A Linguistic Perspective on Reference: Choosing a Feature Set for Generating Referring Expressions in Context

Fahime Same
University of Cologne
f.same@uni-koeln.de

Kees van Deemter
Utrecht University
c.j.vandeemter@uu.nl

Abstract

This paper reports on a structured evaluation of feature-based Machine Learning algorithms for selecting the form of a referring expression in discourse context. Based on this evaluation, we selected seven feature sets from the literature, amounting to 65 distinct linguistic features. The features were then grouped into 9 broad classes. After building Random Forest models, we used Feature Importance Ranking and Sequential Forward Search methods to assess the "importance" of the features. Combining the results of the two methods, we propose a consensus feature set. The 6 features in our consensus set come from 4 different classes, namely grammatical role, inherent features of the referent, antecedent form and recency.

1 Introduction

Various studies have raised the question of which factors play a role in the choice of referring expressions. One of the main ideas in this tradition (henceforth, the linguistic tradition) is that there is a direct relationship between the "prominence" (in a broad sense) of a referent at a given point in the discourse, and the form used to refer to it. If a referent is prominent, a short anaphoric form (e.g. a pronoun) suffices; if it is less prominent, longer forms with more semantic content are used. Prominence has been argued to be influenced by various factors such as recency and frequency of mention (Ariel, 1990), grammatical function (Brennan, 1995), distance (McCoy and Strube, 1999), animacy (Fukumura and van Gompel, 2011), and competition between the referents (Arnold and Griffin, 2007).

Reference production is also one of the most-studied topics in Natural Language Generation (Gatt and Krahmer, 2018), where it is known as Referring Expression Generation [REG] (Krahmer and van Deemter, 2019). A key part of the REG problem is deciding which form (e.g., proper name, definite description or pronoun) to employ to refer to a referent at a given point in the discourse. Henceforth, we call this task Selection of Referential Form (SRF). SRF models come in many shapes and forms, with feature-based Machine Learning (ML) models playing a dominant role. However, the feature sets employed by these models can differ considerably from one model to the next, and although features akin to the ones employed in the linguistic tradition are often used, other types of features, which are harder to interpret linguistically, are frequent as well.

Our aim in this paper is to examine feature-based SRF models from a linguistic perspective. We will conduct a systematic evaluation of these models, asking what features make them work best. Having done this, we propose a "consensus" feature set. Finally, we compare the features in our consensus feature set against the factors considered to be important in the linguistic tradition.

An important question in any systematic evaluation is how the objects of study (in our case, SRF feature sets) are selected. We have proceeded as follows:

- We selected all SRF algorithms submitted to GREC (Belz et al., 2010) and extracted the feature sets used by these algorithms. GREC was a Shared Task Evaluation task that still forms a natural starting point, because it attracted all the main SRF algorithms that existed at the time.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4575–4586
Barcelona, Spain (Online), December 8-13, 2020

- We additionally selected two other feature sets from the papers archived in the ACL anthology. The selection method will be detailed in subsection 3.1.

- We re-implemented the features that were obtained, following the method detailed in subsection 3.2.

Note that our systematic evaluation does not include Deep Learning methods (e.g., Castro Ferreira et al. (2018); Cao and Cheung (2019)) since, at the current state of the art, these do not yet offer much opportunity for linguistic interpretation. Our focus is on interpretable linguistic features.

To perform our evaluation, two further choices need to be explained and motivated: Since the SRF task as such can be defined in different ways, we had to define an exact task, and we had to specify a corpus. These two choices will be further explained in section 3. Section 4 details the feature selection experiments and propose a consensus feature set. The paper concludes with a discussion of the extent to which the features are linguistically interpretable.

2 Related Work

Data-driven models employ different features to make choices similar to those made by humans. For instance, Greenbacker and McCoy (2009a) used linguistically informed features such as recency, subjecthood, parallelism and ambiguity. In a comprehensive study, Kibrik et al. (2016) argued that reference generation is a multifactorial process governed by various linguistically motivated features. Hendrickx et al. (2008), on the other hand, focused more on existing patterns in the text and less on the use of linguistic categories. In a more recent study, Castro Ferreira et al. (2016) used features marking the syntactic position, recency and referential status of the referents to model the referential choice variations.

