Towards more variation in text generation: Developing and evaluating variation models for choice of referential form

Thiago Castro Ferreira and **Emiel Krahmer** and **Sander Wubben**
Tilburg center for Cognition and Communication (TiCC)
Tilburg University
The Netherlands
{tcastrof,e.j.krahmer,s.wubben}@tilburguniversity.edu

Abstract

In this study, we introduce a non-deterministic method for referring expression generation. We describe two models that account for individual variation in the choice of referential form in automatically generated text: a Naive Bayes model and a Recurrent Neural Network. Both are evaluated using the VaREG corpus. Then we select the best performing model to generate referential forms in texts from the GREC-2.0 corpus and conduct an evaluation experiment in which humans judge the coherence and comprehensibility of the generated texts, comparing them both with the original references and those produced by a random baseline model.

1 Introduction

Automatic text generation is the process of converting non-linguistic data into coherent and comprehensible text (Reiter and Dale, 2000). In recent years, interest in text generation has substantially increased, due to the emergence of new applications such as "robot-journalism" (Clerwall, 2014). Even though computers these days are perfectly capable of automatically producing text, the results are arguably often rather rigid, always producing the same kind and style of text, which makes them somewhat "boring" to read, especially when reading multiple texts in succession.

Human-written texts, by contrast, do not suffer from this problem, presumably because human authors have an innate tendency to produce variation in their use of words and constructions. Indeed, psycholinguistic research has shown that when speakers produce referring expressions in comparable contexts, they non-deterministically vary both the form and the contents of their references (Dale and Viethen, 2010; Van Deemter et al., 2012). In this paper, we present and evaluate models of referring expression generation that mimic this human non-determinacy and show that this enables us to generate varied references in texts, which, in terms of coherence and comprehensibility, did not yield significant differences from human-produced references according to human judges.

In particular, in this study we focus on the choice of referential *form*, which is the first decision to be made by referring expression generation models (Reiter and Dale, 2000) and which determines whether a reference takes the form of a proper name, a pronoun, a definite description, etc. Several such models have been proposed (Reiter and Dale, 2000; Henschel et al., 2000; Callaway and Lester, 2002; Krahmer and Theune, 2002; Gupta and Bandopadhyay, 2009; Greenbacker and McCoy, 2009). However, all of these are fully deterministic, always choosing the same referential form in the same context.

The fact that these models are generally based on text corpora which have only one gold standard form per reference (the one produced by the original author) does not help either. When the corpus contains, say, a description at some point in the text, this does not mean that, for example, a proper name could not occur in that position as well (Yeh and Mellish, 1997; Ferreira et al., 2016). Generally, we just don't know. To counter this problem, a recent corpus, called VaREG, was developed in which 20 different writers were asked to produce references for a particular topic in a variety of texts, giving rise to a distribution over forms per reference (Ferreira et al., 2016). This gives us the possibility to distinguish situations where there is more or less agreement between writers in their choices of referential form. But it also enables a new paradigm for choosing referential forms,

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 568–577,
Berlin, Germany, August 7-12, 2016.

where instead of predicting the most likely referential form, we can in fact predict the frequency in which a reference assumes a specific form, allowing us to turn the choice of referential form into a non-deterministic probabilistic model.

In this study, we introduce two different models that take the individual variation into account for the choice of referential form, one based on Naive Bayes and one on Recurrent Neural Networks. Both are evaluated using the VaREG corpus. Furthermore, we use the best performing model to generate referential forms in texts from the GREC-2.0 corpus, based on the roulette-wheel generation process (Belz, 2008), and conduct an evaluation experiment in which humans judge the coherence and comprehensibility of the generated texts, comparing them both with the original references and those produced by a random baseline model.

2 Related Studies

Several models for the choice of referential form have been proposed in the literature. They can roughly be distinguished in two groups: rule-based and data-driven models.

