Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2989–2997
Marseille, 20-25 June 2022

Constructing Distributions of Variation in Referring Expression Type from Corpora for Model Evaluation

T. Mark Ellison, Fahime Same
University of Cologne
Cologne, Germany
{t.m.ellison, f.same}@uni-koeln.de

Abstract

The generation of referring expressions (REs) is a non-deterministic task. However, the algorithms for the generation of REs are standardly evaluated against corpora of written texts which include only one RE per each reference. Our goal in this work is firstly to reproduce one of the few studies taking the distributional nature of the RE generation into account. We add to this work, by introducing a method for exploring variation in human RE choice on the basis of longitudinal corpora - substantial corpora with a single human judgement (in the process of composition) per RE. We focus on the prediction of RE types, *proper name*, *description* and *pronoun*. We compare evaluations made against distributions over these types with evaluations made against parallel human judgements. Our results show agreement in the evaluation of learning algorithms against distributions constructed from parallel human evaluations and from longitudinal data.

Keywords: Referring Expression, Prediction, Evaluation, Variation

1. Introduction

Referring is a fundamental part of communication and has been extensively investigated in theoretical, empirical and computational linguistic studies. Referring Expression Generation (REG), as one of the main components of a Natural Language Generation (NLG) pipeline, tackles two different tasks: the first task, or one-shot REG, is concerned with finding a set of attributes to single out a particular referent from a set of distractors. The second task, which is the focus of this article, involves generating referring expressions (REs) in a discourse context. This task is hereafter referred to as *REG-in-context*. Belz and Varges (2007) defined REG-in-context as follows: "given an intended referent and a discourse context, how do we generate appropriate referential [referring] expressions to refer to the referent at different points in the discourse?" (p. 9). Classic REG-in-context studies often approach REG in two steps. The first step is to decide which form to use in a given context; for example, whether to use a pronoun (e.g. he), a proper name (Barack Obama), or a description (the former president of the United States). The second step determines the content and linguistic realisation of the expressions (Reiter and Dale, 2000; Krahmer and van Deemter, 2012).[1] The focus of this paper is on the first step of the classic REG approach, namely the choice of RE form.

Algorithms for the generation of REs in context are standardly evaluated against corpora of written texts, offering a single 'correct' response in the given context. However, the generation of REs is a probabilistic task and not a deterministic one. Although a certain form might be the dominant option at a certain position and given a certain context, this does not always mean that other forms are ungrammatical or inappropriate. Our goal in this study is to explore the evaluation of non-deterministic predictions of Referring Expression Forms (henceforth REFs). It is worth noting that non-deterministic approaches have been discussed mainly in the context of one-shot REG (Gatt et al., 2013; van Deemter, 2016; van Gompel et al., 2019), and less so in the context of REG in-context.

Two exceptional REG-in-context studies do consider the distributional nature of referring expression choice: Castro Ferreira et al. (2016a) and Castro Ferreira et al. (2016b) (henceforth VaREG studies). These studies analyse individual differences in modelling reference. In the first study, Castro Ferreira et al. (2016a) developed a new corpus, VaREG, which is composed of REs produced by a number of participants in identical contexts. They measured the normalised entropy of selections made by the different participants in the study and demonstrated substantial variation. In a follow-up study, they introduced the use of the Jensen-Shannon Divergence metric to measure the similarity between human-produced and predicted distributions. The appeal of their non-deterministic approach is offset by the human time required to build a corpus of parallel human judgements. Their corpus, VaREG, is composed of 36 different texts in which only references to the main topic of the texts are annotated. With more and larger texts, it is rarely feasible to run an experiment in which many human participants recreate the REs. The goal of the present study is to infer variation in human behaviour through the variation found within a corpus of texts,

[1] With the rapid advancement of neural approaches, end-to-end REG models are also gaining popularity where the decision about form and content is made simultaneously (Castro Ferreira et al., 2018; Cao and Cheung, 2019; Cunha et al., 2020).

gaining the benefits of understanding human variation without the substantial cost of human-intensive studies.