In one of the few studies targeting feature selection, Greenbacker and McCoy (2009b) surveyed the psycholinguistics literature to decide on feature sets for their SRF study. They implemented several linguistically informed rules in their prototyping system, and "examined the incorrect classifications which resulted in an attempt to discover which other factors suggested by psycholinguistic research could explain the patterns they observed." Additionally, they incorporated some of the features used in Hendrickx et al. (2008) and created an extensive set of features which they felt had an impact on the referential choice. Afterwards, they trained C5.0 decision trees on 5 subsets of these features. The interesting result of their study is that the maximum number of features does not necessarily lead to best results. The problem with their study is that firstly, they do not propose any explanation on how they have selected different subsets. Secondly, their feature selection strategy is subjective, since it mostly concerns with the features they find important, and has less to do with the features of other models. Lastly, they have provided no linguistic explanation for their best performing system. Kibrik et al. (2016) also briefly talked about the feature selection in their study, highlighting the importance of recency-related features. Although their study is linguistically informed, the annotation effort behind it is very intense. The question is whether they could achieve a nearly similar performance using a smaller set of features.

3 Corpus and feature sets considered

As stated in section 1, we want to evaluate the features used in various ML SRF studies in a systematic way. So, we need first to select a collection of feature sets. Subsection 3.1 describes the criteria for the selection procedure, and provides an overview of the chosen sets. Afterwards, we explain how we applied the selected feature sets to the OntoNotes corpus (subsection 3.2).

3.1 Feature sets

The criteria to select the feature sets were as follows: we looked for studies that (1) had SRF as their main objective , (2) used a ML method, (3) used an English dataset and (4) applied interpretable features.

We first selected all the feature sets used by the SRF algorithms in the GREC challenges. We excluded the JUNLG set (Gupta and Bandopadhyay, 2009) because their study was rule-based, and the WLV feature set (Orăsan and Dornescu, 2009) because we were not able to interpret all their features.

Since the GREC challenges were conducted several years ago, we also examined the more recent literature to include later SRF feature sets satisfying our criteria. We downloaded the full ACL anthology BibTex file from `https://www.aclweb.org/anthology/` and used regular expressions to search manually the expressions of Table 1 in the title and abstracts of the articles:

[R|r]eferring [E|e]xpression.*[M|m]achine [L|l]earning	title =.*[R|r]eferring [E|e]xpression
[G|g]enerat[ion|ing].*[R|r]eferring [E|e]xpression.*discourse	title =.*[R|r]efer.* [G|g]eneration
[D|d]ata-driven.*[E|e]xpression	

Table 1: Terms used to search for SRF studies

We excluded several results because they did not meet the criteria defined above (Zarrieß and Kuhn, 2013; Siddharthan et al., 2011; Stent, 2011; Castro Ferreira and Paraboni, 2017). Based on the result of the manual search, we included feature sets from the papers by Castro Ferreira et al. (2016) and Kibrik et al. (2016), which together with the GREC feature sets form the seven sets we use in the feature selection experiments. As a naming convention, the GREC feature sets are called with their names from Belz et al. (2010); the other two feature sets are named after the last name of their first authors.

Number	Dataset	Reference	Number of features
1	IS-G	Bohnet (2008)	5
2	Ferreira	Castro Ferreira et al. (2016)	5
3	OSU	Jamison and Mehay (2008)	8
4	ICSI	Favre and Bohnet (2009)	14
5	Kibrik	Kibrik et al. (2016)	17
6	U-Del	Greenbacker and McCoy (2009a)	18
7	CNTS	Hendrickx et al. (2008)	21

Table 2: The datasets used in this study. The first two columns show the number and the name with which the datasets will be referred to.

To provide an overview, we grouped the features into 9 broad categories namely `Grammatical role`, `Inherent features`, `Referential status`, `Recency`, `Competition`, `Antecedent form`, `Surrounding patterns`, `Position` and `Protagonism` explained below. Throughout this article, `REF` refers to the current referent and `ANTE` refers to its coreferential antecedent. In Tables 3-7, the first column, `Feature`, provides the description of each feature. The column `Type` indicates whether the value of the feature is numeric (num), categorical (cat), Boolean (bool) or character (char). Also, the number `[N]` next to the `Type` attribute specifies how many distinct features it encodes. For instance, "Grammatical role of the 2^{nd} and 3^{rd} ANTE" with the type attribute `cat[2]` refers to 2 categorical features namely "Grammatical role of the 2^{nd} ANTE" and "Grammatical role of the 3^{rd} ANTE". The column `DT` shows which data sets contain the feature.

