CatolicaSC at TAC 2011

Paulo C F de Oliveira
Catholic University of Santa Catarina State
Joinville, SC - Brazil
paulooliveira@catolicasc.org.br

Abstract

The purpose of this paper is to report our participation in the Text Analysis Conference 2011. We also analyze the performance of VERT-F, which is the system developed by our team. The results have shown that our system achieved a good performance, compared to the other peer systems.

1 Introduction

In this paper, we report the participation of our institution at the Text Analysis Conference (TAC, 2011). We have participated in the Summarization Track, and specifically in the Automatically Evaluating Summaries of Peers (AESOP) task.

This task evaluates a summary automatically for a given metric, which has been devised by a peer evaluation system. We had only one submission for this task.

2 VERT [1] – our summary evaluation system

VERT-F is based on a graph theory method that performs sentence matching. This method is called the maximum bipartite matching problem (MBMP). The MBMP leads to the well-known precision and recall – the most common metrics used to evaluate Natural Language Processing (NLP) systems (Salton & McGill, 1983) (van Rijsbergen, 1979).

In graph theory, a bipartite graph is a special graph where the set of vertices can be divided into two disjoint sets with two vertices of the same set never sharing an edge (Cormen, Leiserson, Rivest, & Stein, 2001); (The Open University, 2001).

An overall description about our system can be found in (Oliveira, 2005).

3 Analyzing the TAC results (AESOP task)

In this section VERT-F's performance will be analyzed. According to NIST, there has been something new this year. In addition to the metrics that reflect summary content, namely, *the (Modified) Pyramid score*, which measures summary content, and *Overall Responsiveness*, which measures a combination of content and linguistic quality, AESOP focused on *Readability*, as well. NIST used correlation analysis, i.e. correlation comparison between the automatic metric scores produced by a system and human scores (Mani, 2000).

[1] VERT stands for **V**aluation using **E**nhanced **R**ationale **T**echnique

3.1 Test Data

For the AESOP task, NIST has used all test data and summaries produced within the TAC 2011 Guided Summarization task. In fact, the same collection from the newswire portion of the TAC 2010 KBP Source Data (LDC Catalog no. LDC2010E12). It comprises news articles taken from several sources such as the New York Times, the Associated Press, and the Xinhua News Agency newswires. The news collection spans the time-period 2007-2008.

3.2 NIST Evaluation Procedure

The TAC data set consisted of human-authored summaries (i.e. model summaries) and automatic (non-model) summaries. In order to process our run, we have used these model summaries as the reference summaries, and the others as candidate summaries.

According to NIST, 7 participants submitted 22 metrics in the AESOP task, resulting in 25 metrics which were evaluated. These submissions have been labeled by a random number (i.e. 1-25). Submissions 1, 2 and 3 refer to ROUGE-2, ROUGE-SU4, and BE baselines, respectively. Our team submitted only 1 run (ID no. 4).

For each of our automatic metric submitted, NIST computed Pearson's, Spearman's, and Kendall's correlations with *Pyramid*, *Overall Readability*, and *Overall Responsiveness*, as well as the discriminative power of the automatic metric in comparison with these three manual metrics.

Table 1 and Table 2 present the NIST calculations results for all the participants against the 3 scores above, and for the initial and the update summaries, respectively (All Peers case).

Table 1 shows that in terms of correlation with *Pyramid* score, VERT-F obtained a very good degree of correlation for the 3 correlation coefficients (i.e. rank 1 for Pearson, and rank 2 for Spearman and rank 3 for Kendall). It is a remarkable performance, if it is compared to the 24 others metrics.

Meanwhile, the correlations with the *Responsiveness* score, our system achieved rank 1 for the Pearson coefficient, and rank 2 for both Spearman and Kendall coefficients. That was a good performance as well.

And finally, the correlations with the *Readability* score; VERT-F reached rank 1 for Pearson and Kendall coefficients, and rank 2 for Spearman, which is a good performance too.

