Computational Modelling of Plurality and Definiteness in Chinese Noun Phrases

Yuqi Liu♠, Guanyi Chen♡†, Kees van Deemter♠
♡Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer Science, Central China Normal University
♠Department of Information and Computing Sciences, Utrecht University
y.liu30@students.uu.nl, g.chen@ccnu.edu.cn, c.j.vandeemter@uu.nl

Abstract

Theoretical linguists have suggested that some languages (e.g., Chinese and Japanese) are "cooler" than other languages based on the observation that the intended meaning of phrases in these languages depends more on their contexts. As a result, many expressions in these languages are shortened, and their meaning is inferred from the context. In this paper, we focus on the omission of the plurality and definiteness markers in Chinese noun phrases (NPs) to investigate the predictability of their intended meaning given the contexts. To this end, we built a corpus of Chinese NPs, each of which is accompanied by its corresponding context, and by labels indicating its singularity/plurality and definiteness/indefiniteness. We carried out corpus assessments and analyses. The results suggest that Chinese speakers indeed drop plurality and definiteness markers very frequently. Building on the corpus, we train a bank of computational models using both classic machine learning models and state-of-the-art pre-trained language models to predict the plurality and definiteness of each NP. We report on the performance of these models and analyse their behaviours. The code and data used in this paper are available at: https://github.com/andyzxq/chinese_np_def.

Keywords: Chinese Linguistics, Noun Phrase, Plurality, Definiteness

1. Introduction

It has been pointed out that speakers trade-off clarity against brevity (Grice, 1975) and speakers of different languages appear to handle this trade-off differently (Newnham, 1971). Ross (1982) and Huang (1984) elaborated this idea by hypothesising that some languages (especially, Eastern Asian languages, e.g., Chinese and Japanese) are "cooler" than other languages. A language A is considered to be cooler than language B if understanding sentences of A tends to require more work by readers or the hearers than understanding sentences of B. As a consequence, speakers of relatively cool languages often omit pronouns (causing *pro-drop*) and assume that listeners can infer the missing information from the context. Later on, the theory was extended, suggesting that many components in cool language are omittable (Van der Auwera and Baoill, 1998), such as plurality markers, definiteness markers (Huang et al., 2009), discourse connectives (Yu, 1993) and so on.

So far, most works have analysed related language phenomena as built into a language's grammar (e.g., the grammar of Chinese permits pro-drop). Only a few studies focused on the pragmatic aspects of coolness (Chen and van Deemter, 2020, 2022; Chen, 2022). For instance, Chen et al. (2018) investigated the use of pro-drop by modelling the choices of speakers computationally. To the best of our knowledge, no similar study has focused on listeners' understanding.

To fill this gap, we investigate the comprehension of two kinds of omittable information in Chinese noun phrases (NPs)[1], namely, plurality and definiteness, which are two major foci of research on NPs (Iljic, 1994; Bremmers et al., 2022). The corresponding comprehension tasks for English are trivial because the plurality and definiteness of an English NP are always conveyed through explicit markers. In contrast, in Chinese, a bare noun can be either definite or indefinite and either singular or plural. Consider the following examples of the noun "狗" (*dog*) from Huang et al. (2009):

(1) a. 狗 很 聪明 。
gou hen congming .
'Dogs are intelligent.'
b. 我 看到 狗 。
wo kandao gou .
'I saw a dog/dogs.'
c. 狗 跑走了 。
gou paozou le .
'The dog(s) ran away.'

The word "狗" in (1-a) makes a general reference, translated as "*dogs*". In the sentence (1-b), the NP "狗" is indefinite, but whether it refers to a single dog or a set of dogs needs to be decided by wider

†Corresponding Author

[1]Note that we concentrate on Mandarin Chinese in this study.

LREC-COLING 2024, pages 3666–3676
20-25 May, 2024.

contexts. Likewise, the plurality status of the "狗" in the sentence (1-c) is hard to decide without further context, but it is certainly definite.

In this study, we build computational models to understand the following research question:

> *To what extent plurality and definiteness of Chinese NPs are predictable from their contexts*?

To this end, we formalised these two comprehension tasks as two classification tasks, i.e., classifying a Chinese NP as plural or singular and as definite or indefinite. We first built a dataset, in which each NP is annotated with its plurality and definiteness, on the basis of a large-scale English-Chinese parallel corpus. More specially, from the parallel corpus, we did word alignments and designed an algorithm to match NPs in the two languages based on the alignment results. We extracted NPs in Chinese and annotated the plurality and definiteness of each NP according to its matched English NP. To guarantee the quality of the dataset, we conducted two human assessment studies which were then analysed and compared.

We then performed a corpus analysis e.g. to investigate to what proportion plurality and definiteness are implicitly expressed. Subsequently, we tested mainstream classification techniques, from the classic machine learning based classifiers to the more recent pre-trained language model based classifiers, on the dataset to investigate the predictability of plurality and definiteness, and we analysed their behaviour.

