Using Berkeley Coreference System to compare models trained on various genres of Ontonotes 5.0 data

Abstract

Our work with quiz bowl coreference data indicates that performances of models trained for coreference resolution varies when the data they are trained on differ sufficiently in genre to the data in which coreference is to be detected. So we try to build models on the various genres present in OntoNotes. We used OntoNotes 5.0, as opposed to 4.0 which we had used while writing our paper. Not only is version 5.0 newer, it has certain genres of coreference which weren't present in earlier versions.

1 Introduction

We divide OntoNotes 5.0 into its constituent genres, namely, broadcast conversations, broadcast news, magazines (comprising of translated sinorama), newswire data (comprising of Wall Street Journal, translated Xinhua etc.), pivot data (comprising of the Old and the New Testament), telephonic conversations, and weblogs. Out of these, the newswire section and the weblog sections are very sparsely annotated in OntoNotes 5.0 for coreference, hence we used only those documents which had coreference annotations. While none of these genres are qualitatively as different from standard newswire text as quiz bowl data, dividing them and evaluating their performance on each other as well as quiz bowl data proved to be interesting. However, as a point of caution, the quantity of data for each of the genres varies considerably, and so these performances should not be taken as a final indicator of suitability of these genres for difficult coreference resolution task.

2 Observations and Analysis

If we disregard the size of data present in each genre dataset as well as the annotation density of the documents present in those, and by removing the un-annotated documents from newswire and weblog sections, the model trained on broadcast conversations has the highest precision tested on all the other genres while newswire has the highest recall.

In order to tie these observations with our work, we will next try these genres along with word embeddings, as was present in our system. That will tell us if the different genres have different improvements once word embeddings are added to them, and how much that difference compares to the quiz bowl genre as it had a significant improvement on our system as compared to the Berkeley system.

	P								R								F1							
	QB	BC	BN	MZ	NW	PT	TC	WB	QB	BC	BN	MZ	NW	PT	TC	WB	QB	BC	BN	MZ	NW	PT	TC	WB
QB		**48.45**	42.79	45.12	46.74	45.02	38.71	41.99		14.77	17.56	**19.01**	19	13.22	18.96	17.89		22.25	24.46	**26.24**	26	20.21	25	24.6
BC	38.29		49.62	50.86	50.9	43.6	40.96	**51.88**	22.63		48.34	45.32	48.1	33.45	**49.25**	47.59	28.28		48.62	47.8	49	37.3	44.69	**49.36**
BN	44.62	**59.86**		57.49	58.7	49.24	43.45	56.22	22.47	36.55		49.59	**52.31**	29.56	48.47	48.22	29.77	45.13		53.21	**55.22**	36.53	45.81	51.82
MZ	48.41	**65.14**	60.88		64.55	52.71	43.4	62.51	28.13	43.4	58.28		**62.1**	37.08	52.85	42.74	35.3	51.73	59.41		**63.16**	42.88	47.65	60.52
NW	41.24	**60.15**	57.6	59.21		52.53	39.51	58.84	20.2	33.74	51.19	**52.89**		27.64	47.12	50.96	27.01	42.93	54.17	**55.86**		35.59	42.97	54.55
PT	35	44.06	44.04	**46.75**	46.69		40.94	44.81	28.99	41.11	56.55	50.04	54.58		**58.23**	51.57	31.4	41.89	49.37	48.21	**50**		48.06	47.66
TC	0	**53.35**	51.17	51.55	50.88	46.4		50.37	0	46.56	49.6	48.45	**51.7**	42.96		46.13	0	49.7	49.91	49.93	**51.22**	44.24		47.64
WB	40.48	57.57	52.68	**57.73**	58.2	47.04	39.7		26.64	42.52	52.15	54.39	**57.39**	36.01	51.38		32.07	48.75	52.26	55.97	**57.71**	40.27	44.79	

Table 1: Results of training and testing on the various genres. The row represents the genre tested and the column the genre the model was trained on.