu4 dian4xian4gan1 一柱電線杆
ji4 記	yi1 ji4 zuo3gou1quan2 一記右勾拳	suo3 所	yi1 suo3 da4xue2 一所大學	zhu4 炷	yi1 shu4 xiang1 一炷香
ji4 劑	yi1 jie4 qiang2xin1ji4 一劑強心劑	ti2 題	yi1 ti2 xuan3ze2ti2 一題選擇題	zun1 尊	yi1 zun1 fo2xiang4 一尊佛像
jia4 架	yi1 jia4 fei1ji1 一架飛機	tiao2 條	yi1 tiao2 wei2jing1 一條圍巾	zuo4 座	yi1 zuo4 shan1 一座山
jian1 間	yi1 jian1 shu1dian4 一間書店				

Table 3: 21 dual status C/M out of the MDDCC's 173 Classifiers

ba3 把	ba3 把 c	yi1 ba3 dao1zi 一把刀子	ba3 把 m	yi1 ba3 tong2ban3 一把銅板
ban1 班	ban1 班 c	yi1 ban1 fei1ji1 一班飛機	ban1 班 m	yi1 ban1 xue2sheng1 一班學生
ban4 瓣	ban4 瓣 c	yi1 ban4 hua1ban4 一瓣花瓣	ban4 瓣 m	yi1 ban4 ju2zi 一瓣橘子
bu4 部	bu4 部 c	yi1 bu4 qi4che1 一部汽車	bu4 部 m	yi1 bu4 shu1 一部書
dian3 點	dian3 點 c	yi1 dian3 zhi4 一點痣	dian3 點 m	yi1 dian3 qian2 一點錢

fen4 分	fen4 分 c	yi1 fen4 bao4gao4 一分報告	fen4 分 m	yi1 fen4 qing2yi4 一分情意
fen4 份	fen4 份 c	yi1 fen4 bao4gao4 一份報告	fen4 份 m	yi1 fen4 qing2yi4 一份情意
jia1 家	jia1 家 c	yi1 jia1 gong1si1 一家公司	jia1 家 m	yi1 jia1 ao4zhou1ren2 一家澳洲人
jie2 節	jie2 節 c	yi1 jie2 che1xiang1 一節車廂	jie2 節 m	yi1 jie2 gan1zhe4 一節甘蔗
kou3 口	kou3 口 c	yi1 kou3 jing3 一口井	kou3 口 m	yi1 kou3 zhu4ya2 一口蛀牙
kuai4 塊	kuai4 塊 c	yi1 kuai4 zhuan1tou2 一塊磚頭	kuai4 塊 m	yi1 kuai4 di4 一塊地
lun2 輪	lun2 輪 c	yi1 lun2 ming2yue4 一輪明月	lun2 輪 m	yi1 lun2 bi3sai4 一輪比賽
lü3 縷	lü3 縷 c	yi1 lü3 xian4 一縷線	lü3 縷 m	yi1 lü3 qing1yan1 一縷清煙
men2 門	men2 門 c	yi1 men2 da4pao4 一門大砲	men2 門 m	yi1 men2 sheng1yi4 一門生意
pian4 片	pian4 片 c	yi1 pian4 shu4ye4 一片樹葉	pian4 片 m	yi1 pian4 nai3you2 一片奶油
qi3 起	qi3 起 c	yi1 qi3 yi4wai4 一起意外	qi3 起 m	yi1 qi3 ren2ma3 一起人馬
tai2 台	tai2 台 c	yi1 tai2 dian4shi4 一台電視	tai2 台 m	yi1 tai2 ge1zai3xi4 一台歌仔戲
xian4 線	xian4 線 c	yi1 xian4 che1dao4 一線車道	xian4 線 m	yi1 xian4 xi1wang4 一線希望
ye4 葉	ye4 葉 c	yi1 ye4 pian1zhou1 一葉扁舟	ye4 葉 m	yi1 ye4 shu1 一葉書
zhi1 支	zhi1 支 c	yi1 zhi1 ge1 一支歌	zhi1 支 m	yi1 zhi1 chun2mao2sha1 一支純毛紗
zong1 宗	zong1 宗 c	yi1 zong1 yi4wai4 一宗意外	zong1 宗 m	yi1 zong1 huo4wu4 一宗貨物

Table 4: 76 Measure Words out of the MDDCC's 173 Classifiers

ban3 版	yi1 ban3 xin1wen2 一版新聞	hui2 回	ba1shi2hui2 hong2lou2meng4 八十回紅樓夢	piao4 票	yi1 piao4 sheng1yi4 一票生意
bang1 幫	yi1 bang1 gong1ren2 一幫工人	huo3 夥	yi1 huo3 qiang2dao4 一夥強盜	pie3 撇	yi1 pie3 hu2xu1 一撇鬍鬚
cao2 槽	yi1 cao2 ya2	ji2 級	yi1 ji2 shi2jie1	pou2 抔	yi1 pou2 tu3