2. Dataset

One of the major challenges of the present study is the construction of a large-scale dataset in which each NP is annotated with its plurality and definiteness. This is extraordinarily hard not only because building a large-scale human-annotated dataset is expensive, but also because many linguistic studies have demonstrated that deciding plurality and definiteness (especially definiteness) in Chinese NPs is a challenging task for even native speakers (e.g., Robertson (2000)).

Instead, inspired by Wang et al. (2016), in which they focused on pro-drop in machine translation systems, and the "translation mining" in corpus linguistics (Bremmers et al., 2022), since English speakers always convey plurality and definiteness explicitly, we can annotate a Chinese NP automatically if we have its English translation. Such information can be found in any English-Chinese parallel corpus.

More specifically, given a parallel corpus, we first did the word alignments and designed a simple but effective algorithm to extract and match NPs in both languages. Then, we annotated each Chinese NP based on its associated English NP. In what follows, we detail the automatic annotation process, introduce the resulting corpus and how we assess its quality.

2.1. Dataset Construction

Since we are investigating the pragmatics of Chinese NPs, the corpus needs to reflect the everyday use of language. In other words, the corpora that are constructed from news or novels are not appropriate. Therefore, we used the TV episode subtitle corpus, which was constructed and pre-processed by Wang et al. (2018)[2]. It contains 4.39 million English and Chinese sentence pairs in total.

Word Alignment. We used GIZA++ (Och and Ney, 2003) to generate alignment proposals. Note that the alignment proposal is sometimes different when aligning words in English to Chinese and when aligning words in Chinese to English. Therefore, at this step, we recorded the alignments of both "directions" for future use.

NP Identification. We used CoreNLP (Manning et al., 2014) to parse each sentence in each language and extracted all NPs from the parse tree. We also recorded the Part-of-speech (POS) tag for each word in this step.

NP Matching. With the word alignments and identified NPs in hand, we design a simple but effective method in which there are two steps: (1) for each direction, an NP in the source language is paired with the NP in the target language that has the most aligned words with it; (2) a match is done if and only if two NPs are paired in both directions.

Post-processing. Since NPs are often in nested structures and not all NPs interested us, we filtered out some NPs: (1) we removed all NP conjunctions and only kept their constituents. For example, the NP "*Zhangsan and Lisi*" contains two NPs. We remove it and keep only "*Zhangsan*" and "*Lisi*"; (2) apart from NP conjunctions, for each NP, we dropped all its constituents. For example, if all of the "*Lisi's book*", "*Lisi*" and "*book*" are matched in the previous step, we only keep "*Lisi's book*" in our dataset. We also remove all NPs that are pronouns as they are not the focus of this study.

[2]The data came from two subtitle websites in China: http://www.opensubtitles.org and http://weisheshou.com.

	PLURALITY		DEFINITENESS	
	Singular	Plural	Definite	Indefinite
train	79158	24528	48471	55215
dev	7894	2474	4777	5591
test	7925	2444	4844	5525

Table 1: The basic statistics of our dataset.

Annotation. For each Chinese NP, we annotated its matched English NP and used the resulting labels (i.e., plurality and definiteness) as its annotation. Concretely, we annotated an English NP as plural if: (1) it has a plural POS tag (i.e., NNS or NNPS); (2) it is a numeral phrase that specifies a quantity larger than one (i.e., "*two cups of coffee*"). Otherwise, it is a singular NP. For the definiteness of an NP, the annotation was done based on (1) its article; (2) whether it is a demonstrative phrase (i.e., whether it contains a demonstrative (decided based on its POS tag and its surface form), such as "*this*" or "*that*"); and (3) whether it is a proper name (i.e., an NP is definite if it is a proper name).

2.2. The Corpus

Due to the limitation of computing resources, we sampled and annotated 5% of the data from Wang et al. (2018) for further computational modelling (described in the next section). Table 1 charts the basic statistics of the resulting dataset. The dataset contains 124K annotated NPs. More than 3 quarters NPs are marked as singular. The definiteness labels are rather balanced. 58K samples are annotated as indefinite NPs while 66K samples are definite. We then divided the dataset into the training, development and test sets with the ratio 8:1:1.

2.2.1. Is "men" a plural marker?

The inflectional morpheme "们" (men) was considered as a plural marker in Chinese. However, in the past several decades, theoretical linguists argued that it is indeed a collective marker (Iljic, 1994; Li, 1999), highlighting that a referent is a group of people and referring to the group as a whole (translated as "*group of*" or "*set of*") because it is incompatible with number phrase. Later on, Huang et al. (2009) further demonstrated that an NP with "们" *must* be interpreted as definite. Therefore, it would be interesting to look into the labels of NPs with "们" in our dataset.

