Models of reference production: How do they withstand the test of time?

Fahime Same[♡], **Guanyi Chen**[♠], and **Kees van Deemter**[♠]
[♡]Department of Linguistics, University of Cologne
[♠]Department of Information and Computing Sciences, Utrecht University
f.same@uni-koeln.de, g.chen@ccnu.edu.cn, c.j.vandeemter@uu.nl

Abstract

In recent years, many NLP studies have focused solely on performance improvement. In this work, we focus on the linguistic and scientific aspects of NLP. We use the task of generating referring expressions in context (REG-in-context) as a case study and start our analysis from GREC, a comprehensive set of shared tasks in English that addressed this topic over a decade ago. We ask what the performance of models would be if we assessed them (1) on more realistic datasets, and (2) using more advanced methods. We test the models using different evaluation metrics and feature selection experiments. We conclude that GREC can no longer be regarded as offering a reliable assessment of models' ability to mimic human reference production, because the results are highly impacted by the choice of corpus and evaluation metrics. Our results also suggest that pre-trained language models are less dependent on the choice of corpus than classic Machine Learning models, and therefore make more robust class predictions.

1 Introduction

NLP research can have different aims. Some NLP research focuses on developing new algorithms or building practical NLP applications. Another line of NLP work constructs computational models that aim to explain human language and language use; this line of work has been dubbed *NLP-as-Science* (van Deemter, 2023). Among other things, NLP-as-Science demands that we ask ourselves to what extent NLP research findings generalise along a range of dimensions.

In addition to the practical applications of Referring Expression Generation (REG, Reiter, 2017), REG is also one of the typical tasks in NLP-as-Science, where REG algorithms are built to model and explain the reference production of human beings (Krahmer and van Deemter, 2012; van Deemter, 2016). In the computational linguistics and cognitive science community, REG can be divided into two distinct tasks: *one-shot REG*, finding a referring expression (RE) to single out a referent from a set, and *REG-in-context*, generating an RE to refer to a referent at a given point in a discourse.

In a classic setup, REG-in-context is often approached in two steps: The first is to decide on the form of an RE at a given point in the discourse, and the second is to decide on its content. Many researchers have been interested in the first subtask, referential form selection: the task to decide which referential form (e.g., pronoun, proper name, description, etc.) an RE takes (McCoy and Strube, 1999; Henschel et al., 2000; Kibrik et al., 2016). Nearly 15 years ago, Belz et al. (2008) introduced the GREC shared tasks and a number of English REG corpora with two goals: (1) assessing the performance of computational models of reference production (Belz et al., 2009), and (2) understanding the contribution of linguistically-inspired factors to the choice of referential form (Greenbacker and McCoy, 2009b; Kibrik et al., 2016; Same and van Deemter, 2020).

15 years have passed since the GREC challenge was organised, and many new models and corpora have been proposed in the meantime (e.g., Castro Ferreira et al. (2018); Cunha et al. (2020), and Same et al. (2022)). We, therefore, decided that it was time to ask, in the spirit of NLP-as-Science, how well the lessons that GREC once taught our research community hold up when scrutinised in light of all these developments. In other words, we will investigate to what extent the findings from GREC can be *generalised* to other corpora and other models.

To this end, we pursue the following objectives: (1) We extend GREC by testing its REG algorithms not only on the GREC corpora but also on a corpus that was not originally considered and that has a different genre, namely the Wall Street Journal

Proceedings of the 16th International Natural Language Generation Conference, pages 93–105
September 11–15, 2023. ©2023 Association for Computational Linguistics

(WSJ) portion of OntoNotes (Hovy et al., 2006; Weischedel et al., 2013); (2) We fine-tune pre-trained language models on the task of REG-in-context and assess them in the GREC framework.

In Section 2, we detail the GREC shared tasks and introduce the corpora used in GREC. Section 3 spells out our research questions. In Section 4 and Section 5, we introduce the algorithms and corpora that we use. Section 6 reports the performance of each algorithm on each corpus, followed by analyses in Section 7. Section 8 will discuss our findings and draw some lessons.

2 The GREC Shared Tasks

In this section, we summarise the GREC task, the corpora used by GREC, and its conclusions.

2.1 The GREC Task and its Corpora

According to Belz et al., "*the GREC tasks are about how to generate appropriate references to an entity in the context of a piece of discourse longer than a sentence*" (2009, p. 297). The main task was to predict the referential form, namely whether to use a pronoun, proper name, description or an empty reference at a given point in discourse.

The GREC challenges use two corpora, both created from the introductory sections of Wikipedia articles: (1) GREC-2.0 (henceforth MSR, as it was used in the GREC-MSR shared tasks of 2008 and 2009) consists of 1941 introductory sections of the articles across five domains (people, river, mountain, city, and country); and (2) GREC-People (henceforth NEG as it was used in the GREC-NEG shared task in 2009) contains 1000 introductory sections from Wikipedia articles about composers, chefs, and inventors. Here is an example from NEG:

(1) **David Chang** (born 1977) is a noted American chef. **He** is chef/owner of Momofuku Noodle Bar, Momofuku Ko and Momofuku Ssäm Bar in New York City. **Chang** attended Trinity College, where **he** majored in religious studies. In 2003, **Chang** opened **his** first restaurant, Momofuku Noodle Bar, in the East Village.

A key difference between MSR and NEG lies in their RE annotation practices. In MSR, only those REs that refer to the main topic of the article are annotated, while in NEG, mentions of all *human* referents are annotated. For instance, in a document about David Chang, MSR will only annotate REs referring to David Chang, while NEG will include annotations for all human referents, including David Chang and others.