Many rule-based models were created for pronominalization, i.e, to choose whether an object or person should be referred to using a pronoun or not. Reiter and Dale (2000) proposed one of the first rule-based models, which opts for a pronominal reference only if the referent was previously mentioned in the discourse and no mention to an entity of same gender can be found between the reference and its antecedent. Henschel et al. (2000) presented a pronominalization model based on recency, discourse status, syntactic position, parallelism and ambiguity. To decide among a pronoun or a definite description, Callaway and Lester (2002) also proposed a rule-based model which makes the choices based on information about the discourse, rhetorical structure, recency and distance. Krahmer and Theune (2002) extended the Incremental algorithm so that if a referent achieves a level of salience in the discourse (measured by a salience weight), a pronoun is used. Otherwise, a definite description is produced to distinguish the referent from the distractors.

Aiming to make choices similar to humans, some studies proposed machine learning models trained on human choices of referential form. The GREC project (Belz et al., 2010) motivated the development of many of those data-driven models. One of the project's shared tasks aimed to predict the form of the references to the main topics of texts taken from Wikipedia. Among the participants of the task, Gupta and Bandopadhyay (2009) presented a model that combined rules and a machine learning technique based on semantic and syntactic category, paragraph and sentence positions, and reference number. Similarly, Greenbacker and McCoy (2009) proposed a decision tree that, besides the features used in Gupta and Bandopadhyay (2009), was also based on recency and part-of-speech features. For more information on the GREC shared task, see Belz et al. (2010).

One limitation that these models all have in common is that they fail to model individual variation. According to their predictions, a reference will always assume the most likely referential form. For example, a model that takes into account syntactic position will always choose the same referential form for the subject of a sentence, while humans tend to vary in their choices of referential form. One of the reasons for this problem arises from the data these models are trained on. Most corpora only contain one referring expression per reference. Only the newly introduced VaREG corpus takes variation into account, containing 20 different expressions for each reference, allowing us to model distributions over referential slots.

3 The VaREG corpus

The VaREG corpus was collected for the study of individual variation in the choice of referential form (Ferreira et al., 2016). The corpus is based on a number of texts, which were presented to participants in such a way that all references to the main topic of the text had been replaced with gaps. Each participant was asked to fill each of those gaps with a referring expression for the topic.

The resulting corpus consists of 9,588 referring expressions, produced by 78 participants for 563 referential gaps - around 20 referring expressions per reference - in 36 English texts. The texts were equally distributed over 3 genres: news texts, reviews of commercial products and encyclopedic texts. The references were annotated according to their syntactic position (subject, object, etc.), referential status (new or old, in text, paragraph and sentence) and recency (number of words between previous reference to the same object or entity),

and the referring expressions of the participants were classified into 5 referential forms: proper names, pronouns, definite descriptions, demonstratives and empty references.

The analysis of the corpus revealed considerable variation among participants in their choices of referential forms. Various factors influenced the amount of variation that occurred. High amounts of variation, for example, were found in product reviews and also in the object position of sentences. Besides allowing us to distinguish between situations with relatively high and relatively low individual variation in choices of referential form, this corpus introduces a new paradigm for the development and evaluation of models for referential choice. Rather than predicting the most likely form of a reference, as is usually done, the new corpus allows us to develop a model that can predict the frequency with which a particular reference can assume different referential forms. In this study, we explore this possibility.

4 Models

We model the individual variation in the choice of referential form in the following way: each reference consists of a tuple (X, y), where X is the set of feature values that describes the reference and y is a distribution of referential forms that indicates the frequency (in proportion) in which X assumes each form. So given X, we expect to find a distribution $\hat{y}$ similar to y.

Table 1 depicts the features used to describe X. The influence of those discourse factors in the choice of referential form has been often studied in the literature. Concerning syntactic position, Brennan (1995) argued that references in the subject position of a sentence are more likely to be shorter than references in the the object position. In favor of status and recency, Chafe (1994) showed that references to previously mentioned referents in the discourse and ones that are close to their antecedents are more likely to be shorter than references to new referents or ones that are distant from their antecedents.