We extracted all NPs whose head noun has a "们" suffix[3] and did statistics on the label distribution. Regarding plurality, we found that although most extracted NPs were still marked as plural. There are still a remarkable amount of singular NPs (approximately, 9.12%). This suggests that, in line with the linguistic theory, the suffix "们" is not a conclusive marker of plural. Regarding definiteness, inconsistent with what linguists suggested, most extracted NPs were marked as indefinite (approximately, 63.84%). For example, the "大人们" (adults) in the example (2) apparently does not have a definite reading. This embodies the conclusion that says "们" must be interpreted as definite is questioned.

(2) 大人们 会 告诉 你 并 不是 这样 。
darenmen hui gaosu ni bing bushi zheyang
Adults will tell you this is not the case.

2.2.2. How frequently do Chinese speakers express plurality or definiteness explicitly?

For each NP in the dataset, we annotate whether it expresses plurality or definiteness explicitly based on the POS tags and the parsing tree of the sentence in which this NP is located. We marked an NP express plurality explicitly if it contains a numeral or a measure word. We marked an NP express definiteness explicitly if (1) it contains a proper name; (2) it includes a possessive; (3) there is a numeral or measure present, with a preceding demonstrative.

At length, we identified that merely 12.42% utterances convey plurality explicitly and 15.86% utterances contain explicit definiteness markers. This confirms that Chinese, as a "cool" language, its speakers indeed do not use explicit plurality and definiteness markers very often.

2.3. Quality Assessment

Last but not least, it is essential to ensure that the corpus is suitable for use in computational modelling. We manually assess its quality from the aspects of plurality and definiteness annotation as well as NP identification. In what follows, we describe our assessment process.

2.3.1. Assessment 1

We randomly sampled 400 samples for human assessment, the NP in each of which was highlighted. We hired four annotators and ensured that each sample was assessed by 2 annotators. All of them are native speakers of Chinese. Three of them are males and one of them is female. Two of them have backgrounds in engineering, one in statistics, and one in Language study.

[3]We remove NPs in which "们" does not function as suffixes, e.g., "哥们" (brother), and is part of pronouns, e.g., "我们" (we).

	Assessment 1				Assessment 2			
	$Acc_{=2}$	$Acc_{\geq 1}$	IAA (%)	IAA (κ)	$Acc_{=2}$	$Acc_{\geq 1}$	IAA (%)	IAA (κ)
NP Identification	79.50	96.25	0.8325	-	-	-	-	-
Plurality	84.00	96.75	0.8725	0.6477	74.00	85.50	0.8850	0.6679
Definiteness	81.00	97.25	0.8375	0.6731	53.00	77.50	0.7550	0.4755

Table 2: Human Assessment Results, in which IAA (%) is the percentage agreement and IAA (κ) is the Cohen's Kappa.

Concretely, we asked annotators three questions (translated from Chinese): (1) Is the highlighted noun phrase correctly identified? (2) Is this a singular/plural (decided by the annotation in our corpus) phrase? and (3) Is this a definite/indefinite phrase?

After the experiment, we computed the accuracy and inter-annotator agreements (IAA). We computed two types of accuracy based on the number of annotators in agreement with the annotation. $Acc_{=2}$ measures the proportion of accurate annotations agreed upon by both annotators, while $Acc_{\geq 1}$ measures those agreed upon by at least one annotator. For IAA, we computed both the percentage agreement and Cohen's Kappa (Cohen, 1960)[4].

Table 2 charts the human assessment results. All three tasks received $Acc_{=2}$ around 80% and $Acc_{\geq 1}$ higher than 96%. One can ask why NP identification has received lower scores than the other two tasks. One major reason is that most identified incorrect NP identifications are about unsuccessfully including all modifiers (e.g., marking only "*the men*" from the true NP "*the man who is old*").

These results suggest that our corpus is of good quality, on the one hand. On the other hand, disagreements between two annotators exist in all three tasks. The percentage agreements of all three tasks are around 85% and the Kappa values for the plurality and definiteness annotations are approximately 0.65, suggesting substantial agreements between annotators. Nonetheless, we also noticed that the IAAs for the definiteness annotation are surprisingly high. This is counter-intuitive because, as aforementioned, many previous studies suggested that deciding the definiteness is hard for Chinese native speakers (e.g., Robertson (2000)). This may be attributed to the Framing Effects in human evaluation (Schoch et al., 2020). In particular, our use of yes or no questions might have influenced the evaluators' decisions, leading to a bias towards favouring a positive response. Additionally, such an influence may be magnified as disagreements exist by nature in our tasks. Therefore, we conducted assessment 2 as a complement.