Name	GREC ST	ALG	Acc
UDel	MSR '09	C5.0	77.71
ICSI	MSR '09	CRF	75.16
CNTS	MSR '08	MBL	72.61
IS-G	MSR '08	MLP	70.78
OSU	MSR '08	MaxEnt	69.82
JUNLG	MSR '09	Rule	75.40

Table 1: An overview of the algorithms submitted to GREC. The first column contains the name of the respective algorithm. The column GREC ST presents the name of the MSR shared task to which the algorithm was submitted. The third column, ALG, lists the algorithms used, where abbreviations from top to bottom are C5.0 decision tree, conditional random field, memory-based learning, multi-layer perceptron, maximum entropy, and frequency-based rules. The fourth column, Acc, reports the original accuracy of the algorithms, as reported in Belz et al. (2009). Note that UDel, ICSI, and JUNLG were submitted to both the MSR '08 and MSR'09 shared tasks, and we only present the newest results here.

2.2 REG Algorithms Submitted to GREC

Various REG algorithms were submitted to the GREC challenges. These consist of feature-based ML algorithms: CNTS (Hendrickx et al., 2008), ICSI (Favre and Bohnet, 2009), IS-G (Bohnet, 2008), OSU Jamison and Mehay (2008) and UDel Greenbacker and McCoy (2009a), and an algorithm that mixes feature-based ML and rules: JUNLG (Gupta and Bandopadhyay, 2009). Table 1 presents the details of each model, including the ML method, and the original reported accuracy on MSR (cf. Belz et al. (2009) for details).

2.3 Feature Selection

The GREC Tasks were designed to find out *what kind of information is useful for making choices between different kinds of referring expressions in context* (Belz et al., 2009, p. 297). However, the original paper does not consider the factors that contributed to the RE choice in the systems submitted to GREC. In a follow-up study, Greenbacker and McCoy (2009b) conducted a feature selection study informed by psycholinguistics. They experimented with various feature subsets derived from their system, known as UDel, which had previously been submitted to the GREC. Additionally, they incorporated selected features from another

REG system, CNTS (Hendrickx et al., 2008), into their study. They show that features motivated by psycholinguistic studies and certain sentence construction features have a positive impact on the performance of REG models. Follow-up feature-selection studies including Kibrik et al. (2016) and Same and van Deemter (2020) also emphasise the contribution of factors such as recency and grammatical role to the choice of RE form.

3 Research Questions

15 years after the GREC shared tasks, we were curious to know to what extent the conclusions from GREC still "stand". We, therefore, came up with the following research questions.

In the first place, we are interested in *the impact of the choice of corpus on the performance of REG algorithms* ($\mathcal{R}_1$). GREC uses only the introductory part of Wikipedia articles (see Section 2), which represents only one genre of human language use. Considering that a good REG algorithm needs to model the general use of reference, a better evaluation framework should include texts from multiple genres. Therefore, we also include the WSJ corpus in the study (see Section 5 for more details) and conduct a correlation analysis to quantify how the choice of corpus impacts the evaluation results.

Second, previous studies suggested that classic machine learning (ML) based REG algorithms perform on par with most recent neural methods (Same et al., 2022). However, their study has three limitations: (1) they did not incorporate pre-trained language models (PLMs); (2) they focused on the surface forms of REs, which partly depend on the performance of surface realisation; (3) they did not assess the models based on the intuition that a model with good explanatory power should be less influenced by the choice of corpus. Therefore, we adopt PLMs to the task of REG-in-context (see Section 4 for more details) and investigate *how good is the explanatory power of PLM-based REG models compared to classic ML-based models* ($\mathcal{R}_2$) using the enhanced GREC framework.

Finally, as previously mentioned, one of the primary theoretical objectives of GREC was to computationally explore the contribution of factors that originate from linguistic studies to the choice of referential forms. It is reasonable to expect that such contributions may change depending on the choice of corpus. In this study, we conduct an importance analysis to investigate *whether the importance ranking of linguistic factors changes when we use different corpora* ($\mathcal{R}_3$).

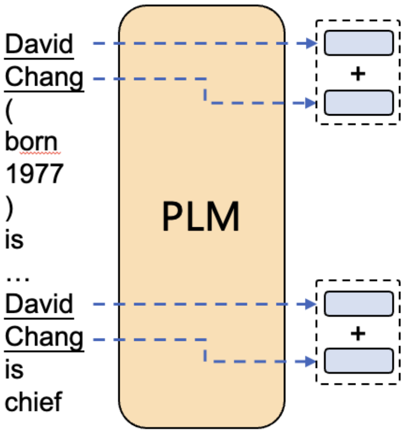

Figure 1: Illustration of the PLM-based REG Algorithm.

4 REG Algorithms

In what follows, we introduce the REG algorithms that are considered in this study.

4.1 ML-based REG

For this study, we have narrowed our focus to feature-based ML algorithms that predict the type of RE. Consequently, we reconstruct five ML-based REG algorithms, namely `UDel`, `ICSI`, `CNTS`, `IS-G`, and `OSU`, along with their respective feature sets, while excluding `JUNLG`. Note that we implement `CNTS` slightly differently from Hendrickx et al. (2008). Concretely, Hendrickx et al. (2008) have mentioned that they have used the TiMBL package (Daelemans et al., 2007) for implementing the Memory Based Learning algorithm. Instead, we implemented the k-Nearest Neighbors algorithm. According to Daelemans et al. (2007), Memory Based Learning is the direct descendant of k-Nearest Neighbors. More information on the implementation of these models can be found in Appendix B.

