Feature Selection for Reference Generation as Informed by Psycholinguistic Research

Charles F. Greenbacker (charlieg@cis.udel.edu)
Department of Computer and Information Sciences, University of Delaware, 101 Smith Hall
Newark, DE 19716 USA

Kathleen F. McCoy (mccoy@cis.udel.edu)
Department of Computer and Information Sciences, University of Delaware, 101 Smith Hall
Newark, DE 19716 USA

Abstract

We present a method of feature selection for the production of appropriate types of referring expressions which attempts to bridge the gap between computational and empirical approaches to reference generation. Our task is to correctly identify the type of reference employed by a human author in a text where such references to the main subject have been tagged and removed. The goal is to provide a natural language generation system with guidelines to determine when the use of different types of reference (names, pronouns, etc.) is most appropriate. Our work combines experimental and computational approaches to develop a set of features capturing certain information below the surface level of text with which to train a classifier system. Ultimately, this classifier could be used to control the production of referring expressions in machine-generated communication. A decision tree trained on the features selected by our method produces results approaching those of the highest-performing purely empirical-based machine learning systems for the given task.

Background

The generation of referring expressions is a major focus within the field of Natural Language Generation (NLG). Part of the task of producing such references is determining when each type of referring expression is appropriate. For example, the choice to use a pronoun rather than a definite description, or vice versa, can be influenced by a number of complex factors. The purpose of this work is to attempt to discover some of the factors that affect humans' decision to select one type of reference over another, so that NLG systems can make similar decisions.

Generation of References in Context (GREC) is a shared task challenge in NLG where participants develop systems that can select appropriate references to entities in a document from a list of alternatives. GREC data consists of introductory sections of Wikipedia articles from which instances of referential expressions have been replaced by a list of possible references of different types (Belz & Varges, 2007). Following the success of the GREC task as part of the Referring Expression Generation Challenge 2008 (Belz, Kow, Viethen, & Gatt, 2008), a second round has been incorporated into Generation Challenges 2009 for the 12th European Workshop on Natural Language Generation. Our present work was initiated to develop a submission for GREC-MSR (Main Subject Reference) '09. We view the GREC task as a two-part problem, the first of which is to correctly identify the reference type, and the second being to select the specific reference. This extended abstract describes our work on the first part of this task.

Figure 1 contains a rather short example article from the GREC corpus. The items in bold have been tagged as referring expressions pointing to the main subject entity (Mount Greylock). In the corresponding GREC data format, the tags contain basic information about the references (e.g., syntactic category) and a list of possible alternative referential expressions. As this example was pulled from the training set, the specific reference appearing in the article is also indicated. The goal of the GREC-MSR '09 task is to accurately predict the type of reference (REG08-TYPE) and exact referential expression used by the original author(s) by selecting from the list of alternative references. A much more thorough description of the GREC data format can be found in Belz et al. (2008).

> **Mount Greylock** is a mountain of 3,491 feet (1,064 m) in elevation, located in northwestern Massachusetts. **It** is the highest point in the state.

Figure 1: Example article from GREC corpus, titled "Mount Greylock." Tagged references to the main subject entity are in boldface.

Although clearly beneficial to the development of techniques for generating appropriate references, the GREC task does differ significantly from traditional referring expression generation in the context of an NLG system. In particular, standard information used for the generation of natural language, is not available in the GREC data because the system must work from the surface text as opposed to the underlying data available in the conventional NLG task. Thus, data typically input to the reference generation task necessary to determine whether interfering antecedents exist or whether a pronoun would be ambiguous or required (e.g., other objects being referred to, attributes of objects, surface syntactic information) must be derived from the surface level. Also missing is information concerning discourse segments which is widely acknowledged (Grosz & Sidner, 1986; Reichman, 1985) as important to the task of generating referring expressions. Similarly, sentence segmentation and access to syntactic information in the text must be derived. In a sense, the generation of referring expressions has been relegated to a post-processing step, without the benefit of potentially vast amounts of information from preceding generation steps.

