Summarization Evaluation Using Transformed Basic Elements

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 new, more sophisticated implementation of the BE framework that uses transformations to match BEs (minimal-length syntactically well-formed units) that are lexically different yet semantically similar. We demonstrate the effectiveness of BEwT-E using DUC and TAC datasets.

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 in 2006 (Hovy et al., 2006). This method facilitates matching of expressive variants of syntactically well-formed units called Basic Elements (BEs). The system achieved achieve fairly good correlation with human evaluation. 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.

During the extraction process, it is possible that several identical BEs may be extracted from the same document. Since duplicate BEs do not, by themselves, convey much information about the content of a summary, we experimented both with and without including duplicates.

```
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.

We experimented with three different weighting methods that use the number of reference summaries in which a BE occurs in order to determine its weight. The three weighting schemes are *binary* (each matched reference BE is worth 1 regardless of the number of summaries containing it), *root* (the BE weight is equal to the square root of the number of references containing it), and *total* (the BE score accrues 1 point for each reference summary 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. 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".

Hypernym/Hyponym: This transformation uses WordNet hypernyms and hyponyms to generalize/specialize nouns and verbs so that BEs like "newspaper|NN" and "press|NN" can match. This transformation treats person, organization, and location entities identified by the NER system as "person|NNP", "organization|NNP", and "location|NNP", respectively. For now, this transformation is limited to just the immediate parent and child sense nodes in the WordNet hierarchy. As with the other WordNet-based transformations, BEwT-E assumes each word is an instance of its most frequent sense.

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 and generalization 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.

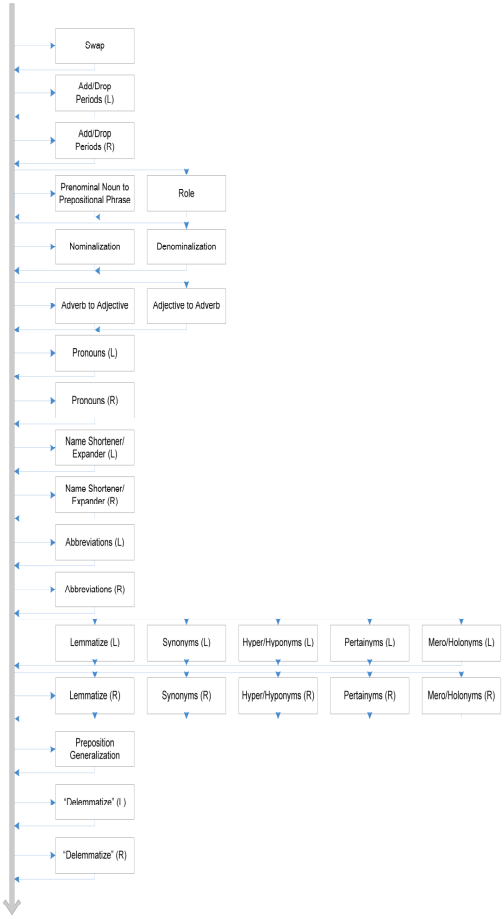

Figure 2. Diagram of pathways through the BE transformation pipeline. 'L' and 'R' indicate whether the transformation is limited to the leftmost or rightmost word in the BE.

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.

$$maximize \sum_{i=0}^{N} \sum_{j=0}^{M} C(i,j) W(j) x_{ij}$$

$$subject\ to$$

$$\sum_{i=0}^{N} x_{ij} \in \{0,1\}\ for all j where\ 0 \le j \le M$$

$$\sum_{j=0}^{M} x_{ij} \in \{0,1\}\ for all i where\ 0 \le i \le N$$

$$x_{ij} \in \{0,1\}$$

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 Evaluation

4.1 Performance on DUC05-07

To evaluate BEwT-E, we conducted a number of experiments using Document Understanding Conference (DUC) datasets from 2005–2007. For the 45–50 topics in each of these DUC evaluations, automated systems generated summaries of at most 250 words. 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 DUCs. We compare our system's performance on these datasets with other systems such as the original BE system, ROUGE, and AutoSummENG. We use the Pearson correlation coefficient to measure agreement with the scores and the Spearman coefficient to measure correlation with the rankings.

