BEwT-E for TAC 2009's AESOP Task

Stephen Tratz and **Eduard Hovy**
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
{stratz, hovy}@isi.edu

Abstract

This paper describes BEwT-E (Basic Elements with Transformations for Evaluation), an automatic system for summarization evaluation. BEwT-E is a newer, more sophisticated implementation of the BE framework that uses transformations to match BEs (minimal-length syntactically well-formed units) that are semantically similar but are lexically and/or syntactically different. We present BEwT-E's results on DUC and TAC datasets from 2005 to 2009 and are pleased to report that, of the systems that participated in AESOP, BEwT-E was one of the strongest performers, achieving the best performance using the Spearman metric when evaluated on the TAC 2009 update summaries.

1 Introduction

Human evaluation for text summarization can be time consuming, costly, and prone to human variability (Teufel and van Halteren, 2004; Nenkova and Passonneau, 2004). In order to more efficiently and objectively evaluate text summarization systems, automated evaluation methods have been developed. ROUGE (Lin and Hovy, 2003) uses lexical n-grams to compare human written summaries with computer-generated summaries. Subsequent automated evaluation systems such as ROSE (Conroy and Dang, 2008) have investigated matching variants and additional parameters for the purpose of bringing human and automated summary scores into better correspondence. AutoSummENG is a summarization evaluation method that evaluates summaries by extracting and comparing graphs of character n-grams (Giannakopoulos et al., 2008). Other n-gram methods such as POURPRE have been successfully applied to question answering evaluation (Lin and Demner-Fushman, 2005).

A problem with all these methods is their reliance on surface-level formulation, and the absence of sensitivity to syntactic structure. This problem arises in several forms. The phrase “large car” in a system summary, for example, would not match “large green car” in a gold standard summary, despite “large” and “green” independently modifying “car”. In an attempt to overcome this, ROUGE employed so-called skip n-grams, namely n-grams that can accommodate a small number of skipped items.

Another variant of the problem is the inability to match alternative phrasings. No automated text summarization evaluation system will match “a massive emerald-colored vehicle” to “a large green car”. A third is the inability to handle multi-word names and name aliases, such as “United States”, “USA”, etc.

To overcome these types of shortcomings, the Basic Element summarization method was developed and tested (Hovy et al., 2005; Hovy et al., 2006). This method facilitates matching of expressive variants of syntactically well-formed units called Basic Elements (BEs). The system achieved fairly good correlation with human evaluation. DEPEVAL(summ), a similar metric, uses a different parser, extraction rule set, etc in order to effectively evaluate automatic summaries (Owczarzak, 2009). However, it still only performed rudimentary matching of alternative phrasings, using a list of paraphrases (Zhou et al., 2006). This paper describes a new implementation of the BE method, called BE with Transformations for Evaluation (BEwT-E), that includes a significantly improved matching capability using a variety of operations to transform and match BEs in various ways. The extended BE method generally performs well against other automated methods for evaluating summaries.

We first outline the BE method and our new implementation of it, including BE weighting. Next we describe the transformations we use for more powerful matching. Finally, we describe the system's performance on previous Document Understanding Conference (DUC) data as well as Text Analysis Conference (TAC) data.

2 The Basic Element Method

The intuition behind Basic Elements is to decompose summaries to lists of minimal-length syntactically well-defined units (BEs) and then to compare the two lists to obtain a similarity score. Five issues must be addressed:

- What is the nature of a minimal unit (BE)?
- How are BEs extracted?
- How should each BE be weighted?
- How should matches be determined?
- How should the matches be combined into an overall score?

As described in (Hovy et al., 2005), each BE is a syntactic unit (a single word or multi-word phrase; a modifier-head pair, etc.). In the new implementation, each BE consists of a list of one to three words and their associated parts-of-speech or NER type. Examples of these include:

- Unigram BEs: all nouns, verbs, and adjectives found in the summary
- Bigram BEs: subject+verb, verb+object, headnoun+headnoun_of_appositive, verb+adverb, adj+noun, verb+adjective, prenominal_noun+head_noun, possessor+head_noun, verb+particle.
- Trigram BEs: two head words connected via a preposition

3 Comparing Summaries

3.1 Extracting BEs

In order to extract the BEs, we first parse the summaries using the Charniak parser (Charniak and Johnson, 2005), identify named entities using the LingPipe NER system (Baldwin and Carpenter), and then extract the BEs using a series of Tregex rules (Levy and Andrew, 2006). Tregex rules can be thought of as regular expressions over trees. Examples of the Tregex rules used by BEwT-E and the BEs they produce for a sample sentence are given in Figure 1.