Method

Our approach to this task features a unique method that combines computational and experimental approaches to determining appropriate reference types. Our intuition was that findings in psycholinguistic research could be utilized to inform feature selection for representing instances of referring expressions in the GREC data, and that a competitively-performing classifier system could be trained on these features. In contrast with automatic feature selection algorithms, our development process involved surveying psycholinguistic literature for potential features, building a rapid prototyping system facilitating the extraction of said features from the data and determining their preliminary efficacy at choosing the correct reference type, iteratively reviewing errors resulting from this prototype system to theorize additional features and patterns governing reference generation, and finally training a decision tree classifier based on the selected features.

A number of authors (Arnold, 1998; Gordon & Hendrick, 1998) identified several factors which psycholinguistic research indicates influence how pronouns are interpreted, including Subjecthood, Parallelism, Recency, and Ambiguity. Due to the lack of standard generation information, Parallelism and Ambiguity offer particular challenges for the GREC task, as the grammatical position (or syntactic category) of named entities other than the main subject are not provided in the GREC-MSR data (although it should be mentioned that such information is provided in GREC-NEG, a related task). We chose to employ Recency as our starting point, and, as recommended by McCoy and Strube (1999), classified long-distance anaphoric expressions (where the most recent reference to the entity occurs more than two sentences ago) as names and short-distance anaphoric expressions (subsequent references in the same sentence) as pronouns. We implemented these rules in our prototyping system, and examined the incorrect classifications which resulted in an attempt to discover which other factors suggested by psycholinguistic research could explain the patterns we observed.

The next features we chose to explore were the syntactic categories of the three most recent references, as well as for the reference at hand (all available in the GREC data). As these categories identify the grammatical position of references (either nouns in subject or object position, or a determiner), we hoped that they could be used to cover the Subjecthood and Parallelism factors discussed in Arnold (1998). The choice of using exactly three most recent references was motivated by the intuition that the impact of earlier reference forms are increasingly smaller, as well as the use of three references by systems performing similar tasks (Hendrickx, Daelemans, Luyckx, Morante, & Van Asch, 2008). Additionally, numerous documents in our data set are quite brief, and many contain only three or fewer references to the main entity.

We also extracted binary features indicating whether or not the main entity was the subject of this sentence, as well as the two previous sentences. Another boolean feature was triggered if the main subject entity was in subject position of the sentence containing the most recent reference. Two additional boolean features were used to represent patterns which seemed apparent during our inspection of erroneous assignments: 1. whether the reference was immediately preceded by the words "and," "but," or "then," and 2. whether the reference fell between a comma and the word "and." Our observations suggested that instances of the entity's name other than those involved in referring expressions, such as the example given in Figure 2, seemed to be a factor in choosing when to use pronouns. So, we added a feature measuring the number of sentences since the most recent "non-referential instance." Finally, as a simple attempt at addressing the Ambiguity factor, we made a basic effort at recognizing whether a possible interfering antecedent (or "distractors" as described in Siddharthan and Copestake (2004)) occurred earlier in the current sentence, or in the intervening text since the last reference.

> The municipality was abolished in 1928, and the name "Mexico City" can now refer to two things.

Figure 2: Example of non-referential instance. In this sentence, "Mexico City" is not a reference to the main entity (Mexico City), but rather to the name "Mexico City."

This process of trial and error was continued until we had collected an extensive list of features which we felt had a high probability of indicating the correct reference type for a given instance. We also added several additional features used by the highest-performing submission from GREC '08, CNTS-Type-g (Hendrickx et al., 2008). These extra features included semantic category of the current document, sentence count, reference count, NP count, local context of the three words and POS tags preceding and following the reference, and the distance to the previous reference measured in NPs. The full list of candidate features is provided in Table 1. We used various subsets of this list of features to train a series of decision tree classifiers with the proprietary C5.0 algorithm developed by Ross Quinlan (RuleQuest Research Pty Ltd, 2008). The features measuring (in number of sentences) distance to the three previous references and the most recent non-referential instance were realized as both continuous values, as well as a version in which all values greater than 2 are reassigned a value of 3. Given the size of our data set, using continuous values for these features resulted in far more branching in the decision trees generated, with several branches followed by a handful of examples out of the thousands in the training set. The benefit of using the discrete variant may simply be to help avoid the potential for overfitting.