4.2 PLM-based REG

Deep learning approaches have been used in many previous works on REG (Castro Ferreira et al., 2019; Cao and Cheung, 2019; Cunha et al., 2020; Chen et al., 2021). Different from previous work[1],

[1] Note that Chen et al. (2021, 2023) also leveraged a PLM, but did not fine-tune it. Instead, they used the word representations from the PLM as static inputs to an RNN and made predictions using the RNN.

we fine-tune PLMs on REG corpora in this study.

To fine-tune PLMs on REG corpora, we began by pre-processing each corpus using the same paradigm as described by Cunha et al. (2020). More precisely, each referent in a given document was replaced with its corresponding proper name. For example, all underlined REs in Example (1) were replaced by "David Chang". Subsequently, as depicted in Figure 1, we fed the data into a PLM, and, for each referent (e.g., "David Chang"), we extracted the representations of its first token and its last token and summed them. The final representations were then sent to a fully connected layer for predicting the RE forms. In this study, we use BERT and RoBERTa (see section 6.1 for more details).

5 REG Corpora

In the following, we explain the corpora used in this work. These corpora are English-language corpora.

5.1 The MSR and NEG Corpora

In the current study, we only use the articles from the training sets of these corpora (see the number of documents in Table 2). Following the same approach as Castro Ferreira et al. (2018), we created a version of the GREC corpora for the End-to-end (E2E) REG modelling. For the classic ML models, we reproduced the models using the feature sets from the studies mentioned in Section 2.2.

5.2 The WSJ Corpus

As mentioned earlier, the WSJ portion of the OntoNotes corpus (Weischedel et al., 2013) is our third data source.[2] We use the version of the corpus that Same et al. (2022) developed for E2E REG modeling.[3] Since empty pronouns are not annotated in WSJ, we decided to also exclude them from the two GREC corpora and focus on a 3-label classification task. The labels considered in this study are *pronoun*, *description*, and *proper name*. Table 2 presents a detailed overview of these corpora.

Data Splits. We have made a document-wise split of the data. We split the WSJ data in accordance with the CoNLL 2012 Shared Task (Pradhan et al., 2012). Our WSJ training, development, and test sets contain 20275, 2831, and 2294 samples, respectively. We did an 85-5-10 split of the GREC datasets in accordance with Belz et al. (2009). After excluding empty pronouns, the MSR training, development, and test sets contain 9413, 519, 1038 instances, and the NEG training, development, and test sets contain 6681, 259, 896 instances.

	MSR	NEG	WSJ
number of documents	1655	808	582
word/doc (mean)	148	129	530
sent/doc (mean)	7.1	5.8	25
par/doc (mean)	2.3	2.2	10.8
referent/doc (mean)	1	2.6	15
number of RE	11705	8378	25400
description %	13.84%	4%	38.29%
proper name %	38.09%	40.79%	34.57%
pronoun %	41.79%	48.75%	27.14%
empty %	6.28%	6.47%	-

Table 2: Comparison of the MSR, NEG, and WSJ corpora in terms of their length-related characteristics and distribution of REs. *Doc*, *sent* and *par* stands for *documents*, *sentences* and *paragraphs*.

Proportion of Referring Expressions As shown in Table 2, pronouns and proper names make up 80% and 89.5% of the referential instances in MSR and NEG, respectively. This implies that the other two referential forms, namely descriptions and empty references, account for approximately 20% of the cases in MSR and about 10% in NEG. Given this imbalance in the frequency of different forms within the two corpora, we question its potential effect on algorithm performance. Specifically, we are wondering if forms with lower frequencies are accurately predicted by the algorithms.

6 Evaluation

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

6.1 Implementation Details

For BERT and RoBERTa, we used *bert-base-cased* and *roberta-base*, both from Hugging Face. For fine-tuning, we set the batch size to 16, the learning rate to 1e-3, the dropout rate to 0.5, and the size of the output layer to 256. We ran each model for 20 epochs and used the one that achieved the highest F1 score on the development set. The implementation details of the classic ML-based models can be found in Appendix B.

6.2 Evaluation Protocol

The main evaluation metric in the GREC-MSR shared tasks was accuracy. In addition to accuracy,

[2] We used Ontonotes 5.0 licensed by the Linguistic Data Consortium (LDC) https://catalog.ldc.upenn.edu/LDC2013T19.

[3] Note that WSJ was used in Same et al. (2022), but no corpus analysis or comparison was provided.

		MSR			NEG			WSJ	
	Acc.	F1	wF1	Acc.	F1	wF1	Acc.	F1	wF1
UDel	66.86	56.76	64.3	**80.80**	55.45	77.9	63.74	64.23	63.2
ICSI	71.19	64.73	70.4	80.36	64.53	78.6	64.62	64.15	63.4
CNTS	68.59	61.39	67.2	78.68	61.62	76.8	64.31	64.59	64.4
OSU	68.02	60.28	66.6	79.24	57.04	76.5	69.20	69.63	68.9
IS-G	67.05	58.83	65.3	77.34	59.52	75.6	69.15	69.35	69.2
BERT	**71.68**	66.70	**71.4**	77.79	72.87	77.7	80.95	80.93	80.9
RoBERTa	70.91	**67.53**	70.7	**80.80**	**77.29**	**80.7**	**82.61**	**82.70**	**82.6**
Average	69.19	62.32	67.99	79.29	64.05	77.69	70.65	70.80	70.37