If a token identified for extraction by a BE extraction rule falls within a string recognized by a Named Entity Recognition (NER) system as an entity, the entire named entity string is extracted in place of the word.

In previous work (Tratz & Hovy, 2008), including several identical BEs extracted from the same document was found to generally be harmful to the overall effectiveness of the system, and, therefore we have only included a single instance of each BE when calculating the results presented in this paper.

```
John's cat drank milk.
Charniak parse:
(S1 (S (NP (NP (NNP John) (POS 's)) (NN cat)) (VP
(VBD drank) (NP (NN milk))) (. .)))

Rule Name: Verb to NPHead
Tregex:VP [<# __=x & < (NP <# !POS=y)]
Tokens to Extract: xy
Extracted BEs: drank|VBD+milk|NN

Rule Name: Possessor of NPHead
Tregex: NP [< (NP <# (POS $- __=x)) & <# __=y]
Tokens to Extract: xy
Extracted BEs: John|Person+cat|NN
```

Figure 1. Example sentence, its Charniak parse, and the output from two BE extraction rules.

3.2 Weighting BEs

In weighting the BEs, a basic assumption to date has been that a fragment of content mentioned in several reference summaries is more important, and should weigh more, than a fragment mentioned in only one. In manual studies, both Teufel and van Halteren (2004) and Nenkova and Passonneau (2005; the Pyramid Method) adopt the 'popularity score' rule: a fragment (called SCU or semantic content unit in the latter) is assigned points equal to the number of reference summaries containing it. Previous work showed that giving additional weight to BEs was, at best, minimally effective and was often detrimental instead (Tratz & Hovy, 2008). Thus, for this paper we only report scores using the *binary* weighting scheme (each matched reference BE is worth 1 regardless of the number of summaries containing it).

3.3 Transformations Definition

The focus of our work is the matching and tallying of BEs from system and human summaries. The original BE system matched primarily by lexical identity and was later expanded by paraphrase substitution using a large list of paraphrase alternatives extracted from a machine translation system (Zhou et al., 2006). However, it is usually possible to express similar information using a wide variety of differences. Recognizing such matches typically requires humans. No automated system today can recognize all variants and know which degrees of semantic similarity they express.

Nonetheless, one can make inroads in addressing this problem automatically. BEwT-E uses a set of transformations to match BEs that convey similar semantic content yet are lexically different. One example of a transformation would be something that allows hypernyms/hyponyms to match (this particular transformation was used in earlier work (Tratz & Hovy, 2008), but was found to be detrimental and therefore was excluded from the present results). What exactly constitutes acceptable similarity is captured by the transformations used by BEwT-E, which are listed below.

Add/Drop Periods: Abbreviations can often occur with or without periods. To handle this, this transformation adds or drops periods. This transformation enables BEs like "U.S.A.|Location" and "USA|Location" to match.

Noun Swapping for IS-A type rules: Some BE extraction rules, such as the rule for handling appositives, extract a pair of nouns that are expected to exhibit an IS-A relationship. Since the order of these nouns is unimportant, this transformation allows the BEs to match even if the nouns are in reverse order. For example, this transformation enables "Phelps|Person+swimmer |NN" to match "swimmer|NN+Phelps|Person".

Prenominal Noun ↔ Prepositional Phrase: This transformation converts BEs such as "Iraq|Location+invasion|NN" into similar BEs such as "invasion|NN_of|IN_Iraq|Location", or vice versa.

Nominalization: This transformation is similar to the denominalization transformation except it operates in the opposite direction. For example, this transformation lets "gerbil|NN_hibernated|VBD" match "hibernation|NN+of|IN+gerbil|NN".

Denominalization: It is common for one reference to an event to occur in the form of a verb while another reference to the same event occurs as a noun. To transform BEs from the noun form back to the verb form, this transformation utilizes the "derivationally related form" relationship links in WordNet (Miller et al., 1990). For example, this transformation enables the BE "rejection|NN+of|IN+John|Person" to match either "John|Person+reject|VB" or "reject|VB+John| Person".