	一槽牙		一級石階		一抔土
ceng2 層	yi1 ceng2 lou2 一層樓	ji2 集	yi1bai3ji2 lian2xu4ju4 一百集連續劇	qi2 期	za2zhi4 di4yi1qi2 雜誌第一期
chong2 重	wan4 chong2 shan1 萬重山	ji2 輯	cong2shu1 di4yi1ji2 叢書第一輯	quan1 圈	yi1 quan2 liu3shu4 一圈柳樹
chuan4 串	yi1 chuan4 fo2zhu1 一串佛珠	jie1 階	yi1 jie1 lou2ti1 一階樓梯	qun2 群	yi1 qun2 peng2you3 一群朋友
cong2 叢	yi1 cong2 ye3cao3 一叢野草	jie2 截	yi1 jie2 zhu2zi 一截竹子	shen1 身	yi1 shen1 yi1shang 一身衣裳
cu4 簇	yi1 cu4 mei2gui1 一簇玫瑰	jin4 進	yi1 jin4 fang2zi 一進房子	shu4 束	yi1 shu4 xian1hua1 一束鮮花
cuo1 撮	yi1 cuo1 mao2fa3 一撮毛髮	juan4 卷	za2zhi4 di4yi1juan4 雜誌第一卷	shuang1 雙	yi1 shuang1 xie2 一雙鞋
da3 打	yi1 da3 qian1bi3 一打鉛筆	ke1 科	ying1wan2yi1ke1 英文一科	si1 絲	yi1 si1 rou4 一絲肉
dai4 代	shang4 yi1 dai4 ren2 上一代人	ke4 客	yi1 ke4 niu2pai2 一客牛排	tai1 胎	yi1 tai1 xiao3gou3 一胎小狗
dai4 帶	yi1 dai4 yu2cun1 一帶漁村	ke4 課	yi1 ke4 shu4xue2 一課數學	tan1 灘	yi1 tan1 shui3 一灘水
di1 滴	yi1 di1 yan3lei4 一滴眼淚	kuan3 款	di4yi1kuan3 gui1ding4 第一款規定	tang2 堂	yi1 tang2 jia1ju4 一堂傢具
die2 疊	yi1 die2 chao1piao4 一疊鈔票	kun3 捆	yi1 kun3 dao4cao3 一捆稻草	tao4 套	yi1 tao4 can1ju4 一套餐具
duan4 段	yi1 duan4 gan1zhe4 一段甘蔗	lan2 欄	yi1 lan2 xin1wen2 一欄新聞	tie4 帖	yi1 tie4 zhong1yao4 一帖中藥
dui1 堆	yi1 dui1 tu3 一堆土	lian2 聯	er4lian2shou1ju4 二聯收據	tuan2 團	yi1 tuan2 shi4bing1 一團士兵
dui4 隊	yi1 dui4 shi4bing1 一隊士兵	lie4 列	yi1 lie4 luo4tuo2 一列駱駝	tuo2 坨	yi1 tuo2 nai3you2 一坨奶油
dui4 對	yi1 dui4 fu1qi1 一對夫妻	liu3 綹	yi1 liu3 tou2fa3 一綹頭髮	wei4 味	hun1cai4 wu3wei4 葷菜五味
fang2 房	yi1 fang2 er2sun1	lu4 路	yi1 lu4 ren2ma3 一路人馬	xiang4 項	xing2fa3 di4ti1xiang4

	一房兒孫				刑法第一項
fu2 服	yi1 fu2 zhong1yao4 一服中藥	luo4 落	yi1 luo4 bao4zhi3 一落報紙	ye4 頁	yi1 ye4 shu1 一頁書
fu4 副	yi1 fu4 kuai4zi 一副筷子	lü3 旅	yi1 lü3 bu4dui4 一旅部隊	zha1 紮	yi1 zha1 zhi3hua1 一紮紙花
gu3 股	yi1 gu3 xiang1qi4 一股香氣	pai2 排	yi1 pai2 shi4bing1 一排士兵	zhang1 章	di4yi1zhang1 nei4rong2 第一章 內容
gua4 掛	yi1 gua4 fo2zhu1 一掛佛珠	peng3 捧	yi1 peng3 sha1 一捧沙	zhen1 針	yi1 xhen1 qiang2xin1ji4 一針強心劑
hang2 行	yi1 hang2 liu3shu4 一行柳樹	pi1 批	yi1 pi1 huo4 一批貨	zhuo1 桌	yi1 zhuo1 cai4 一桌菜
hu4 戶	yi1 hu4 nong2min2 一戶農民	pi3 匹	yi1 pi3 bu4 一匹布	zu3 組	yi1 zu3 ren2yuan2 一組人員
hao4 號’	di4yi1hao4 dao4lu4 第一號道路				

B Results for All Models

The following pages present the raw results for all the models in the following order: RULE, LSTM, MLM and BERT.

RULE	category	precision	recall	f1-score	support
weighted avg		0,58229696	0,61896477	0,58238572	136221
macro avg		0,34872135	0,2049533	0,2338891	136221
个	True classifier	0,66256386	0,86976502	0,7521556	61581
种	Not in lists	0,55422423	0,43822506	0,48944555	10344
次	Not in lists	0,48975069	0,47986104	0,48475543	9211
张	True classifier	0,71839952	0,70097604	0,70958084	3381
件	True classifier	0,67582988	0,83606031	0,74745447	3117
句	True classifier	0,75576471	0,6555102	0,7020765	2450
条	True classifier	0,59349593	0,62101234	0,60694242	2351
位	True classifier	0,45298165	0,18303985	0,26072607	2158
家	Dual classifier	0,4974773	0,28696158	0,36397195	1718
场	Not in lists	0,57236842	0,4619469	0,51126347	1695
只	True classifier	0,50412655	0,47690306	0,49013708	1537
点	Dual classifier	0,3163017	0,08990318	0,14001077	1446
部	Dual classifier	0,64764268	0,55729537	0,59908187	1405
首	True classifier	0,82561078	0,73243647	0,77623762	1338
段	Measure word	0,53775039	0,52402402	0,53079848	1332
篇	True classifier	0,89201878	0,72078907	0,79731431	1318
滴	Measure word	0,50543478	0,16089965	0,24409449	1156
份	Dual classifier	0,35255713	0,28647215	0,31609756	1131
号	Measure word	0,47936508	0,26964286	0,34514286	1120
块	Dual classifier	0,49011178	0,54913295	0,51794639	1038
颗	True classifier	0,68656716	0,62683438	0,65534247	954
名	True classifier	0,55172414	0,2962963	0,38554217	810
群	Measure word	0,17073171	0,00877193	0,01668653	798
款	Measure word	0,47107438	0,14766839	0,22485207	772
片	Dual classifier	0,54451346	0,36027397	0,43363561	730
堆	Measure word	0,23255814	0,04379562	0,07371007	685
本	True classifier	0,65635739	0,60538827	0,62984336	631
些	Not in lists	0,144	0,02857143	0,04768212	630
级	Measure word	0,64726027	0,6009539	0,62324815	629
分	Dual classifier	0,40789474	0,15897436	0,22878229	585
杯	Not in lists	0,43832021	0,58699473	0,50187829	569
起	Dual classifier	0,47154472	0,10469314	0,17134417	554
步	Not in lists	0,875	0,48870637	0,62714097	487
顿	Not in lists	0,42672414	0,42307692	0,4248927	468
套	Measure word	0,31355932	0,0809628	0,12869565	457
把	Dual classifier	0,55457227	0,41685144	0,47594937	451
集	Measure word	0,32934132	0,12585812	0,18211921	437
对	Measure word	0,64853556	0,35550459	0,45925926	436
辆	True classifier	0,46381579	0,65581395	0,5433526	430
碗	Not in lists	0,42628205	0,31294118	0,36092266	425
回	Measure word	0,20930233	0,02137767	0,0387931	421
代	Measure word	0,46103896	0,18393782	0,26296296	386
声	True classifier	0,59574468	0,24633431	0,34854772	341
座	True classifier	0,42780749	0,24316109	0,31007752	329
届	Not in lists	0,42982456	0,29969419	0,35315315	327
阵	Not in lists	0,29896907	0,08923077	0,13744076	325
道	True classifier	0,42265193	0,48113208	0,45	318