We propose to achieve this using what we call *longitudinal* corpora - a standard corpus of collected texts, with only one referring expression form choice per referring expression (the actually used form) - to characterise variation. We take this terminology from the original meanings of the word - *lengthwise* - as opposed to a parallel corpus which has a *latitudinal* dimension of parallel judgements made by different informants. We achieve this goal by identifying classes of linguistic contexts. Forms that occur in the same class of linguistic context are presumed to have arisen from the same distribution over available options. The correct distribution is part of speakers' shared linguistic competence. We hypothesise that these will match the distributions of options seen when multiple speakers are asked to choose REFs for the same referring expression given the referent and the context.

The structure of this paper is as follows. In section 2, we provide an overview of the studies tackling REG-in-context, and highlight their differences. Then, section 3 presents a detailed account of the VaREG studies, since they form the launching pad for the current study. In section 4, we detail the preparatory steps of the current study, namely deciding which corpora, feature set, and Machine Learning (ML) methods to use. Section 5 presents the findings of our study, focusing on the implications of the longitudinal and parallel distributions. Finally, section 6 gives a brief summary and review of the findings.

2. Related work

The choice of referring expressions reflects the semantic prominence of entities in the common ground (von Heusinger and Schumacher, 2019), and so the factors influencing semantic prominence have often featured in theoretical and empirical works on the choice of referring expressions. Less informative expressions such as pronouns are often used to refer to prominent entities, while those with more content are used to refer to less prominent ones (Ariel, 2001; Gundel et al., 1993). The factors influencing prominence have been argued to include grammatical function (Brennan, 1995), animacy (Fukumura and van Gompel, 2011), recency (Ariel, 1990; McCoy and Strube, 1999) and the existence of competing referents (Arnold and Griffin, 2007).

Computational models for generating referring expressions can be classified into three groups: rule-based, feature-based, and end-to-end.

Rule-based models, such as McCoy and Strube (1999), Henschel et al. (2000) and Krahmer and Theune (2002), employ centering or salience rules, and are often proposed as explanations of pronominalization patterns.

GREC, Generating Referring Expressions in Context, was a series of Shared Task Evaluation problems that triggered what can be regarded as one of the first systematic studies on the generation of REs in context (Belz et al., 2009). These shared tasks introduced various data-driven feature-based models, such as Greenbacker and McCoy (2009) and Hendrickx et al. (2008). These models differ from each other in the features they employ. For instance, Greenbacker and McCoy (2009) proposed a psycholinguistically motivated model incorporating various features encoding the subjecthood. The model by Hendrickx et al. (2008), on the other hand, made use of various textual features such as N-gram patterns before and after the target REs. Following the footsteps of the GREC models, Kibrik et al. (2016) regarded referential choice prediction as a multi-factorial process and included a large number of factors in their study. Same and van Deemter (2020) conducted several experiments on linguistically-motivated features taken from previous studies to arrive at a consensus set of features for the referential choice prediction task.

The majority of rule- and feature-based models aim at predicting the form of expressions. End-to-end models such as Castro Ferreira et al. (2018), Cao and Cheung (2019) and Cunha et al. (2020), in contrast, aim for the simultaneous prediction of form and content.

All these models are trained on text corpora which have only one 'gold-standard' form per reference. The choice of referring expressions is a non-deterministic task, meaning more than one form might be feasible for each reference. Furthermore, the distribution over form choice in this situation might well be part of the speaker's linguistic knowledge. In a rare study addressing referential choice variation, Castro Ferreira et al. (2016a) show that there is substantial variation in the choices made by human participants. In response to this situation, they constructed a new corpus, VaREG, which offers more than one form per reference [2]. Additionally, in a follow-up study, Castro Ferreira et al. (2016b) used the human evaluations to assess their ML models' performance. We provide an account of these 2 works in section 3 as they are important background to the current study.

[2]It is worth-mentioning that in earlier work, Belz and Varges (2007) created a somewhat similar corpus using 10% of the MSR (Main Subject Reference) corpus. They stated that in 50.1% of the cases, the participants of their study chose the exact same referring expressions for each slot. We do not discuss their study further because each of their texts was re-created by only 3 participants offering little opportunity to see variation, and they did not directly use their distributions to evaluate ML model performance.

