UDel: Named Entity Recognition and Reference Regeneration from Surface Text

Nicole L. Sparks, Charles F. Greenbacker, Kathleen F. McCoy, and Che-Yu Kuo
Department of Computer and Information Sciences
University of Delaware
Newark, Delaware, USA
[sparks|charlieg|mccoy|kuo]@cis.udel.edu

Abstract

This report describes the methods and results of a system developed for the GREC Named Entity Recognition and GREC Named Entity Regeneration Challenges 2010. We explain our process of automatically annotating surface text, as well as how we use this output to select improved referring expressions for named entities.

1 Introduction

Generation of References in Context (GREC) is a set of shared task challenges in NLG involving a corpus of introductory sentences from Wikipedia articles. The Named Entity Recognition (GREC-NER) task requires participants to recognize all mentions of people in a document and indicate which mentions corefer. In the Named Entity Regeneration (GREC-Full) task, submitted systems attempt to improve the clarity and fluency of a text by generating improved referring expressions (REs) for all references to people. Participants are encouraged to use the output from GREC-NER as input for the GREC-Full task. To provide ample opportunities for improvement, a certain portion of REs in the corpus have been replaced by more-specified named references. Ideally, the GREC-Full output will be more fluent and have greater referential clarity than the GREC-NER input.

2 Method

The first step in our process to complete the GREC-NER task is to prepare the corpus for input into the parser by stripping all XML tags and segmenting the text into sentences. This is accomplished with a simple script based on common abbreviations and sentence-final punctuation.

Next, the files are run through the Stanford Parser (The Stanford Natural Language Processing Group, 2010), providing a typed dependency representation of the input text from which we extract the syntactic functions (SYNFUNC) of, and relationships between, words in the sentence.

The unmarked segmented text is also used as input for the Stanford Named Entity Recognizer (The Stanford Natural Language Processing Group, 2009). We eliminate named entity tags for locations and organizations, leaving only person entities behind. We find the pronouns and common nouns (e.g. "grandmother") referring to person entities that the NER tool does not tag. We also identify the REG08-Type and case for each RE. Entities found by the NER tool are marked as names, and the additional REs we identified are marked as either pronouns or common nouns. Case values are determined by analyzing the assigned type and any type dependency representation (provided by the parser) involving the entity. At this stage we also note the gender of each pronoun and common noun, the plurality of each reference, and begin to deal with embedded entities.

The next step identifies which tagged mentions corefer. We implemented a coreference resolution tool using a shallow rule-based approach inspired by Lappin and Leass (1994) and Bontcheva et al. (2002). Each mention is compared to all previously-seen entities on the basis of case, gender, SYNFUNC, plurality, and type. Each entity is then evaluated in order of appearance and compared to all previous entities starting with the most recent and working back to the first in the text. We apply rules to each of these pairs based on the REG08-Type attribute of the current entity. Names and common nouns are analyzed using string and word token matching. We collected extensive, cross-cultural lists of male and female first names to help identify the gender of named entities, which we use together with SYNFUNC values for pronoun resolution. Separate rules govern gender-neutral pronouns such as "who." By the end of this stage, we have all of the resources

	MUC-6			CEAF			B-CUBED		
Corpus	**F**	**prec.**	**recall**	**F**	**prec.**	**recall**	**F**	**prec.**	**recall**
Entire Set	71.984	69.657	74.471	68.893	68.893	68.893	72.882	74.309	71.509
Chefs	71.094	65.942	77.119	65.722	65.722	65.722	71.245	69.352	73.244
Composers	68.866	66.800	71.064	68.672	68.672	68.672	71.929	73.490	70.433
Inventors	76.170	77.155	75.210	72.650	72.650	72.650	75.443	80.721	70.812

Table 1: Self-evaluation scores for GREC-NER.

necessary to complete the GREC-NER task.

As a post-processing step, we remove all extra (non-GREC) tags used in previous steps, re-order the remaining attributes in the proper sequence, add the list of REs (ALT-REFEX), and write the final output following the GREC format. At this point, the GREC-NER task is concluded and its output is used as input for the GREC-Full task.

To improve the fluency and clarity of the text by regenerating the referring expressions, we rely on the system we developed for the GREC Named Entity Challenge 2010 (NEG), a refined version of our 2009 submission (Greenbacker and McCoy, 2009a). This system trains decision trees on a psycholinguistically-inspired feature set (described by Greenbacker and McCoy (2009b)) extracted from a training corpus. It predicts the most appropriate reference type and case for the given context, and selects the best match from the list of available REs. For the GREC-Full task, however, instead of using the files annotated by the GREC organizers as input, we use the files we annotated automatically in the GREC-NER task. By keeping the GREC-NER output in the GREC format, our NEG system was able to successfully run unmodified and generate our output for GREC-Full.

3 Results

Scores calculated by the GREC self-evaluation tools are provided in Table 1 for GREC-NER and in Table 2 for GREC-Full.

Corpus	**NIST**	**BLEU-4**
Entire Set	8.1500	0.7953
Chefs	7.5937	0.7895
Composers	7.5381	0.8026
Inventors	7.5722	0.7936

Table 2: Self-evaluation scores for GREC-Full.

4 Conclusions

Until we compare our results with others teams or an oracle, it is difficult to gauge our performance. However, at this first iteration of these tasks, we're pleased just to have end-to-end RE regeneration working to completion with meaningful output.

5 Future Work

Future improvements to our coreference resolution approach involve analyzing adjacent text, utilizing more of the parser output, and applying machine learning to our GREC-NER methods.

References

Kalina Bontcheva, Marin Dimitrov, Diana Maynard, Valentin Tablan, and Hamish Cunningham. 2002. Shallow Methods for Named Entity Coreference Resolution. In *Chaînes de références et résolveurs d'anaphores, workshop TALN 2002*, Nancy, France.

Charles Greenbacker and Kathleen McCoy. 2009a. UDel: Extending reference generation to multiple entities. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 105–106, Suntec, Singapore, August. Association for Computational Linguistics.

Charles F. Greenbacker and Kathleen F. McCoy. 2009b. 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)*, Amsterdam, July.

Shalom Lappin and Herbert J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

The Stanford Natural Language Processing Group. 2009. Stanford Named Entity Recognizer. `http://nlp.stanford.edu/software/CRF-NER.shtml`.

The Stanford Natural Language Processing Group. 2010. The Stanford Parser: A statistical parser. `http://nlp.stanford.edu/software/lex-parser.shtml`.