类	Not in lists	0,16666667	0,01298701	0,02409639	308
班	Dual classifier	0,29268293	0,07868852	0,12403101	305
层	Measure word	0,3487395	0,27666667	0,30855019	300
项	Measure word	0,35609756	0,24914676	0,29317269	293
口	Dual classifier	0,48	0,24742268	0,32653061	291
番	Not in lists	0,55357143	0,21305842	0,30769231	291
双	Measure word	0,50207469	0,4245614	0,46007605	285
支	Dual classifier	0,37055838	0,25886525	0,30480167	282
台	Dual classifier	0,32596685	0,21611722	0,25991189	273
朵	True classifier	0,44	0,572	0,4973913	250
瓶	Not in lists	0,36440678	0,17622951	0,23756906	244
等	Not in lists	0,60191083	0,79079498	0,6835443	239
股	Measure word	0,45783133	0,16450216	0,24203822	231
丝	Measure word	0,15060241	0,10964912	0,12690355	228
趟	Not in lists	0,375	0,02643172	0,04938272	227
根	True classifier	0,53043478	0,26872247	0,35672515	227
封	True classifier	0,68837209	0,65777778	0,67272727	225
包	Not in lists	0,22222222	0,06635071	0,10218978	211
轮	Dual classifier	0,44117647	0,07352941	0,12605042	204
头	True classifier	0,39772727	0,18229167	0,25	192
幅	True classifier	0,47368421	0,18947368	0,27067669	190
遍	Not in lists	0,0952381	0,01058201	0,01904762	189
副	Measure word	0,31182796	0,16477273	0,21561338	176
门	Dual classifier	0,45454545	0,1910828	0,2690583	157
粒	True classifier	0,26229508	0,10526316	0,15023474	152
枚	True classifier	0,44680851	0,14189189	0,21538462	148
组	Measure word	0,32	0,11034483	0,16410256	145
盘	Not in lists	0,33333333	0,06206897	0,10465116	145
批	Measure word	0,17391304	0,02797203	0,04819277	143
间	True classifier	0,21052632	0,08695652	0,12307692	138
笔	True classifier	0,35526316	0,2	0,25592417	135
身	Measure word	0,52459016	0,26890756	0,35555556	119
棵	True classifier	0,57142857	0,60504202	0,5877551	119
波	Not in lists	0,4	0,13793103	0,20512821	116
样	Not in lists	0,33333333	0,03448276	0,0625	116
桌	Measure word	0,16666667	0,02727273	0,046875	110
楼	Not in lists	0,16666667	0,03809524	0,0620155	105
季	Not in lists	0,390625	0,24271845	0,2994012	103
盒	Not in lists	0,22222222	0,01980198	0,03636364	101
下	Not in lists	0,06896552	0,02061856	0,03174603	97
盆	Not in lists	0,38461538	0,15625	0,22222222	96
面	True classifier	0,53947368	0,44086022	0,4852071	93
箱	Not in lists	0,37037037	0,10989011	0,16949153	91
页	Dual classifier	0,29411765	0,05617978	0,09433962	89
处	True classifier	0,30612245	0,16853933	0,2173913	89
节	Dual classifier	0,52884615	0,64705882	0,58201058	85
串	Measure word	0,22727273	0,06097561	0,09615385	82
排	Measure word	0,11111111	0,02469136	0,04040404	81
栋	True classifier	0,33333333	0,08974359	0,14141414	78
系列	Not in lists	0	0	0	76
袋	Not in lists	0,11111111	0,01351351	0,02409639	74
锅	Not in lists	0,55555556	0,06944444	0,12345679	72

盏	True classifier	0,51282051	0,55555556	0,53333333	72
束	Measure word	0,35714286	0,07142857	0,11904762	70
通	Not in lists	0,85714286	0,08823529	0,16	68
团	Measure word	0,55	0,1641791	0,25287356	67
圈	Measure word	0	0	0	66
所	True classifier	0,63636364	0,109375	0,18666667	64
发	True classifier	0	0	0	62
世	Not in lists	0,41666667	0,16949153	0,24096386	59
扇	True classifier	0,46153846	0,21052632	0,28915663	57
桶	Not in lists	0,32258065	0,17857143	0,22988506	56
壶	Not in lists	0,3125	0,08928571	0,13888889	56
堂	Measure word	0,77777778	0,12962963	0,22222222	54
餐	Not in lists	0	0	0	54
则	True classifier	0	0	0	52
艘	True classifier	0,58490566	0,59615385	0,59047619	52
架	True classifier	0,19736842	0,3	0,23809524	50
线	Dual classifier	0	0	0	49
曲	True classifier	0,2	0,04081633	0,06779661	49
匹	True classifier	0,39506173	0,65306122	0,49230769	49
户	Measure word	0,45283019	0,54545455	0,49484536	44
肚子	Not in lists	0,5	0,04651163	0,08510638	43
笼	Not in lists	0,8	0,0952381	0,17021277	42
手	Not in lists	0,28571429	0,04761905	0,08163265	42
伙	Measure word	0	0	0	38
枝	True classifier	0,57142857	0,10810811	0,18181818	37
罐	Not in lists	0	0	0	36
卷	True classifier	0,22222222	0,05555556	0,08888889	36
幕	Not in lists	0,5	0,08571429	0,14634146	35
行	Measure word	0,52941176	0,26470588	0,35294118	34
码	Not in lists	0	0	0	34
刀	Not in lists	0,23076923	0,1	0,13953488	30
任	Not in lists	0,27777778	0,16666667	0,20833333	30
株	True classifier	0	0	0	30
脸	Not in lists	0	0	0	28
环	Not in lists	0,33333333	0,11111111	0,16666667	27
辈	Not in lists	0,46153846	0,22222222	0,3	27
幢	True classifier	0	0	0	27
般	Not in lists	0,5	0,11538462	0,1875	26
桩	Not in lists	0	0	0	26
顶	True classifier	0,33333333	0,04166667	0,07407407	24
尾	True classifier	0	0	0	23
尊	True classifier	0,42857143	0,13043478	0,2	23
列	Measure word	0,26190476	0,5	0,34375	22
册	True classifier	0	0	0	22
章	Measure word	0	0	0	21
路	Measure word	0	0	0	21
宗	Dual classifier	0	0	0	20
版	Measure word	0	0	0	20
杆	True classifier	0,5	0,05263158	0,0952381	19
拨	Not in lists	0,5	0,05555556	0,1	18
袭	True classifier	0	0	0	18
记	True classifier	0	0	0	17