3. The VaREG studies

Castro Ferreira et al. (2016a) (hereafter CF+) developed a corpus capturing variation in the use of REs. Their corpus consists of 36 texts with 563 selected REs in three different genres: news texts, reviews of commercial products, and Wikipedia texts. All references to the topic of a text were replaced by gaps, and participants in their study filled the gaps with referring expressions for the text topic. The texts were divided into 4 lists of nine texts each. Approximately 20 participants were assigned to each list. Hence, for each slot, there are about 20 human-produced REs. Table 1 shows the original version of an excerpt from the text entitled "Google stock drops 12 percent", along with the embedded referring expressions produced by two participants.

CF+ classified the human-generated referential forms of the produced expressions into five categories, shown here with examples in parentheses: name (Google), pronoun (it), description (the giant tech), demonstrative (this company), and empty (-).

They also annotated the following linguistic features for each referent in the corpus: grammatical role, referential status, i.e., whether the RE is given/new, on the levels of sentence, paragraph, and text, and recency - the number of words between the current and the preceding mention. These linguistic annotations let them assess variation relative to the linguistic factors. To assess the variation in participants' choice of REFs, they measured the entropy of choices normalised according to (1), where x corresponds to references in the current gap, and n=5 is the number of different forms annotated (pronoun, name, description, empty and demonstrative). The probability of having a referring expression form i express token x is given by $p(x_i)$. The entropy measure takes values between 0 and 1, where 0 shows complete agreement between participants' choices and 1 shows no visible bias towards one class of representation or another. This quantity is the Kullback-Liebler distance from the uniform distribution.

$$H(x) = -\sum_{i=1}^{n=5} \frac{p(x_i)\log(p(x_i))}{\log(n)} \quad (1)$$

CF+ found substantial variation between participants in their choices of referring expressions. They explored how three linguistic factors (recency, referential status, and grammatical role) impacted relative entropy, determining which factors have the most impact on variation. They found, for instance, that participant responses were more variable when the referring expression occurred in object position. Likewise, more variation was observed when referring expressions were relatively far from their antecedents.

In a follow-up study, Castro Ferreira et al. (2016b) used the same corpus to develop a REG model where instead of choosing the form with the highest likelihood, the model can predict the frequency with which a particular model can assume different REFs. After running various models, they evaluated to what extent each model captures the variation in human referring expression choice, comparing the distribution over referring expression class predicted by the model against the corresponding distribution made of human choices. To compare the distributions, they used the Jensen-Shannon Divergence (JSD) which is based on the Kullback-Liebler Divergence (KLD).

The Kullback-Liebler Divergence (Kullback, 1997) is an information-theoretic measure which expresses degrees of difference between distributions. It can be thought of as the average amount of extra information which must be supplied to represent an item x occurring with relative frequency $p(x)$, if it was expected with frequency $q(x)$. The divergence (measured in bits) is shown in (2).

$$KL(p||q) = \sum_x p(x)\log_2 \frac{p(x)}{q(x)} \quad (2)$$

Given two distributions p and q, the Jensen-Shannon divergence metric (3) is the average of the KLD measures from a midpoint distribution $r = \frac{p+q}{2}$ to p and to q. This measure has the desirable property of being 0 for identical distributions, and 1 for maximal divergence (Lin, 1991).

$$JSD(p||q) = \frac{1}{2}(KL(p||r) + KL(q||r)) \quad (3)$$

Castro Ferreira et al. (2016b) used this metric to evaluate how well a variety of feature choices captured human variation when combined with a Naive Bayes learner, on the one hand, and a Recurrent Neural Network on the other.

In this paper, we use the same corpus, and the same evaluation measure, to see how important the collection of human variation data is for evaluating competing algorithms.

4. Preparatory steps for the new study

In this section, we detail the choice of corpora, feature sets, and the ML methods exploited in the current study.

4.1. Corpora

We use two corpora. One is the VaREG corpus described above (Castro Ferreira et al., 2016a). From the VaREG corpus, we use both original referring expressions in the corpus and the varied choices of expressions made by human participants.

The other corpus is the Wall Street Journal [henceforth WSJ] section of OntoNotes (Hovy et al., 2006;

Text version	Sentence
Corpus text	After astounding Wall Street with its incredible growth, Google learned on Tuesday the perils of high expectations.
Participant 60	After astounding Wall Street with Google's incredible growth, they learned on Tuesday the perils of high expectations.
Participant 63	After astounding Wall Street with its incredible growth, the giant tech learned on Tuesday the perils of high expectations.

