Generating flexible proper name references in text: Data, models and evaluation

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

This study introduces a statistical model able to generate variations of a proper name by taking into account the person to be mentioned, the discourse context and variation. The model relies on the REGnames corpus, a dataset with 53,102 proper name references to 1,000 people in different discourse contexts. We evaluate the versions of our model from the perspective of how human writers produce proper names, and also how human readers process them. The corpus[1] and the model[2] are publicly available.

1 Introduction

In automatic text generation, Referring Expression Generation (REG) is the task responsible for generating references to discourse entities, addressing, for example, the question whether the text should refer to an entity using a definite description (*the West Coast poet and patron saint of drinking writers*), a pronoun (*he*) or a proper name (*Henry Charles Bukowski*). REG is among the tasks which have received most attention in text generation (see Krahmer and van Deemter (2012), for a survey), but the vast majority of the research has concentrated on the generation of descriptions, while proper name generation has received virtually no attention, albeit with notable exceptions (Siddharthan et al., 2011; van Deemter, 2016) to which we return below.

Still, proper names occur frequently in texts. For instance, Ferreira et al. (2016a) showed that human writers use proper names in 91% of the cases to initially refer to persons. Indeed, some earlier research on text generation has stated that discourse-new references should be generated by using the strategy to "simply give the name of the object (if it has a name)" (Reiter and Dale, 2000). However, the *Bukowski* example already indicates that this is not as straightforward as Reiter and Dale suggest - the poet's full name is *Henry Charles Bukowski* and his birth name is *Heinrich Karl Bukowski*, but he is more commonly known as simply *Charles Bukowski*; see also van Deemter (2016), for a discussion of this and other complicating factors in proper name generation. In addition, Reiter and Dale (2000) do not address how repeated references using a name in a text should be generated. For instance, should our discourse-old example-writer be referred to as *Charles*, *Bukowski* or some combination of these and other attributes (e.g., using a modifier like *the poet Bukowski*)?

Imagine, for the sake of argument, that we would generate proper name references in a text by initially generating the full name, after which repeated references only consist of the last name (a.k.a. the family or surname). Intuitively, it is not difficult to come up with counterexamples to this "rule". Above we already discussed the difficulties of deciding what the most appropriate full name reference is for *Henry Charles Bukowski*, which (like *Keith Rupert Murdoch* and *Walter Bruce Willis*) seems to be the combination of middle and last names (as opposed to *Oprah Gail Winfrey* and *Serena Jameka Williams*, for who it is more common the combination of first and last names). Moreover, using the last name for repeated references may work well for the likes of *Winston Churchill* and *Angela Merkel*, but seems less suitable for *Napoleon Bonaparte* or *Madonna Ciccone*, to mention just two. Moreover, our example rule cannot account for the occurrence of modifiers. And, finally, it seems highly unlikely

[1] http://ilk.uvt.nl/~tcastrof/regnames/
[2] http://github.com/ThiagoCF05/ProperName

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 655–664,
Valencia, Spain, April 3-7, 2017.

that human writers would adhere to such a strict rule. Rather, one might expect writers to vary in their choices of which name to use, depending on stylistic and discourse factors, much like the choice of referential form varies as a function of such factors (Ferreira et al., 2016a; Ferreira et al., 2016b).

In general, we know very little about how proper names should be generated in text – as far as we know, there have been hardly any systematic corpus studies and only very little concrete proposals on how to automatically generate proper name references. In this paper, we therefore present a large scale corpus analysis, and, based on this, two versions of a new probabilistic model of proper name generation: one that always chooses the most likely proper name form and one that relies on a 'roulettewheel' selection model and hence will generate more varied references. These models rely both on the nature of the entity referred to (what is the likelihood that a given person will be referred to using, say, the first or last name?) and on the discourse context for generating proper name references in text. In an intrinsic evaluation experiment, we compare the performance of the two versions of this model with our implementations of the two proposals that have been made before (Siddharthan et al., 2011; van Deemter, 2016). We also describe a human evaluation experiment where we compare original texts with alternative versions that include proper names generated by our model.

2 Related work

Even though proper name references occur frequently in written text, their generation remains seriously understudied. A recent survey of REG models (Krahmer and van Deemter, 2012) has essentially nothing to say about the topic, and general surveys of automatic text generation such as Reiter and Dale (2000) only briefly mention a very basic rule (use a proper name, if available, for first references), without further specifying or evaluating it.