剂	True classifier	0,35714286	0,29411765	0,32258065	17
具	True classifier	0,61538462	0,47058824	0,53333333	17
帖	Measure word	0	0	0	13
队	Measure word	0,1	0,07692308	0,08695652	13
味	Measure word	1	0,16666667	0,28571429	12
例	Not in lists	0	0	0	12
局	Not in lists	0	0	0	12
席	True classifier	0	0	0	9
管	True classifier	0	0	0	9
档子	Not in lists	0	0	0	8
人次	Not in lists	0	0	0	6
缸	Not in lists	0	0	0	6
缕	Dual classifier	0	0	0	4
遭	Not in lists	1	0,25	0,4	4
拳	Not in lists	0	0	0	2
棒	Not in lists	0	0	0	1
员	True classifier	0	0	0	1
堵	True classifier	0	0	0	1
accuracy	0,61896477				

LSTM	**category**	**precision**	**recall**	**f1-score**	**support**
weighted avg		0,67901959	0,70442149	0,68124239	136221
macro avg		0,33113958	0,20121202	0,22475054	136221
个	True classifier	0,80900092	0,88652344	0,84598994	61581
种	Not in lists	0,68719232	0,79621036	0,73769537	10344
次	Not in lists	0,61865222	0,76842905	0,68545419	9211
张	True classifier	0,7004104	0,75717243	0,72768619	3381
件	True classifier	0,76374808	0,79756176	0,78028876	3117
句	True classifier	0,64379666	0,73918367	0,68820065	2450
条	True classifier	0,63698338	0,63589962	0,63644104	2351
位	True classifier	0,51045179	0,35078777	0,41581983	2158
家	Dual classifier	0,44444444	0,49359721	0,46773304	1718
场	Not in lists	0,49780381	0,60176991	0,54487179	1695
只	True classifier	0,44936709	0,55432661	0,49635887	1537
点	Dual classifier	0,49748111	0,27316736	0,35267857	1446
部	Dual classifier	0,70364624	0,6455516	0,67334818	1405
首	True classifier	0,72922252	0,81315396	0,76890459	1338
段	Measure word	0,59775281	0,5990991	0,5984252	1332
篇	True classifier	0,94882914	0,83004552	0,88547147	1318
滴	Measure word	0,59127625	0,316609	0,41239437	1156
份	Dual classifier	0,4801061	0,32007073	0,38408488	1131
号	Measure word	0,49760766	0,74285714	0,59598854	1120
块	Dual classifier	0,48755365	0,54720617	0,51566046	1038
颗	True classifier	0,73459119	0,61215933	0,66781018	954
名	True classifier	0,50266667	0,4654321	0,48333333	810
群	Measure word	0,33501259	0,16666667	0,22259414	798
款	Measure word	0,32482993	0,49481865	0,39219713	772
片	Dual classifier	0,51964286	0,39863014	0,45116279	730
堆	Measure word	0,44067797	0,30364964	0,35955056	685
本	True classifier	0,70143885	0,61806656	0,65711879	631
些	Not in lists	0,36631016	0,21746032	0,27290837	630
级	Measure word	0,86917563	0,77106518	0,81718618	629

分	Dual classifier	0,34250765	0,38290598	0,36158192	585
杯	Not in lists	0,46924177	0,57644991	0,51735016	569
起	Dual classifier	0,35457516	0,39169675	0,37221269	554
步	Not in lists	0,91666667	0,51950719	0,66317169	487
顿	Not in lists	0,57289003	0,47863248	0,52153667	468
套	Measure word	0,42696629	0,16630197	0,23937008	457
把	Dual classifier	0,39085239	0,41685144	0,40343348	451
集	Measure word	0,44278607	0,40732265	0,42431466	437
对	Measure word	0,79925651	0,49311927	0,60992908	436
辆	True classifier	0,50877193	0,53953488	0,52370203	430
碗	Not in lists	0,36430318	0,35058824	0,35731415	425
回	Measure word	0,46478873	0,0783848	0,13414634	421
代	Measure word	0,59562842	0,28238342	0,3831283	386
声	True classifier	0,61184211	0,27272727	0,37728195	341
座	True classifier	0,41504178	0,45288754	0,43313953	329
届	Not in lists	0,6015625	0,47094801	0,52830189	327
阵	Not in lists	0,45945946	0,15692308	0,23394495	325
道	True classifier	0,52742616	0,39308176	0,45045045	318
类	Not in lists	0,47457627	0,09090909	0,15258856	308
班	Dual classifier	0,40944882	0,1704918	0,24074074	305
层	Measure word	0,34108527	0,29333333	0,31541219	300
项	Measure word	0,29508197	0,24573379	0,26815642	293
口	Dual classifier	0,26963351	0,35395189	0,30609212	291
番	Not in lists	0,57377049	0,24054983	0,33898305	291
双	Measure word	0,62765957	0,41403509	0,49894292	285
支	Dual classifier	0,54098361	0,11702128	0,19241983	282
台	Dual classifier	0,35454545	0,14285714	0,20365535	273
朵	True classifier	0,47328244	0,496	0,484375	250
瓶	Not in lists	0,27857143	0,15983607	0,203125	244
等	Not in lists	0,90140845	0,80334728	0,84955752	239
股	Measure word	0,44897959	0,19047619	0,2674772	231
丝	Measure word	0,32926829	0,11842105	0,17419355	228
趟	Not in lists	0,46296296	0,11013216	0,17793594	227
根	True classifier	0,35	0,09251101	0,14634146	227
封	True classifier	0,79787234	0,66666667	0,72639225	225
包	Not in lists	0,15384615	0,00947867	0,01785714	211
轮	Dual classifier	0,41269841	0,12745098	0,19475655	204
头	True classifier	0,50617284	0,21354167	0,3003663	192
幅	True classifier	0,54166667	0,13684211	0,21848739	190
遍	Not in lists	0,31818182	0,07407407	0,12017167	189
副	Measure word	0,44444444	0,11363636	0,18099548	176
门	Dual classifier	0,53125	0,21656051	0,30769231	157
粒	True classifier	0	0	0	152
枚	True classifier	1	0,08783784	0,16149068	148
组	Measure word	0,60526316	0,15862069	0,25136612	145
盘	Not in lists	0	0	0	145
批	Measure word	0	0	0	143
间	True classifier	0,5	0,02898551	0,05479452	138
笔	True classifier	0,5	0,02222222	0,04255319	135
身	Measure word	0,52307692	0,28571429	0,36956522	119
棵	True classifier	0,6	0,50420168	0,54794521	119
波	Not in lists	0	0	0	116