Table 1: Three different variations of one of the news texts from the VaREG corpus.

Weischedel et al., 2013).[3] This corpus contains news articles. The total number of referring expression instances that we use is 30517. We randomly selected 50% of these REs as training data, the remainder as test data. We chose the 50/50 split in order to balance the needs of training size with having a large enough test corpus to have reasonable sample sizes of the distributions corresponding to particular combinations of feature values.

Referring expression classes: 5 classes, namely: pronoun, proper name, description, demonstrative NPs, and null references (called *empty*), are annotated in the VaREG corpus. In the WSJ corpus, empty cases are not annotated, and demonstrative NPs are grouped with other descriptions. Hence, referring expressions are classified into three types of forms: pronoun, description and proper name. To compare the two corpora, we reduced the number of VaREG labels to 3 excluding the null references, and merging demonstratives with the descriptions.

4.2. Feature set

We are interested in looking at the distribution of RE classes within individual combinations of feature values. We therefore only include categorical features with a maximum of 4 values in order to limit the number of distinct combinations.

The current paper makes use of the VaREG corpus but we have decided not to use the feature set described by its creators, as reported in Castro Ferreira et al. (2016b). A preliminary analysis showed poor results with these features when they were applied to the WSJ corpus. We decided, instead, to use the feature set introduced in Same and van Deemter (2020) as it yielded significantly better performance. To limit combinatorial complexity, these features were revised to fit within the above-mentioned constraints on numbers of distinct values. We chose the following features and values as input to the learning algorithms.

- `Grammatical Role`: subject, object, possessive determiner
- `Form of the previous mention`: pronoun, name, description, first_mention
- `Animacy`: human, other
- `Sentence recency` (sentential distance to the previous mention): same sentence, different sentence, first_mention[4]
- `Paragraph recency` (paragraph distance to the previous mention): same paragraph, different paragraph, first mention

We use X to denote the set of feature values that describe the context in which a referring expression occurs. An example of X is the following:

$$\left\{\begin{array}{l}\texttt{grammatical role}\texttt{:}\ \text{subject}\\ \texttt{previous mention form}\texttt{:}\ \text{pronoun}\\ \texttt{animacy}\texttt{:}\ \text{human}\\ \texttt{sentence recency}\texttt{:}\ \text{different}\\ \texttt{paragraph recency}\texttt{:}\ \text{same}\end{array}\right\}$$

There are 216 distinct possible combinations of feature values. 93 of these occur in the WSJ, while only 49 combinations occur in the VaREG corpus.

4.3. ML methods

Referring expression class prediction systems combine feature sets with ML algorithms. One goal of this paper is to explore the evaluation of non-deterministic prediction systems against evidence of human variation. We therefore only consider algorithms that infer distributions over possible REFs, rather than deterministically predict one.

We chose three algorithms popular in the current literature for handling tabular data. These are:

- Random Forests (Biau, 2012),
- XGBoost (Friedman, 2001; Chen et al., 2015), and
- CatBoost (Prokhorenkova et al., 2018).

Each algorithm was trained on the WSJ and VaREG corpora with the features described above. The model parameters are presented in table 2.

[3] OntoNotes Release 5.0 (Ralph Weischedel, et al., 2013) is licensed under the Linguistic Data Consortium: `https://catalog.ldc.upenn.edu/LDC2013T19`.

[4] Since features such as recency are not defined for the first mentions of a referent, we assign the value 'first_mention' to those expressions.

ML Method	Parameters
`Random Forest`	num of trees: `500`, num of variables to split at each node: `4`
`XGBoost`	learning rate: `0.1`, maximum depth of a tree: `6`, sub-sample ratio: `0.75`
`CatBoost`	learning rate: `0.03`, L2 regularization: `3.0`, Bootstrap type: `Bayesian`

Table 2: Parameter setting of the models

5. Distribution experiments

A key aim of the current paper is to propose a longitudinal method for exploring variation in human REF choice, and to evaluate that proposal in comparison with parallel human judgements. This requires us to exploit the only corpus of parallel REF judgements, the VaREG corpus introduced in section 3.