[4]We did not calculate Cohen's Kappa on raw human decisions because, for each question, the marginal probability of one answer is much greater than the other (i.e., there are much more "yes" than "no"), which makes Cohen's Kappa inaccurate (Brennan and Prediger, 1981; Maclure and Willett, 1987; Donker et al., 1993). Instead, we first translated their decisions in accordance with our labels. For example, if the label in our corpus is "plural" and the annotator provided a positive response, we assumed that the annotator annotated this sample as "plural".

2.3.2. Assessment 2

To minimise the bias introduced by the framing effects, in assessment 2, we gave each annotator samples from our dataset in which NPs were highlighted while labels were removed. We asked them to directly annotate the plurality and definiteness of each NP. This time, we sampled another 200 samples and, again, ensured that each sample is annotated by 2 annotators. The results are also reported in Table 2.

Although $Acc_{=2}$ for plurality reduced from 84% to 74% while that for definiteness dramatically reduced from 81% to 51%, there are still approximately 80% of our annotations agreed by at least one human annotator for both tasks. In this new assessment, the IAA for plurality stays high while the IAA for definiteness decreases. The kappa value for definiteness drops from 0.67 (assessment 1) to 0.48, indicating a moderate agreement. These results cohere with what linguists suggested.

2.3.3. Summary

We found that, for all three tasks, disagreements (between two annotators and between the annotators' annotation and our annotation) exist and differ with respect to how the questions are framed. Despite the disagreements, the assessment results indicate that our corpus is of acceptable quality. In both assessments, at worst, approximately 80% of our annotations can be agreed upon by at least one human annotator.

2.4. Limitations

Regarding our annotation and assessment processes, our corpus exhibits the following limitations: **First**, as shown in the assessment experiments,

disagreements exist in the human annotation. This is true for many pragmatic tasks (Poesio et al., 2019). However, our automatic annotation strategy cannot take such agreements into consideration. **Second**, Chinese does not distinguish countable and uncountable nouns. By looking into the human annotation results from assessment 2, we found that since Chinese is a classifier language (i.e., numerals obligatorily appear with classifiers when they modify nouns) many NPs with uncountable nouns were considered plural NPs. Because we annotate NPs in Chinese using the information from their English translations, we failed to annotate these uncountable nouns correctly. **Thrid**, both our automatic annotation and human assessments are precision-oriented. For example, we dropped the Chinese NP that did not match with any English NPs and, during the assessments, we only used NPs that had been matched. This makes our corpus overlook some Chinese NPs and our assessments ignore recall. **Last**, in the assessments, we did not evaluate how the decisions of annotators would be influenced by providing them with additional contexts for each sample. This limitation was recognised because, as mentioned in Section 1, the meaning of a Chinese NP relies more on its context compared to its English counterpart.

3. Models

In this section, we introduce models we built for predicting plurality and definiteness. We tried a large variety of models: from classic machine learning (ML) based models to the most recent pre-trained language model (PLM) based models.

3.1. ML-based Models

We tried a number of classic ML-based classifiers on our plurality and definition prediction tasks. To this end, we first used '*' to mark the target NP in each sample. For example, "我 的 母亲" (my mom) is the target NP in the following sentence.

(3) 我 爱 * 我 的 母亲 * 。
wo ai * wo de muqin * .
I love * my mom * .

We used N-gram ($N = 1, 2, 3, 4$) as features for classification[5]. As for the algorithm, we tried Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM).

[5]The performance of our classifiers can be further boosted using advanced features, e.g., POS tags or syntactic structures. Since, in this study, we were investigating the predictability of plurality and definiteness of NPs from their contexts, we used only raw features from the contexts.

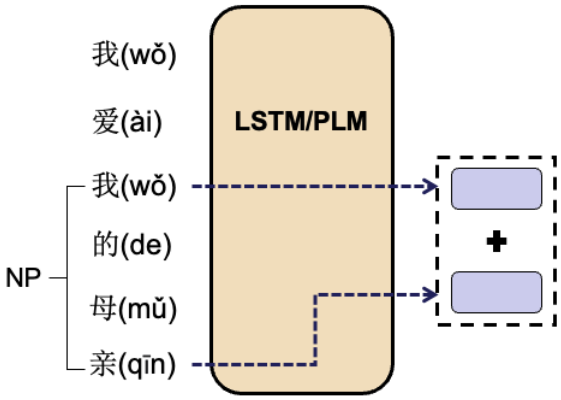

Figure 1: Illustration of the PLM-based Models.

3.2. PLM-based Models

Recently, the developments in NLP to a large extent attributed to the introduction of PLMs. This contribution stems from two perspectives: utilising the knowledge acquired through large-scale pre-training and leveraging a broader context. Recall that we are investigating whether the plurality and definiteness of an NP can be predicted from its context. Therefore, it is plausible to assume that such predictions also benefit from using (contextual) PLMs.

To this end, we fine-tune PLMs on our dataset. As depicted in Figure 1, we fed the raw text into a PLM, and, for each NP, we extracted the representations of its first token and its last token. The prediction was made by a dense output layer based on the summation of these two representations.