Grammatical Role This category contains information about the syntactic position of REF and ANTE.

Feature	Type[N]	DT	Symbol
Grammatical role of REF	cat[1]	1-7	gm
Grammatical role of ANTE	cat[1]	5,6	gm_p1
Grammatical role of the 2^{nd} and 3^{rd} ANTE	cat[2]	6	gm_p2, gm_p3
Trigram grammatical roles of the 3 antecedents	cat[1]	7	gm_tri
Is REF subject of the current & 2 prev sentences?	bool[3]	6	subj_S, subj_prevS, subj_prev2S
Is ANTE in the subject position?	bool[1]	6	ante_subj
Are REF and ANTE prepositional phrases?	bool[2]	5	ref_pp, ante_pp

Table 3: Grammatical features encoded in different datasets. REF and ANTE respectively refers to the current mention and its antecedent.

Inherent features of a referent It contains features marking the inherent properties of the referents such as semantic category or animacy (`anim`) [datasets 3, 4, 5 & 7], gender (`gender`) [5] and plurality (`plur`), i.e. whether the referent is plural or singular [5].

Antecedent form This feature is concerned with the form of ANTE. As Bohnet (2008) notes, most of the times, this feature is determined based on the prediction of its predecessor, hence is regarded as insecure information. This feature, referred to as `ante_form`, is used in datasets 1 & 5.

Position This category contains information about the position of REF.

Feature	Type[N]	DT	Symbol
Sentence Number	num[1]	6,7	sent_num
NP number	num[1]	7	np_num
Mention number	num[1]	1,5,6	ment_num
Referent number	num[1]	6	ref_num
How many times has REF occurred since the beginning? (1,2,3,4+)	cat[1]	4	count_bef
How many times does REF occur since the last change? (1,2,3,4+)	cat[1]	4	count_aft
Mention order (first, second, middle, last)	cat[1]	3	ment_ord
Does the REF appear in the first sentence?	bool[1]	7	first_sent
Does the REF appear in the beginning of a paragraph?	bool[1]	4	firstS_par

Table 4: Positional features of different feature sets

Recency This notion refers to the distance between REF and its ANTE.

Feature	Type[N]	DT	Symbol
Distance in number of words	num[1]	1,5	dist_w
Distance in number of NPs	num[1]	7	dist_np
Distance in number of markables	num[1]	5	dist_mark
Distance in number of sentences	num[1]	5,7	dist_sent
Distance in number of paragraphs	num[1]	5	dist_par
Distance to the nearest non-pronominal antecedent	num[1]	5	dist_full
Word distance (5 bins of 0-10, 11-20, 21-30, 31-40 and 40+)	cat[1]	2	bin5_w
Word distance (3 bins of 0-5, 6-12 and 13+)	cat[1]	3	bin3_w
Sentence distance (+/-2 sentences)	cat[1]	6	bin2_sent
Sentence distance (3 bins of 0, 1, 2+ sentences)	cat[1]	3	bin3_sent

Table 5: Recency features of different feature sets

Competition Features in this category encode the competition between other referents and REF.

Feature	Type[N]	DT	Symbol
Is the previous referring expression about the same entity?	bool[1]	4	same_ante
Does REF have a competitor in the whole text?	bool[1]	3	compet_txt
Does REF have a competitor since the beginning of the text?	bool[1]	3	compet_beg
Is there a competitor between REF and ANTE?	bool[1]	3,6	compet_prev
Are there other referents in the same sentence?	bool[1]	6	compet_sent
Does the previous sentence contain another referent?	bool[1]	7	compet_prevS

Table 6: Competition features of different feature sets

Referential status Features such as whether the referent is new in the sentence (`same_sent`) [datasets 1 & 2], in the paragraph (`new_in_par`) [dataset 2] or in the text (`new_in_text`) [dataset 2] mark the referential status of REF. The sentence-level feature can, in theory, belong also to the recency category.

Surrounding patterns It has information about the lexical and POS tag of the surrounding tokens.