Table 3: Overall accuracy (Acc.), macro-averaged F1 (F1), and weighted-macro F1 (wF1) scores of the algorithms depicted in Section 4. For instance, MSR-UDel refers to a C5.0 classifier trained on the MSR corpus, using the feature set mentioned in Greenbacker and McCoy (2009a).

we also report macro-F1 and weighted-macro F1. We argue that different metrics evaluate algorithms from different perspectives and provide us with different meaningful insights. For pragmatic tasks like REG, it makes sense to ask how well an algorithm performs on naturally distributed data which is often imbalanced. For these cases, reporting accuracy and weighted F1 are logical. Furthermore, analogous to other classification tasks, minority categories should not be overlooked. Take as an example the class *description* in the NEG corpus, which occurs only 4%. If a model fails to produce this class, the produced document might sound unnatural. Therefore, it is important to ensure that an algorithm is not over- or under-generating certain classes. Looking into accuracy and macro-F1 together provides insights into such cases.

6.3 Performance of the Models

The overall accuracy of the models, their macro F1, and their weighted-macro F1 are presented in Table 3. We also present the ranking of the models based on these scores in Appendix A.

PLM-based Models. The best-performing models across all corpora and metrics are PLM-based models. In six out of nine rankings, BERT and RoBERTa are ranked as the top two models. The sole exception is NEG, where BERT is the second worst model. The benefit of using PLMs is the largest on the WSJ corpus. For example, RoBERTa improves the macro F1 score from 69.63 (i.e., the performance of the best ML-based model) to 82.70.

ML-based Models. In contrast to the robust performance of the PLM models, the performance of the classic ML models is more corpus-dependent. In the case of MSR and NEG, ICSI is the best-performing model, while in the case of WSJ, it is at the bottom section of the rankings. Another interesting observation is the performance of the UDel models. In terms of accuracy, UDel has the highest performance in NEG, while it has the lowest performance in both MSR and WSJ. In terms of macro-F1 rankings, the NEG UDel model dropped from first to last place, whereas BERT improved from penultimate place to second place. In general, our ML models yielded lower scores than the original models used in the GREC study (Belz et al., 2009). This could be attributed to a variety of factors, including differences in feature engineering and model parameters.

Comparing Different Metrics. Upon comparing average scores across the three metrics, we observe that for MSR and NEG, PLMs are clear winners only when macro-F1 is the metric in question. However, for WSJ, PLMs are winners on all three metrics. This may be because the distribution of categories in WSJ is much more balanced than in the other two corpora.

7 Analysis

To further compare the different models and investigate the impact of the choice of corpus, we conduct (1) a Bayes Factor (BF) analysis to determine whether the accuracy rates reported in Section 6 come from similar or different distributions, (2) a per-class evaluation of predictions to assess the success of each model in predicting individual classes, (3) a correlation analysis to quantify how the evaluation results change with respect to the choice of a corpus, and (4) a feature selection study to check how the importance of each feature changes as a function of the choice of corpus.

		MSR			NEG			WSJ		
Model	Category	P	R	F	P	R	F	P	R	F
Udel	description	55.36	19.38	28.71	0.00	0.00	0.00	60.29	62.95	61.59
	name	72.39	62.21	66.92	76.65	80.32	78.44	60.42	49.44	54.38
	pronoun	64.53	88.51	74.64	84.06	92.14	87.91	71.00	83.44	76.72
ICSI	description	51.69	38.12	43.88	100.00	17.74	30.13	81.92	40.53	54.22
	name	80.33	66.82	72.95	81.85	73.14	77.25	55.12	86.40	67.37
	pronoun	69.41	87.39	77.37	79.05	94.76	86.19	72.17	69.61	70.86
CNTS	description	53.68	31.88	40.00	75.00	14.52	24.33	64.31	63.67	63.30
	name	76.79	61.75	68.45	77.84	72.87	75.27	60.34	66.75	63.38
	pronoun	66.16	88.51	75.72	79.32	92.14	85.25	71.90	62.54	66.89
OSU	description	53.57	28.12	36.88	100.00	4.84	9.23	72.70	56.91	63.84
	name	69.39	68.43	68.91	79.01	72.07	75.38	63.56	73.30	68.08
	pronoun	69.20	81.98	75.05	79.27	95.20	86.51	73.43	80.87	76.97
ISG	description	57.97	25.00	34.93	77.78	11.29	19.72	73.88	63.41	68.25
	name	71.46	65.21	68.19	71.77	79.79	75.57	62.19	76.64	68.66
	pronoun	65.10	84.01	73.36	82.30	84.28	83.28	75.36	67.36	71.14
BERT	description	52.86	46.25	49.33	62.71	59.68	61.16	82.63	79.37	80.97
	name	74.35	72.81	73.57	77.32	75.27	76.28	79.64	82.69	81.14
	pronoun	74.84	79.73	77.21	80.04	82.31	81.16	80.48	80.87	80.67
RoBERTa	description	56.33	55.62	55.97	76.47	62.90	69.02	86.19	77.40	81.56
	name	76.50	64.52	70.00	78.70	80.59	79.63	77.22	89.25	82.80
	pronoun	71.40	82.66	76.62	83.04	83.41	83.22	86.47	81.19	83.75

Table 4: Per-class precision, recall and F1 score of each label. The results report on training seven different algorithms on three corpora for predicting three labels, namely description, name, and pronoun.