Parallel human judgements are those found in the VaREG corpus, where multiple annotators made independent choices about which referring expression form to use in particular contexts. Sufficient annotators allow the inference of the shared distribution over REFs for any given RE. The parallel annotations define a gold standard for evaluating distributional predictions: an algorithm for REF choice will succeed optimally when it reproduces the human range of choices, and assigns the same probabilities to each.

The creation of such corpora of parallel human REF choices is expensive, especially for large corpora. The subset of the WSJ corpus that we use contains 30517 referring expressions. If annotators achieve the optimistic goal of one REF choice per 10 seconds, the total annotation time for 20 annotations of the corpus would be 1700 hours.

We propose an approach that permits a picture of human variation, without the cost of parallel annotation. Instead, this approach (hereafter longitudinal distribution) makes use of the length of the corpus itself to offer the variation.

5.1. Longitudinal Distribution

As mentioned earlier, we use the distributions over possible REFs for constructing longitudinal distributions. Following is a sentence from the WSJ corpus (`document WSJ0020`).

- [Ex. 1] Saudi Arabia, for its part, has vowed to enact a copyright law compatible with international standards and to apply the law to computer software as well as to literary works, Mrs. Hills said.

Table 3 shows the linguistic context, i.e. the feature–value combinations for the underlined RE, *the law*.
Table 4 shows the probability distribution of each REF in the case of the underlined RE of example 1 inferred by each model.
We construct the longitudinal distribution over REFs by aggregating all the RE instances in the corpus which share the same feature value combinations. If $y(i)$ is the form of RE i, and $X(i)$ is its combination of feature values, then we can express the longitudinal distributions over feature values as in (4).

$$P(y|X) = \frac{|\{i|y = y(i), X = X(i)\}|}{|\{i|X = X(i)\}|} \quad (4)$$

In order to avoid problems of division by zero, or logarithms of zero, all probability values are incremented by $\varepsilon = 0.000001$ and the distributions renormalized.

`grammatical role`: object
`previous mention form`: description
`animacy`: other
`sentence recency`: same sentence
`paragraph recency`: same paragraph

Table 3: Feature-value pairs of example 1.

	DESCRIPTION	NAME	PRONOUN
`RF`	0.424	0.139	0.436
`XGBoost`	0.432	0.114	0.452
`CatBoost`	0.446	0.109	0.443

Table 4: Probabilities of REFs inferred by each model.

Figure 1 shows a bimodal pattern of relationships between the rank number of instances of a feature combination in the corpus and the count of instances. For low-frequency feature combinations, it seems that frequency drops off sharply. Exploring the reasons for this sudden fall off in the frequency of combinations is left for future work.

Figure 2 shows the distribution of entropy in constructed longitudinal distributions as well as the human parallel annotations, by plotting entropy against rank entropy. There is a discontinuity in the gradient of the graph, just as we saw in figure 1. Particularly interesting however is that the growth in entropy in the distributions constructed from human annotations on the VaREG corpus parallels the entropy pattern of the longitudinal distributions in the WSJ.

5.2. VaREG distributional results

Table 5 compares how well the three learning algorithms approximate the distribution of variation in the parallel and longitudinal human distributions of the VaREG corpus. The ranking of divergence values is the same for both the distribution of REFs in the parallel annotation - the gold standard for variation -

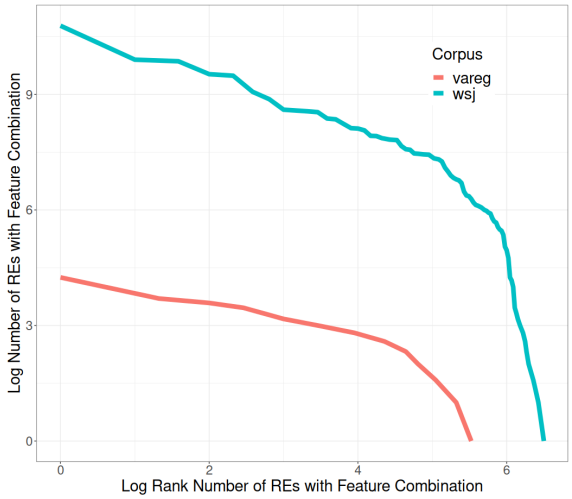

Figure 1: Plot of feature combinations. For any feature combination z, the vertical axis gives the log (base 2) absolute frequency of z - e.g. $\log_2$ 142 if there are 142 instances with this feature combination in the labelled corpus. The horizontal axis shows the log of z's rank on a list of combinations by decreasing frequency. This value could be $\log_2$ 7 if z is the 7th most frequent combination of features in terms of instances. Note the parallel structures in the graphs from the two corpora. The graph shows a combination of two stretches of Zipf-like power law.