Recently, van Deemter (2016) has highlighted the importance of proper name generation. After discussing why a simple rule like the one proposed by Reiter and Dale cannot account for the complexities of proper name references in text, he argues that names could just be treated like other attributes in the generation of descriptions. Put differently, the name of an object can be modelled just like its color or size (typical attributes used in REG examples) – just as a description like *the tall man* rules out men that are not tall, so does a proper name like *Charles* rule out other people not named Charles. A standard REG algorithm, such as, for example, the Incremental Algorithm (Dale and Reiter, 1995) can then be used to compute when a name should be used and in which form. Van Deemter's work is of a theoretical nature; he has not implemented or tested this idea, so we cannot tell how well it can account for proper name references in text. In addition, in this form, his proposal cannot account for possible variations in proper name form throughout a text.

The most detailed study of proper name generation, as far as we know, is the seminal study by Siddharthan et al. (2011), which (re-)generates references to people in news summaries. For their algorithm(s), the authors present two manually constructed rules, based on earlier theories of reference, one for discourse-new references (including the full name) and one for discourse-old references (which in full says: "Use surname only, remove all pre- and post-modifiers."). They discuss, based on corpus analyses, how notions like discourse-new and discourse-old can be learned without manual annotation, and how they co-determine whether additional attributes such as role and affiliation should be included. Finally, they show that their model leads to improved (more coherent) summaries. While the approach offers a very interesting solution for the generation of discourse-new proper name references with modifiers for major characters in a news story (*Former East German leader Erich Honecker*), the proper name generation rule itself is very similar to the example rule discussed in the introduction (use the full name for discourse-new references and only the surname for discourse-old references). It is not specified how the full name should be realised (remember the *Henry Charles Bukowski*-example), and neither can the approach deal with exceptions to the surname-only rule (remember the *Madonna Ciccone*-example) or with intratext variation.

3 REGnames

For our explorations, we relied on the REGnames corpus (Ferreira et al., 2016c). REGnames is a corpus of 53,102 proper names referring to 1,000

people in 15,241 texts. The corpus consists of webpages extracted from the Wikilinks corpus (Singh et al., 2012), which was initially collected for the study of cross-document coreference and consists of more than 40 million references to almost 3 million entities in around 11 million webpages. All the references annotated in Wikilinks were grouped according to the Wikipedia page of the entity. This procedure enables easy identification of the mentioned entity and facilitates the extraction of more information about it.

To build the REGnames corpus, Ferreira et al. (2016c) selected the 1,000 most frequently mentioned people in the Wikilinks corpus. Then for each person, they selected random webpages from Wikilinks which mention the person at least once. On all selected webpages, part-of-speech tagging, lemmatization, named entity recognition, dependency parsing, syntactic parsing, sentiment analysis and coreference resolution was performed by using the Stanford CoreNLP software (Manning et al., 2014).

All extracted proper names were automatically annotated with their syntactic position (subject, object or genitive noun phrase in a sentence) and referential statuses in the text (discourse-new or discourse-old) and in the sentence (sentence-new or sentence-old). The extracted proper names were also annotated according to their form, i.e. which kind(s) of name (first, middle and/or last names), and modifier(s) (title and/or appositive) were part of the proper name. To check for the presence of first, middle and last names, a Proper Name Knowledge Base was extracted from DBpedia (Bizer et al., 2009) with all the names of the people in the corpus. Then, to check for the presence of a title or an appositive, named entity recognition information and the dependency tree were used respectively.

In the corpus analysis, Ferreira et al. (2016c) noticed that proper name references generally decrease in lengths across the text. They also concluded that a discourse-old or sentence-new proper name reference in the object position of a sentence tends to be shorter than a discourse-new or sentence-old proper name reference in the subject position of a sentence. In general, the corpus is a valuable resource which can be used to train a statistical model for proper name generation, as we show in the next section.

4 A model for proper name generation

Similarly to the generation of definite descriptions, our model produces a proper name reference in two sequential steps: content selection and linguistic realization.

4.1 Content Selection

The content selection discussed here is analogous to the selection of semantic attributes (type, color, size, etc) when generating a description of an entity (Dale and Haddock, 1991; Dale and Reiter, 1995). However, instead of attributes, the content selection step in our model aims to choose the *form* of a proper name reference (which kind(s) of name and modifier(s) are part of the proper name reference).