7.1 Bayes Factor Analysis

Given that the accuracy scores are provided for all GREC systems in Belz et al. (2009), we chose to focus our analysis on the raw distributions of these scores. Our aim is to determine if there are significant differences between the accuracies of our models by comparing these distributions. We conduct a Bayes Factor analysis with a beta distribution of 0.01 (henceforth: the threshold). This analysis aims to assess, for each pair of accuracies, how strong the evidence is that they come from a common distribution, or from different ones. A difference below the threshold indicates that accuracy rates come from similar distributions; whereas, a difference above the threshold indicates that they come from different distributions, thus signalling that they differ evidentially. We interpret the strength of the evidence in favour of/against similar/different distributions according to Kass and Raftery (1995). Therefore, based on this approach, we expect that the raw accuracy distributions of the best- and worst-performing models for each corpus differ evidentially.

For MSR, the comparison between the best- and worst-performing models, namely BERT and UDel, provides no evidence that their accuracy rates are evidentially different from each other (BF = 1.4). The same holds for NEG, where the comparison of the best (UDel and RoBERTa) and worst (IS-G) models appear to have similar probability distributions; therefore, these models are not evidentially different from each other. Conversely, in the case of WSJ, the BF analysis provides strong evidence that the accuracy distributions of the top-performing models, BERT and RoBERTa, are different from those of the classic ML models.

To summarise, we only observed significant differences in the WSJ-based models; the GREC models show more or less the same accuracy distributions. A reason might be that the aggregated calculation of accuracy loses the specificity of the classes being calculated.

7.2 Per-class Evaluation

As mentioned earlier, the NEG models demonstrate high accuracy (e.g. the highest average accuracy), but we observe a sharp decline in their macro-F1 values. In this analysis, we want to investigate whether the accuracy scores reported in Table 3 truly reflect the success of these algorithms or if they are merely the by-product of over-generating the dominant label or under-generating the less frequent label. Table 4 presents the *per-class* precision, recall, and F1 scores of these models.

Upon comparing the F1 scores for the class *description* across the three corpora, we observe that the WSJ models consistently achieve the highest scores, with all algorithms exceeding an F1 score of 50. In contrast, the F1 scores for both MSR and NEG are considerably lower than those of WSJ. The F1 scores for NEG are particularly low, with two notable instances, `UDel` and `OSU`, scoring 0 and below 10 respectively. The poor prediction of the class description by the classic ML NEG models is likely due to an insufficient number of instances in the training dataset, thereby hindering the proper training of the algorithms. In contrast, the two PLM models demonstrate acceptable performance in predicting the class description (`BERT` = 61.16 & `RoBERTa` = 69.02). This could indicate that pre-trained language models are advantageous where there is a class imbalance.

Another interesting observation concerns the high recall of the “pronoun" prediction in the NEG models. Four of the classic models have a recall of over 92. In the case of `OSU`, for example, the recall is 95, which means that of all the cases that are pronouns, 95% are labelled correctly. This is possibly an indication that pronouns have been over-generated in this system. In the PLM models, the recall is below 84.

In sum, the results of our per-class evaluation show the difficulties that the classic ML-based NEG models had in predicting the class *description*. The MSR models also had poor performance in predicting descriptions, yet they were more successful than NEG. These results tentatively suggest that feature-based classification models need to be trained on an adequate and relatively balanced number of instances to reliably predict all classes. The results of this study suggest that the PLM models are less dependent on the choice of corpus, and therefore predict classes more robustly.

7.3 Correlation Analysis

To quantify how the evaluation results change with respect to corpora, we compute the Spearman correlation coefficient between every pair of corpora, indicating how the rank of the models changes. Table 5 shows the computed coefficients along with the p-values of the tests. It is noteworthy that only the results evaluated by the macro-weighted F1 on MSR and NEG are significantly correlated ($p < .001$).

		acc	F1	wF1
MSR/NEG	r_s	-0.1081	0.9643	0.4643
	p	0.8175	0.0005	0.2939
MSR/WSJ	r_s	0.2857	0.5357	0.4643
	p	0.5345	0.2152	0.2939
NEG/WSJ	r_s	-0.1261	0.5000	-0.0357
	p	0.7876	0.2532	0.9394

Table 5: Spearman correlation coefficient r_s and the p-value between every pair of corpora in terms of accuracy, macro-averaged F1, and weighted F1.

The lack of correlation between the results on MSR/WSJ and those on NEG/WSJ suggests that using a corpus of a different genre could greatly influence the ranking of the models and, therefore, make the conclusions difficult to generalise. Additionally, these results are in line with the fact that MSR and NEG are from the same source, both being the introductory part of Wikipedia articles, and a higher correlation is to be expected. Also, we may conclude that macro-averaged F1 is a more reliable evaluation metric (see the discussions in Section 6, Section 7.1, and Section 7.2).

7.4 Feature Selection Study

We performed a feature importance analysis to check whether the contribution of linguistic factors changes depending on the choice of the corpus. We used XGBoost from the family of Gradient Boosting trees (Chen and Guestrin, 2016) and then computed the permutated variable importance for each model. Data were analysed in two ways: firstly, we used the complete dataset, as outlined in Section 5; secondly, we excluded first-mention REs to concentrate only on subsequent mentions. Considering that the choice of a referent’ first mention is less context-dependent, we only report on the latter dataset below:

As expected, the ranking of feature importance varies across different corpora. However, a substantial overlap is observed when considering the most important features across the three corpora. An example is the semantic category of the REs that is used in various MSR and WSJ REG models.[4] In the case of MSR, the REs belong to five semantic categories: human, city, country, river, and mountain. In the case of WSJ, the REs are annotated for a wide

[4] Only human referents are annotated in NEG; therefore, this feature is not applicable.