All features were defined categorically, including the recency. This latter is treated by describing if a reference's antecedent is 10 or less words away, between 11 and 20 words, between 21 and 30 words, between 31 and 40 words and more than 40 words away.

To predict a distribution $\hat{y}$ based on X, we propose two models: a Naive Bayes and a Recurrent Neural Network.

4.1 Naive Bayes

Given a set of referential forms F, the probability that a reference assumes a particular form $f \in F$ according to this model is given by:

$$P(f \mid X) \propto \frac{P(f) \prod_{x \in X} P(x \mid f)}{\sum_{f' \in F} P(f') \prod_{x \in X} P(x \mid f')} \quad (1)$$

To avoid zero probabilities, we used additive smoothing with $\alpha = 2e^{-308}$. So given a reference described by X, $\hat{y}$ is the distribution over F:

$$\hat{y} = \begin{bmatrix} P(f_1 \mid X) \\ ... \\ P(f_{|F|} \mid X) \end{bmatrix} \quad (2)$$

4.2 Recurrent Neural Network

Some referential theories support the idea that a referential form is chosen based on previous choices to the same referent. Arnold (1998) argued that subjects of a sentence are more likely to be later pronominalized, as well as references in parallel syntactic position with their antecedents. Chafe (1994) sustained that referents mentioned in recent clauses also tend to be pronominalized. Since Naive Bayes does not take into account the sequential nature of text, we use a Recurrent Neural Network (RNN) to be able to take context into account.

RNN is a powerful structure to handle sequences of data. It can map a sequence of references $(X_1, ..., X_t)$ to their referential forms distributions $(y_1, ..., y_t)$ based on the previous steps.

Our approach here is similar to the one presented by Mesnil et al. (2013). But instead of word continuous representations, a referential embedding is created for each combination of feature values in X. So given a reference X_t and a context window size win, the embeddings of the references $X_{t-win/2}^{t-1}$, X_t and $X_{t+1}^{t+win/2}$ are merged to form a representation e_t. This representation is used in equations 3 and 4 to find a distribution over the referential forms that X_t could assume.

$$h_t = sigmoid(W^{hx} e_t + W^{hh} h_{t-1}) \quad (3)$$

$$\hat{y}_t = softmax(W^{yh} h_t) \quad (4)$$

Feature	Description
Syntactic position	Subject, object or a genitive noun phrase in the sentence.
Referential Status	First mention to the referent (new) or not (old) at the level of text, paragraph and sentence.
Recency	Distance between a given reference and the last, previous reference to the same referent.

Table 1: Features used to describe the references.

We assume a sequence of tuples $\{(X_1, y_1)..., (X_t, y_t)\}$ as all the references to a referent throughout a text.

We trained our RNN using Backpropagation Through Time. To measure the error among y and $\hat{y}$, we use cross entropy as a cost function. The values for the remaining parameters of the RNN are introduced in Table 2. We chose them based on an ad-hoc analysis, where we searched for an optimal combination to obtain the best predictions.

Batch Size	10
Context Window Size	3
Epochs	15
Embedding Dimension	50
Hidden Layer Size	50
Learning Rate	0.1

Table 2: RNN Settings

5 Individual Variation Experiments

For each reference slot encountered in the VaREG corpus, we evaluated how well a model takes the individual variation into account in the choice of referential form by comparing its predicted distribution of referential forms ($\hat{y}$) with the real distribution (y). We performed this comparison through two experiments.

In the first, the models were trained and tested with VaREG corpus. In the second, we aimed to check to what extent the referring expressions from the GREC-2.0 corpus are similar in form to the referring expressions from VaREG corpus by training the models with the first corpus and testing with the second.

5.1 Method

4-fold-cross-validation was used to train the models in the first experiment. The number of folds was chosen based on the set-up of the VaREG corpus, which consists of 4 groups of texts. Given the structure of the corpus, we decided that training our model with 3 groups of texts and testing it on the held-out group was the most natural solution to avoid overfitting. Each fold has the same amount of texts per genre.

