JU_CSE_GREC10: Named Entity Generation at GREC 2010

Amitava Das[1], Tanik Saikh[2], Tapabrata Mondal[3], Sivaji Bandyopadhyay[4]
Department of Computer Science and Engineering
Jadavpur University,
Kolkata-700032, India
amitava.santu@gmail.com[1], tanik4u@gmail.com[2], tapabratamondal@gmail.com[3], sivaji_cse_ju@yahoo.com[4]

Abstract

This paper presents the experiments carried out at Jadavpur University as part of the participation in the GREC Named Entity Generation Challenge 2010. The Baseline system is based on the SEMCAT, SYNCAT and SYNFUNC features of REF and REG08-TYPE and CASE features of REFEX elements. The discourse level system is based on the additional positional features: paragraph number, sentence number, word position in the sentence and mention number of a particular named entity in the document. The inclusion of discourse level features has improved the performance of the system.

1 Baseline System

The baseline system is based on the following linguistic features of REF elements: SEMCAT (Semantic Category), SYNCAT (Syntactic Category) and SYNFUNC (Syntactic Function) (Anja Belz, 2010) and the following linguistic features of REFEX elements: REG08-TYPE (Entity type) and CASE (Case marker). The baseline system has been separately trained on the training set data for the three domains: chefs, composers and inventors. The system has been tested on each development set by identifying the most probable REFEX element among the possible alternatives based on the REF element feature combination. The probability assigned to a REFEX element corresponding to a certain feature combination of REF element is calculated as follows:

$$p(R_v) = \frac{N_{REFEX}^{D_i}}{N_{REF}^{D_i}}$$

where $p(R_v)$ is the probability of the targeted REFEX element to be assigned, $N_{REF}^{D_i}$ is the total number of occurrences of REF element feature combinations, D_i denotes the domain i.e., Chefs, Composers and Inventors and $N_{REF}^{D_i}$ denotes the total number of occurrences of the REFEX element corresponding to the REF feature combination.

It has been observed that many times the most probable REFEX element as identified from the training set is not present among the alternative REFEX elements. In these cases the system assigns the next highest probable REFEX element learnt from the training set that matches with one of the REFEX elements among the alternatives. In some cases more than one REFEX element get same probability in the training set. In these cases, the REFEX element that occurs earlier in the alternative set is assigned. The experimental result of Baseline system is reported in Table 1.

	Chefs	**Composers**	**Inventors**
Precision	0.63	0.68	0.70
Recall	0.69	0.60	0.64
F-Measure	0.66	0.64	0.68

Table 1: Result of Baseline System

2 Discourse Level System

The discourse level features like paragraph number, sentence number and position of a particular word in a sentence have been added with the features considered in the baseline system. As mentioned in Section 1, more than one REFEX element can have the same probability value. This happens as REFEX elements are identified by two features only REG08-TYPE and CASE.

	Name	Pronoun	Common	Empty
Chefs	2317	3071	55	646
Composers	2616	4037	92	858
Inventors	1959	2826	75	621

Table 2: Distribution of REFEX Types among three domains.

The above problem occurs mainly for *Name* type. Pronouns are very frequent in all the three domains but they have small number of variations as: he, her, him, himself, his, she, who, whom and whose. Common type REFEX elements are too infrequent in the training set and they are very hard to generalize. *Empty* type has only one REFEX value as: “_”. The distribution of the various REFEX types among the three domains in the training set is shown in Table 2.

2.1 Analysis of *Name* type entities

Table 2 shows that *name* types are very frequent in all the three domains. *Name* type entities are further differentiated by adding more features derived from the analysis of the *name* type element.

Firstly, the full name of each named entity has been identified by Entity identification number (id), maximum length among all occurrences of that named entity and case marker as plain. For example, in Figure 1, the REFEX element of id 3 has been chosen as a full name of entity “0” as it has the longest string with case “plain”.

After identification of full name of each REFEX entity, the following features are identified for each occurrence of an entity:: Complete Name Genitive (CNG), Complete Name (CN), First Name Genitive (FNG), First Name (FN), Last Name Genitive (LNG), Last Name (LN), Middle Name Genitive (MNG) and Middle Name (MN). These features are binary in nature and for each occurrence of an entity only one of the above features will be true.

Pronouns are kept as the REFEX element feature with its surface level pattern as they have only 9 variations. Common types are considered with tag level “common” as they hard to generalize. Empty types are tagged as “empty” as they have only one tag value “_”.

1	<REFEX ENTITY="0" REG08-TYPE="name" CASE="genitive">Alain Senderens's</REFEX>	CNG
2	<REFEX ENTITY="0" REG08-TYPE="name" CASE="genitive">Senderens's</REFEX>	LNG
3	<REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Alain Senderens</REFEX>	CN
4	<REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Senderens</REFEX>	LN

Figure 1: Example of Full Name Identification

3 Experimental Results

The experimental results of the discourse level system on the development set are reported in the Table 3 and Table 4 respectively. Table 3 reports the results when the system has been trained separately with domain specific training set and Table 4 reports the results when the training has been carried out on the complete training set.

The comparison of the results of the baseline and the discourse level system shows an overall improvement. But there are some interesting observations when comparing the results in Table 3 and Table 4. Currently detailed analyses of the results are being carried out.

	Chefs			Composers			Inventors		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Name	0.69	0.74	0.71	0.78	0.61	0.69	0.77	0.67	0.71
Pronoun	0.81	0.76	0.79	0.70	0.84	0.76	0.76	0.87	0.81
Common	0.76	0.87	0.81	0.37	0.44	0.40	0.44	0.65	0.68
Empty	0.92	0.88	0.90	0.86	0.92	0.89	0.72	0.65	0.68

Table 3: Experimental Results of Discourse Level System on the Development Set (Training with Domain Specific Training Set)

	Reg08 Type		**String Accuracy**	**BLEU**				**NIST**	**String Edit Distance**	
	Precision	**Recall**							**Mean**	**Mean Normalized**
				1	**2**	**3**	**4**			
Chefs	0.66	0.70	0.57	0.68	0.70	0.76	0.81	3.70	0.77	0.38
Composers	0.63	0.67	0.56	0.61	0.57	0.54	0.50	3.34	1.07	0.40
Inventors	0.60	0.62	0.50	0.55	0.54	0.52	0.49	2.90	1.25	0.47
Total	0.63	0.66	0.54	0.61	0.58	0.57	0.55	3.83	1.03	0.42

Table 4: Table 4: Experimental Results of Discourse Level System on the Development Set (Training with Complete Training Set)

References

Anja Belz. 2010. GREC Named Entity Generation Challenge 2010: Participants’ Pack.