Feature	Type[N]	DT	Symbol
Word unigram and bigram before and after the target	char[4]	4,7	w_(uni|bi)_(bef|aft)
Word trigram before and after the target	char[2]	7	w_tri_(bef|aft)
3 POStags before and after the target	char[6]	7	pos_(1|2|3)_(bef|aft)
Punctuation type before and after the target	cat[2]	4	punct_(bef|aft)
Morphology of the previous and next words (-ed, -ing, -s, -)	cat[2]	4	morph_(bef|aft)
Is the target immediately followed by `and`, `but`, `then`?	bool[3]	6	w_(and|but|then)
Is the target between `comma` and `and`?	bool[1]	6	w_command

Table 7: Surrounding pattern features of different feature sets

Protagonism Kibrik et al. (2016) used 2 measurements of protagonism. One measures the ratio of the chain length of REF to the maximal chain length in the text (`protagonism1`). The second one measures the chain length of REF to the sum of all markables in the text (`protagonism2`).

3.2 Applying feature sets to the OntoNotes corpus

This section begins by giving a brief overview of OntoNotes, the corpus used in this study. Afterwards, the referring expression types used in the prediction task will be explained.

We use the Wall Street Journal portion of OntoNotes (Pradhan et al., 2013) in this study. One of the reasons why we use this data is that it is annotated with structural information (syntax and predicate argument structure) and shallow semantics. Also, in order to extract paragraph information, we incorporated the information from the PDTB parser (`https://github.com/WING-NUS/pdtb-parser/tree/master/external/aux_data/paragraphs`).

Applying the feature sets to the OntoNotes corpus, 65 distinct features were attained[1]. It is noteworthy that applying the features was not always straightforward. Particular difficulty was posed for instance by recency features. To find the distance in words between two mentions, two different approaches were possible: either to keep the punctuation in the counting or to ignore it. The word distance features in this study take punctuation into the consideration. After excluding the first and second person referents, we ended up using 30500 referring expressions, divided into 70% training and 30% test sets.

For the referential choice prediction task, we took the intersection of the referring expression categories used in Belz et al. (2010), Kibrik et al. (2016) , and Castro Ferreira et al. (2016). Hence, the task in the current study is to predict the referential form being `pronoun`, `proper name` or `definite description`.

In this section, we described the feature sets which will be used in our feature selection studies, and explained how we applied them to the OntoNotes corpus. In section 4, we explain the feature selection experiments for the assessment of the features.

4 Feature selection experiments for assessing the features

We start by briefly explaining the classification algorithm and the feature selection methods we use in our experiments (subsection 4.1). Afterwards, we elaborate on the classification models trained on the OntoNotes data using the proposed feature sets (subsection 4.2). The section continues with two feature selection experiments with which we assess the importance of the features (subsection 4.3 and subsection 4.4). The next step is to use different subsets of the features and to re-run the classification algorithms, to see how this will affect the accuracy of the models.

4.1 The classification algorithm and the feature selection methods explained

We use the Random Forest algorithm (RF), an ensemble learning method, as our classifier in this study. The classification is based on the results achieved from the myriad of decision trees it generates while

[1]There are a few features that we did not apply to the corpus. Examples of such cases are elementary discourse unit (EDU) and rhetorical distance (RhD) measurements, both from Kibrik et al. (2016).

training (Nayak and Natarajan, 2016; Biau, 2012). We employ RF because it also reliably computes the permutation importance of the variables while training the classification models.

Afterwards, we use two automatic feature selection methods to assess the features used in each model: "Rank Features by Importance" [henceforth RFI] and "Sequential Forward Search" [henceforth SFS]. These methods will be detailed in subsection 4.3 and subsection 4.4.

4.2 Building RF models for predicting the referential choice

To implement RF, we used *ranger* (Wright and Ziegler, 2015), which is a fast implementation of Random Forests in R. Table 8 presents the results of the models with the original features of each feature set.

Dataset	IS-G	Ferreira	OSU	ICSI	Kibrik	UDeL	CNTS
Accuracy	0.68	0.601	0.697	0.69	0.793	0.624	0.723

Table 8: Accuracy rates of the RF models with their original features

According to Table 8, the model trained by the Kibrik feature set has the highest accuracy (henceforth, best performing model), followed respectively by the CNTS and OSU models. In the next section, we evaluate the features of each set to see which contributed the most to the predictive success of the models.