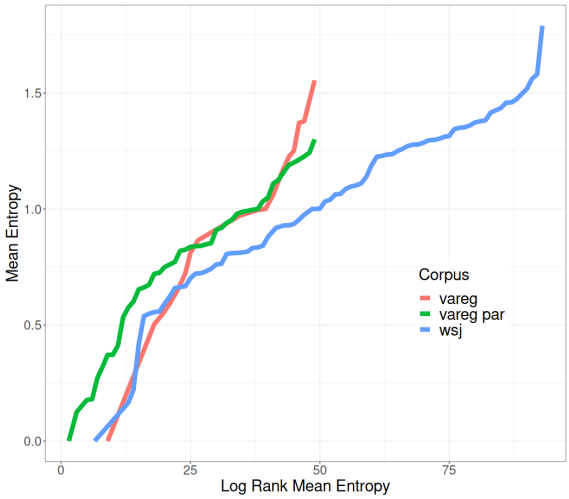

Figure 2: Plot of entropies associated with different feature–value combinations across 3 corpora. For a given feature combination z, the y-axis shows the average entropy over REFs found in RE instances having this combination of feature values. If all forms are equally likely, then the average entropy will be 1.58. The x-axis shows the log of the rank entropy - e.g. $\log_2$ 7 for the feature combination with the 7th highest mean entropy. These values are shown for the longitudinal distributions (vareg and wsj), and for the human parallel annotations of the VaREG corpus (vareg par).

and in the longitudinal aggregation based on feature combinations. The parallel ranking supplies suggestive (though not statistically profound) evidence that the longitudinal aggregation of REFs in distributions conditioned by the feature values (based on their relative frequencies in the test data) offers a potential alternative to parallel human annotation.

VaREG	Parallel	Longitudinal
RF	0.094	0.065
XGBoost	0.086	0.061
CatBoost	**0.076**	**0.059**

Table 5: JSD divergences between the trained algorithms and human parallel and longitudinal REF distributions. Lower divergence values correspond to more-similar distributions.

The average JSD of the human parallel and longitudinal distributions at each referring expression in the VaREG corpus is 0.096. That both CatBoost and XGBoost have smaller distances than these to both the parallel and the longitudinal distributions suggests that these have homed in on distributions that fall between these two.

5.3. WSJ distributional results

Table 6 shows the results of constructing longitudinal human variation distributions on the much larger WSJ corpus.

WSJ	Longitudinal
RF	0.00304
XGBoost	0.00317
CatBoost	**0.00296**

Table 6: JSD divergences between the trained algorithms and longitudinal REF distributions. Lower divergence values correspond to more-similar distributions.

Where we see approximately 8-10% differences between algorithms in the VaREG longitudinal comparison, this number drops to around 3-4% in the WSJ data. At the same time, the divergences have become an order of magnitude smaller. This is paralleled by the divergences between the predictions made by the different algorithms. In application to the VaREG corpus, the difference in the predicted distributions of RF and CatBoost (the most different) was 0.022, while for WSJ the corresponding difference was 0.00082. This was the largest difference between predicted distributions. As the corpus size increases, the added training data results in greater similarity of the predictions made by the three algorithms. At the same time, they all converge more closely on the longitudinal distributions, which, we suggest, reflect real human variation.

6. Discussion

This study has made use of distributional information in longer corpora to infer variation in human REF

choices. Such an option obviates the need for expensive parallel human annotation of corpora, such as was undertaken for a small corpus by Castro Ferreira et al. (2016a).