Unlike VaREG, GREC-2.0 corpus does not have a set of referring expressions for the exact same reference. So, in the second experiment, the referential form distributions y were defined globally by grouping the references by X and computing the frequency of each referential form.

We also re-annoted the GREC-2.0 corpus to make it compatible with the VaREG corpus. In particular, we added features for status and recency to the GREC-2.0 corpus and made the terminology consistent beween the two corpora[1]. Both the VaREG corpus and the re-annotated GREC-2.0 corpus are publicly available[2].

5.2 Metrics

For each reference, Jensen-Shannon divergence (Lin, 1991) was used to measure the similarity between y and $\hat{y}$:

$$JSD(y||\hat{y}) = \frac{1}{2}D(y||m) + \frac{1}{2}D(\hat{y}||m) \quad (5)$$

$$\text{where } m = \tfrac{1}{2}(y + \hat{y})$$

In this measure, D is the Kullback-Leibler divergence (Kullback, 1968). The Jensen-Shannon divergence ranges from 0 to 1, in which 0 indicates full convergence of the two distributions and 1 full divergence. Therefore, a lower number indicates a better individual variation modeling.

To check the behaviour of $\hat{y}$ based on y in each reference, the referential forms of both distributions were ranked and their relation were analysed with the Spearman's rank correlation coefficient. This measure ranges between -1 and 1, where -1 indicates a fully opposed behaviour among the variables and 1 the exact same behaviour among them. 0 indicates a non-linear correlation among the involved variables.

[1] Texts also used in VaREG had their references removed from the GREC-2.0 version used in here.

[2] http://ilk.uvt.nl/~tcastrof/acl2016

5.3 Baselines

We considered two baseline models in the experiments. The first, called *Random*, assumes $\hat{y}$ as a random distribution of forms for each reference.

The second model, called *ParagraphStatus*, always chooses a proper name when the reference is to a new topic in the paragraph (the distribution will assume the value 1 to the proper name form and 0 to the others), and a pronoun otherwise (value 1 to the pronoun form and 0 to the others).

5.4 Results

5.4.1 Cross-validation on VaREG corpus

Models	JSD	$\rho_{y,\hat{y}}$
Random	0.63	-0.01
ParagraphStatus	0.43	0.66
NB+Syntax−Status−Recency	0.39	0.69
NB−Syntax+Status−Recency	0.32	**0.75**
NB−Syntax−Status+Recency	0.41	0.68
NB+Syntax+Status−Recency	**0.31**	**0.75**
NB+Syntax−Status+Recency	0.38	0.70
NB−Syntax+Status+Recency	0.33	0.73
NB+Syntax+Status+Recency	**0.31**	0.74
RNN+Syntax−Status−Recency	0.37	0.71
RNN−Syntax+Status−Recency	0.36	0.72
RNN−Syntax−Status+Recency	0.40	0.70
RNN+Syntax+Status−Recency	0.33	0.73
RNN+Syntax−Status+Recency	0.37	0.71
RNN−Syntax+Status+Recency	0.36	0.72
RNN+Syntax+Status+Recency	0.33	0.72

Table 3: Average Jensen-Shannon divergence and Spearman's correlation coefficient of the models in Experiment 1.

Table 3 depicts the Jensen-Shannon divergence and Spearman's correlation coefficient of the models cross-validated on VaREG corpus. All our models outperformed the baselines.

Considering the models in which the references are described by only one kind of feature, it seems that the status features (+Status) are the ones that best contributed to model the individual variation in the choice of referential form, whereas the recency (+Recency) is the worst. Syntactic position is sandwiched among the previous two.