Features By analysing the REGnames corpus, Ferreira et al. (2016c) observed that proper names vary in their forms throughout a text. Moreover, as discussed in the Introduction (Section 1), a proper name form can also be influenced by the person to be mentioned. Thus, we conditioned the choice of a specific proper name form by a set of discourse features that describe the reference as well as to the person to be mentioned.

Table 1 depicts the discourse features used to describe the proper name references. We choose them based on the analysis of the REGnames corpus (Section 3).

Forms Our model selects a proper name form over all forms annotated on the REGnames corpus, i.e. a total of 28 possible ones. Table 2 depicts the most frequent ones. The complete list can be found at the webpage that describes the REGnames corpus[3].

Notation Given a person p to be referred to by his/her proper name and the set of discourse features D that describe the reference, we aim to predict the form $f \in F$ of a proper name as Equation 1 shows.

$$P(f \mid D, p) = \frac{P(f \mid p) \prod_{d \in D} P(d \mid f, p)}{\sum_{f' \in F} P(f' \mid p) \prod_{d \in D} P(d \mid f', p)} \quad (1)$$

To account for unseen data, the conditional probabilities are computed using the additive

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

Feature	Description
Syntactic Position	Subject, object or a genitive noun phrase in the sentence.
Referential Status	First mention of the referent (new) or not (old) at the level of text and sentence.

Table 1: Discourse features that describe the references.

smoothing technique with $\alpha = 1$. Equations 2 and 3 summarize the procedure.

$$P(f \mid p) = \frac{count(f \cap p) + \alpha}{count(p) + \alpha|F|} \quad (2)$$

$$P(d \mid f, p) = \frac{count(d \cap f \cap p) + \alpha}{count(f \cap p) + \alpha|D|} \quad (3)$$

Variation Besides the fact that proper name references may vary in their forms throughout a text and according to the person to be referred to, they may also vary in similar situations of a text. In an extrinsic evaluation comparing human- and machine-generated summaries, for instance, Siddharthan et al. (2011) reported that the lack of variation in the form of discourse-old proper names references was one of the disadvantages of their summarization system in the cases where human summaries were chosen. Our model fills this gap by performing Equation 1 over all the proper name forms given a set of similar references. That is proper name references to the same person and described by the same set of discourse feature values. This procedure results in a frequency distribution over all relevant proper name forms. Then, similar to the rouletewheel selection of Ferreira et al. (2016b) for the choice of referential forms, we can randomly apply the frequencies into a group of similar references in such a way that their forms will be representative of the distribution predicted by the model. For instance, given a group of 5 references and a frequency distribution of 0.8 for the *first+last* form and 0.2 for the *last* form, 4 references would assume the first form, whereas 1 reference would assume the other one.

4.2 Linguistic Realization

Once we select the form of a proper name reference to a person in a particular discourse context, we linguistically realize this reference by choosing the most likely words - including titles and proper nouns - to be part of it. The process is analogous to the linguistic realization of a set of attribute-values into a description (Bohnet, 2008; Zarriess and Kuhn, 2013). Equation 4 summarizes it.

Form	Frequency
First+Last	46.2%
Last	34.9%
First	8.5%
Middle+Last	2.8%
First+Middle+Last	2.3%
Middle	1.5%
Others	3.5%

Table 2: Most popular proper name forms in REGnames corpus and their frequencies.

$$P(n_1 \ldots n_t \mid f, p) = \prod_t P(n_t \mid n_{t-1}, \{e_i\}_{i=1}^{|f|}, p) \quad (4)$$

The vocabulary used in the linguistic realization step consists of all the titles found in REGnames, all the possible names of the given person present in the corpus' proper name knowledge base, and an *end* token, present at the end of all proper name references in the training set. The process finishes when this token is predicted ($n_t = END$). The choice of a word n_t is conditioned to the previous generated word in the proper name reference (n_{t-1}), the elements present in the given form ($\{e_i\}_{i=1}^{|f|}$: constrained to first, middle and last name; plus title and appositive) and the person to be referred to (p). If $P(n_t \mid n_{t-1}, \{e_i\}_{i=1}^{|f|}, p) = 0$, we drop the less frequent element from the given proper name form. If all the elements were dropped and the probability would still be 0, we conditioned the choice only to the person ($P(n_t \mid p)$). Regarding the cases in which the original proper name form indicates the presence of an appositive, we add a description - obtained from Wikidata (Vrandečić and Krötzsch, 2014) - at the end of the generated proper name reference.