There is an important reason for wanting human distributional information in the evaluation of REF prediction systems: Language is inherently non-deterministic. Speakers do not produce the same REF for the same contexts. While we have no direct evidence for this, we would speculate that the same speaker would not necessarily give the same RE the same form if asked to annotate it twice. Viethen and Dale (2006) express it this way, although we would also add the qualification *in the same context*:

> Not only do different people use different referring expressions for the same object, but the same person may use different expressions for the same object on different occasions. (Viethen and Dale, 2006, p. 119)

To the extent that this is true, it suggests a failure in how we perform evaluation of REF prediction systems. Feature-based models such as the ones considered in this paper produce non-deterministic predictions. It is appropriate, therefore, that they be evaluated by comparison with a distribution, rather than by accuracy as if they were deterministic predictors.

If we continue to use an evaluation measure in which the maximum a posteriori probability is used as a single deterministic prediction from an algorithm, then certain problems with evaluation will persist. Inherent natural variation means that there is a ceiling of accuracy which no predictive system can transcend. In contrast, if we evaluate by comparing distributions, this problem does not arise.

We have taken up the choice of the JSD as a metric for comparing distributions as used by Castro Ferreira et al. (2016b). This metric has firm foundations with information theory, being a symmetrisation of the Kullback-Liebler distance, which is tied to both Shannon Information on the one hand and Fisher Information (Ly et al., 2017) on the other.

Ideally, we would evaluate the longitudinal distributional corpus against a VaREG-style parallel corpus. However, this only makes sense if the corpora are of comparable genres. VaREG contains a combination of non-technical news articles, product and media reviews, and wikipedia articles. In contrast the WSJ contains solely financial news reports. In future work, we propose to construct a VaREG like corpus of parallel human judgements over a subset of the WSJ, to enable a more coherent comparison with other analyses using the WSJ.

If it is important to evaluate algorithms against distributions of human choices, then we need appropriate corpora in which such choices can be found. These will be corpora ideally like the VaREG corpus, with multiple choices made by informants of the same reference in the same context. There are, unfortunately, a few problems with VaREG which make it less than ideal in diverse contexts. The size of the corpus is relatively small. The corpus itself has only 563 instances of annotated referring expressions. The second issue with VaREG is that only references to the main topic of the text are annotated in this corpus. One might wonder, whether the referential form choices of the participants would have been different if they were selecting types for all the referring expressions in the text, much as authors need to do. However, substantial corpora making up for the shortcomings of VaREG are unlikely to be available anytime in the near future.

The goal of the current paper, and future work in this area, is to create ersatz parallel annotations by aggregating referring expressions in similar contexts within a large corpus. We have dubbed the distributions resulting from such an analysis as *longitudinal*.

One indicator of the importance of larger corpora for evaluation can be seen in the fact that the algorithms trained on half of the WSJ did well in predicting the longitudinal distributions in the other half. As we saw in section 5.3 the divergence from the longitudinal distribution was approximately 30 times larger in the small VaREG corpus than in the much larger WSJ.

We have seen a growing interest and attention to the human evaluation of the output of the NLG systems (Howcroft et al., 2020; van der Lee et al., 2019). These studies have human participants judge various quality criteria (e.g. fluency, clarity, and naturalness) of the system outputs. There is a difference between these human judgements and the ones addressed in this study (to which less attention has been paid). While the former evaluate the output of the ML systems, the latter offer alternative outputs to the original corpus' data, against which the non-deterministic predictions of the ML models are evaluated.

One concern that may be raised is that the use of longitudinal distributions for evaluation may appear circular. The longitudinal distributions we construct for comparison with algorithm predictions are defined in terms of the combinations of feature values. These feature values themselves condition the distributions over REF predicted by the learning algorithms. However, there is no circularity for two reasons. The first reason is that we only use this approach for the evaluation of machine learning algorithms, and not for the evaluation of feature sets. The second reason is that while the learning algorithms are trained on the training set, the longitudinal distributions for

evaluation are constructed from the test set.

7. Conclusion

In this paper, we have approximated parallel REF selection by longitudinal construction of distributions. It is, of course, still necessary to validate this approach in an experiment where human participants reproduce a small section of a corpus, such as the WSJ. Our hypothesis is that the evaluations of algorithms based on such parallel annotations will match those obtained by comparison with longitudinal distributions.