"Role" Transform: In some sentences, the role a person plays appears as a prenominal noun next to his/her name while in other sentences the person is observed performing the action associated with the role. This transformation was created to handle these situations. For example, this transformation enables BEs "Barry_Bonds|Person+hit |VBD" and "hitter|NN+Barry_Bonds|Person" to match. In order to do this, it uses the "derivationally related form" relationship links in WordNet.

Adjective to Adverb: This transformation converts BEs with an adjective and an event word such as "quick|JJ+at|IN+coronating|VBG", "quick|JJ+coronation|NN", into similar BEs with a verb and adverb such as "quickly|RB+coronate| VB". Derivationally related form WordNet links are used to obtain the new verb part.

Adverb to Adjective: This transformation performs the opposite function as the Adjective to Adverb transformation. To map from adverbs to adjectives, it uses pertainym WordNet links.

Pronoun Transform: Pronouns are commonly used in place of more specific references, presenting problems for NLP systems. This transform allows personal pronouns to match person names and the plural pronouns "they" and "them" to match organization names and plural nouns. Thus, "Alcoa|Organization" could match "they|PRP" and "John" could match "he|PRP".

Name Shortener/Expander: This transformation transforms entity names so that BEs like "John_ B_Smith|Person" can match BEs like "Smith| Person", "John|Person" or "John_Smith|Person" and organization names like "Google|Organization" can match "Google_Inc|Organization".

Abbreviations/Acronyms: BEwT-E has a transformation that enables matching abbreviations with their expanded form. This transformation consists of two parts. This first part is simply a lookup list of common abbreviations that includes lists of person titles, street names, states, provinces, measurements, and countries. The second part is a block of code capable of generating some of the most likely abbreviations for persons, organizations, and locations. This transformation enables "UN|NNP" to match "United_ Nations|Organization".

Lemmatization/Delemmatization: Words in BEs can be transformed so that they match regardless of tense and number. For example, this transformation enables “green|JJ+plants|NNS” to match “green|JJ+plant|NN”.

Synonyms: This transformation matches nouns, verbs, and adjectives to their synonyms using WordNet. Words are assumed to be instances of their most frequent sense. For example, this transformation enables “drink|VB+potion|NN” to match “imbibe|VB+potion|NN”.

Pertainyms Transform: Using pertainym and “derivationally related form” relationship links in WordNet, this transform enables BEwT-E to match BEs like “America|Location” to “American|JJ” and “biological|JJ+instruments|NNS” to “biology|NN+instruments|NNS”.

Membership Meronym/Holonym Transform: Unfortunately, due to limitations of WordNet, there are cases when the “pertainyms” transformation does not perform as many transformations as one would expect. By using membership meronym and holonym links from capitalized entries in WordNet, this transformations enables BEwT-E to match BEs like “China|Location+people|NNS” and “Chinese|JJ+people|NNS”.

Preposition Generalization: The Preposition Project has produced a sense inventory of English prepositions (Litkowski and Hargraves, 2005). This was used to create a list of all legal preposition mappings so that prepositions could be expanded. For example, this transformation enables “man|NN+from|IN+La_Mancha|Location” to match “man|NN+of|IN+La_Mancha|Location”. If BEwT-E utilized a preposition sense disambiguation system, this transformation could be further restricted.

Many of these transformations can be applied more or less aggressively. For example, synonym lookups could be limited only to bigram and trigram BEs and/or could use all available WordNet senses instead of just the most frequent sense. Exploring the potential and risks of such degrees is an interesting subject for future research. In the system to date, we have tried to keep the transformations simple.

3.4 Transformations Implementation

The application of the transformations occurs during a step between BE extraction and the overall score computation. Each summary is processed separately.

First, a reference BE pool of all the BEs extracted from the references for a particular summary is constructed. This pool is the complete set of BEs that other BEs may be mapped to.

Before a summary's BEs are passed individually through the pipeline, the summary's BEs and the reference BEs are passed into a reinitialization method of each of the transformations. The purpose of this method call is to give the name shortener/expander, abbreviation, and pronoun replacement transformations a chance to build up a set of legal term substitutions so that they will operate more efficiently on the individual BEs.

After the transformations have been reinitialized, each BE for the current summary is passed through the transformation pipeline. A diagram of the transformation pipeline is given in Figure 2. Any transformed versions of the BE are passed into the subsequent transformation. The transformed versions of the original BEs that match at least one of the BEs in the reference set are saved along with the list of transformations used to produce them.