Table 1: Full list of candidate features indicating which decision trees they were used in.

		Used in Which Decision Trees				
Feature	**Possible Values**	**DT_1**	**DT_2**	**DT_3**	**DT_4**	**DT_5**
distance in number of sentences since last three references & last non-referential instance	continuous	X		X		
distance in number of sentences since last three references & last non-referential instance	3, 2, 1, or 0		X		X	X
syntactic category of current & three most recent references	null, np-obj, np-subj, or subj-det	X	X	X	X	X
for current & 2 previous sentences, was main entity the subject?	boolean		X	X	X	X
was main entity subject of sentence containing last reference?	boolean		X	X	X	X
reference follows and, but, or then?	boolean		X	X	X	X
reference between comma & and?	boolean		X	X	X	X
possible interfering antecedent in current sentence?	boolean				X	X
possible interfering antecedent since last reference?	boolean		X	X	X	X
semantic category of document	city, country, mountain, person, or river	X	X	X	X	X
sentence count	continuous				X	X
NP count	continuous					X
reference count	continuous				X	X
3 words preceding & following reference	{entire set of words appearing in corpus}					X
3 POS tags preceding & following reference	{POS tags assigned by Brill Tagger (Brill, 1994)}					X
distance in number of NPs since last reference	continuous					X

Table 2: Comparison of our best classifiers (in bold) to GREC '08 submissions with respect to type accuracy, as reported in Belz et al. (2008).

System	REG08-Type Accuracy for Development Set					
	All	Cities	Countries	Rivers	People	Mountains
CNTS-Type-g	76.52	64.65	75	65	85.37	75.42
CNTS-Prop-s	73.93	65.66	69.57	70	79.51	74.58
DT_X	**71.98**	**64.95**	**68.48**	**65**	**81.37**	**68.75**
DT_2	**70.9**	**59.79**	**67.39**	**60**	**81.37**	**68.75**
DT_4	**70.9**	**64.95**	**68.48**	**65**	**79.9**	**67.08**
DT_5	**67.69**	**59.79**	**64.13**	**50**	**78.92**	**64.17**
IS-G	66	54.5	64	80	66.8	65
OSU-n-nonRE	62.5	53.54	63.04	65	67.32	61.67
OSU-b-all	58.54	53.54	57.61	75	65.85	49.58
OSU-b-nonRE	51.07	51.52	52.26	40	57.07	45.83

Results

The accuracy of each trained decision tree, as computed via ten-fold cross-validation on the training data, is presented in Table 3. Surprisingly, the highest-performing decision tree did not use the full feature set, in fact, it used a relatively limited subset.

Table 3: Accuracy of each decision tree, calculated via ten-fold cross-validation on the training set.

Decision Tree	Accuracy
DT_1	68.2%
DT_2	72.6%
DT_3	72.2%
DT_4	72.0%
DT_5	71.7%

In order to compare our method to the GREC '08 participants' submissions, we computed referring expression type accuracy for the development set of 97 texts (included as part of the GREC data package). The set of features resulting in our two best performing classifiers (DT_2 & DT_4), trained on the 1658-text training set, each yield an overall REG08-Type accuracy of 70.90%. By combining these classifiers, using whichever one produced better results for each semantic category, we were able to build a new classifier (DT_X) with an overall accuracy of 71.98%. This is very competitive when compared to the scores reported for GREC '08. In fact, our classifier scored higher than each of the participants' systems, except for the two variants from the winning team, the better of which reported a remarkable accuracy of 76.52% (Belz et al., 2008). Table 2 provides a detailed comparison between the type accuracy achieved by our system and those for GREC '08. Details for each GREC '08 submission can be found in the system reports produced by the participants (for CNTS in Hendrickx et al. (2008), IS-G in Bohnet (2008), and OSU in Jamison and Mehay (2008)).