While we see a lot of advantages in using longitudinal corpora as proxies for parallel human corpora, some caveats must be born in mind. If there are feature–value combinations present in only a few examples in the long corpus, giving only an approximate picture of the distribution reflected by those combinations, then it may not be possible to make good predictions about the referring expressions with those feature–value combinations. Usually, however, the longer the corpus, the more likely we are to obtain an adequate sample of referring expression form for each feature–value combination.

8. Acknowledgements

This work was funded by the German Research Foundation (DFG)– Project-ID 281511265 – SFB 1252 "Prominence in Language".

9. Bibliographical References

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. Routledge.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87.

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

Belz, A. and Varges, S. (2007). Generation of repeated references to discourse entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 9–16.

Belz, A., Kow, E., Viethen, J., and Gatt, A. (2009). Generating referring expressions in context: The GREC task evaluation challenges. In *Empirical methods in natural language generation*, pages 294–327. Springer.

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

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

Cao, M. and Cheung, J. C. K. (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.

Castro Ferreira, T., Krahmer, E., and Wubben, S. (2016a). Individual variation in the choice of referential form. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 423–427, San Diego, California, June. Association for Computational Linguistics.

Castro Ferreira, T., Krahmer, E., and Wubben, S. (2016b). 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.

Castro Ferreira, T., Moussallem, D., Kádár, Á., Wubben, S., and Krahmer, E. (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.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4).

Cunha, R., Castro Ferreira, T., Pagano, A., and Alves, F. (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), December. International Committee on Computational Linguistics.

van Deemter, K. (2016). *Computational models of referring: a study in cognitive science*. MIT Press.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Fukumura, K. and van Gompel, R. P. (2011). The effect of animacy on the choice of referring expression. *Language and cognitive processes*, 26(10):1472–1504.

Gatt, A., van Gompel, R. P., van Deemter, K., and Krahmer, E. (2013). Are we bayesian referring expression generators. Cognitive Science Society.

Greenbacker, C. and McCoy, K. (2009). UDel: Generating referring expressions guided by psycholinguistc findings. In *Proceedings of the 2009 Work-*

shop on Language Generation and Summarisation (UCNLG+Sum 2009), pages 101–102, Suntec, Singapore, August. Association for Computational Linguistics.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Hendrickx, I., Daelemans, W., Luyckx, K., Morante, R., and Van Asch, V. (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.

Henschel, R., Cheng, H., and Poesio, M. (2000). Pronominalization revisited. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 306–312. Association for Computational Linguistics.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland, December. Association for Computational Linguistics.

Kibrik, A. A., Khudyakova, M. V., Dobrov, G. B., Linnik, A., and Zalmanov, D. A. (2016). Referential choice: Predictability and its limits. *Frontiers in psychology*, 7(1429).

Krahmer, E. and Theune, M. (2002). Efficient context-sensitive generation of referring expressions. *Information sharing: Reference and presupposition in language generation and interpretation*, 143:223–263.

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

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., and Wagenmakers, E.-J. (2017). A tutorial on Fisher Information. *Journal of Mathematical Psychology*, 80:40–55.

McCoy, K. E. and Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description? In *The Relation of Discourse/Dialogue Structure and Reference*.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, pages 6638–6648.

Reiter, E. and Dale, R. (2000). *Building natural language generation systems*. Cambridge university press.

Same, F. and van Deemter, K. (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), December. International Committee on Computational Linguistics.

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., and Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, October–November. Association for Computational Linguistics.

van Gompel, R. P. G., van Deemter, K., Gatt, A., Snoeren, R., and Krahmer, E. J. (2019). Conceptualization in reference production: Probabilistic modeling and experimental testing. 126(3):345–373.

Viethen, J. and Dale, R. (2006). Towards the evaluation of referring expression generation. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 115–122, Sydney, Australia, November.

von Heusinger, K. and Schumacher, P. B. (2019). Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

10. Language Resource References

Castro Ferreira, Thiago and Krahmer, Emiel and Wubben, Sander. (2016). *Individual Variation in the Choice of Referential Form*. Association for Computational Linguistics.

Ralph Weischedel, et al. (2013). *OntoNotes Release 5.0 LDC2013T19*. Philadelphia: Linguistic Data Consortium, ISLRN 151-738-649-048-2.