In this study, we tried the following PLMs: (1) Chinese `BERT` and `RoBERTa` (Devlin et al., 2019; Liu et al., 2019). (2) `BERT-wwm`: vanilla Chinese BERT was pre-trained as a fully character-based model, but Cui et al. (2021) proved that the performance can be boosted if Whole Word Masking (WWM; rather than character level masking) is done during pre-training. (3) `mBERT`: since in addition to Chinese, there are multiple other "cool" languages (e.g., Japanese, Korean and Arabic), we, therefore, wanted to validate whether the predictions can benefit from multilingual pre-training or not. (4) `BiLSTM`: In addition to PLMs, we also tested bi-directional LSTM (Schuster and Paliwal, 1997) initialised by the Glove embeddings (Pennington et al., 2014). The architecture is the same as how we used PLMs.

4. Experiments

In this section, we introduce the evaluation protocol and report the performance of the models.

	Plurality						Definiteness					
	MACRO AVG			WEIGHTED AVG			MACRO AVG			WEIGHTED AVG		
	P	R	F	P	R	F	P	R	F	P	R	F
RF	81.08	58.19	58.53	80.26	79.69	74.19	68.63	67.24	67.10	68.51	68.09	67.47
LR	76.08	67.39	69.79	80.11	81.58	79.77	71.73	71.53	71.58	71.78	71.82	71.75
SVM	75.56	67.37	69.69	79.88	81.40	79.65	71.34	71.04	71.10	71.37	71.40	71.29
BiLSTM	79.31	70.94	73.59	82.49	83.50	82.14	76.78	76.88	76.80	76.95	76.84	76.87
BERT	80.88	77.96	79.24	85.23	85.73	85.37	81.60	81.66	81.63	81.71	81.69	81.69
BERT-wwm	80.94	**78.34**	**79.50**	**85.38**	85.83	**85.52**	81.95	81.82	81.87	81.98	81.98	81.97
mBERT	80.07	76.96	78.30	84.58	85.15	84.74	80.70	80.41	80.50	80.68	80.66	80.62
RoBERTa	81.21	77.53	79.09	85.22	85.79	85.35	**82.27**	**82.10**	**82.16**	**82.28**	**82.28**	**82.26**
RoBERTa (large)	**81.72**	77.37	79.17	**85.38**	**85.98**	85.46	81.80	81.58	81.66	81.79	81.79	81.76

Table 3: The performance of our models for plurality and definiteness predictions depicted in Section 3. "P", "R" and "F" stand for precision, recall and F-score respectively. The best results are **boldfaced**, whereas the second best are underlined. The PLMs that do not mark '(large)' use their base version. For many Chinese PLMs, only the base models are publicly available.

4.1. Evaluation Protocol

We tuned the hyper-parameters of each of our models on the development set and chose the setting with the best macro F1 score. We report the macro/weighted averaged precision, recall, and F1 on the test set.

4.2. Experimental Results

Table 3 depicts the results of the plurality and definiteness classifications. The results suggest that all models can learn useful information for both plurality and definiteness predictions. Similar to human beings, models also face more challenges when making predictions about definiteness compared to plurality, as evidenced by the lower weighted scores in definiteness predictions compared to plurality predictions.

For model performance, as expected, PLM-based models outperformed their ML-based counterparts. Among ML-based models, we found that LR is very effective, achieving weighted-averaged F-scores of 79.77 for plurality predictions and 71.75 for definiteness predictions. BiLSTM with Glove embeddings defeated all ML-based models but lost to Models with BERT. This embodies that context plays an important role in the prediction of plurality and definiteness, which is consistent with the definition of "cool" (see Section 1).

Among BERT-based models, we had the following observations: (1) BERT-wwm performed remarkably well. It generally performed the best for plurality prediction and was the second-best model for definiteness prediction. This demonstrated that, on pragmatics tasks (e.g., our tasks), BERT does benefit from whole word mask pre-training probably because the intended meaning of a word (noun in our situation) is mainly inferred from its context rather than its inner structure. (2) BERT did not benefit from multilingual pre-training as mBERT received 84.74 weighted F-score on plurality predictions and 80.62 on definiteness predictions though mBERT was pre-trained on typical "cool" languages, including Arabic, Japanese and Korean. This is probably attributed to the fact that speakers of these "cool" languages use contexts differently and, therefore, multi-lingual pre-training may not yield substantial benefits to downstream tasks that rely on context. This makes supervision signals become needed. In the future, it would be valuable to build an NP corpus in multiple "cool" languages and see whether the predictions can benefit or not. (3) Interestingly, on our tasks, the amount of parameters is not the more the better. RoBERTa-large performed worse than the vanilla BERT on plurality predictions and worse than RoBERTa-base on definiteness predictions. Further probing experiments are needed to explain what happens.