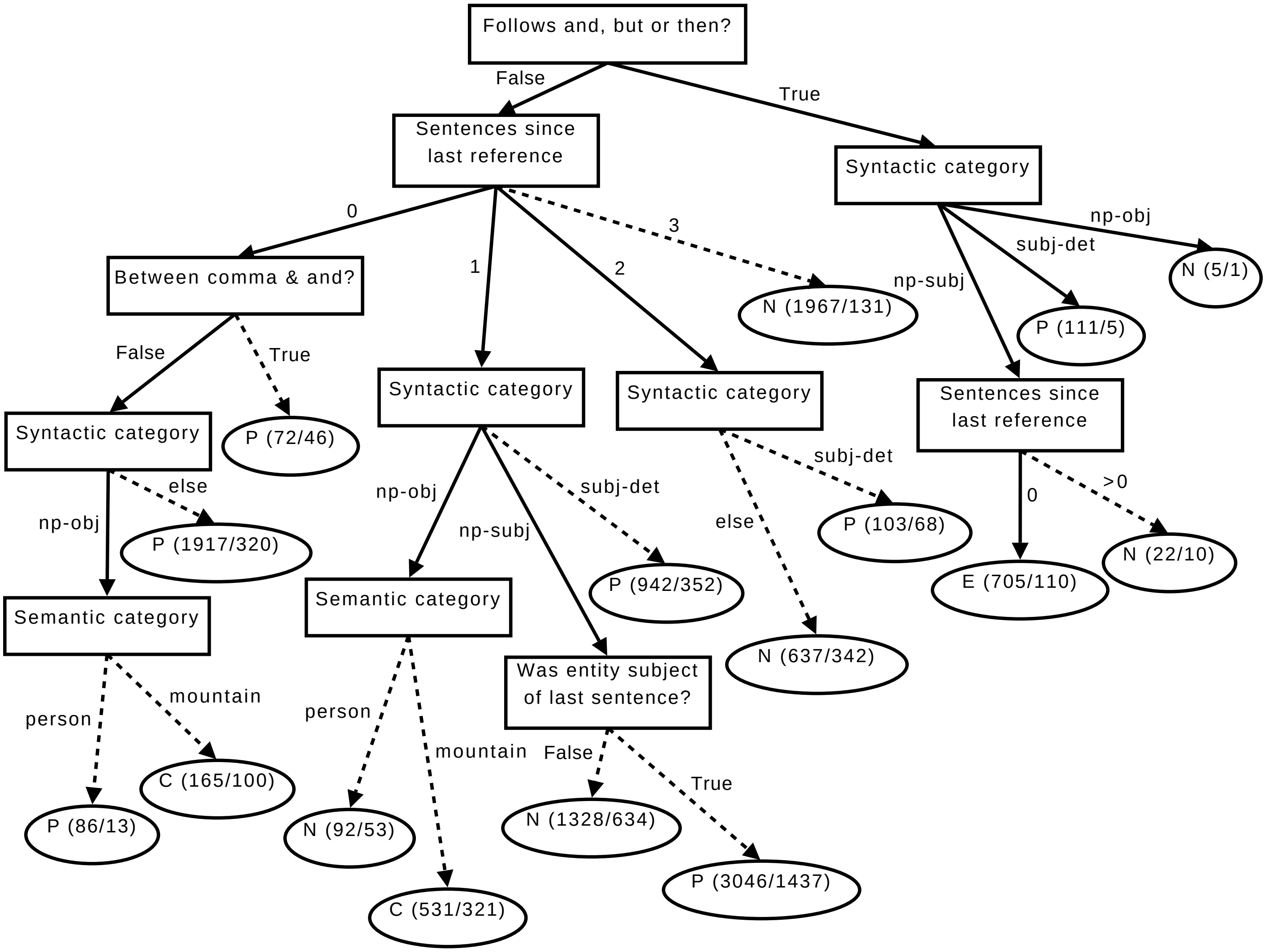
Discussion

While at first glance it is a bit troubling that the best performing system using traditional empirically-based features is outperforming our system, part of the reason for this could very well be our use of very basic means of NP-chunking and named entity recognition (NER). Due to time constraints and other factors, we were unable to incorporate more sophisticated methods, a deficiency that might explain the relatively poor performance of DT_5 which sought to utilize the features employed by the best team from GREC '08. Indeed, a classifier trained on only those features produced an overall type accuracy of only 57.89% on the development set. We suspect that this score is the result of our rudimentary approach to NP-chunking and NER, and the fact that we were unable to effectively identify the main verb in each sentence, which was another feature employed by the leading GREC '08 participant. This points to an immediate future area of research which would focus on including more sophisticated methods for these subtasks in order to further validate our feature selection. If it can be shown that these features play an important role in the production of referring expressions, then improving the accuracy of feature identification would certainly be beneficial.

The ideal solution to this shortcoming would be to incorporate more robust generation knowledge into the data set, as mentioned earlier. The use of NP-chunking and other natural language *processing* techniques would no longer be necessary as this information would be readily available, having been determined through earlier phases of the generation process. The relegation of NER is even more apparent. Given the sentence, "Jack and Joe went to the store, and he bought a loaf of bread," where "he" is intended to refer to "Jack," an NER tool should correctly recognize that "Joe" is an interfering antecedent causing the pronoun to be ambiguous. Instead, if the sentence is about "Jack and Jill," the tool may not be able to determine that the difference in gender makes the use of the pronoun "he" perfectly clear. Adding a gender-neutral name into the mix, for example, "Jack and Sam," further demonstrates the dilemma, especially if there are no other clues available in the surface text from which a tool can infer the gender of "Sam." However, deeper knowledge about each of these entities would likely be directly available from the generation information used during the earlier formation of these sentences, and there would be no need to compute it post hoc.