5 Baselines

In order to evaluate the performance of our model, we developed three baseline models. All the models have their outputs constrained to three choices: given name, surname and full name of a person.

Given name and surname are determined by the values of the following attributes in the person's DBpedia page: *foaf:givenName* and *foaf:surname*. Full name was defined as the combination of both values. If these attributes are missing, we use the birth name of the person, also extracted from DBpedia (*dbp:birthName*). In this situation, the full name of a person will be the proper birth name, whereas given and surnames will be the first and last tokens from the birth name, respectively.

The first baseline, called *Random*, is a baseline that randomly chooses one of the three options to generate a proper name.

The second baseline is an adaptation of the model proposed by van Deemter (2016) and will be called *Deemter*. Among the full name, given name and surname of a person, our adaptation chooses the shortest name that distinguishes the mentioned person from all other entities in the current and previous 3 sentences in the text. It is important to stress that this model is our adaptation, since the proposal of van Deemter (2016) only applies for initial references, not for repeated ones in a text.

Finally, the third system we compare against is based on Siddharthan et al. (2011) and will be called *Siddharthan*. This baseline chooses the full name of a person for discourse-new references; and his/her surname otherwise.

6 Automatic Evaluation

We intrinsically evaluate the models by training and testing them on a subset of the REGnames corpus. This evaluation aims to investigate how close our model can produce proper name references to the ones generated by human writers.

6.1 Data

We considered a subset of the REGnames corpus as our evaluation data. From the 1,000 people in the corpus, we first filtered the ones whose birth names were not mentioned, or for whom the values of the DBpedia's attributes *foaf:name*, *foaf:givenName* and *foaf:surname* were missing. This measure was taken in order to have a consistent vocabulary to linguistically realize the proper name references, as well as to make sure that our baselines would always have a consistent output. Then, from the remaining people, we only selected the ones with at least 50 proper name references in the REGnames corpus such that we could train and test our model properly. In total, we used 43,655 proper names references to 432 people as our evaluation data.

In order to investigate the influence of the text domain in the generation of proper names, we classified the webpages from where our evaluation data were extracted according to 3 domains: Blog, News and Wiki. All the webpages whose the url contained the substrings *blog*, *tumblr* or *wordpress* were classified as part of the blog domain. If the substrings were *new* or *article*, the webpage was classified as a news. Finally, we classified as Wiki all the webpages whose the url contained the substring *wiki*. All the other webpages were grouped into a *Other domains* category.

6.2 Method

10-fold-cross-validation was performed to evaluate the models. We made sure that the number of references per person was uniform among the folds. To measure the models performance in the choice of the proper name form, accuracy was used. To check the similarity among the realized proper name reference and the gold standard one, we used the string edit distance.

6.3 Models

We evaluated the three proposed baselines (*Random*, *Deemter* and *Siddharthan*) and two versions of our model: *PN-Variation* and *PN+Variation*.

PN-Variation does not take the variation into account in the content selection. In other words, this model always chooses the most likely proper name form for the references in the test set which refer to the same person and are described by the same combination of discourse feature values. On the other hand, *PN+Variation* takes variation into account by applying the distribution of proper name forms obtained from the training set to the similar references in the test set, as explained in Section 4.1.

6.4 Results

Table 3 summarizes the accuracy-scores of the models in the prediction of the proper name forms. Both versions of our model outperform the baselines for all the domains. *PN-Variation* is the model with the highest accuracy.

Figure 1 depicts the string edit distance among the gold standard proper names and the ones generated by the proposed models. A Repeated Measures ANOVA determined that the string edit dis-

Model	Blog	News	Wiki	Other domains	Overall
Random	0.25	0.22	0.22	0.25	0.25
Deemter	0.33	0.30	0.28	0.33	0.33
Siddharthan	0.52	0.48	0.42	0.45	0.48
PN-Variation	0.66	0.63	0.66	0.70	**0.68**
PN+Variation	0.58	0.55	0.59	0.63	0.60

Table 3: Proper name form accuracies of our two models (PN-Variation and PN+Variation) as a function of text genre and compared to three baseline models (Random, Deemter, Siddharthan).