5. Analysis

In what follows, we analyse the model behaviour concerning three questions.

5.1. What is the impact of Context Size?

According to what Huang (1984) hypothesised, the interpretation of the plurality and definiteness of an NP relies on its context and such context is not necessarily only the current sentence but also the whole discourse. For example, without more context, it is hard to decide the plurality of the NP in example (1-c). However, in the current experimental setting, we only fed the models with only one sentence, namely the target sentence.

	4-way						2-way (merged)					
	MACRO AVG			WEIGHTED AVG			MACRO AVG			WEIGHTED AVG		
	P	R	F	P	R	F	P	R	F	P	R	F
BERT	67.37	64.26	65.53	70.72	71.20	70.79	65.62	63.35	64.34	69.49	69.91	69.61
BERT-wwm	67.94	65.74	66.72	71.54	71.86	71.62	66.51	**64.23**	**65.24**	70.03	70.40	70.14
mBERT	67.73	64.58	65.69	71.12	71.46	71.01	64.19	61.51	62.62	68.11	68.59	68.21
RoBERTa	68.25	**66.42**	**67.24**	72.03	72.36	72.14	67.08	63.89	65.23	**70.29**	**70.74**	**70.36**
RoBERTa (large)	**68.73**	65.51	66.87	**72.09**	**72.55**	**72.18**	**67.11**	63.36	64.90	69.90	70.35	69.92

Table 4: The results of 4-way prediction and the merged results of 2 binary predictions.

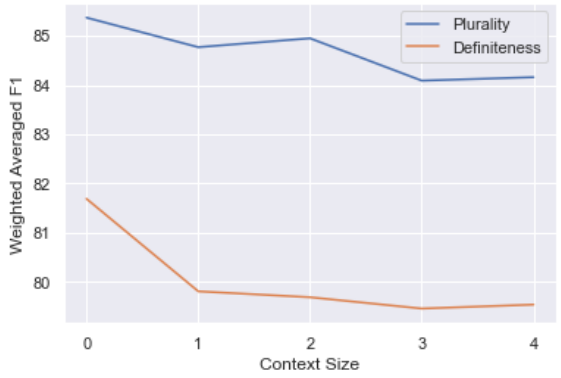

Figure 2: Weighted F1 concerning different context sizes. The size is measured by the number of sentences around the target sentence.

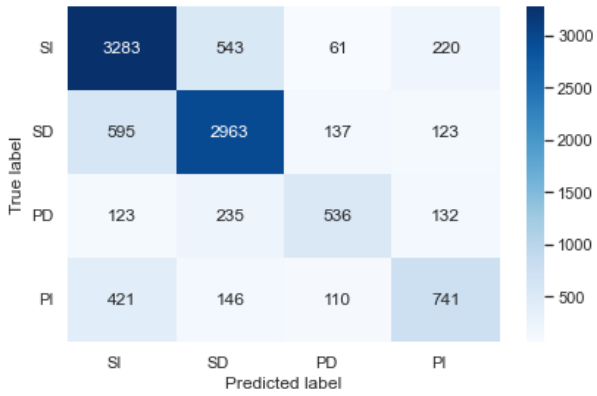

Figure 3: The confusion matrix for 4-way prediction of RoBERTa-large, in which S, P, I and D mean "singular", "plural", "indefinite" and "definite", respectively.

Therefore, it is plausible to expect that if we increase the size of contexts, the predictions become more accurate. To validate this idea, we increased the size and assessed BERT with the inputs with different amounts of contexts. Figure 2 prints the evaluation results of the two tasks, in which 1 means both the previous sentence and the preceding sentence are seen as the context and be fed to the models together with the target sentence.

Nevertheless, different from the expectation, the performance of both tasks decreases with the increase of the context size. The decrease in performance is more pronounced in definiteness prediction compared to plurality prediction. A possible explanation is that although wider contexts add useful information to the prediction, it also adds confusion as our focus is only a small part (i.e., the NP) of the target sentence. This makes it hard for the model to extract useful information from the representation of a wide context, and add it to the representation of the target NP (which is often a few words; recall that we only used the representation of the target NP for prediction), and make predictions. It is worth noting that similar phenomena are observed in other pragmatics tasks (Joshi et al., 2019; Baruah et al., 2020; Same et al., 2022; Chen et al., 2023).

5.2. Do the plurality and definiteness predictions help each other?

Since both plurality and definiteness are information carried by NPs. One could expect that the information that is needed for predicting the plurality of an NP might help determine the definiteness of the same NP and vice versa. In other words, we might benefit from predicting plurality and definiteness simultaneously. Rather than employing multi-task learning, we opted to fine-tune the models for 4-way predictions. Specifically, given an NP, the models classify it into one of four categories: indefinite singular, indefinite plural, definite singular, or definite plural. To fairly compare the model performance for 4-way prediction and 2 separate binary predictions, we merged the predictions obtained in Section 4 and re-computed each score.