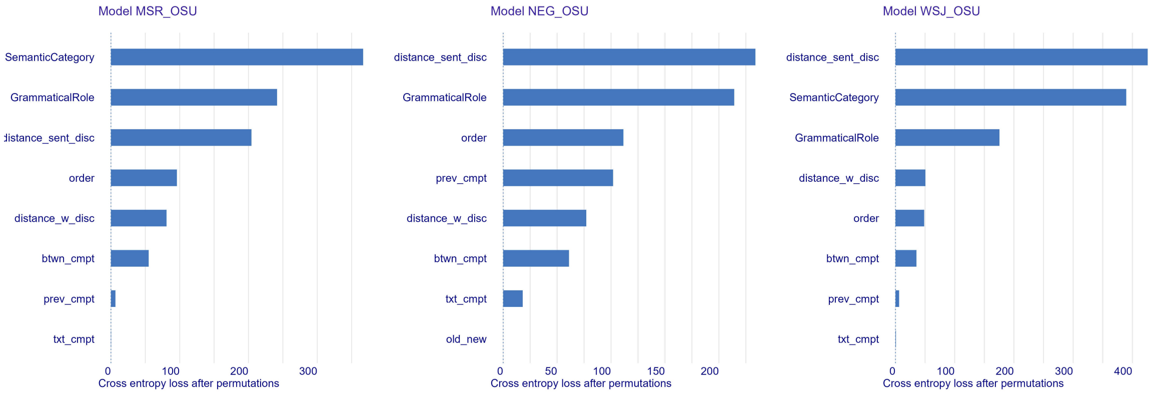

Figure 2: Different rankings of the features in MSR, NEG, and WSJ OSU models.

range of categories including human, city, country, organisation, objects, etc. Notably, in every model that employs semantic category information, this feature has either the highest or second-highest importance ranking. A plausible explanation could be that humans use different referencing strategies to refer to different categories of referents.

In addition to the semantic category, the grammatical role of the RE and the categorical sentential distance to the antecedent consistently have a high importance ranking. The grammatical role marks the distinction between subject, object, and determiner roles. The categorical distance in the number of sentences provides information on how far an RE is to its nearest coreferential antecedent. For instance, whether they are both in the same sentence or are separated by one or more sentences. Figure 2 illustrates the importance rankings of the OSU features in the three corpora. Other importance ranking graphs are available in Appendix C. For a comprehensive description of all features employed in classic ML models and the feature importance analysis, refer to Same and van Deemter (2020).

8 Discussion

In this paper, we have conducted a series of reproductions, evaluations, and analyses to check whether the conclusions of GREC are still true after 15 years. Below, we summarise and discuss our findings in accordance with our three research questions in Section 3. We also report our post-hoc observations on the choice of evaluation metric.

Performance of REG Algorithms. To answer research question $\mathcal{R}_2$, we extended the GREC by introducing a corpus of a different genre, WSJ, and two pre-trained (PLM-based) REG models. We found that, on MSR, PLM-based and ML-based models perform similarly, as confirmed by both the BF and per-class analyses. With regards to NEG, PLM-based and ML-based models have similar accuracy scores, as confirmed by the BF analysis, but there are large differences when micro-F1 is used, as confirmed by the per-class evaluation (i.e., ML-based models have difficulty predicting descriptions). On WSJ, PLM-based models are the clear winners.

These results suggest that, in terms of explanatory power, PLM-based models have good performance and good "direct support", i.e., a good ability to generalise to different contexts (see van Deemter (2023) for further discussion). Whether they have good "indirect support" (e.g., whether their predictions are in line with linguistic theories) needs to be investigated in further probing studies.

Impact of the Choice of Corpus. As our evaluations and analyses demonstrate, the choice of corpus plays a crucial role in assessing REG algorithms. This role is twofold. Firstly, the choice of corpus strongly influences the evaluation results, pertaining to the research question $\mathcal{R}_1$. Secondly, in addition to the score differences discussed in Section 6, we found that: (1) the difference between PLM-based and ML-based models on WSJ is larger (and evidentially different) than on MSR and NEG models (as evidenced by the BF analysis); (2) the correlations of the evaluation results between WSJ and both MSR and NEG are not significant.

For $\mathcal{R}_3$, we conducted feature selection analyses across the three corpora, discovering that the importance of the features ranks differently for each corpus. This suggests that when investigating the "indirect support" for a model, one needs to aggregate findings from multiple corpora with different genres.

The Use of Evaluation Metrics. As we discussed in Section 6.2, different metrics evaluate different aspects of a model. This was further ascertained by the inconsistency of the BF analysis and per-class analysis. One lesson we have learned is that it is not enough to report or do analyses on a single metric. Another lesson is that the evaluation results by macro-F1 are more reliable than other metrics because (1) they are consistent across corpora with similar genres (i.e., MSR and NEG; see the Correlation analysis results); (2) the differences identified by using macro-F1 can be confirmed by the per-class evaluation.

9 Conclusion

We are now in a position to address the question that we raised in the Introduction: Can the conclusions from the GREC shared tasks still be trusted? By examining a wider class of corpora, models, and evaluation metrics than before, we found that the answer to this question is essentially negative since the GREC conclusions are prone to drastic change once a different corpus or a different metric is employed.

