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

Fahime (Fafa) Same[1] Guanyi Chen[2] Kees van Deemter[3]

f.same@uni-koeln.de g.chen@ccnu.edu.cn c.j.vandeemter@uu.nl

[1]University of Cologne [2]Central China Normal University [3]Utrecht University

INLG 2023

Introduction and Task Definition

Introduction

NLP research pursues diverse goals:

1. Developing new models and building practical applications

2. Constructing computational models to explain human language use

Introduction

NLP research pursues diverse goals:

❶ Developing new models and building practical applications

❷ Constructing computational models to explain human language use → NLP as Science

NLP-as-Science requires us to question
how broadly NLP research findings generalise across various dimensions.

Referring Expression Generation in Context (REG-in-context)

Given an intended referent and a discourse context, how do we generate appropriate referring expressions (REs) to refer to the referent at different points in the discourse? (Belz and Varges, 2007)

1. What form (e.g., pronoun, proper name) should the RE take?
2. What content should be included in the RE?

Referring Expression Generation in Context (REG-in-context)

Given an intended referent and a discourse context, how do we generate appropriate referring expressions (REs) to refer to the referent at different points in the discourse? (Belz and Varges, 2007)

1. What form (e.g., pronoun, proper name) should the RE take?
2. What content should be included in the RE?

REFERENT: HOMER SIMPSON

> **Homer Jay Simpson** (born May 12 1956) is the main protagonist and one of the five main characters of The Simpsons series (or show). **He** is the spouse of Marge Simpson and father of Bart, Lisa and Maggie Simpson. **Homer** is overweight (said to be 240 pounds), lazy, and often ignorant to the world around **him**.

Referring Expression Generation in Context (REG-in-context)

Given an intended referent and a discourse context, how do we generate appropriate referring expressions (REs) to refer to the referent at different points in the discourse? (Belz and Varges, 2007)

1. What **form** (e.g., pronoun, proper name) should the RE take? → Our focus in this talk
2. What **content** should be included in the RE?

REFERENT: HOMER SIMPSON

> **Homer Jay Simpson** (born May 12 1956) is the main protagonist and one of the five main characters of The Simpsons series (or show). **He** is the spouse of Marge Simpson and father of Bart, Lisa and Maggie Simpson. **Homer** is overweight (said to be 240 pounds), lazy, and often ignorant to the world around **him**.

A Bit of History: GREC Shared Tasks (Belz et al., 2009)

Generating **R**eferring **E**xpressions in **C**ontext (GREC): A series of shared tasks (2008-2009)

Task definition: How to generate appropriate references to an entity in the context of a piece of discourse longer than a sentence?
Another goal: What kind of information is useful for making choices between different kinds of referring expressions in context?

A Bit of History: GREC Shared Tasks (Belz et al., 2009)

Generating **R**eferring **E**xpressions in **C**ontext (GREC): A series of shared tasks (2008-2009)

Task definition: How to generate appropriate references to an entity in the context of a piece of discourse longer than a sentence?
Another goal: What kind of information is useful for making choices between different kinds of referring expressions in context?

Corpora: introductory sections of Wikipedia articles

1. GREC-2.0 ($\sim$ 2000 docs in 5 domains)
2. GREC-People ($\sim$ 1000 docs about people)

Models: Various feature-based and rule-based models.

Study Outline

In This Talk:

- We replicate the GREC study AND
- We then extend it along different dimensions.
  1. We include a corpus from a different genre:
     - Wall Street Journal (**WSJ**) portion of OntoNotes (Weischedel et al., 2013)
  2. In addition to the classic ML models, we fine-tune Pre-trained Language Models (PLMs):
     - BERT
     - RoBERTa
  3. We employ diverse evaluation methods:
     - Accuracy, macro-F1, weighted-macro F1
     - Per-class evaluation
     - Bayes Factor analysis
     - Correlation analysis
     - Feature Selection experiments

Our Goals

1. **Choice of Corpus:** What impact does the choice of corpus have on the performance of REG algorithms?
2. **Model Comparison:** How does the explanatory power of PLM-based REG models compare to classic ML-based models?
3. **Evaluation Metrics:** What insights do different evaluation metrics provide about the performance of the models?
4. **Linguistic Features:** Does the importance ranking of linguistic factors vary when using different corpora?