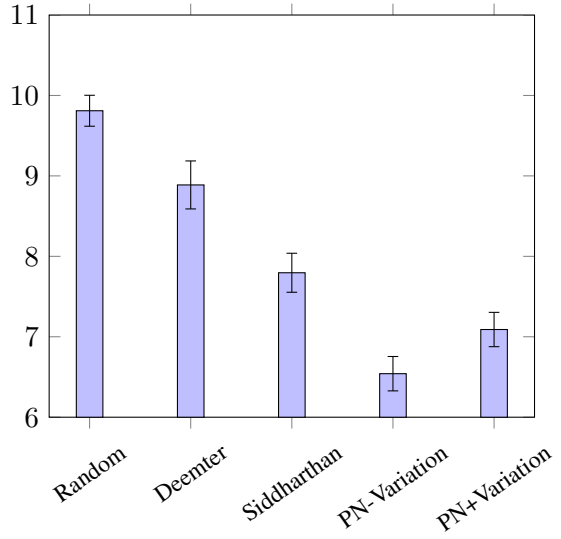


Figure 1: String edit distance in the overall corpus. Error bars represent 95% confidence intervals.

tances of the models were significantly different ($F(4, 36) = 1630, p < .001$). We performed a post hoc analysis with paired t-test using Bonferroni adjusted alpha levels of 0.005 per test $(0.05/10)$. Both versions of our model significantly outperform the baselines with all pairwise comparisons significant at $p < .001$. Regarding the comparison of our models, *PN-Variation* is significantly better than *PN+Variation* ($t(9) = -38.14, p < .001$).

Figure 2 shows the evaluation of our models by domain. A Repeated Measures ANOVA shows that the string edit distances of the models are significantly different in all domains (Blog: $F(4, 36) = 718.8, p < .001$; News: $F(4, 36) = 308.2, p < .001$; Wiki: $F(4, 36) = 118.5, p < .001$; Other domains: $F(4, 36) = 2213, p < .001$).

We also performed a post hoc analysis for the results by domain in the same style we did for the general results. In the blog and news domains, both versions of our model significantly outperform all the baselines with all pairwise comparisons significant at $p < .005$. Among our models, *PN-Variation* is significantly better than *PN+Variation* (Blog: $t(9) = -26.33, p < .001$; News: $t(9) = -7.45, p < .001$).

In the wiki domain and in texts which are not part of the blog, news and wiki domain, both versions of our model also significantly outperform all the baselines with all pairwise comparisons significant at $p < .001$. The difference in the results of *PN-Variation* and *PN+Variation* is also significant (Wiki: $t(9) = -4.91, p < .001$; Other domains: $t(9) = -27.14, p < .001$)

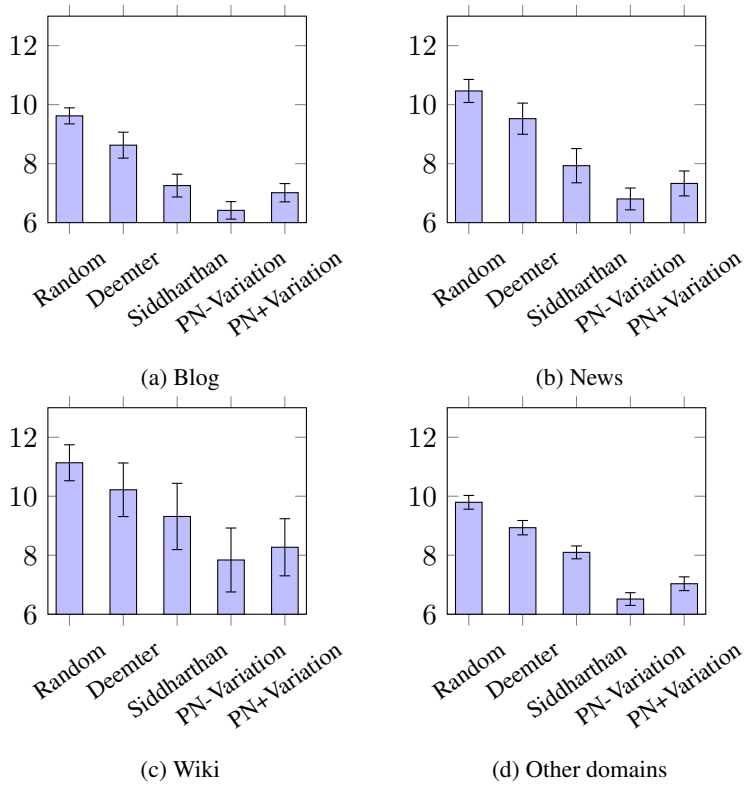
7 Human Evaluation

We also performed a human evaluation aiming to compare original texts with alternative versions whose proper name references were generated by our model. This evaluation aims to investigate the quality of the proper name references from the perspective of the human reader.