In the comparison within Naive Bayes and RNN models, the ones in which the references are described by syntactic position and referential status (+Syntax+Status−Recency) obtained the best results for both measures. Figure 1 depicts the average Jensen-Shannon divergences by genre of Naive Bayes and RNN models in which the references are described by this combination of features. Both models presented the best results in

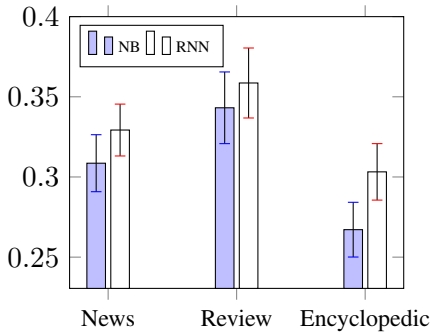


Figure 1: Jensen-Shannon divergence of NB+Syntax+Status−Recency (NB) and RNN+Syntax+Status−Recency (RNN) by genre in Experiment 1. Error bars represent 95% confidence intervals.

encyclopedic texts, and the worst in product reviews.

Although RNNs are able to model the individual variation in a reference based on its antecedents, they did not introduce significantly better results than Naive Bayes. In fact, NB+Syntax+Status−Recency is significantly better than RNN+Syntax+Status−Recency in modeling the individual variation in news (Wilcoxon $Z = 11574.5$, $p < 0.01$) and encyclopedic texts (Wilcoxon $Z = 4232.5$, $p < 0.001$).

5.4.2 Training on GREC-2.0 and evaluating on VaREG corpus

Models	JSD	$\rho_{y,\hat{y}}$
Random	0.63	-0.01
ParagraphStatus	0.43	0.66
NB+Syntax+Status−Recency	**0.36**	**0.67**
NB+Syntax+Status+Recency	0.37	0.64
RNN+Syntax+Status−Recency	0.37	0.62
RNN+Syntax+Status+Recency	0.37	0.64

Table 4: Average Jensen-Shannon divergence and Spearman's correlation coefficient of the models in Experiment 2.

Table 4 shows the results of models trained with GREC-2.0 and tested with VaREG corpus. These models are the two versions of Naive Bayes, and the two versions of RNN which were best evaluated in the previous experiment.

The results of this experiment follow the results of the previous one. Our models outperformed the baselines and NB+Syntax+Status−Recency was the model that obtained the best results for both measures.

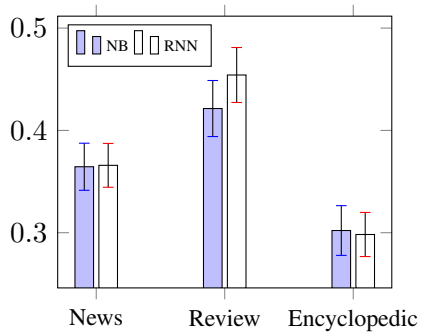


Figure 2: Average Jensen-Shannon divergence of NB+Syntax+Status−Recency (NB) and RNN+Syntax+Status-Recency (RNN) by genre in Experiment 2. Error bars represent 95% confidence intervals.

Figure 2 depicts the Jensen-Shannon divergence measures of models NB+Syntax+Status-Recency and RNN+Syntax+Status-Recency by text genre. As in the previous experiment, both Naive Bayes and RNN models best modeled the individual variation in encyclopedic texts. Moreover, there was not significant difference among NB+Syntax+Status-Recency and RNN+Syntax+Status-Recency in the three text genres.

In general, the models trained with VaREG corpus seemed to model the individual variation in the choice of referential form better than the models trained with GREC-2.0 corpus.

6 Coherence and comprehensibility of the texts

In this section, we investigate to what extent texts generated by our method, including variation of referential form, are judged coherent and comprehensible by readers. We do this by comparing texts from the GREC-2.0 corpus in which all references were (re)generated using our method, with the original text and with a variant that includes random variation of referential form.