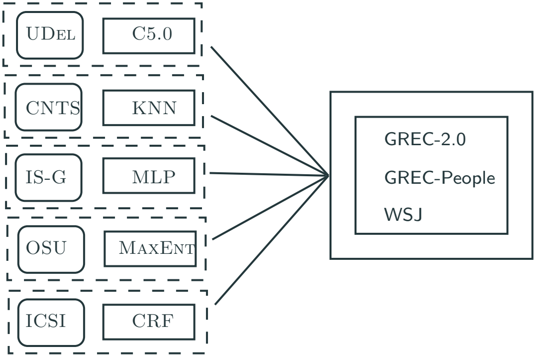
Our Goals

1. **Choice of Corpus** What impact does the choice of corpus have on the performance of REG algorithms?
2. **Model Comparison** How does the explanatory power of PLM-based REG models compare to classic ML-based models?
3. **Evaluation Metrics** What insights do different evaluation metrics provide about the performance of the models?
4. **Linguistic Features** Does the importance ranking of linguistic factors vary when using different corpora?

Methodology and Data Overview

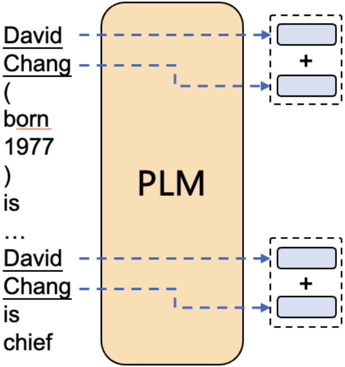
REG Algorithms

Feature-based ML models



5 different ML models, based on those submitted to GREC.

PLM-based models



Same paradigm as Cunha et al. (2020)

2 PLMs (BERT & RoBERTa)

REG Corpora & RE Classes

Corpora used: 2 Wikipedia corpora (GREC-2.0 & GREC-Peope) and 1 news corpus (WSJ)

	GREC-2.0	GREC-People	WSJ
words/doc	148	129	530
sentences/doc	7.1	5.8	25
referents/doc	1	2.6	15
total number of REs	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%	-

REG Corpora & RE Classes

Corpora used: 2 Wikipedia corpora (GREC-2.0 & GREC-Peope) and 1 news corpus (WSJ)

	GREC-2.0	GREC-People	WSJ
words/doc	148	129	530
sentences/doc	7.1	5.8	25
referents/doc	1	2.6	15
total number of REs	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%	-

REG Corpora & RE Classes

Corpora used: 2 Wikipedia corpora (GREC-2.0 & GREC-Peope) and 1 news corpus (WSJ)

	GREC-2.0	GREC-People	WSJ
words/doc	148	129	530
sentences/doc	7.1	5.8	25
referents/doc	1	2.6	15
total number of REs	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%	-

REG Corpora & RE Classes

Corpora used: 2 Wikipedia corpora (GREC-2.0 & GREC-Peope) and 1 news corpus (WSJ)

	GREC-2.0	GREC-People	WSJ
words/doc	148	129	530
sentences/doc	7.1	5.8	25
referents/doc	1	2.6	15
total number of REs	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%	-

Referring Expression classes considered:

1. Proper name (e.g., Lewis Hamilton)
2. Description (e.g., the F1 driver)
3. Pronoun (e.g., he)

Model Performance

Performance of the Models

	GREC-2.0			GREC-People			WSJ		
	Acc.	macroF1	wF1	Acc.	macroF1	wF1	Acc.	macroF1	wF1
UDel	66.86	56.76	64.3	**80.80**	55.45	77.9	63.74	64.23	63.2
ICSI	<u>71.19</u>	64.73	70.4	80.36	64.53	<u>78.6</u>	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**	<u>66.70</u>	**71.4**	77.79	<u>72.87</u>	77.7	<u>80.95</u>	<u>80.93</u>	<u>80.9</u>
RoBERTa	70.91	**67.53**	<u>70.7</u>	**80.80**	**77.29**	**80.7**	**82.61**	**82.70**	**82.6**

❶ PLMs perform best across all corpora, but the lead isn't always large.