7.1 Materials

We used 9 abstracts from English Wikipedia pages whose topic is one of the people studied in the REGnames corpus. They were extracted from DBpedia and have at least 10 proper name references to the topic.

Although our model did not yield its best results for this domain, it was chosen based on the relatively short length of the texts and the large amount of proper name references they have. Moreover, the proper name references in Wikipedia abstracts are similar to the ones generated by our *Siddharthan* baseline, i.e. a full name to discourse-new people, and surname to discourse-old people.



(a) Blog (b) News

(c) Wiki (d) Other domains

Figure 2: String edit distances of the models in the (2a) blog, (2b) news, (2c) wiki and (2d) in other domains which are not the previous ones. Error bars represent 95% confidence intervals.

7.2 Method

For each abstract, we designed 3 trials. In the first, we presented participants with the original text next to the version with the proper name references generated by the *PN-Variation* model (Original vs. No Variation). In the second, we presented the original text next to the version with the proper name references generated by the *PN+Variation* (Original vs. Variation). Finally, the third trial consists of the text versions with the proper name references produced by both versions of our model (No Variation vs. Variation). The trials of a text were distributed in different lists such that we obtained 3 lists with 9 texts - 3 trials of each type in a list. In all the texts, the proper name references were highlighted in yellow. For each trial, we asked participants to choose which text they preferred, taking into account the highlighted references. The experiment is publicly available[4].

We recruited 60 participants through Crowdflower – 20 per list. Of the participants, 44 were female and their average age was 36 years. All participants reported to be proficient in the English language (58 were native speakers).

[4] http://ilk.uvt.nl/~tcastrof/eacl2017

7.3 Results

The texts of the "Original" version were the favourite of 69% of the participants in comparison with texts of the "No Variation" version (Chi-square $\chi^2(2, 180) = 25.69, p < .001$), and 75% participants with the "Variation" version (Chi-square $\chi^2(2, 180) = 45; p < .001$). Regarding the "No Variation vs. Variation" trials, texts of the "No Variation" version were the favourite of the participants in 59% of the cases (Chi-square $\chi^2(2, 180) = 6.42; p < .05$).

8 General Discussion

Proper name generation is a seriously understudied phenomenon in automatic text generation. There are many different ways in which a person can be referred to in a text using their name (*Barack Hussein Obama II*, *Barack Obama*, *Obama*, *President Obama*, etc.) and arguably a text that uses different naming formats in different conditions is more human-like than one that relies on a fixed strategy (e.g., always use the full name).

This paper introduced a new statistical model for the generation of proper names in text, taking into account three different factors: (1) who

the person is, (2) in which discourse context the proper name reference should be generated and (3) the different forms that a proper name can assume in similar situations (variation). The model was developed based on the REGnames corpus (Ferreira et al., 2016c), which contains a large number of proper name references in various discourse situations. We also implemented two other systems for the sake of comparison: one based on the Siddharthan et al. (2011) model and one based on the ideas for proper name reference proposed by van Deemter (2016).

We developed two versions of our model: one that deterministically generated the best proper name form in a given setting (*PN-variation*), and one that relied on a probabilistic distribution over different forms, allowing for more variation in the output (*PN+Variation*). Both models were systematically compared to a random baseline and the two alternative models due to Siddharthan et al. (2011) and van Deemter (2016).

Automatic Evaluation We first conducted an automatic evaluation investigating to what extent the evaluated models produced proper name references similar to the ones generated by human writers, using a held-out subset of the REGnames corpus. In general, we found that both versions of our model were able to outperform a random baseline and the two reference systems, where the version without variation (*PN-Variation*) yielded the best results. Across text domains, there was variation in the performance of both versions of our model. The worst results were registered in the Wiki domain, suggesting that text domain is a factor that may be taken into account in the task of generating proper names.