6.1 Our model for choice of referential form

To generate the referring expressions for the topic of a given text of GREC-2.0, we first group all references by syntactic position and referential status values. Then for each group, we shuffle the references and choose their forms according to the distribution predicted by our best performing model (the NB+Syntax+Status−Recency trained on VaREG). The choice of referential forms follows the roulette-wheel generation process (Belz, 2008). This process entails that if a group has 5 references and our model predicts a distribution of 0.75 proper names and 0.25 pronouns, 4 references of the group will be proper names and 1 a pronoun.

This covers the selection of referential forms (deciding which form to use at which particular point in the text). To deal with their linguistic realisation, we implemented the following heuristics. For the cases in which a proper name reference is selected, we choose a realization depending on referential status. If the reference is the first mention to the topic in the text, the reference is realized with the topic's longest proper name. Otherwise, the reference is realized with its shortest proper name. For the cases in which a definite description is selected, but where the original GREC-2.0 corpus does not provide a description for the topic, we select the shortest predicate adjective of the first sentence of the text, immediately following the main verb. For instance, for the sentence *"Alan Mathison Turing was an English mathematician, logician, and cryptographer."*, the selected definite description would be "**The English mathematician**". In the cases where a reference should assume the form of a demonstrative, the definite article of the definite description is replaced by the demonstrative "this" (In the previous example, "**This English mathematician**").

6.2 Evaluation Method

We evaluated three versions of each text. The *Original* is the original text in the corpus, including the original referring expressions selected by the author. We compare this with a *Random* variant, which does include variation of referential forms, but selects them in a fully random way. Finally, in the third, *Generated* version, all references are generated according to the method outlined at Section 6.1. Table 5 depicts an example of text in the three versions.

In total, we make 3 versions of 9 pseudo-randomly selected texts (5 covering animate topics and 4 inanimate ones, varying in length) from the GREC-2.0 corpus, yielding 27 texts in total. These were distributed over 3 lists, such that each list contains one variant of each text, and there is an equal number of texts from the 3 conditions (*Original*, *Random*, *Generated*). In all texts, all

Version	Text
Original	**Spain**, officially the Kingdom of Spain, is a country located in Southern Europe, with two small exclaves in North Africa (both bordering Morocco). **Spain** is a democracy **which** is organized as a parliamentary monarchy. **It** is a developed country with the ninth-largest economy in the world. **It** is the largest of the three sovereign nations that make up the Iberian Peninsula–the others are Portugal and the microstate of Andorra.
Random	**It**, officially the Kingdom of Spain, is a country located in Southern Europe, with two small exclaves in North Africa (both bordering Morocco). **The country** is a democracy **that** is organized as a parliamentary monarchy. **It** is a developed country with the ninth-largest economy in the world. **This country** is the largest of the three sovereign nations that make up the Iberian Peninsula–the others are Portugal and the microstate of Andorra.
Generated	**Spain**, officially the Kingdom of Spain, is a country located in Southern Europe, with two small exclaves in North Africa (both bordering Morocco). **Spain** is a democracy **that** is organized as a parliamentary monarchy. **The country** is a developed country with the ninth-largest economy in the world. **It** is the largest of the three sovereign nations that make up the Iberian Peninsula–the others are Portugal and the microstate of Andorra.

Table 5: Example of text in the Original, Random and Generated version.

references to the topic were highlighted in yellow. The experiment was run on CrowdFlower and is publicly available[3].

The experiment was performed by 30 participants (10 per list). Their average age was 36 years, and 22 were female. All were proficient in English (the language of the experiment), 26 participants were native speakers. They were asked to rate each text in terms of how coherent and comprehensible they considered it, on a scale from 1 (Very Bad) to 5 (Very Good).

6.3 Results

Figure 3 depicts the average coherence and comprehensibility of the texts where their topics are described by the *Original*, *Random* and *Generated* approaches, respectively. Inspection of this Figure clearly shows that the *Random* texts are rated lower than both the *Original* and the *Generated* texts, and that the latter are rated very similarly on both dimensions.