样	Not in lists	0,57142857	0,27586207	0,37209302	116
桌	Measure word	0,33333333	0,01818182	0,03448276	110
楼	Not in lists	0,25333333	0,18095238	0,21111111	105
季	Not in lists	0,5	0,33980583	0,40462428	103
盒	Not in lists	0,16666667	0,00990099	0,01869159	101
下	Not in lists	0	0	0	97
盆	Not in lists	1	0,02083333	0,04081633	96
面	True classifier	0,7	0,22580645	0,34146341	93
箱	Not in lists	0,4	0,02197802	0,04166667	91
页	Dual classifier	1	0,02247191	0,04395604	89
处	True classifier	0	0	0	89
节	Dual classifier	0,62195122	0,6	0,61077844	85
串	Measure word	0	0	0	82
排	Measure word	0	0	0	81
栋	True classifier	0,61290323	0,24358974	0,34862385	78
系列	Not in lists	0,2	0,05263158	0,08333333	76
袋	Not in lists	0	0	0	74
锅	Not in lists	0,375	0,04166667	0,075	72
盏	True classifier	0,675	0,375	0,48214286	72
束	Measure word	0	0	0	70
通	Not in lists	0	0	0	68
团	Measure word	0,88235294	0,2238806	0,35714286	67
圈	Measure word	0	0	0	66
所	True classifier	0,5	0,046875	0,08571429	64
发	True classifier	0	0	0	62
世	Not in lists	0	0	0	59
扇	True classifier	0	0	0	57
桶	Not in lists	0	0	0	56
壶	Not in lists	0	0	0	56
堂	Measure word	0	0	0	54
餐	Not in lists	0	0	0	54
则	True classifier	0	0	0	52
艘	True classifier	0,55555556	0,19230769	0,28571429	52
架	True classifier	0	0	0	50
线	Dual classifier	0	0	0	49
曲	True classifier	0,33333333	0,02040816	0,03846154	49
匹	True classifier	0,76923077	0,20408163	0,32258065	49
户	Measure word	0,41176471	0,15909091	0,2295082	44
肚子	Not in lists	0	0	0	43
笼	Not in lists	0,75	0,14285714	0,24	42
手	Not in lists	1	0,45238095	0,62295082	42
伙	Measure word	0	0	0	38
枝	True classifier	0	0	0	37
罐	Not in lists	0	0	0	36
卷	True classifier	1	0,13888889	0,24390244	36
幕	Not in lists	0	0	0	35
行	Measure word	0	0	0	34
码	Not in lists	0,42857143	0,08823529	0,14634146	34
刀	Not in lists	0	0	0	30
任	Not in lists	0	0	0	30
株	True classifier	0	0	0	30
脸	Not in lists	0	0	0	28

环	Not in lists	1	0,03703704	0,07142857	27
辈	Not in lists	0	0	0	27
幢	True classifier	0	0	0	27
般	Not in lists	0,33333333	0,03846154	0,06896552	26
桩	Not in lists	0	0	0	26
顶	True classifier	0	0	0	24
尾	True classifier	0	0	0	23
尊	True classifier	0	0	0	23
列	Measure word	0	0	0	22
册	True classifier	0	0	0	22
章	Measure word	0	0	0	21
路	Measure word	0	0	0	21
宗	Dual classifier	0	0	0	20
版	Measure word	0	0	0	20
杆	True classifier	0	0	0	19
拨	Not in lists	0	0	0	18
袭	True classifier	0	0	0	18
记	True classifier	0	0	0	17
剂	True classifier	0	0	0	17
具	True classifier	0	0	0	17
帖	Measure word	0	0	0	13
队	Measure word	0	0	0	13
味	Measure word	0	0	0	12
例	Not in lists	0	0	0	12
局	Not in lists	0	0	0	12
席	True classifier	0	0	0	9
管	True classifier	0	0	0	9
档子	Not in lists	0	0	0	8
人次	Not in lists	0	0	0	6
缸	Not in lists	0	0	0	6
缕	Dual classifier	0	0	0	4
遭	Not in lists	0	0	0	4
拳	Not in lists	0	0	0	2
棒	Not in lists	0	0	0	1
员	True classifier	0	0	0	1
堵	True classifier	0	0	0	1
accuracy	0,704421492				

MLM	**category**	**precision**	**recall**	**f1-score**	**support**
weighted avg		0,77280731	0,62226088	0,6821145	136221
macro avg*		51,91	33,4	37,68	136221
个	True classifier	0,83262864	0,7163086	0,770101	61581
种	Not in lists	0,85707269	0,67478732	0,75508438	10344
次	Not in lists	0,83471188	0,65736619	0,73549954	9211
张	True classifier	0,85561682	0,73439811	0,79038676	3381
件	True classifier	0,83745583	0,68431184	0,75317797	3117
句	True classifier	0,9	0,69795918	0,7862069	2450
条	True classifier	0,74180564	0,54870268	0,63080685	2351
位	True classifier	0,49654776	0,39990732	0,44301848	2158
家	Dual classifier	0,68191721	0,54656577	0,60678514	1718
场	Not in lists	0,76747967	0,55693215	0,64547009	1695
只	True classifier	0,64980237	0,53480807	0,58672377	1537