Table 4 reports the performance of each model on 4-way and merged binary predictions. The results suggest that models can significantly benefit from predicting plurality and definiteness simultaneously compared to predicting them separately. For example, in joint prediction, RoBERTa achieved a weighted average F1 score of 72.14. However, when doing binary predictions, the merged

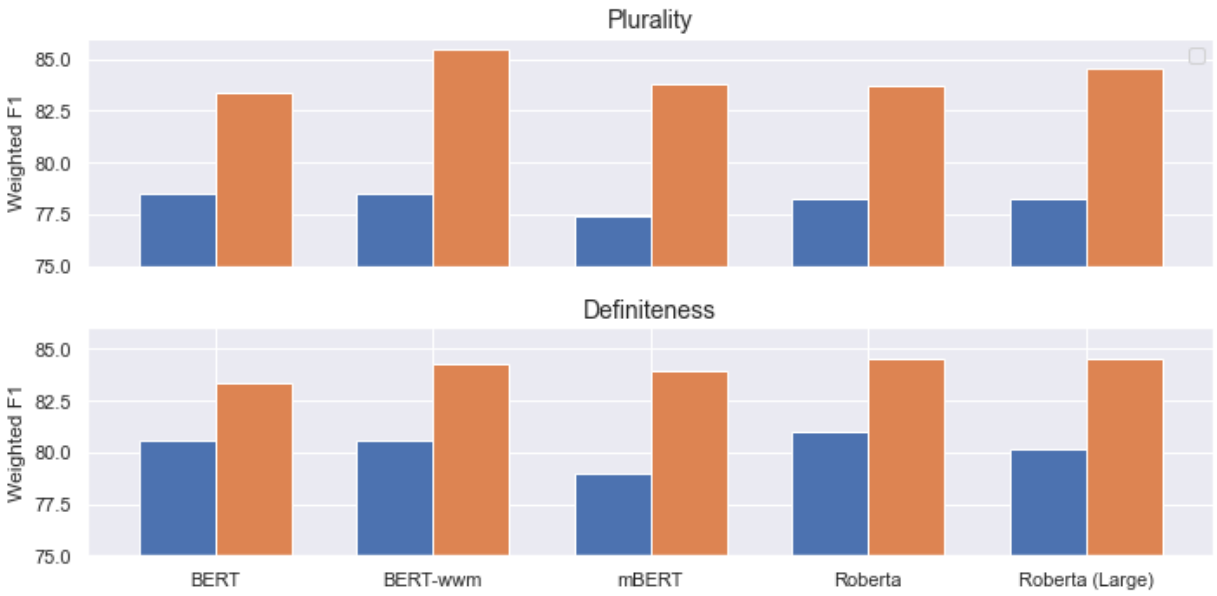

Figure 4: Macro F-scores of BERT-based models on implicit and explicit expressions of plurality and definiteness. The blue bars indicate the performance of models on implicit expressions while the orange bars indicate that on explicit expressions.

weighted F1 score dropped to 70.36.

Focusing on the 4-way prediction results, we found that akin to the binary predictions, RoBERTa had the best performance. It achieved a weighted F1 score of 72.14 and a micro F1 score of 67.24. It was followed by RoBERTa-large, who had an on-par weighted F1 and lower micro F1. BERT-wwm performed slightly worse than them, but still remarkably well. Figure 3 is the confusion matrix of Roberta-large for joint prediction, which further ascertains the theory that deciding definiteness is hard in Chinese as although the labels of the plurality are way more imbalanced than that of the definiteness (see Table 1) the model is still much easier to confuse between "definite" and "indefinite" than between "singular" and "plural".

5.3. How does the explicitness impact the model's behaviours?

In the corpus analysis, we identified that NPs in 12.42% and 15.86% samples from our dataset explicitly express plurality and definiteness respectively. Since these explicit expressions provide clear markers, we expected that the predictions of both tasks on explicit expressions are easier than on implicit expressions. Thus, models would receive higher scores on the portion of explicit expressions. To examine this, we assessed BERT-based models on implicit and explicit expressions respectively and report the results in Figure 4[6]. As expected, for both tasks, all models performed better on explicit expressions and implicit expressions.

Besides, we also have some interesting observations: (1) the difference between the performance on explicit expressions and on implicit expressions is larger on plurality prediction than definiteness prediction. (2) For plurality prediction, except mBERT, all other models have similar performance on implicit expressions. BERT-wwm performed significantly better on explicit expressions than other models. (3) For definiteness prediction, RoBERTa performed the best on both implicit and explicit expressions.