To reduce the number of computations performed, a list of the transformed versions of a BE is maintained along with the list of set(s) of transformations used to produce each transformed version. If a transformed version of a BE is identical to a previous production and uses a superset of the transformations used in the previous production, the new production will be ignored and not passed to the next transformation.

Many possible transformation orderings exist. The current order is based upon human intuition. The noun swap and period modification transformations, which are unlikely to make mistakes but may positively affect the outcome of later transformations are first. Following these are the transformations that affect a BE's structure, including the transformations that may result in a combination of added/removed central preposition, changed parts-of-speech, and/or changed word position. These were placed before the simple term substitution transformations under the assumption that the reverse order would be more error prone. The remaining transformations only affect individual terms within the BEs. These transformations start with ones related to names, including the name shortener/expander, pronoun, and abbreviations and then lead into the transformations that use simple WordNet or preposition substitutions. Finally, the “delemmatize” transformation ends the pipeline. The impact of transformation order is an area for future research.

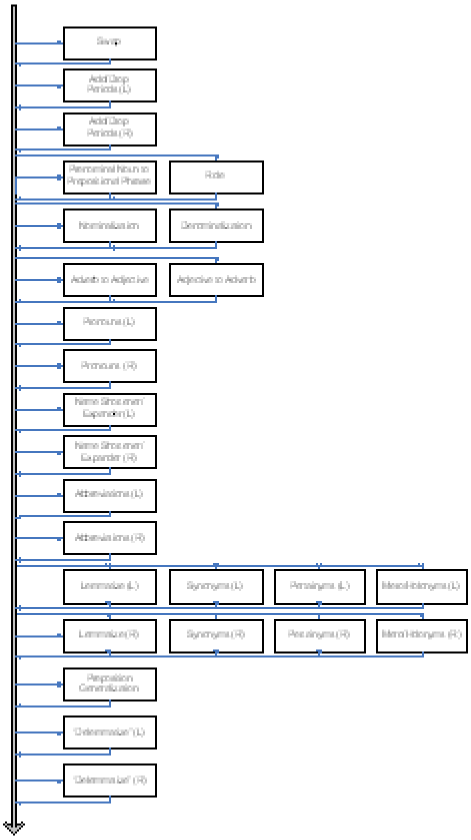

Figure 2. Diagram of pathways through the BE transformation pipeline. 'L' and 'R' indicate whether the transformation is limited to the leftmost or rightmost word in the BE.

3.5 Computing the Overall Score

After undergoing several transformations, a single BE may match several of the reference summary’s BEs. These reference summary BEs may have different weights based upon their frequency in the reference summaries and, in future versions of BEwT-E, the matching may have a value less than 1.0 if a transformation was required to perform the match. This complicates the scoring process because, in computing the comparison score between two summaries, no BE is allowed to match or be matched multiple times.

The BE matching problem is essentially an instance of the weighted assignment problem and the unnormalized formula is expressed mathematically in Figure 3. The BE weighting function W determines the weight of the reference BE and is discussed in Section 3.2. The comparison function C returns a measure of how similar a pair of BEs is. Currently, C always returns 1.0 even though parameters exist for adjusting the similarity of the match based upon the set of transforms used to produced it. In the future, these parameters may be tuned.

BEwT-E implements a successive shortest paths (also know as shortest augmenting paths) algorithm to find the optimal BE matching. For more information regarding using successive shortest paths for solving assignment problems see (Enquist, 1982).

The total value of the matching is normalized by the total weight of the reference summary's BEs. Thus, BEwT-E score is essentially a recall-oriented measure.

$$maximize \sum_{i=0}^{N} \sum_{j=0}^{M} C(i,j)\, W(j)\, x_{ij}$$
$$subject \text{ to}$$
$$\sum_{i=0}^{N} x_{ij} \in \{0,1\}\, forall j where\, 0 \leq j \leq M$$
$$\sum_{j=0}^{M} x_{ij} \in \{0,1\}\, forall i where\, 0 \leq i \leq N$$
$$x_{ij} \in \{0,1\}$$

Figure 3. Problem of calculating unnormalized comparison score between two BE sets using comparison and weighting functions C and W.