4.3 Experiment one: evaluating the importance of the features using RFI

To assess the features, we use the built-in permutation importance (Breiman, 2001) of RF, ranking the "importance" of the features. According to Strobl et al. (2008), to measure the importance of the feature X_i, first the model is built, and its accuracy is computed in Out-of-bag (OOB) observations. Afterwards, any link between the values of X_i and the outcome of the model is broken by the permutation of all the values of X_i, and the accuracy of the model with the permuted values is re-computed. The difference between the accuracy of the new model and the original score is defined as the permutation importance of X_i. Hence, if a feature has noise or random values, it is likely that the permutation does not influence the accuracy. Instead, a high difference between the two rates signal the importance of the feature for the prediction task. Figure 1 shows the importance of different variables in the seven models. The higher the value of the *Mean Decrease in Accuracy* on the x-axis, the greater the importance of the feature.

Also, we computed the p-values for the variables following the method of Altmann et al. (2010) under the null hypothesis that the permutation of the variable has no impact on the accuracy. Out of the 65 distinct features, the null hypothesis was confirmed for 4 features from the UDeL dataset (`w_and`, `w_but`, `w_then` & `w_command`) and one feature from OSU (`compet_txt`). The rest of features contributed to the models with varying degrees of importance.

4.4 Experiment two: evaluating the importance of the features using SFS

Our second method is Sequential Forward Search. The algorithm starts with an empty set and adds the features until this no longer yields much improvement in accuracy. The algorithm stops if the improvement is below the minimum required value of improvement (alpha=0.01) that we set. We used the R package mlr (Bischl et al., 2016) for the implementation of the SFS algorithm. The learner we used in this model is `classif.randomForest` and the resampling strategy is `Holdout`. Each box in Figure 2 shows the selected features of each feature set.

IS-G	Ferreira	OSU	ICSI	Kibrik	UDeL	CNTS
gm	gm	gm	gm	anim	gm	gm
ante_form	same_sent	anim	anim	plur	gm_p1	anim
same_sent	bin5_w	bin3_sent	same_ante	dist_sent	bin2_sent	dist_sent
	new_in_par	bin3_w	count_aft	dist_par	ent_num	dist_np
			punct_bef	ante_form		pos_1_bef

