UDel: Generating Referring Expressions Guided by Psycholinguistic Findings

Charles Greenbacker and Kathleen McCoy
Dept. of Computer and Information Sciences
University of Delaware
Newark, Delaware, USA
[charlieg|mccoy]@cis.udel.edu

Abstract

We present an approach to generating referring expressions in context utilizing feature selection informed by psycholinguistic research. Features suggested by studies on pronoun interpretation were used to train a classifier system which determined the most appropriate selection from a list of possible references. This application demonstrates one way to help bridge the gap between computational and empirical means of reference generation.

1 Introduction

This paper provides a system report on our submission for the GREC-MSR (Main Subject References) Task, one of the two shared task competitions for Generation Challenges 2009. The objective is to select the most appropriate reference to the main subject entity from a given list of alternatives. The corpus consists of introductory sections from approximately 2,000 Wikipedia articles in which references to the main subject have been annotated (Belz and Varges, 2007). The training set contains articles from the categories of cities, countries, mountains, people, and rivers. The overall purpose is to develop guidelines for natural language generation systems to determine what forms of referential expressions are most appropriate in a particular context.

2 Method

The first step of our approach was to perform a literature survey of psycholinguistic research related to the production of referring expressions by human beings. Our intuition was that findings in this field could be used to develop a useful set of features with which to train a classifier system to perform the GREC-MSR task. Several common factors governing the interpretation of pronouns were identified by multiple authors (Arnold, 1998; Gordon and Hendrick, 1998). These included Subjecthood, Parallelism, Recency, and Ambiguity. Following (McCoy and Strube, 1999), we selected Recency as our starting point and tracked the intervals between references measured in sentences. Referring expressions which were separated from the most recent reference by more than two sentences were marked as long-distance references. To cover the Subjecthood and Parallelism factors, we extracted the syntactic category of the current and three most recent references directly from the GREC data. This information also helped us determine if the entity was the subject of the sentence at hand, as well as the two previous sentences. Additionally, we tracked whether the entity was in subject position of the sentence where the previous reference appeared. Finally, we made a simple attempt at recognizing potential interfering antecedents (Siddharthan and Copestake, 2004) occurring in the current sentence and the text since that last reference.

Observing the performance of prototyping systems led us to include boolean features indicating whether the reference immediately followed the words “and,” “but,” or “then,” or if it appeared between a comma and the word “and.” We also found that non-annotated instances of the entity’s name, which actually serve as references to the name itself rather than to the entity, factor into Recency. Figure 1 provides an example of such a “non-referential instance.” We added a feature to measure distance to these items, similar to the distance between references. Sentence and reference counters rounded out

Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLP 2009, pages 101–102,
Suntec, Singapore, 6 August 2009.

the full set of features.

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

Figure 1: 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.”

3 System Description

A series of C5.0 decision trees (RuleQuest Research Pty Ltd, 2008) were trained to determine the most appropriate reference type for each instance in the training set. Each tree used a slightly different subset of features. It was determined that one decision tree in particular performed the best on mountain and person articles, and another tree on the remaining categories. Both of these trees were incorporated into the submitted system.

Our system first performed some preprocessing for sentence segmentation and identified any non-referential instances as described in Section 2. Next, it marshalled all of the relevant data for the feature set. These data points were used to represent the context of the referring expression and were sent to the decision trees to determine the most appropriate reference type. Once the type had been selected, the list of alternative referring expressions were scanned using a few simple rules. For the first instance of a name in an article, the longest non-emphatic name was chosen. For subsequent instances, the shortest non-emphatic name was selected. For the other 3 types, the first matching option in the list was used, backing off to a pronoun or name if the preferred type was not available.

4 Results

The performance of our system, as tested on the development set and scored by the GREC evaluation software, is offered in Table 1.

5 Conclusions

We’ve shown that psycholinguistic research can be helpful in determining feature selection for generating referring expressions. We suspect the performance of our system could be improved by employing more sophisticated means of sentence segmentation and named entity recognition for identifying interfering antecedents.

Table 1: Scores from GREC evaluation software.

Component Score	Value
total pairs	656
reg08 type matches	461
reg08 type accuracy	0.702743902439024
reg08 type precision	0.702743902439024
reg08 type recall	0.702743902439024
string matches	417
string accuracy	0.635670731707317
mean edit distance	0.955792682926829
mean normalised edit distance	0.338262195121951
BLEU 1 score	0.6245
BLEU 2 score	0.6103
BLEU 3 score	0.6218
BLEU 4 score	0.6048

References

Jennifer E. Arnold. 1998. *Reference Form and Discourse Patterns*. Doctoral dissertation, Department of Linguistics, Stanford University, June.

Anja Belz and Sabastian Varges. 2007. Generation of repeated references to discourse entities. In *Proceedings of the 11th European Workshop on NLG*, pages 9–16, Schloss Dagstuhl, Germany.

Peter C. Gordon and Randall Hendrick. 1998. The representation and processing of coreference in discourse. *Cognitive Science*, 22(4):389–424.

Kathleen F. McCoy and Michael Strube. 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*, pages 63 – 71, College Park, Maryland. Association for Computational Linguistics.

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

Advaith Siddharthan and Ann Copestake. 2004. Generating referring expressions in open domains. In *Proceedings of the 42th Meeting of the Association for Computational Linguistics Annual Conference*, pages 408–415, Barcelona, Spain.