In Tables 1 to 3, we present the results for the DUC 2005–2007 datasets. The *d* and *nd* suffixes indicate whether or not duplicate BEs were included. The *bin*, *root*, and *tot* suffixes indicate the BE weighting scheme (*binary*, *root*, or *total*) used. The *off*, *on*, and *aeh* suffixes indicate whether the transformations are off, on, or if all transformations except the hypernym/hyponym transformation are on. The reason for leaving out the hypernym/hyponym transformation is given in Section 4.3. Bold scores are statistically significant at .01 while scores in italics are significant at .05. Scores with grayed cells are significantly different from those for the original BE scorer at .05 or better. AutoSummENG05 and AutoSummENG06 use parameters estimated from DUC05 and DUC06, respectively.

DUC2007	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E.nd.bin.aeh	**0.937**	**0.873**	0.480	**0.944**	**0.881**	0.567
BEwT-E.nd.rt.aeh	**0.937**	**0.874**	0.480	**0.963**	**0.873**	*0.604*
BEwT-E.nd.tot.aeh	**0.932**	**0.858**	0.560	**0.968**	**0.864**	*0.640*
BEwT-E.d.bin.aeh	**0.932**	**0.866**	0.351	**0.915**	**0.875**	0.499
BEwT-E.d.rt.aeh	**0.923**	**0.846**	0.351	**0.941**	**0.865**	0.504
BEwT-E.d.tot.aeh	**0.923**	**0.843**	0.412	**0.951**	**0.856**	0.503
BEwT-E.nd.bin.on	**0.931**	**0.860**	0.505	**0.941**	**0.867**	*0.581*
BEwT-E.nd.rt.on	**0.931**	**0.856**	0.560	**0.960**	**0.862**	*0.611*
BEwT-E.nd.tot.on	**0.923**	**0.839**	0.560	**0.967**	**0.856**	*0.641*
BEwT-E.d.bin.on	**0.921**	**0.841**	0.375	**0.915**	**0.867**	0.499
BEwT-E.d.rt.on	**0.919**	**0.835**	0.369	**0.939**	**0.861**	0.484
BEwT-E.d.tot.on	**0.921**	**0.841**	0.345	**0.948**	**0.855**	0.463
BEwT-E.nd.bin.off	**0.937**	**0.872**	0.480	**0.942**	**0.882**	*0.560*
BEwT-E.nd.rt.off	**0.936**	**0.867**	0.560	**0.964**	**0.873**	*0.605*
BEwT-E.nd.tot.off	**0.932**	**0.860**	0.560	**0.969**	**0.862**	*0.651*
BEwT-E.d.bin.off	**0.929**	**0.860**	0.332	**0.917**	**0.867**	0.490
BEwT-E.d.rt.off	**0.928**	**0.855**	0.394	**0.946**	**0.854**	0.499
BEwT-E.d.tot.off	**0.921**	**0.841**	0.369	**0.954**	**0.842**	0.499
Original BE	**0.942**	**0.885**	0.424	**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**	-.135	**0.877**	**0.831**	0.259

Table 1. System-level of correlation of BEwT-E and average content for DUC 2007 by peer type.

DUC2006	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E.nd.bin.aeh	**0.932**	**0.868**	0.475	**0.943**	**0.877**	0.497
BEwT-E.nd.rt.aeh	**0.931**	**0.865**	0.475	**0.965**	**0.878**	0.498
BEwT-E.nd.tot.aeh	**0.928**	**0.860**	0.475	**0.975**	**0.876**	0.498
BEwT-E.d.bin.aeh	**0.904**	**0.808**	0.445	**0.901**	**0.829**	0.460
BEwT-E.d.rt.aeh	**0.893**	**0.789**	0.340	**0.922**	**0.817**	0.440
BEwT-E.d.tot.aeh	**0.890**	**0.781**	0.309	**0.930**	**0.806**	0.402
BEwT-E.nd.bin.on	**0.930**	**0.862**	0.475	**0.936**	**0.875**	0.483
BEwT-E.nd.rt.on	**0.924**	**0.850**	0.475	**0.961**	**0.877**	0.498
BEwT-E.nd.tot.on	**0.929**	**0.862**	0.475	**0.974**	**0.877**	0.516
BEwT-E.d.bin.on	**0.904**	**0.807**	0.463	**0.893**	**0.833**	0.450
BEwT-E.d.rt.on	**0.899**	**0.797**	0.426	**0.917**	**0.821**	0.435
BEwT-E.d.tot.on	**0.895**	**0.793**	0.309	**0.926**	**0.812**	0.400
BEwT-E.nd.bin.off	**0.929**	**0.860**	0.475	**0.944**	**0.884**	0.520
BEwT-E.nd.rt.off	**0.929**	**0.861**	0.475	**0.968**	**0.883**	0.526
BEwT-E.nd.tot.off	**0.926**	**0.855**	0.475	**0.977**	**0.879**	0.533
BEwT-E.d.bin.off	**0.905**	**0.809**	0.482	**0.899**	**0.833**	0.506
BEwT-E.d.rt.off	**0.893**	**0.784**	0.482	**0.928**	**0.815**	0.507
BEwT-E.d.tot.off	**0.885**	**0.770**	0.327	**0.938**	**0.800**	0.492
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. System-level of correlation of BEwT-E and average content for DUC 2006 by peer type.