Figure 2: SFS optimal features of each feature set

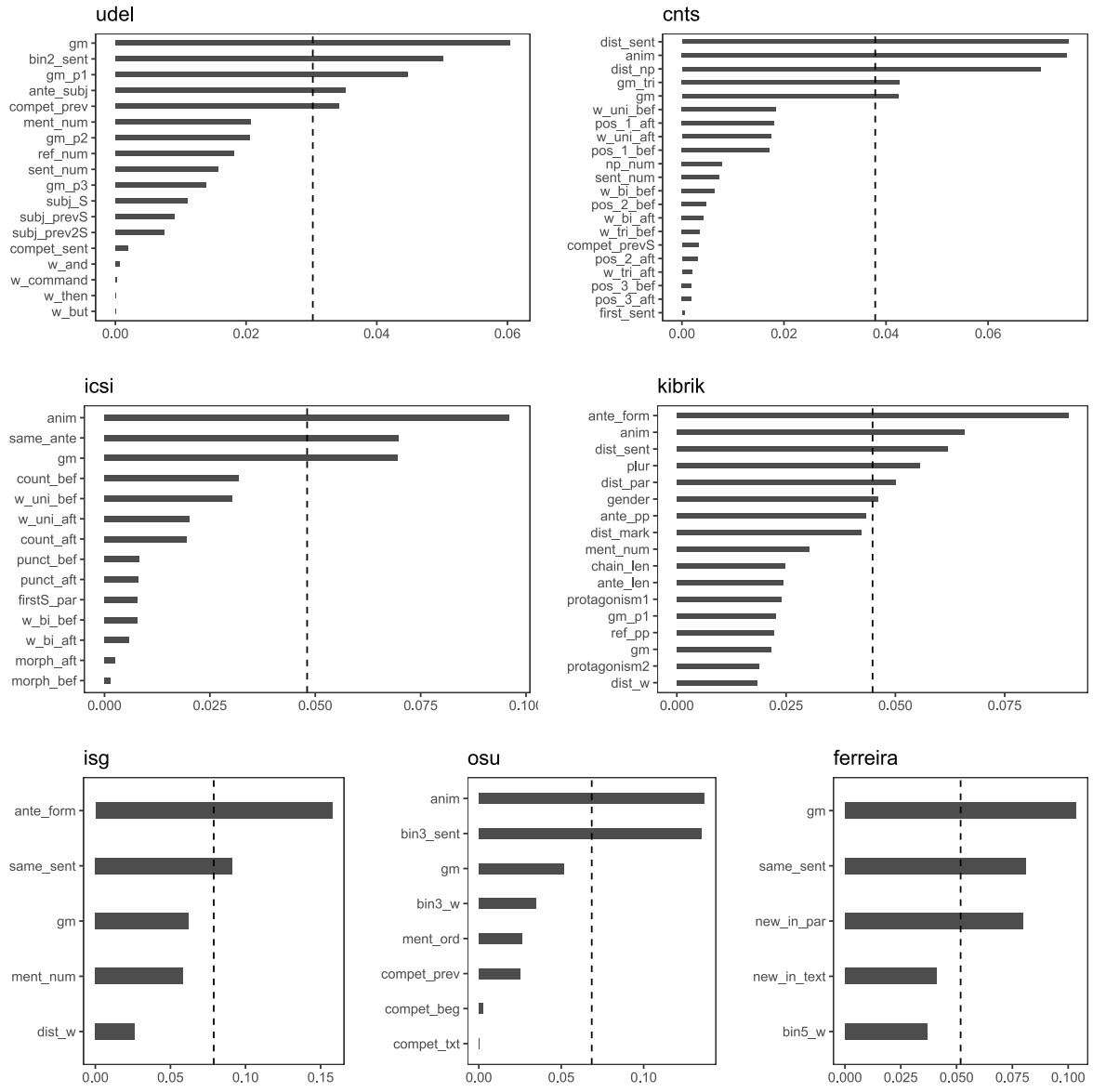

Figure 1: Variable importance plot of the RF models. The y-axis lists the features of each model; the x-axis shows the permutation importance (Mean Decrease in Accuracy) of each feature.

4.5 Experiment three: exploring different feature subsets based on their importance

Firstly, in 4.5.1, we explore the accuracy of different subsets of each feature set based on the results of the RFI and SFS experiments. In 4.5.2, we explore the subsets of all the features combined, trying to come up with an optimal consensus set of features.

4.5.1 Subsets of each feature set

The first row of Table 9 with the name `original` reports on the original accuracy of each model. Rows `top1`, `top2` and `top3` report respectively on performing RF on the first, the first two and the first three features of each feature set with the highest permutation importance (according to Figure 1). For instance, in the case of OSU, the first two features with the highest permutation importance are {`anim`, `bin3_sent`}. The `top50%` row reports on applying the RF to the top 50% features of each feature set. The dashed line in Figure 1 shows which features are among the top 50%. For instance, in case of Kibrik, the top 50% features are {`ante_form`, `anim`, `dist_sent`, `plur`, `dist_par`, `gender`}. Finally, the row `sfs` reports on applying RF to the subsets chosen in the SFS

experiment (Figure 2). In case of UDeL, for instance, the RF is applied to the feature set {gm, gm_p1, bin2_sent, ent_num}. Surprisingly, in case of IS-G and OSU, the accuracy of the models solely incorporating SFS features are slightly higher than the original algorithms.

Row name	IS-G	Ferreira	OSU	ICSI	Kibrik	UDeL	CNTS
original	0.68	0.601	0.697	0.69	0.793	0.624	0.723
top1	0.56	0.492	0.556	0.556	0.56	0.492	0.556
top2	0.66	0.577	0.665	0.593	0.632	0.560	0.661
top3	0.684	0.593	0.694	0.661	0.712	0.585	0.670
top50%	0.66	0.593	0.665	0.659	0.761	0.597	0.696
sfs	0.684	0.601	0.701	0.681	0.761	0.608	0.701

Table 9: The accuracy of various subset models

4.5.2 Subsets of all features

After trying out different subsets of each feature set, we are now going to explore different combinations of all features based on the results of the RFI and SFS experiments.

1. We first applied RF to the set of those features with the highest permutation importance in the RFI experiment. According to Figure 1, the set is {gm, anim, dist_sent, ante_form}. The accuracy of this model is 0.728.

2. This time, we applied RF to the union of the two most important features of each feature set: {gm, bin2_sent, dist_sent, bin3_sent, same_ante, ante_form, same_sent, anim}. The accuracy of this model is 0.723. Against our expectation, the accuracy of this model is lower than the accuracy of the union of the single most important features.

