JUNLG-MSR: A Machine Learning Approach of Main Subject Reference Selection with Rule Based Improvement

Samir Gupta
Department of Computer Science and Engineering, Jadavpur University.
Kolkata-700032, India.
samir.ju@gmail.com

Sivaji Bandopadhyay
Department of Computer Science and Engineering, Jadavpur University.
Kolkata-700032, India.
sivaji_cse_ju@yahoo.com

Abstract

The GREC-MSR task is to generate appropriate references to an entity in the context of a piece of discourse longer than a sentence. In MSR '09 run of this task, the main aim is to select the actual main subject reference (MSR) from a list of given referential expressions that is appropriate in context. We used a machine learning approach augmented with some rules to select the most appropriate referential expression. Our approach uses the training set for learning and then combines some of the rules found by observation to improve the system.

1 Introduction

In this paper we provide a description of our system for the GREC MSR task of Generation Challenges 2009. GREC-2.0 Corpus of 2,000 Wikipedia introduction sections in which references to the main subject of the Wikipedia article have been annotated was provided to us by the organizers. The corpus was divided into five different domains like cities, countries, mountains, people and rivers.

The basic approach we used was to develop a baseline system first by training the system on the training set. This system then selects the most frequent referential expression based on a number of parameters of the corresponding reference. After evaluation on the development set we used the development set to deduce certain rules based on observation and iteratively added these rules to the system and evaluated resulting performance. Thus the system development can be divided into two phases which are discussed in sections 2 and 3.

2 Baseline System: Training and Classification

The machine learning approach we used for the baseline system was domain independent and hence was build by populating a single database with the training set data. First we parsed the contents of the XML files of the training sets using a Java DOM XML Parser. Then we inserted the training set data into the database named grec which had two tables: parsed_ref and possible_refex. There is a one to many mapping from possible_refex to parsed_ref. The possible_refex contains all possible REFEX elements i.e. referential expressions possible while parsed_ref contains all the parsed references of the training set with attributes such as syncat, semcat, paragraph number, reference number (with respect to a paragraph), sentence number and a foreign key refex id referring to the possible_refex table.

The prediction of the referential expression was done based on features such as the semantic category, syntactic category, paragraph number, reference number with respect to a paragraph and sentence number of the referent. One example from the database is, if the semcat of the reference is cities, syncat is np-subj, paragraph number is 2, ref number is 1 and sentence number equals 1 then in 74% of the cases of the training set the referential expression was with refex id=1 (i.e. type=common, emphatic=no, head=nominal and

Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLP 2009, pages 103–104,
Suntec, Singapore, 6 August 2009.

case= plain) and refex id = 4 (i.e. type=name, emphatic=no, head=nominal and case= plain) had the second highest count (19.6%). Thus we selected the most frequent refex from the possible referential expressions corresponding to the feature set of the reference, based on their count in the training set populated database. These decision rules with their associated probabilities are stored in a table which served as our model for classification. When a number of referential expressions from the alt_refex match from the list of the given refexes then we select the refex with the longest surface form. In certain case when the refex was not in the alt_refex element we select the second best case from our decision model. Results of this intermediate baseline system are given in Table 1.

Domain	String Acc.	Reg 08 type Acc.	Mean Edit Distance	Norm. mean edit distance
Cities	0.404	0.495	1.657	0.575
Countr.	0.468	0.576	1.467	0.471
Mount.	0.567	0.646	1.192	0.380
People	0.576	0.673	0.902	0.379
Rivers	0.6	0.6	1.06	0.36
Overall	0.532	0.62	1.205	0.421

Table 1: Baseline Results

3 Rule based Improvement

After the baseline system was evaluated on the development set we iteratively added some rules to optimize the system output. These rules are applied only when a reference matches the below stated condition, otherwise the result from the baseline system was used.

The different rules that we deduced are as follows:

- The referential expression is empty if its immediate preceding word is a conjunction and the referent's syncat is np-subj. Thus the surface form of the refex is null.
- In the people domain if the best case output from the baseline results in Reg-type = "name" and if earlier in the paragraph the person's full name has been referred to, then subsequent references will have a shorter version of the referential expression i.e. shorter surface form (example: Zinn's instead of Howard Zinn's)
- If the same sentence spans two or more references then generally a pronoun form is used if a noun has been used earlier.
- Generally common form of the noun is used instead of the baseline pronoun output if words like in, for, to, of, in precedes the reference (maximum distance 3 words). This rule is applied to all domains except people.

The first and the last rules had some effect to the system but the improvement from the other rules was very negligible. Final results are tabulated in Table 2.

4 Results

We provide final results of our system in Table 2 Script geval.pl was provided by the organizers for this purpose. We see that inclusion of the above rules in the system increased it's accuracy by almost 4-5%. More rules can be added to system by studying cases of the training set which do not get classified correctly by the best case baseline system. Overall reg08 accuracy, precision and recall were 66.4 %.

Domain	String Acc.	Reg 08 type Acc.	Mean Edit Dist.	Norm. mean edit Dist.
Cities	0.434	0.525	1.596	0.544
Countr.	0.5	0.619	1.381	0.431
Mount.	0.583	0.663	1.158	0.363
People	0.659	0.756	0.746	0.296
Rivers	0.65	0.65	0.95	0.31
Overall	0.575	0.664	1.12	0.377

Table 2: Final Results

References

Anja Belz and Albert Gatt. 2008. *Grec Main Subject Reference Generation Challenge 2009: Participants' Pack.* http://www.nltg.brighton.ac.uk/research/genchal09

Anja Belz, Eric Kow, Jette Viethen, Albert Gatt. 2008. The GREC Challenge 2008: Overview and Evaluation Results. *In Proceedings of the Fifth International Natural Language Generation Conference (INLG-2008)* pages 183-192.