点	Dual classifier	0,36758321	0,35131397	0,3592645	1446
部	Dual classifier	0,70927093	0,56085409	0,6263911	1405
首	True classifier	0,89068826	0,82212257	0,85503304	1338
段	Measure word	0,76847291	0,58558559	0,66467831	1332
篇	True classifier	0,91118421	0,84066768	0,87450671	1318
滴	Measure word	0,79908676	0,15138408	0,25454545	1156
份	Dual classifier	0,58492823	0,43236074	0,49720386	1131
号	Measure word	0,82251908	0,38482143	0,5243309	1120
块	Dual classifier	0,71195652	0,37861272	0,49433962	1038
颗	True classifier	0,83201058	0,65932914	0,73567251	954
名	True classifier	0,69855072	0,59506173	0,64266667	810
群	Measure word	0,51526718	0,16917293	0,25471698	798
款	Measure word	0,68023256	0,45466321	0,54503106	772
片	Dual classifier	0,74914676	0,60136986	0,66717325	730
堆	Measure word	0,66528926	0,2350365	0,34735707	685
本	True classifier	0,79918864	0,62440571	0,70106762	631
些	Not in lists	0,09869803	0,37301587	0,15609432	630
级	Measure word	0,90356394	0,68521463	0,77938517	629
分	Dual classifier	0,35313531	0,54871795	0,42971888	585
杯	Not in lists	0,5487106	0,67311072	0,60457774	569
起	Dual classifier	0,61024845	0,70938628	0,65609349	554
步	Not in lists	0,86092715	0,26694045	0,40752351	487
顿	Not in lists	0,63186813	0,49145299	0,55288462	468
套	Measure word	0,59171598	0,21881838	0,31948882	457
把	Dual classifier	0,73462783	0,50332594	0,59736842	451
集	Measure word	0,65502183	0,34324943	0,45045045	437
对	Measure word	0,80246914	0,44724771	0,57437408	436
辆	True classifier	0,67094017	0,73023256	0,69933185	430
碗	Not in lists	0,56647399	0,46117647	0,50843061	425
回	Measure word	0,72222222	0,33966746	0,46203554	421
代	Measure word	0,79761905	0,52072539	0,63009404	386
声	True classifier	0,63513514	0,55131965	0,59026688	341
座	True classifier	0,66019417	0,41337386	0,50841121	329
届	Not in lists	0,83882784	0,70030581	0,76333333	327
阵	Not in lists	0,54601227	0,27384615	0,3647541	325
道	True classifier	0,50151057	0,52201258	0,51155624	318
类	Not in lists	0,62121212	0,26623377	0,37272727	308
班	Dual classifier	0,49112426	0,27213115	0,35021097	305
层	Measure word	0,67788462	0,47	0,55511811	300
项	Measure word	0,68911917	0,45392491	0,5473251	293
口	Dual classifier	0,65486726	0,50859107	0,57253385	291
番	Not in lists	0,70748299	0,35738832	0,47488584	291
双	Measure word	0,68965517	0,63157895	0,65934066	285
支	Dual classifier	0,47945205	0,24822695	0,3271028	282
台	Dual classifier	0,42718447	0,32234432	0,36743215	273
朵	True classifier	0,62857143	0,528	0,57391304	250
瓶	Not in lists	0,405	0,33196721	0,36486486	244
等	Not in lists	0,91584158	0,77405858	0,83900227	239
股	Measure word	0,59217877	0,45887446	0,51707317	231
丝	Measure word	0,61458333	0,25877193	0,36419753	228
趟	Not in lists	0,49382716	0,17621145	0,25974026	227
根	True classifier	0,50769231	0,43612335	0,46919431	227

封	True classifier	0,85	0,45333333	0,59130435	225
包	Not in lists	0,43181818	0,09004739	0,14901961	211
轮	Dual classifier	0,67692308	0,43137255	0,52694611	204
头	True classifier	0,48695652	0,29166667	0,36482085	192
幅	True classifier	0,61607143	0,36315789	0,45695364	190
遍	Not in lists	0,578125	0,1957672	0,29249012	189
副	Measure word	0,59230769	0,4375	0,50326797	176
门	Dual classifier	0,62857143	0,28025478	0,3876652	157
粒	True classifier	0,5	0,10526316	0,17391304	152
枚	True classifier	0,55555556	0,2027027	0,2970297	148
组	Measure word	0,38596491	0,15172414	0,21782178	145
盘	Not in lists	0,57142857	0,13793103	0,22222222	145
批	Measure word	0,39344262	0,16783217	0,23529412	143
间	True classifier	0,49333333	0,26811594	0,34741784	138
笔	True classifier	0,64367816	0,41481481	0,5045045	135
身	Measure word	0,432	0,45378151	0,44262295	119
棵	True classifier	0,68421053	0,65546218	0,6695279	119
波	Not in lists	0,41025641	0,13793103	0,20645161	116
样	Not in lists	0,08860759	0,42241379	0,14648729	116
桌	Measure word	0,5	0,1	0,16666667	110
楼	Not in lists	0,45045045	0,47619048	0,46296296	105
季	Not in lists	0,40298507	0,52427184	0,4556962	103
盒	Not in lists	0,4	0,03960396	0,07207207	101
下	Not in lists	0,05918058	0,40206186	0,1031746	97
盆	Not in lists	0,32142857	0,1875	0,23684211	96
面	True classifier	0,3045977	0,56989247	0,39700375	93
箱	Not in lists	0,5	0,04395604	0,08080808	91
页	Dual classifier	0,44262295	0,30337079	0,36	89
处	True classifier	0,375	0,20224719	0,26277372	89
节	Dual classifier	0,59210526	0,52941176	0,55900621	85
串	Measure word	0,33333333	0,03658537	0,06593407	82
排	Measure word	0,45	0,11111111	0,17821782	81
栋	True classifier	0,51851852	0,35897436	0,42424242	78
系列	Not in lists	0	0	0	76
袋	Not in lists	0,1875	0,04054054	0,06666667	74
锅	Not in lists	0,5483871	0,23611111	0,33009709	72
盏	True classifier	0,6557377	0,55555556	0,60150376	72
束	Measure word	0,52941176	0,25714286	0,34615385	70
通	Not in lists	0,31372549	0,23529412	0,26890756	68
团	Measure word	0,46875	0,2238806	0,3030303	67
圈	Measure word	0,41666667	0,22727273	0,29411765	66
所	True classifier	0,58208955	0,609375	0,59541985	64
发	True classifier	0,27118644	0,25806452	0,26446281	62
世	Not in lists	0,45454545	0,33898305	0,38834951	59
扇	True classifier	0,78431373	0,70175439	0,74074074	57
桶	Not in lists	0,66666667	0,21428571	0,32432432	56
壶	Not in lists	0,75	0,10714286	0,1875	56
堂	Measure word	0,36	0,16666667	0,2278481	54
餐	Not in lists	0,09090909	0,05555556	0,06896552	54
则	True classifier	0,39285714	0,21153846	0,275	52
艘	True classifier	0,81132075	0,82692308	0,81904762	52
架	True classifier	0,46875	0,3	0,36585366	50