Perhaps this should come as no surprise. According to a widely accepted view of scientific progress (e.g., Jaynes (2002); applied to NLP in (van Deemter, 2023)), theories should be updated again and again in light of new data (i.e., indirect Support), and when new models are proposed, the plausibility of existing models should be compared against the plausibility of these new models (as well as pre-existing ones). New metrics deserve a place in this story as well, even though they are often overlooked. In other words, what we have seen in the present study is nothing more than science in progress – something we are bound to see more of as the enterprise called NLP-as-Science matures.

Ethics Statement: Regarding potential biases, in addition to the biases present in text-based datasets, biases can also be introduced by the pre-trained language models (Bender et al., 2021) used in this work. In other words, the REG algorithms we developed in this study may make different predictions with respect to different genders, for instance. In the future, we plan to investigate this phenomenon and find ways to mitigate it.

Supplementary Materials Availability Statement: All associated data, source code, output files, scripts, documentation, and other relevant material to this paper are publicly available and can be accessed on our GitHub repository: `https://github.com/fsame/REG_GREC-WSJ`, DOI: 10.5281/zenodo.8182689.

Acknowledgements: We thank the anonymous reviewers for their helpful comments. Fahime Same is supported by the German Research Foundation (DFG)– Project-ID 281511265 – SFB 1252 "Prominence in Language".

References

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2008. The GREC challenge 2008: Overview and evaluation results. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 183–193, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2009. Generating referring expressions in context: The GREC task evaluation challenges. In *Empirical methods in natural language generation*, pages 294–327. Springer.

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*.

Bernd Bohnet. 2008. IS-G: The comparison of different learning techniques for the selection of the main subject references. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 192–193, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Meng Cao and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. 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 3163–3172, Hong Kong, China. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. 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 552–562, Hong Kong, China. Association for Computational Linguistics.

Guanyi Chen, Fahime Same, and Kees van Deemter. 2021. What can neural referential form selectors learn? In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 154–166, Aberdeen, Scotland, UK. Association for Computational Linguistics.

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

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Alves. 2020. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2261–2272, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2007. Timbl: Tilburg memory-based learner.

Benoit Favre and Bernd Bohnet. 2009. ICSI-CRF: The generation of references to the main subject and named entities using conditional random fields. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 99–100, Suntec, Singapore. Association for Computational Linguistics.

Charles Greenbacker and Kathleen McCoy. 2009a. UDel: Generating referring expressions guided by psycholinguistc findings. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 101–102, Suntec, Singapore. Association for Computational Linguistics.

Charles F Greenbacker and Kathleen F McCoy. 2009b. Feature selection for reference generation as informed by psycholinguistic research. In *Proceedings of the CogSci 2009 Workshop on Production of Referring Expressions (PRE-Cogsci 2009)*.

Samir Gupta and Sivaji Bandopadhyay. 2009. Junlgmsr: A machine learning approach of main subject reference selection with rule based improvement. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 103–104. Association for Computational Linguistics.

Iris Hendrickx, Walter Daelemans, Kim Luyckx, Roser Morante, and Vincent Van Asch. 2008. CNTS: Memory-based learning of generating repeated references. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 194–95, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Renate Henschel, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 306–312. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Emily Jamison and Dennis Mehay. 2008. OSU-2: Generating referring expressions with a maximum entropy classifier. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 196–197, Salt Fork, Ohio, USA. Association for Computational Linguistics.

E.T. Jaynes. 2002. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Robert E Kass and Adrian E Raftery. 1995. Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Andrej A Kibrik, Mariya V Khudyakova, Grigory B Dobrov, Anastasia Linnik, and Dmitrij A Zalmanov. 2016. Referential choice: Predictability and its limits. *Frontiers in psychology*, 7(1429).

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Max Kuhn, Steve Weston, Mark Culp, Nathan Coulter, and Ross Quinlan. 2018. Package 'c50'.

Kathleen E. McCoy and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *The Relation of Discourse/Dialogue Structure and Reference*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Ehud Reiter. 2017. A commercial perspective on reference. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 134–138, Santiago de Compostela, Spain. Association for Computational Linguistics.

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.

Fahime Same and Kees van Deemter. 2020. A linguistic perspective on reference: Choosing a feature set for generating referring expressions in context. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4575–4586, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Kees van Deemter. 2023. Dimensions of explanatory value in nlp models. *Computational Linguistics*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

A Ranking of the Models

Accuracy-based Ranking

MSR: BERT > ICSI > RoBERTa > CNTS > OSU > IS-G > UDel
NEG: UDel = RoBERTa > ICSI > OSU > CNTS > BERT > IS-G
WSJ: RoBERTa > BERT > OSU > IS-G > ICSI > CNTS > UDel

Macro-F1 Ranking

MSR: RoBERTa > BERT > ICSI > CNTS > OSU > IS-G > UDel
NEG: RoBERTa > BERT > ICSI > CNTS > IS-G > OSU > UDel
WSJ: RoBERTa > BERT > OSU > IS-G > CNTS > UDel > ICSI

Macro-weighted F1 Ranking

MSR: BERT > RoBERTa > ICSI > CNTS > OSU > IS-G > UDel
NEG: RoBERTa > ICSI > UDel > BERT > CNTS > OSU > IS-G
WSJ: RoBERTa > BERT > IS-G > OSU > CNTS > ICSI > UDel

B Implementation Details for ML-based Models

The R programming language was used mostly for running the classic ML models. The specification of the models can be found below:

Conditional Random Field [CRF]. The R Package CRF (https://cran.r-project.org/web/packages/crfsuite/) was used to train these models. The iterations are set to 3000, and the learning method is Stochastic Gradient Descent with L2 regularization term (l2sgd).