Performance of the Models

	GREC-2.0			GREC-People			WSJ		
	Acc.	macroF1	wF1	Acc.	macroF1	wF1	Acc.	macroF1	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**

❶ PLMs perform best across all corpora, but the lead isn't always large.

❷ The advantage of using PLMs is most pronounced with WSJ.

Performance of the Models

	GREC-2.0			GREC-People			WSJ		
	Acc.	macroF1	wF1	Acc.	macroF1	wF1	Acc.	macroF1	wF1
UDel	$\boxed{66.86}$	56.76	64.3	$\boxed{80.80}$	55.45	77.9	$\boxed{63.74}$	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**

❶ PLMs perform best across all corpora, but the lead isn't always large.

❷ The advantage of using PLMs is most pronounced with WSJ.

❸ ML-based models are more corpus-dependent.

Performance of the Models

	GREC-2.0			GREC-People			WSJ		
	Acc.	macroF1	wF1	Acc.	macroF1	wF1	Acc.	macroF1	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**

1. PLMs perform best across all corpora, but the lead isn't always large.
2. The advantage of using PLMs is most pronounced with WSJ.
3. ML-based models are more corpus-dependent.
4. Only with WSJ are PLMs the clear winners across all metrics.

Evaluation

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes.

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes. $\boxed{\textbf{Evaluation Metrics}}$

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes. **Evaluation Metrics**
2. **Bayes Factor Analysis:** To determine whether pairs of raw accuracies come from similar or different distributions (i.e., whether they differ significantly).

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes. **Evaluation Metrics**
2. **Bayes Factor Analysis:** To determine whether pairs of raw accuracies come from similar or different distributions (i.e., whether they differ significantly). **Model Comparison**

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes. **Evaluation Metrics**
2. **Bayes Factor Analysis:** To determine whether pairs of raw accuracies come from similar or different distributions (i.e., whether they differ significantly). **Model Comparison**
3. **Correlation Analysis:** To quantify how evaluation results vary with respect to corpora (using the Spearman correlation coefficient)

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes. **Evaluation Metrics**
2. **Bayes Factor Analysis:** To determine whether pairs of raw accuracies come from similar or different distributions (i.e., whether they differ significantly). **Model Comparison**
3. **Correlation Analysis:** To quantify how evaluation results vary with respect to corpora (using the Spearman correlation coefficient) **Choice of Corpus**

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes. **Evaluation Metrics**
2. **Bayes Factor Analysis:** To determine whether pairs of raw accuracies come from similar or different distributions (i.e., whether they differ significantly). **Model Comparison**
3. **Correlation Analysis:** To quantify how evaluation results vary with respect to corpora (using the Spearman correlation coefficient) **Choice of Corpus**
4. **Feature Selection analysis:** To determine whether the contribution of linguistic factors varies depending on the choice of corpus.

Evaluations

1. **Per-class Analysis:** To determine the success of each model in predicting individual classes. **Evaluation Metrics**
2. **Bayes Factor Analysis:** To determine whether pairs of raw accuracies come from similar or different distributions (i.e., whether they differ significantly). **Model Comparison**
3. **Correlation Analysis:** To quantify how evaluation results vary with respect to corpora (using the Spearman correlation coefficient) **Choice of Corpus**
4. **Feature Selection analysis:** To determine whether the contribution of linguistic factors varies depending on the choice of corpus. **Linguistic Features**

Evaluation Results: Per-class Analysis

Model	Category	GREC-2.0 Recall	GREC-2.0 macroF1	GREC-People Recall	GREC-People macroF1	wsj Recall	wsj macroF1
`Udel`	description	19.38	28.71	0.00	0.00	62.95	61.59
	pronoun	88.51	74.64	92.14	87.91	83.44	76.72
`RoBERTa`	description	55.62	55.97	62.90	69.02	77.40	81.56
	pronoun	82.66	76.62	83.41	83.22	81.19	83.75

Class: Description

Classic ML models

- Performance: Low on the GREC corpora & above 0.5 on WSJ.
- Issue: Sensitive to class imbalance.