线	Dual classifier	0,61538462	0,32653061	0,42666667	49
曲	True classifier	0,45454545	0,30612245	0,36585366	49
匹	True classifier	0,69387755	0,69387755	0,69387755	49
户	Measure word	0,75862069	0,5	0,60273973	44
肚子	Not in lists	0	0	0	43
笼	Not in lists	1	0,30952381	0,47272727	42
手	Not in lists	0,24074074	0,30952381	0,27083333	42
伙	Measure word	0,75	0,15789474	0,26086957	38
枝	True classifier	0,2	0,05405405	0,08510638	37
罐	Not in lists	0	0	0	36
卷	True classifier	0,5625	0,25	0,34615385	36
幕	Not in lists	0,25925926	0,2	0,22580645	35
行	Measure word	0,22222222	0,23529412	0,22857143	34
码	Not in lists	1	0,08823529	0,16216216	34
刀	Not in lists	0,61538462	0,26666667	0,37209302	30
任	Not in lists	0,63636364	0,46666667	0,53846154	30
株	True classifier	0,25	0,06666667	0,10526316	30
脸	Not in lists	0,10638298	0,17857143	0,13333333	28
环	Not in lists	0,92307692	0,44444444	0,6	27
辈	Not in lists	0,16666667	0,11111111	0,13333333	27
幢	True classifier	0,5	0,03703704	0,06896552	27
般	Not in lists	0,04494382	0,15384615	0,06956522	26
桩	Not in lists	0	0	0	26
顶	True classifier	0,48	0,5	0,48979592	24
尾	True classifier	0,90909091	0,43478261	0,58823529	23
尊	True classifier	0,70588235	0,52173913	0,6	23
列	Measure word	0,22222222	0,18181818	0,2	22
册	True classifier	1	0,18181818	0,30769231	22
章	Measure word	0,03067485	0,23809524	0,05434783	21
路	Measure word	0,04347826	0,19047619	0,07079646	21
宗	Dual classifier	0,77777778	0,35	0,48275862	20
版	Measure word	0,25	0,1	0,14285714	20
杆	True classifier	1	0,05263158	0,1	19
拨	Not in lists	1	0,05555556	0,10526316	18
袭	True classifier	0,5	0,22222222	0,30769231	18
记	True classifier	0,18181818	0,11764706	0,14285714	17
剂	True classifier	0,44444444	0,23529412	0,30769231	17
具	True classifier	0,7	0,41176471	0,51851852	17
帖	Measure word	0,5	0,07692308	0,13333333	13
队	Measure word	0	0	0	13
味	Measure word	0,2	0,41666667	0,27027027	12
例	Not in lists	0,5	0,25	0,33333333	12
局	Not in lists	0	0	0	12
席	True classifier	0,5	0,11111111	0,18181818	9
管	True classifier	0	0	0	9
档子	Not in lists	0	0	0	8
人次	Not in lists	0	0	0	6
缸	Not in lists	1	0,16666667	0,28571429	6
缕	Dual classifier	0,07692308	0,5	0,13333333	4
遭	Not in lists	0	0	0	4
拳	Not in lists	0	0	0	2
棒	Not in lists	0	0	0	1

员	True classifier	0	0	0	1
堵	True classifier	0	0	0	1
accuracy	0,622260885				
**These values have been calculated manually due to problems in the automatic system*					

BERT	category	precision	recall	f1-score	support
weighted avg		0,80703749	0,81709869	0,80771142	136221
macro avg		0,52860653	0,38099189	0,40771484	136221
个	True classifier	0,8959591	0,92785112	0,91162627	61581
种	Not in lists	0,84989331	0,88563418	0,86739573	10344
次	Not in lists	0,82882603	0,8684182	0,84816032	9211
张	True classifier	0,80915395	0,8627625	0,83509877	3381
件	True classifier	0,85958372	0,88771254	0,87342172	3117
句	True classifier	0,82828685	0,84857143	0,83830645	2450
条	True classifier	0,74743326	0,77413866	0,76055161	2351
位	True classifier	0,66324435	0,5987025	0,62932294	2158
家	Dual classifier	0,70411568	0,73690338	0,72013652	1718
场	Not in lists	0,70685841	0,7539823	0,72966029	1695
只	True classifier	0,6610499	0,66363045	0,66233766	1537
点	Dual classifier	0,6240115	0,60027663	0,61191399	1446
部	Dual classifier	0,78370584	0,77366548	0,7786533	1405
首	True classifier	0,86278736	0,89760837	0,87985348	1338
段	Measure word	0,73333333	0,71021021	0,72158658	1332
篇	True classifier	0,91896145	0,8861912	0,90227887	1318
滴	Measure word	0,74494949	0,76557093	0,75511945	1156
份	Dual classifier	0,58733032	0,57382847	0,58050089	1131
号	Measure word	0,8019884	0,86428571	0,8319725	1120
块	Dual classifier	0,65233303	0,68689788	0,6691694	1038
颗	True classifier	0,75653083	0,75890985	0,75771847	954
名	True classifier	0,73435504	0,70987654	0,72190835	810
群	Measure word	0,51097561	0,52506266	0,51792336	798
款	Measure word	0,66111772	0,72020725	0,68939864	772
片	Dual classifier	0,72294372	0,68630137	0,70414617	730
堆	Measure word	0,55520995	0,52116788	0,5376506	685
本	True classifier	0,75195008	0,76386688	0,75786164	631
些	Not in lists	0,62020906	0,56507937	0,59136213	630
级	Measure word	0,88760331	0,85373609	0,87034036	629
分	Dual classifier	0,72186495	0,76752137	0,74399337	585
杯	Not in lists	0,62234795	0,77328647	0,68965517	569
起	Dual classifier	0,6873065	0,80144404	0,74	554
步	Not in lists	0,77346939	0,77823409	0,77584442	487
顿	Not in lists	0,5862069	0,69017094	0,63395486	468
套	Measure word	0,5467128	0,34573304	0,42359249	457
把	Dual classifier	0,64853556	0,68736142	0,66738428	451
集	Measure word	0,5754717	0,6979405	0,63081696	437
对	Measure word	0,83950617	0,62385321	0,71578947	436
辆	True classifier	0,64944649	0,81860465	0,72427984	430
碗	Not in lists	0,55514019	0,69882353	0,61875	425
回	Measure word	0,75206612	0,43230404	0,54901961	421
代	Measure word	0,68421053	0,67357513	0,67885117	386
声	True classifier	0,64637681	0,65395894	0,65014577	341
座	True classifier	0,67462687	0,68693009	0,68072289	329