3.6 Multiple References

In order to calculate a BEwT-E score when multiple references are available, we compare the peer summary against each of the reference summaries and consider the highest score to be the multi-reference score. However, to account for the fact that comparing a reference summary against itself would result in a perfect score and not comparing it against itself would mean the summary was compared against fewer references than the automatic peers, jackknifing was used and is enabled by default. This involves creating N subsets of the N reference summaries, each of which is missing one reference. The score for each peer summary is then calculated by taking the average of the multi-reference scores produced by using these N different subsets.

4 Results

4.1 Performance on DUC05-07, TAC08

BEwT-E has previously been evaluated on a number of text summarization datasets including those from Document Understanding Conferences (DUC) 2005-2007 and the Text Analysis Conference (TAC) 2008. For the 45–50 topics in each of the DUC evaluations, automated systems generated summaries of at most 250 words. For the 48 topics in TAC 2008, the participating systems produced 2 summaries of at most 100 words each. The first TAC summary was created from a base set of documents representing the topic. The TAC second summary was created using an additional "update" set of documents and was supposed to summarize the information in the "update" set that was not present in the base document set. For both the DUC and TAC conferences, human judges then assigned a score to each system-generated summary by comparing it to the four or more gold standard reference summaries created by humans for each topic.

Our aim is to produce scores that correlate well with average human-produced score and/or rankings of the systems that participated in the DUC/TAC evaluations. We use the Pearson correlation coefficient to measure correlation with the scores and the Spearman coefficient to measure correlation with the rankings. We compare our system's performance on these datasets with other systems such as the original BE system, ROUGE, and AutoSummENG.

In Tables 1 to 7, we present the results for the DUC 2005–2007 and TAC 2008 datasets. The (T on) and (T off) labels indicate whether the transformations are off or on. AutoSummENG05 and AutoSummENG06 use parameters estimated from DUC05 and DUC06, respectively. The BEwT-E system was not significantly changed after TAC 2008 and any differences between the results presented below and the results presented in 2008's paper (Tratz and Hovy, 2008) is due to bug fixes, minor updates, etc.

DUC2007	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.940	0.880	0.480	0.949	0.884	0.567
BEwT-E (T off)	0.938	0.875	0.480	0.946	0.880	0.560
Original BE	0.942	0.885	0.425	0.906	0.861	0.551
AutoSummENG05	0.925	0.842	0.659	0.966	0.871	0.673
AutoSummENG06	0.935	0.864	0.615	0.964	0.880	0.649
ROUGE2	0.929	0.869	0.031	0.911	0.878	0.412
ROUGESU4	0.908	0.827	-0.14	0.877	0.831	0.259

Table 1. Correlation versus average content for DUC 2007 by peer type.

DUC2006	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.934	0.872	0.475	0.947	0.878	0.497
BEwT-E (T off)	0.925	0.852	0.475	0.948	0.884	0.520
Original BE	0.898	0.797	0.432	0.884	0.782	0.571
AutoSummENG05	0.937	0.871	0.759	0.967	0.891	0.715
AutoSummENG06	0.935	0.870	0.648	0.966	0.904	0.684
ROUGE2	0.885	0.767	0.469	0.897	0.836	0.642
ROUGESU4	0.898	0.790	0.741	0.877	0.850	0.695

Table 2. Correlation versus average content for DUC 2006 by peer type.

DUC2005	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.941	0.876	0.709	0.982	0.892	0.562
BEwT-E (T off)	0.943	0.880	0.709	0.981	0.890	0.552
Original BE	0.926	0.840	0.758	0.976	0.882	0.656
AutoSummENG05	0.929	0.840	0.936	0.977	0.885	0.878
AutoSummENG06	0.957	0.906	0.857	0.985	0.908	0.830
ROUGE2	0.951	0.906	0.430	0.972	0.930	0.444
ROUGESU4	0.942	0.876	0.721	0.958	0.919	0.488

Table 3. Correlation versus responsiveness for DUC 2005 by peer type.

TAC2008-Base	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.879	0.823	0.539	0.887	0.857	0.561
BEwT-E (T off)	0.894	0.844	0.659	0.881	0.869	0.513
Original BE	0.873	0.814	0.467	0.887	0.817	0.595
ROUGE2	0.903	0.867	0.539	0.851	0.829	0.645
ROUGESU4	0.884	0.833	0.874	0.852	0.802	0.846
Modified Pyramid	0.917	0.878	0.611	0.968	0.899	0.509

Table 4. Correlation versus overall responsive scores on the TAC 2008 base summaries by peer type.