Decision Tree [C5.0]. The R Package C5.0 (Kuhn et al., 2018) was used to build the decision trees. The number of boosting iterations (trials) is set to 3, and the splitting criterion is information gain (entropy).

Memory-Based Learning [MBL]. As mentioned before, we implemented the k-Nearest Neighbors [KNN] algorithm instead of MBL. The R package caret with the method KNN was used to implement this model.

Maximum Entropy [MaxEnt]. The multinom algorithm from the nnet R package was used to implement this model.

Multi-Layer Perceptron [MLP]. The Keras package was used to implement MLP. The model consists of two hidden layers with 16 and 8 units, respectively. The hidden layers use the rectified linear activation function (ReLU), and the output layer uses the Sigmoid activation function. The model is fitted for 50 training epochs. In addition, 50 samples (batch size) are propagated through the network.

eXtreme Gradient Boosting [XGBoost]. XGBoost was used for the feature selection experiments. We used the R packages xgboost and DALEXtra for the analysis. We set the learning rate to 0.05, the minimum split loss to 0.01, the maximum depth of a tree to 5, and the sub-sample ratio of the training instances to 0.5.

C Feature Importance Rankings

The graphs in Figure 3 show the rankings across MSR, WSJ, and WSJ. A maximum number of eight features is depicted in the graphs.

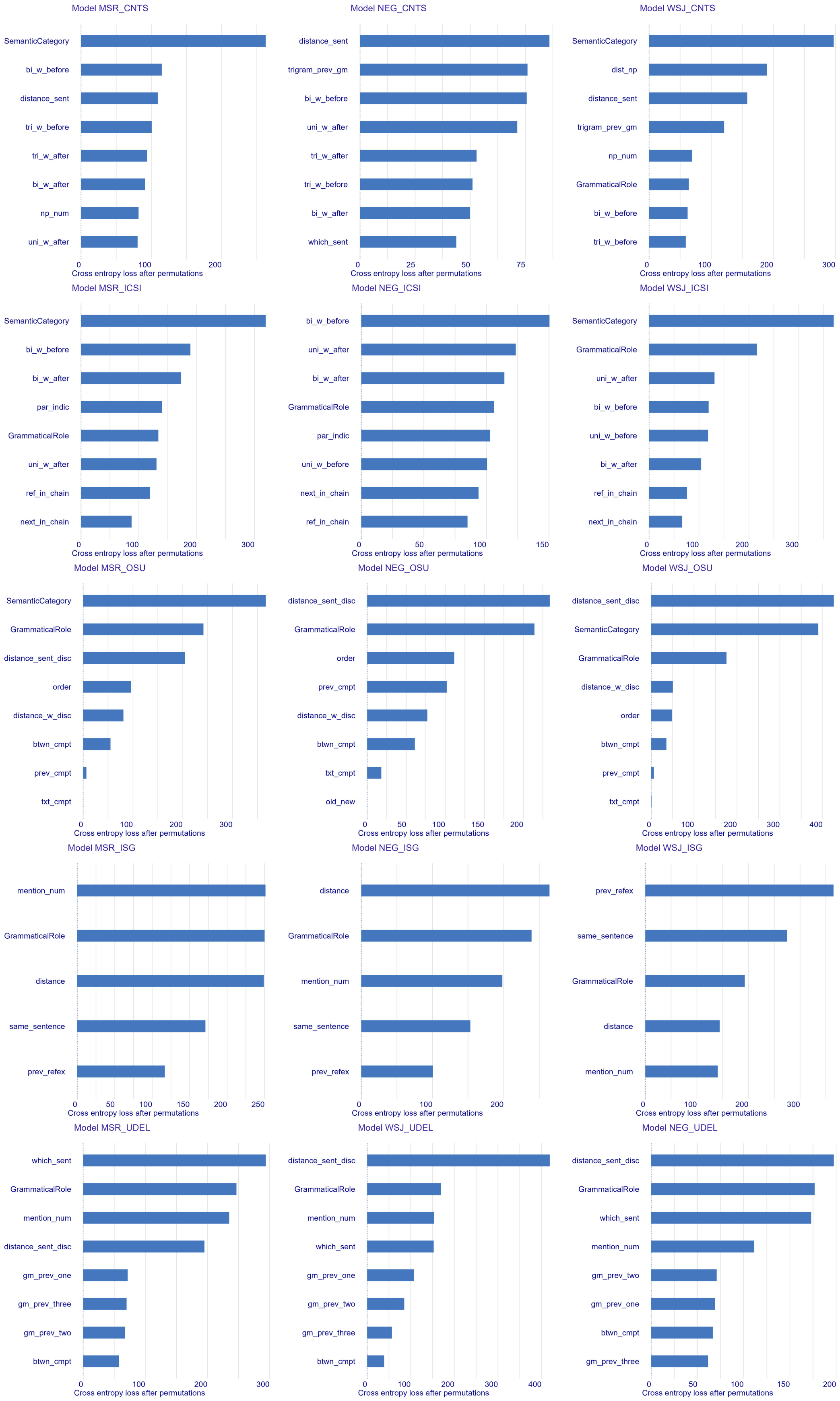

Figure 3: Importance ranking of the features in MSR, NEG, and WSJ models.