DUC2005	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E.nd.bin.aeh	**0.941**	**0.875**	*0.709*	**0.982**	**0.890**	*0.554*
BEwT-E.nd.rt.aeh	**0.943**	**0.878**	**0.758**	**0.985**	**0.893**	*0.629*
BEwT-E.nd.tot.aeh	**0.945**	**0.882**	**0.782**	**0.986**	**0.895**	*0.692*

BEwT-E.d.bin.aeh	**0.936**	**0.865**	*0.673*	**0.977**	**0.899**	0.508
BEwT-E.d.rt.aeh	**0.929**	**0.853**	*0.564*	**0.982**	**0.899**	*0.551*
BEwT-E.d.tot.aeh	**0.924**	**0.839**	*0.588*	**0.983**	**0.897**	*0.578*
BEwT-E.nd.bin.on	**0.940**	**0.876**	*0.648*	**0.977**	**0.880**	0.528
BEwT-E.nd.rt.on	**0.941**	**0.875**	*0.697*	**0.983**	**0.885**	*0.609*
BEwT-E.nd.tot.on	**0.943**	**0.879**	**0.758**	**0.985**	**0.889**	*0.684*
BEwT-E.d.bin.on	**0.928**	**0.851**	0.527	**0.972**	**0.894**	0.492
BEwT-E.d.rt.on	**0.928**	**0.850**	*0.564*	**0.979**	**0.897**	0.544
BEwT-E.d.tot.on	**0.921**	**0.837**	0.479	**0.981**	**0.896**	*0.566*
BEwT-E.nd.bin.off	**0.944**	**0.882**	*0.709*	**0.981**	**0.892**	*0.550*
BEwT-E.nd.rt.off	**0.940**	**0.872**	**0.758**	**0.984**	**0.894**	*0.614*
BEwT-E.nd.tot.off	**0.936**	**0.864**	**0.758**	**0.984**	**0.895**	*0.678*
BEwT-E.d.bin.off	**0.936**	**0.867**	*0.648*	**0.977**	**0.894**	0.506
BEwT-E.d.rt.off	**0.931**	**0.859**	0.539	**0.982**	**0.889**	0.547
BEwT-E.d.tot.off	**0.911**	**0.814**	0.455	**0.982**	**0.883**	*0.561*
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. System-level of correlation of BEwT-E and responsiveness for DUC 2005 by peer type.

4.2 Performance on TAC 2008

We next applied BEwT-E to the TAC 2008 Update Summarization task system-level results. For each of 48 topics, the participating systems produced 2 summaries of at most 100 words each. The first summary was created from a base set of documents representing the topic. The 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. Human judges assigned scores to summaries by comparing them against human-written summaries. Tables 4-6 present correlation results indicating how well BEwT-E correlated with overall responsiveness and modified Pyramid scores.