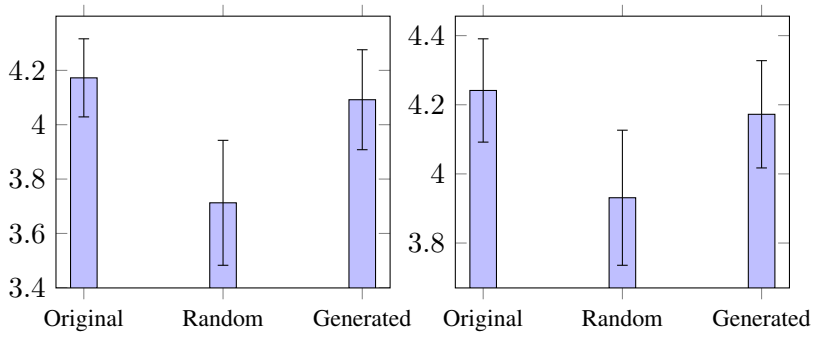
This is confirmed by the statistical analysis. According to a Friedman test, there is statistically significant difference in the coherence ($\chi^2 = 11.79$, $p < 0.005$) and comprehensibility ($\chi^2 = 8.98$, $p = 0.01$) for the three kinds of texts. We then conducted a post hoc analysis with Wilcoxon signed-rank test corrected for multiple comparisons using the Bonferroni method, resulting in a significance level set at $p < 0.017$. Texts of the *Original* approach are statistically more coherent ($Z = 322$, $p < 0.017$) and comprehensible ($Z = 407.5$, $p < 0.017$) than texts of the *Random* one. Texts of the *Generated* approach are also statistically more coherent ($Z = 275$, $p < 0.017$), but not more comprehensible ($Z = 378$, $p < 0.05$) than texts of the *Random* one. Finally, and crucially, comparing *Original* and *Generated* texts revealed no significant differences for coherence ($Z = 540$, $p < 0.5$) nor for comprehensibility ($Z = 391.5$, $p < 0.5$).

[3] http://ilk.uvt.nl/~tcastrof/acl2016

7 Discussion

In this paper we explored the possibilities of introducing more variation in automatically generated texts, by trying to model individual variation in the selection of referential form. We relied on a new corpus (VaREG (Ferreira et al., 2016)), which does not contain a single expression for each reference in a text, but rather a distribution of referential forms produced by 20 different people. In contrast to earlier models for referential choice which always deterministically choose the most likely form of a reference, we proposed a Naive Bayes and a Recurrent Neural Network model which aimed to predict the frequency distribution with which a reference can assume a specific referential form, based on discourse features including syntactic position, referential status and recency. Given a reference, we evaluated how well each different model could capture the individual variation found in the VaREG corpus by comparing its predicted distribution of referential forms with the real one in the corpus. We trained the models in two different ways: first using the VaREG, and second using the GREC-2.0 corpus. The Naive Bayes model, trained on VaREG corpus, in which the references were described by syntactic position and referential status features was the one that best modeled the individual variation in the choice of referential form.

Features Referential status features were the most helpful for modeling the individual variation in the choice of referential form. They were fol-



(a) Coherence (b) Comprehensibility

Figure 3: Average coherence (3a) and comprehensibility (3b) of the texts with the original, randomized and generated referring expressions. Error bars represent 95% confidence intervals.

lowed by the syntactic position feature. Both of these findings are consistent with the observations about human variation in the selection of referential forms, as discussed by Ferreira et al. (2016). This study argued that writers are more likely to vary in their choices when a reference is in the object position, and when it is an old mention in the text, but new in the sentence. Recency was not a helpful feature for our models, and this may be due to the way the feature was represented - i.e., as a categorical rather than a continuous feature. Moreover, the recency feature was measured in terms of words between the current reference and the most recent previous one to the same referent. Perhaps, it would be better to measure recency in terms of different discourse entities mentioned between two references to the same referent.

Genre In agreement with Ferreira et al. (2016), we also found that genre mattered. For modeling variation, our models performed best when applied to encyclopedic texts, and worst in product reviews, with news sandwiched in between.