6. Conclusion

We investigated one pragmatic aspect of the "coolness" hypothesis by Huang (1984): in a "cool" language, whether the meaning of an omittable component is predictable or not. To this end, we studied the predictability of plurality and definiteness in Chinese NPs, which, syntactically, are omittable. We first constructed a Chinese corpus where each NP is marked with its plurality and definiteness. Two assessment studies showed that our corpus is of good quality. A corpus analysis suggests that Chinese speakers frequently drop plural and definiteness markers.

Based on the corpus, we built computational models using both classic ML-based models and the most recent PLM-based models. The experimental results showed that both ML-based models and PLM-based models can learn information for

[6]To highlight the differences, we report macro-F this time.

predicting the meaning of plurality and definiteness of NPs from their contexts and that BERT-wwm generally performed the best due to its good ability to extract information from contexts in Chinese. Further analyses of the models suggested that the information for predicting plurality and definiteness benefits from each other.

Regarding "coolness", through computational modelling, we confirmed that the plurality and definiteness of Chinese NPs are predictable from their contexts. Furthermore, these predictions can be improved if the model's ability to capture contexts is enhanced. Nonetheless, in addition to the research question presented in the current study (see Section 1), another crucial question remains unanswered: to what extent do these computational models mimic listeners' way of comprehending plurality and definiteness? To address this question in the future, we intend to create a corpus in which disagreements among listeners are annotated, which is then used for assessing computational models.

7. Ethical Considerations and Limitations

In this work, potential biases may have two sources. One is the dataset. Our dataset was built from TV episodes, which have never been filtered with respect to toxic content. The other is the pre-trained language models we used, which have been widely discussed in, e.g., Bender et al. (2021).

In addition to the limitations we discussed in Section 2.4, another key limitation is that our corpus analyses and computational modelling were done on the data from a single source, namely, conversations in TV episodes. It is not fully clear whether our findings can be generalised to data in other genres. Moreover, the data we used is a parallel corpus, where the Chinese texts were translated from English. While these texts maintain a natural tone, there's a risk that translations may diverge from everyday language use.

8. Acknowledgement

We are grateful for comments from reviewers and people in Utrecht's Natural Language Processing group.

9. Bibliographical References

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Context-aware sarcasm detection using BERT. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 83–87, Online. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT*.

David Bremmers, Jianan Liu, Martijn van der Klis, and Bert Le Bruyn. 2022. Translation mining: Definiteness across languages (a reply to jenks 2018). *Linguistic Inquiry*, 53(4):735–752.

Robert L Brennan and Dale J Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699.

Guanyi Chen. 2022. *Computational generation of Chinese noun phrases*. Ph.D. thesis, Utrecht University.

Guanyi Chen, Fahime Same, and Kees van Deemter. 2023. Neural referential form selection: Generalisability and interpretability. *Computer Speech & Language*, 79:101466.

Guanyi Chen and Kees van Deemter. 2020. Lessons from computational modelling of reference production in Mandarin and English. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 263–272, Dublin, Ireland. Association for Computational Linguistics.

Guanyi Chen and Kees van Deemter. 2022. Understanding the use of quantifiers in Mandarin. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 73–80, Online only. Association for Computational Linguistics.

Guanyi Chen, Kees van Deemter, and Chenghua Lin. 2018. Modelling pro-drop with the rational speech acts model. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 159–164, Tilburg University, The Netherlands. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

DK Donker, A Hasman, and HP Van Geijn. 1993. Interpretation of low kappa values. *International journal of bio-medical computing*, 33(1):55–64.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry*, pages 531–574.

C-T James Huang, Y-H Audrey Li, and Yafei Li. 2009. The syntax of chinese. *(No Title)*.

Robert Iljic. 1994. Quantification in mandarin chinese: Two markers of plurality.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Yen-hui Audrey Li. 1999. Plurality in a classifier language. *Journal of East Asian Linguistics*, pages 75–99.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Malcolm Maclure and Walter C Willett. 1987. Misinterpretation and misuse of the kappa statistic. *American journal of epidemiology*, 126(2):161–169.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Richard Newnham. 1971. *About Chinese*. Penguin Books Ltd.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Robertson. 2000. Variability in the use of the english article system by chinese learners of english. *Second language research*, 16(2):135–172.

John Ross. 1982. Pronoun deleting processes in german. In *Annual Meeting of the Linguistics Society of America*, San Diego, California.

Fahime Same, Guanyi Chen, and Kees Van Deemter. 2022. Non-neural models matter: a re-evaluation of neural referring expression generation systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5554–5567, Dublin, Ireland. Association for Computational Linguistics.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Johan Van der Auwera and Dónall Ó Baoill. 1998. *Adverbial constructions in the languages of Europe*. Walter de Gruyter.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–993, San Diego, California. Association for Computational Linguistics.

Ning Yu. 1993. Chinese as a paratactic language. *Journal of Second Language Acquisition and Teaching*, 1:1–15.