TAC2008-Base	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.954	0.932	0.857	0.917	0.955	0.684
BEwT-E (T off)	0.954	0.933	0.905	0.912	0.955	0.691
Original BE	0.934	0.903	0.762	0.917	0.913	0.663
ROUGE2	0.936	0.909	0.857	0.869	0.907	0.544
ROUGESU4	0.921	0.885	0.857	0.871	0.886	0.543
Responsiveness	0.917	0.878	0.611	0.968	0.899	0.509

Table 5. Correlation of BEwT-E and modified Pyramid scores on the TAC 2008 base summaries by peer type.

TAC2008-Update	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.932	0.900	0.743	0.877	0.928	0.521
BEwT-E (T off)	0.931	0.898	0.755	0.886	0.932	0.718
Original BE	0.917	0.878	0.683	0.905	0.912	0.464
ROUGE2	0.922	0.886	0.587	0.882	0.909	0.579
ROUGESU4	0.929	0.896	0.898	0.835	0.901	0.796
Modified Pyramid	0.948	0.925	0.695	0.980	0.949	0.741

Table 6. Correlation versus overall responsiveness scores on the TAC 2008 update summaries by peer type.

TAC2008-Update	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.974	0.963	0.476	0.901	0.957	0.439
BEwT-E (T off)	0.973	0.962	0.381	0.907	0.958	0.424
Original BE	0.956	0.938	0.190	0.915	0.943	0.054
ROUGE2	0.960	0.944	-0.02	0.896	0.942	-0.01
ROUGESU4	0.954	0.934	0.357	0.859	0.925	0.333
Responsiveness	0.948	0.925	0.695	0.980	0.949	0.741

Table 7. Correlation versus modified Pyramid scores on the TAC 2008 update summaries by peer type.

4.2 Performance on TAC 2009

The rules for TAC 2009 were more or less the same as for TAC 2008. Unlike TAC 2008, however, TAC 2009 had an associated evaluation task named AESOP for evaluating evaluation software such as BEwT-E. A total of 44 topics were used for the evaluation.

Tables 4-6 present correlation results indicating how well BEwT-E correlated with overall responsiveness and modified Pyramid scores.

TAC2009-Base	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.890	0.843	0.286	0.493	0.663	0.302
BEwT-E (T off)	0.893	0.847	0.238	0.455	0.635	0.336
Original BE	0.851	0.842	0.190	0.458	0.692	0.214
ROUGE2	0.890	0.843	0.095	0.589	0.758	0.302
ROUGESU4	0.866	0.804	0.095	0.619	0.767	0.295
Responsiveness	0.910	0.866	0.690	0.972	0.902	0.688

Table 8. Correlation versus overall responsiveness Pyramid scores for TAC09 base summaries by peer type.

TAC2009-Base	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.951	0.931	0.357	0.618	0.830	0.412
BEwT-E (T off)	0.955	0.939	0.238	0.581	0.804	0.337
Original BE	0.197	0.932	0.333	0.588	0.856	0.241
ROUGE2	0.961	0.949	0.143	0.707	0.911	0.257
ROUGESU4	0.945	0.923	0.143	0.735	0.920	0.298
Responsiveness	0.910	0.866	0.690	0.972	0.902	0.688

Table 9. Correlation versus modified Pyramid scores for TAC09 base summaries by peer type.

TAC2009-Update	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.876	0.822	0.383	0.474	0.667	0.462
BEwT-E (T off)	0.878	0.827	0.311	0.428	0.640	0.478
Original BE	0.851	0.816	0.323	0.448	0.695	0.391
ROUGE2	0.824	0.755	0.29	0.535	0.718	0.43
ROUGESU4	0.795	0.718	0.299	0.564	0.729	0.355
Responsiveness	0.886	0.833	0.359	0.953	0.861	0.486

Table 10. Correlation versus overall responsiveness scores for TAC09 update summaries by peer type.

TAC2009-Update	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E (T on)	0.964	0.951	-0.02	0.656	0.907	0.179
BEwT-E (T off)	0.961	0.950	-0.17	0.614	0.886	0.130
Original BE	0.934	0.934	-0.02	0.630	0.924	0.092
ROUGE2	0.924	0.899	-0.02	0.707	0.939	0.274
ROUGESU4	0.897	0.865	0.048	0.727	0.939	0.195
Responsiveness	0.886	0.833	0.359	0.953	0.861	0.486

Table 11. Correlation versus modified Pyramid scores for TAC09 base summaries by peer type.