届	Not in lists	0,76627219	0,79204893	0,77894737	327
阵	Not in lists	0,57876712	0,52	0,54781199	325
道	True classifier	0,63481229	0,58490566	0,60883797	318
类	Not in lists	0,65680473	0,36038961	0,46540881	308
班	Dual classifier	0,56478405	0,55737705	0,56105611	305
层	Measure word	0,59793814	0,58	0,58883249	300
项	Measure word	0,62030075	0,56313993	0,59033989	293
口	Dual classifier	0,65503876	0,58075601	0,61566485	291
番	Not in lists	0,58436214	0,48797251	0,53183521	291
双	Measure word	0,70322581	0,76491228	0,73277311	285
支	Dual classifier	0,50543478	0,32978723	0,39914163	282
台	Dual classifier	0,49246231	0,35897436	0,41525424	273
朵	True classifier	0,55752212	0,756	0,6417657	250
瓶	Not in lists	0,39393939	0,4795082	0,43253235	244
等	Not in lists	0,92990654	0,83263598	0,8785872	239
股	Measure word	0,60515021	0,61038961	0,60775862	231
丝	Measure word	0,55665025	0,49561404	0,52436195	228
趟	Not in lists	0,52427184	0,47577093	0,49884527	227
根	True classifier	0,50922509	0,60792952	0,55421687	227
封	True classifier	0,78995434	0,76888889	0,77927928	225
包	Not in lists	0,34188034	0,18957346	0,24390244	211
轮	Dual classifier	0,6119403	0,40196078	0,4852071	204
头	True classifier	0,65853659	0,421875	0,51428571	192
幅	True classifier	0,61068702	0,42105263	0,49844237	190
遍	Not in lists	0,54658385	0,46560847	0,50285714	189
副	Measure word	0,49333333	0,63068182	0,55361596	176
门	Dual classifier	0,56692913	0,45859873	0,50704225	157
粒	True classifier	0,43283582	0,19078947	0,26484018	152
枚	True classifier	0,64285714	0,24324324	0,35294118	148
组	Measure word	0,70454545	0,2137931	0,32804233	145
盘	Not in lists	0,5	0,10344828	0,17142857	145
批	Measure word	0,52	0,09090909	0,1547619	143
间	True classifier	0,44186047	0,27536232	0,33928571	138
笔	True classifier	0,62135922	0,47407407	0,53781513	135
身	Measure word	0,57142857	0,47058824	0,51612903	119
棵	True classifier	0,61805556	0,74789916	0,67680608	119
波	Not in lists	0,625	0,12931034	0,21428571	116
样	Not in lists	0,59633028	0,56034483	0,57777778	116
桌	Measure word	0,66666667	0,18181818	0,28571429	110
楼	Not in lists	0,51041667	0,46666667	0,48756219	105
季	Not in lists	0,50537634	0,45631068	0,47959184	103
盒	Not in lists	0,31578947	0,05940594	0,1	101
下	Not in lists	1	0,01030928	0,02040816	97
盆	Not in lists	0,42857143	0,15625	0,22900763	96
面	True classifier	0,65217391	0,48387097	0,55555556	93
箱	Not in lists	0,44444444	0,13186813	0,20338983	91
页	Dual classifier	0,48979592	0,26966292	0,34782609	89
处	True classifier	0,85714286	0,06741573	0,125	89
节	Dual classifier	0,62637363	0,67058824	0,64772727	85
串	Measure word	0	0	0	82
排	Measure word	0	0	0	81
栋	True classifier	0,52702703	0,5	0,51315789	78

系列	Not in lists	0,54285714	0,25	0,34234234	76
袋	Not in lists	0	0	0	74
锅	Not in lists	0,59090909	0,18055556	0,27659574	72
盏	True classifier	0,66176471	0,625	0,64285714	72
束	Measure word	0,66666667	0,08571429	0,15189873	70
通	Not in lists	0,61904762	0,19117647	0,29213483	68
团	Measure word	1	0,20895522	0,34567901	67
圈	Measure word	0,79166667	0,28787879	0,42222222	66
所	True classifier	0,66071429	0,578125	0,61666667	64
发	True classifier	1	0,01612903	0,03174603	62
世	Not in lists	0,48387097	0,25423729	0,33333333	59
扇	True classifier	0,78431373	0,70175439	0,74074074	57
桶	Not in lists	1	0,07142857	0,13333333	56
壶	Not in lists	0	0	0	56
堂	Measure word	0,83333333	0,09259259	0,16666667	54
餐	Not in lists	0	0	0	54
则	True classifier	0,625	0,19230769	0,29411765	52
艘	True classifier	0,70491803	0,82692308	0,76106195	52
架	True classifier	0,64	0,32	0,42666667	50
线	Dual classifier	0	0	0	49
曲	True classifier	0,625	0,10204082	0,1754386	49
匹	True classifier	0,81395349	0,71428571	0,76086957	49
户	Measure word	0,75757576	0,56818182	0,64935065	44
肚子	Not in lists	0,71428571	0,23255814	0,35087719	43
笼	Not in lists	0,57142857	0,57142857	0,57142857	42
手	Not in lists	0,92	0,54761905	0,68656716	42
伙	Measure word	0	0	0	38
枝	True classifier	0,125	0,10810811	0,11594203	37
罐	Not in lists	0	0	0	36
卷	True classifier	1	0,13888889	0,24390244	36
幕	Not in lists	0	0	0	35
行	Measure word	0	0	0	34
码	Not in lists	0,66666667	0,29411765	0,40816327	34
刀	Not in lists	0	0	0	30
任	Not in lists	0,58333333	0,23333333	0,33333333	30
株	True classifier	0	0	0	30
脸	Not in lists	0	0	0	28
环	Not in lists	0,93333333	0,51851852	0,66666667	27
辈	Not in lists	0	0	0	27
幢	True classifier	0	0	0	27
般	Not in lists	0,625	0,19230769	0,29411765	26
桩	Not in lists	0	0	0	26
顶	True classifier	0,83333333	0,20833333	0,33333333	24
尾	True classifier	1	0,43478261	0,60606061	23
尊	True classifier	0	0	0	23
列	Measure word	1	0,09090909	0,16666667	22
册	True classifier	0	0	0	22
章	Measure word	0	0	0	21
路	Measure word	0	0	0	21
宗	Dual classifier	0	0	0	20
版	Measure word	1	0,05	0,0952381	20
杆	True classifier	0	0	0	19

拨	Not in lists	0	0	0	18
袭	True classifier	1	0,11111111	0,2	18
记	True classifier	0	0	0	17
剂	True classifier	1	0,05882353	0,11111111	17
具	True classifier	0,77777778	0,41176471	0,53846154	17
帖	Measure word	0	0	0	13
队	Measure word	0	0	0	13
味	Measure word	1	0,16666667	0,28571429	12
例	Not in lists	0	0	0	12
局	Not in lists	0	0	0	12
席	True classifier	0	0	0	9
管	True classifier	0	0	0	9
档子	Not in lists	0	0	0	8
人次	Not in lists	0	0	0	6
缸	Not in lists	0	0	0	6
缕	Dual classifier	0	0	0	4
遭	Not in lists	1	0,5	0,66666667	4
拳	Not in lists	0	0	0	2
棒	Not in lists	0	0	0	1
员	True classifier	0	0	0	1
堵	True classifier	0	0	0	1
accuracy	0,817098685				