One of the advantages of employing a decision tree-based approach, as opposed to other machine learning techniques, is the ability to examine the tree and observe exactly how the features intertwine. As our best classifier (DT_X) actually uses two decision trees (DT_2 and DT_4), we can compare these trees to get a better sense of how each feature contributes to the classification. Figure 3 shows the first few levels of the decision tree for DT_2. Although the feature set for DT_4 contains only three additional features, its decision tree is far more complex. The new features, reference count and sentence count, are responsible for much of the growth in complexity over DT_2, and the fact that both of these are continuous values may be the direct cause. However, a non-trivial identical segment appears in both trees, indicating that the reference types for training instances following these paths are not affected by the additional features. This identical portion is shown in Figure 3, and includes the root node plus everything descending from its right branch (where the boolean feature, "Follows and, but or then," is true). The features appearing higher in both decision trees, due to their higher calculated information gain, include several of the linguistically-motivated factors, such as the number of sentences since the last reference (Recency), the syntactic category of the current & previous references (Subjecthood & Parallelism), whether the entity was the subject of the previous sentence (Subjecthood), and the existence of interfering antecedents (Ambiguity). The positive results achieved using these features reinforce the psycholinguistic research regarding their roles in

Figure 3: Partial representation of decision tree used in DT_2 classifier.



Note: Dotted lines indicate additional branching occurs below this level. Capital letters in leaf nodes stand for assigned reference type: N for name, P for pronoun, C for common noun, E for ellipsis. Values in leaf nodes reflect the total number of instances in the training set mapped to this leaf and the number classified incorrectly. Leaf nodes attached to dotted lines represent the dominant class assigned via subsequent branching, and the values give the set size and accuracy achieved by assigning the dominant class to all instances mapping to this point.

pronominalization. The fact that other boolean features we identified during the iterative development process, "Follows and, but or then" and "Between comma & and," appear high in the trees as well suggests certain sentence constructions also affect reference type.