PLMs

- Performance: Above 0.5 across all corpora.
- Advantage: Effectively handle class imbalance.

Evaluation Results: Per-class Analysis

Model	Category	GREC-2.0		GREC-People		wsj	
		Recall	macroF1	Recall	macroF1	Recall	macroF1
Udel	description	19.38	28.71	0.00	0.00	62.95	61.59
	pronoun	88.51	74.64	92.14	87.91	83.44	76.72
RoBERTa	description	55.62	55.97	62.90	69.02	77.40	81.56
	pronoun	82.66	76.62	83.41	83.22	81.19	83.75

Class: Description

Classic ML models

- Performance: Low on the GREC corpora & above 0.5 on WSJ.
- Issue: Sensitive to class imbalance.

PLMs

- Performance: Above 0.5 across all corpora.
- Advantage: Effectively handle class imbalance.

Evaluation Results: Per-class Analysis

		GREC-2.0		GREC-People		wsj	
Model	Category	Recall	macroF1	Recall	macroF1	Recall	macroF1
`Udel`	description	19.38	28.71	0.00	0.00	62.95	61.59
	pronoun	88.51	74.64	92.14	87.91	83.44	76.72
`RoBERTa`	description	55.62	55.97	62.90	69.02	77.40	81.56
	pronoun	82.66	76.62	83.41	83.22	81.19	83.75

Class: Description

Classic ML models

- Performance: Low on the GREC corpora & above 0.5 on WSJ.
- Issue: Sensitive to class imbalance.

PLMs

- Performance: Above 0.5 across all corpora.
- Advantage: Effectively handle class imbalance.

Class: Pronoun

Classic ML models

- Performance: Very high recall on GREC-People (above 0.9 in most cases)
- Issue: Tendency to over-generate pronouns.

Evaluation Results: Per-class Analysis

Model	Category	GREC-2.0		GREC-People		wsj	
		Recall	macroF1	Recall	macroF1	Recall	macroF1
`Udel`	description	19.38	28.71	0.00	0.00	62.95	61.59
	pronoun	88.51	74.64	$\boxed{92.14}$	87.91	83.44	76.72
`RoBERTa`	description	55.62	55.97	62.90	69.02	77.40	81.56
	pronoun	82.66	76.62	$\boxed{83.41}$	83.22	81.19	83.75

Class: Description

Classic ML models

- Performance: Low on the GREC corpora & above 0.5 on WSJ.
- Issue: Sensitive to class imbalance.

PLMs

- Performance: Above 0.5 across all corpora.
- Advantage: Effectively handle class imbalance.

Class: Pronoun

Classic ML models

- Performance: Very high recall on GREC-People (above 0.9 in most cases)
- Issue: Tendency to over-generate pronouns.

Evaluation Results (Cont.)

Bayes Factor Analysis:

1. **GREC Corpora:** No evidential difference between PLMs and classic ML models.
2. **WSJ:** PLMs differ evidentially from classic MLs (strong support).

Evaluation Results (Cont.)

Bayes Factor Analysis:

1. **GREC Corpora:** No evidential difference between PLMs and classic ML models.
2. **WSJ:** PLMs differ evidentially from classic MLs (strong support).

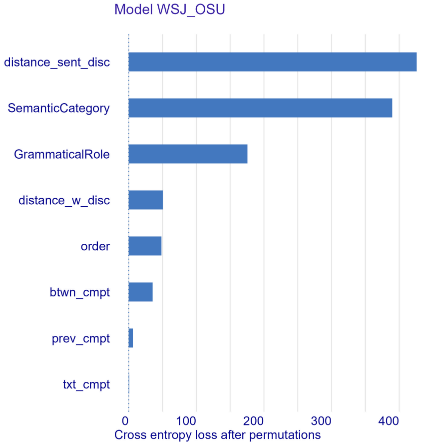
Correlation Analysis:

1. Only the macro-weighted F1 scores on the two GREC corpora are significantly correlated ($p < .001$).
2. Expected correlation between the two GREC corpora, as they share the same genre.
3. No correlation between either of the GREC corpora and WSJ $\rightarrow$ different corpus genres can significantly influence model rankings, making conclusions less generalisable.