3. We applied RF to the union of all the SFS feature sets shown in Figure 2. The set has 19 distinct features, and the accuracy is 0.779. So far, this is the highest accuracy obtained from the subsetting of the features.

4. Since the subset outlined in item 3 led to the highest accuracy compared to the other subsets, we ran SFS on this feature set to end up with a smaller set of features. The idea here is to reach an optimal compromise between the number of features (which should be small) and the performance (which should be high). The features selected by SFS are {gm, ante_form, bin3_sent, anim, plur, dist_par}. The accuracy of the prediction with the selected subset is 0.776. We used Bayes Factor analysis with a Bernoulli distribution to determine whether there is evidence for a difference in accuracy levels of less than or greater than 0,05 between the best performing model, Kibrik with the accuracy of 0.793, and the new model. There is very strong evidence of the accuracies being closer than the threshold, hence being evidentially indistinguishable.

5 Conclusion

The aim of this study was to systematically examine feature-based SRF models, fleshing out what features make the models work best. By evaluating different feature sets of the computational SRF studies from a linguistic perspective, we tried to build a bridge between the features employed by computational models and the linguistic explanations behind those features.

Based on the results of the two feature selection experiments conducted on seven feature sets, and the approach outlined in section 4.5.2, we composed a consensus set consisting of six features from 4 classes: animacy and plurality [Inherent features of a referent], grammatical role of the current mention, form of the antecedent, and categorical distance in number of sentences [recency].

Comparing the consensus set with the previously proposed feature sets has interesting implications for both feature-based SRF research and linguistics.

Implications for feature-based SRF studies: We found that by using a smaller set of features, models can often achieve nearly identical performance. For example, as shown in Table 9, the performance of the Kibrik model using 17 features is 0.793, but we can achieve an accuracy of 0.776 using only 6 features. Furthermore, using the largest number of features does not guarantee the best possible performance: the results in Table 9 show that the performance of a subset is often similar to that of a superset. In the case of 2 feature sets, OSU and IS-G, the performance of one of the subsets was even better than the original.

We saw in Table 5 that all systems except ICSI encode recency in one way or the other. Some encode recency in terms of "lower-level" units such as counting words, NPs and markables; others focus on "higher-level" units such as sentences and paragraphs. The two experiments show the higher-level metrics are always ranked more highly than the lower-level ones. Clearly, features encoding similar concepts (e.g. distance to the antecedent) do not always contribute equally to a model. The same holds for referential status, where sentence-level features play a much more important role than the others.

Implications for linguistics: In section 3.1, we grouped the features that were used by the systems that we studied into 9 broad categories. The 6 features that we chose (above) as our consensus set are all from only 4 of these 9, namely `grammatical role`, `inherent features` of the referent, `antecedent form` and `recency`.

- `Inherent features`: We showed that two inherent features of a referent, namely animacy and plurality, i.e. whether the referent is plural or singular, play major roles in predicting the referential choice. Given the linguistics literature, the importance of animacy is no great surprise; for example, Fukumura and van Gompel (2011) reported that pronouns were more frequent for referring to animate than inanimate referents. More of a surprise is the role of plurality for SRF which has attracted less attention. The psycholinguistics literature suggests that when conjoined noun phrases are introduced into the discourse, they are treated as a group, and the group is in focus, which makes it prominent (Patson and Warren, 2011). According to Gordon et al. (1999)'s *repeated name penalty*, referring to a prominent referent with a proper name instead of a pronoun increases the processing time. An implication of this is the possibility that including the plurality feature in a prediction task might facilitate the pronoun detection.

- `Grammatical role`: Various studies, including centering-based research, tend to emphasize that referents in subject position have a higher tendency to be pronominalized in the subsequent sentence (Brennan et al., 1987; Brennan, 1995; Kaiser, 2010). The focus of previous research has usually been on the subjecthood of the *antecedent*, and less so on the subjecthood of the current mention. Our analysis suggests that the grammatical role of the current mention is more important than that of the antecedent in predicting the choice of referential form.

- `Antecedent form`: In this case, our findings match those in the linguistic tradition (Gundel and Hedberg, 2008). This factor could be important either because people tend to avoid consecutive uses of the same expression (Bohnet, 2008), or because having a pronominal antecedent enhances the prominence of the referent (Kaiser, 2003).