A lingering question remains as to what the highest possible score can be realistically expected for this task. In essence, the task is not to select the "correct" reference type, but simply the one chosen by the human author. Needless to say, human authors are certainly not infallible, and Wikipedia is not without errors. Additionally, it is often the case that the use of either a definite description or a pronoun is completely acceptable, and two different humans may make different choices for seemingly arbitrary reasons. Yeh and Mellish (1996) found only about 75% agreement between persons on whether or not to use a pronoun for reach reference, although previous studies gave a figure as high as 90%. In a experiment described by Belz and Varges (2007), a set of three humans agreed on the specific referring expression (not just the type) only 50.1% of the time, while in only 64.9% of the cases did the subjects agree on whether or not to use a pronoun. While this has been described as a "considerable amount of agreement," we feel these numbers indicate that there are many cases where there is more than one "acceptable solution."

Future Work

Beyond the use of more advanced NLP tools and/or augmenting the data set with deeper generation knowledge, another possible area of future work would be to develop additional features based on other parts of Arnold's psycholinguistic model. Two factors, Focus and Goal Status, have not been given a treatment at all. Focus is defined by the mention of the referent in a cleft construction, and Goal Status is determined by the appearance of the entity as the goal argument of a verb (Arnold, 1998). Focus seems to naturally lend itself to our method of feature selection, as one can imagine writing some routine allowing a system to detect prototypical cleft sentences. However, Goal Status would likely prove to be much more difficult. Without substantial generation information or detailed semantic models, it may be extremely challenging to develop a system capable of reliably speculating about the intentions behind an author's usage of a verb.

Conclusions

We've shown that findings in psycholinguistic research regarding the production of referring expressions can be useful in determining proper feature selection for the task of selecting appropriate reference types. With more time, additional iterations of our method, involving further review of psycholinguistic literature coupled with additional examination of incorrect assignments and tuning of the machine learning system, as well as the utilization of more sophisticated NP-chunking and named entity recognition methods, could yield even better results.

References

Arnold, J. E. (1998). *Reference form and discourse patterns*. Doctoral dissertation, Department of Linguistics, Stanford University.

Belz, A., Kow, E., Viethen, J., & Gatt, A. (2008). The GREC challenge 2008: Overview and evaluation results. In *Proceedings of the 5th international natural language generation conference* (pp. 183–191). Salt Fork, OH, USA.

Belz, A., & Varges, S. (2007). Generation of repeated references to discourse entities. In *Proceedings of the 11th european workshop on NLG* (pp. 9–16). Schloss Dagstuhl, Germany.

Bohnet, B. (2008). IS-G: The comparison of different learning techniques for the selection of the main subject references. In *Proceedings of the 5th international language generation conference.* Ohio, USA.

Brill, E. (1994). Some advances in rule-based part-of-speech tagging. In *Proceedings of the 12th national conference on artificial intelligence, AAAI* (p. 722-727).

Gordon, P. C., & Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cognitive Science*, *22*(4), 389-424.

Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics Journal*, *12*(3), 175-204.

Hendrickx, I., Daelemans, W., Luyckx, K., Morante, R., & Van Asch, V. (2008). CNTS: Memory-based learning of generating repeated references. In *Proceedings of the 5th international language generation conference.* Ohio, USA.

Jamison, E., & Mehay, D. (2008). OSU-2: Generating referring expressions with a maximum entropy classifier. In *Proceedings of the 5th international language generation conference.* Ohio, USA.

McCoy, K. F., & Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description. In *Proceedings of workshop on the relation of discourse/dialogue structure and reference, held in conjunction with the 38th annual meeting* (p. 63 - 71). College Park, Maryland.

Reichman, R. (1985). *Getting computers to talk like you and me: Discourse context, focus, and semantics*. Cambridge: MIT Press.

RuleQuest Research Pty Ltd. (2008). *Data mining tools See5 and C5.0.* http://www.rulequest.com/see5-info.html.

Siddharthan, A., & Copestake, A. (2004). Generating referring expressions in open domains. In *Proceedings of the 42th meeting of the association for computational linguistics annual conference* (p. 408-415). Barcelona, Spain.

Yeh, C.-L., & Mellish, C. (1996). An evaluation of anaphor generation in chinese. In *Proceedings of the eighth international generation workshop.* Herstmonceux Castle, Sussex, UK.