Evaluation Results (Cont.)

Feature Selection Analysis

1. Method: Excluded first-mention referring expressions and computed the permutated variable importance for each model trained using XGBoost.
2. Feature rankings vary across corpora, but there's a significant overlap when considering the most important features.
3. Most important features include: Semantic category, grammatical role, and sentential distance.



Discussion and Final Thoughts

Discussion

1. **Model Comparison**
   - PLMs exhibit good performance and can generalise across different contexts.
   - Unlike classic ML models, PLMs are less influenced by the choice of corpus, leading to more robust class predictions.

Discussion

❶ **Model Comparison**

- PLMs exhibit good performance and can generalise across different contexts.
- Unlike classic ML models, PLMs are less influenced by the choice of corpus, leading to more robust class predictions.

❷ **Choice of Corpus**

- The choice of corpus strongly impacts the evaluation outcomes.
- The performance difference between PLMs and classic ML models is more pronounced on WSJ than on MSR and NEG.
- Correlation is observed exclusively when the corpora share the same genre.

Discussion

❶ Model Comparison

- PLMs exhibit good performance and can generalise across different contexts.
- Unlike classic ML models, PLMs are less influenced by the choice of corpus, leading to more robust class predictions.

❷ Choice of Corpus

- The choice of corpus strongly impacts the evaluation outcomes.
- The performance difference between PLMs and classic ML models is more pronounced on WSJ than on MSR and NEG.
- Correlation is observed exclusively when the corpora share the same genre.

❸ Evaluation Metrics

- Different metrics assess different aspects of a model.
- Evaluation outcomes based on macro-F1 are deemed more reliable than those based on accuracy.

Discussion

❶ **Model Comparison**

- PLMs exhibit good performance and can generalise across different contexts.
- Unlike classic ML models, PLMs are less influenced by the choice of corpus, leading to more robust class predictions.

❷ **Choice of Corpus**

- The choice of corpus strongly impacts the evaluation outcomes.
- The performance difference between PLMs and classic ML models is more pronounced on WSJ than on MSR and NEG.
- Correlation is observed exclusively when the corpora share the same genre.

❸ **Evaluation Metrics**

- Different metrics assess different aspects of a model.
- Evaluation outcomes based on macro-F1 are deemed more reliable than those based on accuracy.

❹ **Linguistic Features**

- The importance of features varies for each corpus; however, there is a considerable overlap for linguistically-informed features, such as grammatical role and recency.

Discussion

1. Can earlier REG models withstand the test of time?
   - After examining a range of corpora, models, and metrics, the answer is essentially negative.
   - Earlier models are prone to significant changes once new corpora and metrics are employed.
2. Why is NLP-as-Science as crucial as application-oriented NLP?
   - Theories and practices need to be updated in light of new data.
   - It is essential to evaluate the validity of existing models against new ones to ensure continuous improvement and progress.
   - Metrics, often overlooked, are crucial for progress.
   - Our study: A snapshot of science in progress.

Final Thoughts

1. Both NLP-as-Science and application-oriented NLP have a stake in generalisability
   - NLP-as-Science aims to learn general lessons about language.
   - Application-oriented NLP aims to build software that's versatile across multiple applications.
2. However, our study suggests that results can be heavily influenced by one's choice of corpora and metrics.
3. What happens when the task is more complex?
4. What are the implications for our field of research?

Many Thanks!

Code: `https://github.com/fsame/REG_GREC-WSJ`

References i

Anja Belz and Sebastian Varges. Generation of repeated references to discourse entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 9–16. Association for Computational Linguistics, 2007.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. Generating referring expressions in context: The grec task evaluation challenges. In *Conference of the European Association for Computational Linguistics*, pages 294–327. Springer, 2009.

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Alves. 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), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.205. URL `https://aclanthology.org/2020.coling-main.205`.

References ii

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