Naive Bayes model vs. RNNs Although the RNNs were able to model individual variation in the choice of referential form to some extent, they did not perform significantly better than the Naive Bayes models, which might have to do with the relatively small dataset. However, we think the size of the corpus matches the relatively low complexity of the problem we address. In the most complex case (i.e., when a reference is described by its syntactic position, status and recency), an input can be represented in 120 different ways to predict a multinomial distribution of size 5 (number of referential forms). This complexity is much smaller than other problems typically modeled by RNNs. In text production, for instance, an input may be represented by thousands of words to predict a large multinomial distribution over a vocabulary (Sutskever et al., 2014). Additionally, it is important to stress that we actually have a real multinomial distribution to compare with the distribution predicted by the RNN in each situation. We observed that it is possible to compute more fine-grained error costs in our case, which makes the RNN converge faster when it is backpropagated. In sum, we believe that those two factors combined compensate for the size of the dataset. A possible explanation for the non-difference among the Naive Bayes model and RNNs is the use of the referential status features, which perhaps are already enough to model the relation among a reference and its antecedents.

VaREG corpus vs. GREC-2.0 corpus Interestingly, our proposed models yielded better performance when trained on the VaREG than on the GREC-2.0 corpus. This shows a difference among the referential choices of both corpora. We conjecture this difference is partly due to differences in text genres, since the VaREG corpus contains texts from three different genres, whereas the GREC-2.0 corpus only has encyclopedic texts. Earlier work has also highlighted the influence of text genre on the amount of individual variation in writers' choices for referential forms (Ferreira et al., 2016).

Coherence and comprehensibility In the second part of the study, we used the best performing model to generate referential forms in texts from the GREC-2.0 corpus, using a roulette-based model sampling from the predicted distributions over referential forms. We evaluated the texts gen-

erated in this way in an experiment in which humans were asked to judge the coherence and comprehensibility of the generated texts, comparing them both with the original references and those produced by a random baseline model. In terms of coherence and comprehensibility, we found that the texts in which the references were generated by our model were not significantly different than the human generated ones, and significantly better than the randomly generated ones. This shows that our solution does not only model the individual variation in the choice of referential form, but that this also does not negatively affect the quality of the texts. This is an important step towards developing new models for automatic text generation that are less predictable and more varied.

Acknowledgments

This work has been supported by the National Council of Scientific and Technological Development from Brazil (CNPq).

References

Jennifer E Arnold. 1998. *Reference form and discourse patterns*. Ph.D. thesis, Stanford University Stanford, CA.

Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Nat. Lang. Eng.* 14(4):431–455.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Empirical methods in natural language generation. Springer-Verlag, Berlin, Heidelberg, chapter Generating Referring Expressions in Context: The GREC Task Evaluation Challenges, pages 294–327.

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

Charles B. Callaway and James C. Lester. 2002. Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 88–95.

Wallace L. Chafe. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Christer Clerwall. 2014. Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice* 8(5):519–531.

Robert Dale and Jette Viethen. 2010. Attribute-centric referring expression generation. In *Empirical methods in natural language generation*, Springer, pages 163–179.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. 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*. Association for Computational Linguistics, San Diego, California.

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

Samir Gupta and Sivaji Bandopadhyay. 2009. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, UCNLG+Sum '09, pages 103–104.

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

Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information sharing: Reference and presupposition in language generation and interpretation*, CSLI, Stanford, CA, pages 223–264.

Solomon Kullback. 1968. *Information theory and statistics*. Courier Corporation.

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

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and

learning methods for spoken language understanding. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. pages 3771–3775.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge University Press, New York, NY, USA.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. pages 3104–3112.

Kees Van Deemter, Albert Gatt, Roger PG van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in cognitive science* 4(2):166–183.

Ching-Long Yeh and Chris Mellish. 1997. An empirical study on the generation of anaphora in chinese. *Comput. Linguist.* 23(1):171–190.