- `Recency`: Recency, in the linguistic tradition, has often been emphasized, but often without a clear definition. Our study suggests that recency is best defined in terms of the number of *sentences* that intervene between the antecedent and the current mention; next best is recency metric defined in terms of the number of paragraphs. This finding is in line with the linguistic tradition, which tend to focus on "higher level" measures (Fox, 1987; Tomlin, 1987; McCoy and Strube, 1999; Henschel et al., 2000; Arnold et al., 2009).

This concludes our comparison of different feature-based computational studies of referential choice. In view of the above observations, we hope that computational and theoretical studies of language will continue to provide inspiration to each other.

References

André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Mira Ariel. 1990. *Accessing Noun-Phrase Antecedents*. Routledge.

Jennifer E Arnold and Zenzi M Griffin. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of memory and language*, 56(4):521–536.

Jennifer E Arnold, Loisa Bennetto, and Joshua J Diehl. 2009. Reference production in young speakers with and without autism: Effects of discourse status and processing constraints. *Cognition*, 110(2):131–146.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Generating referring expressions in context: The task evaluation challenges. In *Empirical methods in natural language generation*, pages 294–327. Springer.

Gérard Biau. 2012. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095.

Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M Jones. 2016. mlr: Machine learning in R. *The Journal of Machine Learning Research*, 17(1):5938–5942.

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. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Susan E Brennan, Marilyn W Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162.

Susan E Brennan. 1995. Centering attention in discourse. *Language and Cognitive processes*, 10(2):137–167.

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, November. Association for Computational Linguistics.

Thiago Castro Ferreira and Ivandré Paraboni. 2017. Improving the generation of personalised descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 233–237, Santiago de Compostela, Spain, September. Association for Computational Linguistics.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany, August. 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, July. Association for Computational Linguistics.

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, August. Association for Computational Linguistics.

Barbara A. Fox. 1987. *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge Studies in Linguistics. Cambridge University Press.

Kumiko Fukumura and Roger P. G. van Gompel. 2011. The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, 26(10):1472–1504.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Peter C Gordon, Randall Hendrick, Kerry Ledoux, and Chin Lung Yang. 1999. Processing of reference and the structure of language: An analysis of complex noun phrases. *Language and Cognitive Processes*, 14(4):353–379.

Charles Greenbacker and Kathleen McCoy. 2009a. Udel: generating referring expressions guided by psycholinguistic findings. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 101–102. 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)*.

Jeanette K Gundel and Nancy Hedberg. 2008. *Reference: interdisciplinary perspectives*. Oxford University Press.

Samir Gupta and Sivaji Bandopadhyay. 2009. JUNLG-MSR: A machine learning approach of main subject reference selection with rule based improvement. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 103–104, Suntec, Singapore, August. 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, June. 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.

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, June. Association for Computational Linguistics.

Elsi Kaiser. 2003. Word order, grammatical function, and referential form: On the patterns of anaphoric reference in finnish. *Nordlyd*, 31(1).

Elsi Kaiser. 2010. Effects of contrast on referential form: Investigating the distinction between strong and weak pronouns. *Discourse Processes*, 47(6):480–509.

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, 2019. *Computational Generation of Referring Expressions: An Updated Survey*. Oxford University Press.

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

Anmol Nayak and D Natarajan. 2016. Comparative study of naive bayes, support vector machine and random forest classifiers in sentiment analysis of twitter feeds. *Int. J. Adv. Stud. Comput. Sci. Eng*, 5:14–17.

Constantin Orăsan and Iustin Dornescu. 2009. WLV: A confidence-based machine learning method for the GREC-NEG'09 task. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 107–108, Suntec, Singapore, August. Association for Computational Linguistics.

Nikole D Patson and Tessa Warren. 2011. Building complex reference objects from dual sets. *Journal of memory and language*, 64(4):443–459.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, August. Association for Computational Linguistics.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Amanda J Stent. 2011. Computational approaches to the production of referring expressions: Dialog changes (almost) everything. In *PRE-CogSci Workshop*.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307.

Russell S Tomlin. 1987. *Coherence and grounding in discourse: outcome of a symposium, Eugene, Oregon, June 1984*, volume 11. John Benjamins Publishing.

Marvin N Wright and Andreas Ziegler. 2015. ranger: A fast implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Sina Zarrieß and Jonas Kuhn. 2013. Combining referring expression generation and surface realization: A corpus-based investigation of architectures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1547–1557.