BEwT-E in the TAC 2010 AESOP Task

Stephen Tratz
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
stratz@isi.edu

Abstract

This paper presents the TAC 2010 AESOP Task results for the BEwT-E (Basic Elements with Transformations for Evaluation) summarization evaluation system. It also includes results using the 2009 version of BEwT-E and describes the changes made for 2010.

1 Introduction

BEwT-E (Tratz & Hovy, 2008) is a newer, more sophisticated implementation of the BE (Basic Element) framework (Hovy et al., 2005; Hovy et al., 2006) that uses transformations to match BEs (minimal-length content units) that are semantically similar but lexically and/or syntactically different. BEwT-E has previously shown high correlation with human judgments in earlier text summarization tasks, including an excellent showing in TAC 2009's AESOP text summarization evaluation task (Tratz & Hovy, 2009). The purpose of this brief document is to explain the changes made to BEwT-E this year and present the correlation results for its participation in TAC 2010's AESOP task.

2 Changes for 2010

The central change for 2010 was to replace the parser (Charniak, 2000) that was being used with an in-house dependency parser. There were three motivations for doing this. First, the dependency parser can parse over 60 sentences per second, which is substantially faster than the default installation of the Charniak parser. Also, since the parser is written in Java (the same language as BEwT-E), it is more easily ported across operating systems than the Charniak parser, which is written in C. Finally, it is simpler to write rules for extracting BEs from dependency trees than from constituent trees. Though both the dependency parser and the version of BEwT-E that uses it have not yet been publicly released, they will be released in the near future.

3 Results

The correlations results for BEwT-E's participation in TAC 2010's AESOP task are presented in Table 1. For comparison, results produced using the previous version of BEwT-E are given in Table 2.

		Pearson		**Spearman**	
		All	Auto	All	Auto
A	Mod Pyramid	.927	.929	.902	.842
	Responsiveness	.928	.892	.879	.805
B	Mod Pyramid	.868	.905	.880	.804
	Responsiveness	.850	.871	.845	.752

Table 1: Correlation results versus Modified Pyramid and Overall Responsiveness scores for the base documents (A) and update documents (B), both including (All) and excluding (Auto) human written summaries.

		Pearson		**Spearman**	
		All	Auto	All	Auto
A	Mod Pyramid	.908	.934	.915	.862
	Responsiveness	.911	.899	.893	.828
B	Mod Pyramid	.859	.909	.883	.807
	Responsiveness	.841	.879	.849	.759

Table 2: Same as Table 1, but using the previous version of BEwT-E.

4 Discussion

Overall, the impact of using the new dependency parser on the correlation results was limited.

While BEwT-E continues to show strong correlation with human judgments, the overall correlation with human judgments was lower than what was anticipated based upon prior experience. One source of this difference may be the differences in the task itself, such as the lower word limit than previous text summarization tasks. It is also possible that some particular component that BEwT-E relies on, such as the sentence splitter or named entity recognizer, failed to perform well due to some oddities in the data. Conceivably, there may also be cases where a particular system employs a summarization strategy substantially different from others and, as a consequence, receives a substantially higher or lower score than might otherwise be expected. While taking a quick look at some results from systems that were substantially misranked by BEwT-E, we noticed at least one case where the summarization system produced many *very* short sentences; this might result in an increase (or decrease) of content words and, thus, somehow throw BEwT-E off.

5 Future Work

We would like to examine in detail the summaries and summarization engines that BEwT-E misranked. We are actively working on the dependency parser and plan to release both it and a newer version of BEwT-E later in 2011.

References

Charniak, E. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*.

Hovy, E.H., C.Y. Lin, and L. Zhou. 2005. Evaluating DUC 2005 using Basic Elements. In *Proceedings of DUC-2005 workshop*.

Hovy, E.H., C.Y. Lin, L. Zhou, and J. Fukumoto. 2006. Automated Summarization Evaluation with Basic Elements In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 06)*.

Tratz, S. and E. Hovy. 2008. Summarization Evaluation Using Transformed Basic Elements. In *Proceedings of TAC-2008*.

Tratz, S. and E. Hovy. 2009. BEwT-E for TAC 2009's AESOP Task. In *Proceedings of TAC-2009*.