Human Evaluation In the automatic evaluation experiment, the differences between the system with and without variation were small, so in a second study we asked whether human readers preferred the output from one of these systems over the other. For this purpose, we conducted an experiment consisting of pairwise comparisons based on texts taken from the Wikipedia domain, where we compared the output produced by the *PN-variation* and the *PN+variation* system with the original text and also among them. Interestingly, we found that people had a general preference for the no-variation model over the one that non-deterministally generated varied texts. This suggests that readers prefer consistency in proper name references to the same topic in similar situations, which is different from the choice of referential *form* (Ferreira et al., 2016b).

Additionally, we found that participants preferred the original over the regenerated texts. We suspect that this preference was due to the initial discourse-new proper name reference, which in the Wikipedia texts has a special status. Usually, the initial reference to the topic is not the most common proper name reference in other domains, but a specific Wikipedia format which our system does not produce. For example, the original text about Magic Johnson starts with *Earvin "Magic" Johnson Jr.* in the discourse-new proper name reference, while our system simply produced *Magic Johnson*.

Semantic web Earlier work on REG models has concentrated on the generation of descriptions, typically assuming the existence of a knowledge base of entities (Dale and Haddock, 1991; Dale and Reiter, 1995) or introducing one to small domains (Gatt and Belz, 2010). Our REG models for proper names, however, strongly rely on the semantic web as an information resource of the entities to be referred to. Databases like DBpedia (Bizer et al., 2009) and Wikidata (Vrandečić and Krötzsch, 2014) provide information about thousands of entities and can be used in different domains.

Baselines We developed two powerful baselines based on proposals that have been made before. *Deemter* (van Deemter, 2016) relies on the criteria of the first developed REG models (Dale and Haddock, 1991; Dale and Reiter, 1995): given a target, produce a reference that distinguishes it from the distractors in the context. Our model as presented does not make this assumption (it does not always produce a proper name reference that distinguishes the target from the distractors). However, this could be incorporated into our model as well. For instance, given a list of the most likely proper name references produced by our model in a situation, we can choose the one with the highest likelihood that distinguishes the target from all other entities in the current and previous 3 sentences in the text (as in the *Deemter* model).

Regarding performance, *Siddharthan* is the baseline that performed best. The original version, proposed in Siddharthan et al. (2011), is

even able to decide whether to include a modifier in a discourse-new reference based on the global salience of the entity mentioned. However, the model is arguably more limited in the production of a proper name itself. By always generating a surname in discourse-old references for instance, the Siddharthan model is not able to generate at least 10% of the references in the REGnames corpus (8.5% consist of *first name* references, and 1.5% of *middle name* ones).

Conclusion In sum, we conclude that our model is able to generate proper name references similar to the ones produced by human writers. In future research, it would be interesting to further investigate the role of text genre in proper name references as well as the influence of variation on proper name forms.

Acknowledgments

This work has been supported by the National Council of Scientific and Technological Development from Brazil (CNPq). We would also like to thank the members of the Language Production group at TiCC, specially Ákos Kádár, for insightful comments on the manuscript.

References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165. The Web of Data.

Bernd Bohnet. 2008. The fingerprint of human referring expressions and their surface realization with graph transducers. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 207–210, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Dale and Nicholas Haddock. 1991. Generating referring expressions involving relations. In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*, EACL '91, pages 161–166, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016a. Individual variation in the choice of referential form. In *Proceedings of NAACL-HLT*, pages 423–427, San Diego, California. Association for Computational Linguistics.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 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*, pages 568–577, Berlin, Germany. Association for Computational Linguistics.

Thiago Castro Ferreira, Sander Wubben, and Emiel Krahmer. 2016c. Towards proper name generation: a corpus analysis. In *Proceedings of the 9th International Natural Language Generation conference (INLG)*, Edinburgh, Scotland.

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to nlg: The tuna shared task evaluation challenges. In *Empirical methods in natural language generation*, pages 264–293. Springer.

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

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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

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.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015.

Kees van Deemter. 2016. Designing algorithms for referring with proper names. In *Proceedings of the 9th International Natural Language Generation conference*, pages 31–35, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September.

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, Sofia, Bulgaria, August. Association for Computational Linguistics.