5 Discussion

The results show that BEwT-E outperforms ROUGE and the original BE system on most of the recent DUC and TAC datasets. We are particularly heartened by the fact that BEwT-E had the highest overall performance on the Spearman metric for the TAC 2009 update documents and was one of the best performing entries for the base documents. BEwT-E's performance on the Pearson metric was relatively low, however, and this appears to be at least partly caused by a single outlier system entry that, presumably, was entered by NIST and which contains the same set of sentences as one of the gold standard summaries. BEwT-E gives this entry a very high score, which throws off the Pearson correlation but does not affect the Spearman rank correlation metric.

As with previous datasets, the use of BEwT-E's transformations only had a minimal, and not always positive, effect on the TAC 2009 results. While the transformations do not appear to have much affect on the aggregate score, they have been shown to help for a significant proportion of the individual topics (Tratz and Hovy, 2008).

6 Conclusions and Future Work

BEwT-E continues to be one of the best systems for automatic text summary evaluation. We are especially heartened by its top performance on the TAC 2009 update summaries using the Spearman metric. However, we continue to be perplexed at the limited effect of the transformations.

In the future, better agreement with human scores can be achieved in two principal ways. One way is to implement a system that automatically learns optimal values for the various parameters that determine BE weights, BE match score combination coefficients, etc., discussed in Section 3. Parameters can be created for the BE extraction rules to determine which extraction rules produce the most predictive BEs as well as enable us to examine whether different domains or genre require different rule weights.

The second way is to improve the various components of the BE system. Examples include additional transformations, integrating a top-of-the-line NER system, and anaphora resolution capability. Other parsers, including dependency parsers, may produce significantly different results and are worth investigating.

BEwT-E is available to the public for download via http://www.isi.edu/natural-language/#research.

7 Acknowledgments

Stephen Tratz is supported by a NDSEG fellowship. We would also like to thank NIST for allowing us to use the DUC and TAC corpora.

References

Baldwin, B. and B. Carpenter. LingPipe. http://www.alias-i.com/lingpipe/.

Charniak, E. and M. Johnson. 2005. Coarse-to-find n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173-180, Ann Arbor, MI.

Conroy, J.M. and H.Trang Dang. 2008. Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality. *Proceedings of the COLING conference.* Manchester, UK.

DUC conferences. http://duc.nist.org.

Enquist, M. 1982. A Successive Shortest Path Algorithm for the Assignment Problem, *IFOR*, 20(4): 370–384.

Giannakopoulos, G., V. Karkaletsis, G. Vouros, P. Stamatopoulos. 2008. Summarization System Evaluation Revisited: N-gram Graphs. *ACM Transactions on Speech and Language Processing* (to appear).

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

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

Levy, R. and G. Andrew. 2006. Tregex and Tsurgeon: Tools for Querying and Manipulating Tree Data Structures. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06).* Genoa, Italy

Lin, C.Y. and E.H. Hovy. 2003. Automatic Evaluation of Summaries using n-Gram Co-occurrence Statistics. *Proceedings of the HLT-2003 conference.*

Lin, J. and D. Demner-Fushman. 2005. Evaluating Summaries and Answers: Two Sides of the Same Coin? *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Arbor, MI. 41–48.

Litkowski, K.C. and O. Hargraves. 2005. The Preposition Project. *Proceedings of ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications.* University of Essex-Colchester, UK. 171–179.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexiocography,* 2(4): 235–245.

Nenkova, A. and R. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. *Proceedings of the HLT-NAACL 2004 conference*.

Owczarzak, K. DEPEVAL(summ): 2009. Dependency-based Evaluation for Automatic Summaries. *Proceedings of the ACL 2009 conference.* Suntec, Singapore.

Teufel, S. and H. van Halteren. 2004. Evaluating Information Content by Factoid Analysis: Human Annotation and Stability. *Proceedings of the EMNLP 2004 conference.* Barcelona, Spain.

Tratz, S. and E. Hovy. 2008. Summarization Evaluation Using Transformed Basic Elements. *Proceedings of the 1st Text Analysis Conference (TAC)*. Gaithersburg, Maryland, USA.

Zhou, L, C.Y. Lin, D.S. Munteanu, and E.H. Hovy. 2006. ParaEval: Using Paraphrases to Evaluate Summaries Automatically. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*.