As can be seen in Table 2, our system has obtained a satisfactory performance with the *Pyramid* score, that is, rank 3 for Pearson and rank 5 for both Spearman and Kendall coefficients.

For the *Responsiveness* score, VERT-F achieved rank 2 for all 3 coefficients, which can be considered a good performance.

For the *Readability* score, VERT-F achieved rank 1 for all 3 coefficients. That was a very good performance.

4 Conclusions

TAC 2011 results have shown that our automatic evaluation metric accomplished an excellent and robust performance, especially when compared to the other 24 participants.

We can then conclude that our system scores correlated highly and positively in relation to the official NIST scores, as (Mani, 2000) has pointed out.

We believe that our system, VERT-F, can be used as an automatic summary evaluation metric.

References

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. 2001. Introduction to Algorithms (2nd ed.). MIT Press.

Mani, I. 2000. Automatic Summarization. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Oliveira, P. C. F. 2005. How to evaluate the 'goodness' of summaries automatically. PhD Thesis, University of Surrey, Department of Computing, Guildford, UK.

Salton, G., & McGill, M. J. 1983. Introduction to Modern Information Retrieval. McGraw Hill.

TAC. 2011. The Text Analysis Conference. Retrieved from http://www.nist.gov/tac/

The Open University. 2001. Networks - assignment and transportation (Vol. 3).

van Rijsbergen, C. 1979. Information Retrieval (2nd ed.). London: Butterworths.

Initial Summaries									
Run no.	Correlations with Pyramid			Correlations with Responsiveness			Correlations with Readability		
4	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
	r	*rho*	*tau*	*r*	*rho*	*tau*	*r*	*rho*	*tau*
	0.974	0.933	0.785	0.972	0.894	0.740	0.926	0.672	0.519
Rank	**2**	**1**	**3**	**1**	**2**	**2**	**1**	**2**	**1**
Baseline 1	0.752	0.864	0.703	0.779	0.948	0.609	0.663	0.498	0.374
Baseline 2	0.763	0.886	0.723	0.810	0.966	0.629	0.682	0.533	0.400
Baseline 3	0.781	0.878	0.720	0.784	0.941	0.590	0.683	0.531	0.387
Min	0.113	0.179	0.078	0.093	0.187	0.090	0.049	0.083	0.036
Mean	0.720	0.750	0.609	0.703	0.708	0.552	0.653	0.505	0.374
Max	0.975	0.933	0.799	0.972	0.899	0.748	0.926	0.674	0.674
Std.Dev.	0.325	0.298	0.279	0.330	0.272	0.243	0.329	0.219	0.177

Table 1: The 3 correlations of the AESOP metrics with the Pyramid. Responsiveness and Readability scores for the initial summaries

Updated Summaries									
Run no.	Correlations with Pyramid			Correlations with Responsiveness			Correlations with Readability		
4	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
	r	*rho*	*tau*	*r*	*rho*	*tau*	*r*	*rho*	*tau*
	0.950	0.873	0.695	0.974	0.911	0.762	0.934	0.663	0.507
Rank	**3**	**5**	**5**	**2**	**2**	**2**	**1**	**1**	**1**
Baseline 1	0.775	0.851	0.684	0.717	0.869	0.710	0.712	0.550	0.399
Baseline 2	0.730	0.883	0.720	0.675	0.903	0.743	0.686	0.558	0.405
Baseline 3	0.740	0.848	0.686	0.649	0.808	0.637	0.611	0.415	0.287
Min	0.017	0.159	0.103	-0.058	0.063	0.018	-0.043	-0.101	-0.116
Mean	0.680	0.697	0.552	0.652	0.693	0.554	0.626	0.426	0.307
Max	0.953	0.891	0.731	0.975	0.912	0.764	0.934	0.663	0.663
Std.Dev.	0.358	0.282	0.237	0.392	0.331	0.284	0.366	0.281	0.225

Table 2: The 3 correlations of the AESOP metrics with the Pyramid. Responsiveness and Readability scores for the updated summaries