TAC2008-Base	Spearman			Pearson		
Peers included	All	Auto	Hu	All	Auto	Hu
BEwT-E.nd.bin.aeh	**0.878**	**0.821**	0.539	**0.875**	**0.856**	0.561
BEwT-E.nd.rt.aeh	**0.864**	**0.802**	0.539	**0.925**	**0.840**	0.549
BEwT-E.nd.tot.aeh	**0.859**	**0.793**	0.491	**0.943**	**0.825**	0.537
BEwT-E.d.bin.aeh	**0.893**	**0.846**	0.539	**0.843**	**0.843**	0.544
BEwT-E.d.rt.aeh	**0.889**	**0.838**	0.599	**0.890**	**0.824**	0.511
BEwT-E.d.tot.aeh	**0.887**	**0.835**	0.455	**0.909**	**0.807**	0.474
BEwT-E.nd.bin.on	**0.878**	**0.823**	0.431	**0.869**	**0.854**	0.541
BEwT-E.nd.rt.on	**0.869**	**0.809**	0.491	**0.919**	**0.839**	0.530
BEwT-E.nd.tot.on	**0.859**	**0.794**	0.563	**0.939**	**0.823**	0.526
BEwT-E.d.bin.on	**0.897**	**0.852**	0.515	**0.832**	**0.840**	0.522
BEwT-E.d.rt.on	**0.894**	**0.845**	0.515	**0.880**	**0.821**	0.500
BEwT-E.d.tot.on	**0.886**	**0.834**	0.575	**0.900**	**0.805**	0.468
BEwT-E.nd.bin.off	**0.894**	**0.845**	*0.659*	**0.869**	**0.867**	0.513
BEwT-E.nd.rt.off	**0.882**	**0.828**	*0.659*	**0.930**	**0.853**	0.501
BEwT-E.nd.tot.off	**0.871**	**0.811**	*0.731*	**0.950**	**0.838**	0.489
BEwT-E.d.bin.off	**0.891**	**0.849**	0.599	**0.839**	**0.854**	0.469
BEwT-E.d.rt.off	**0.895**	**0.847**	0.575	**0.901**	**0.833**	0.442
BEwT-E.d.tot.off	**0.889**	**0.839**	0.575	**0.923**	**0.812**	0.414
Original BE	**0.873**	**0.815**	0.467	**0.887**	**0.817**	0.595
ROUGE2	**0.905**	**0.867**	0.539	**0.851**	**0.829**	*0.645*
ROUGESU4	**0.884**	**0.832**	**0.874**	**0.852**	**0.802**	**0.846**
Modified Pyramid	**0.917**	**0.878**	0.611	**0.968**	**0.900**	0.509

Table 4. Correlation of BEwT-E and 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.nd.bin.aeh	**0.955**	**0.935**	**0.857**	**0.907**	**0.955**	*0.684*
BEwT-E.nd.rt.aeh	**0.955**	**0.935**	**0.833**	**0.950**	**0.950**	*0.665*
BEwT-E.nd.tot.aeh	**0.953**	**0.931**	0.810	**0.964**	**0.943**	*0.640*
BEwT-E.d.bin.aeh	**0.961**	**0.944**	**0.952**	**0.879**	**0.945**	*0.746*
BEwT-E.d.rt.aeh	**0.957**	**0.937**	**0.905**	**0.921**	**0.935**	*0.722*
BEwT-E.d.tot.aeh	**0.953**	**0.932**	**0.833**	**0.936**	**0.923**	*0.680*
BEwT-E.nd.bin.on	**0.959**	**0.941**	**0.881**	**0.906**	**0.954**	*0.774*
BEwT-E.nd.rt.on	**0.958**	**0.939**	**0.833**	**0.949**	**0.950**	*0.752*
BEwT-E.nd.tot.on	**0.954**	**0.932**	**0.833**	**0.964**	**0.944**	*0.724*
BEwT-E.d.bin.on	**0.957**	**0.938**	**0.952**	**0.875**	**0.943**	**0.825**
BEwT-E.d.rt.on	**0.960**	**0.941**	**0.929**	**0.917**	**0.933**	**0.792**
BEwT-E.d.tot.on	**0.953**	**0.932**	**0.881**	**0.932**	**0.922**	*0.745*
BEwT-E.nd.bin.off	**0.954**	**0.933**	**0.905**	**0.901**	**0.954**	*0.691*
BEwT-E.nd.rt.off	**0.958**	**0.938**	**0.881**	**0.955**	**0.952**	*0.675*
BEwT-E.nd.tot.off	**0.958**	**0.939**	**0.905**	**0.970**	**0.947**	*0.660*
BEwT-E.d.bin.off	**0.952**	**0.933**	**0.905**	**0.878**	**0.945**	*0.735*
BEwT-E.d.rt.off	**0.958**	**0.939**	**0.881**	**0.933**	**0.935**	*0.715*
BEwT-E.d.tot.off	**0.952**	**0.930**	**0.881**	**0.949**	**0.922**	*0.691*
Original BE	**0.934**	**0.904**	*0.762*	**0.917**	**0.913**	*0.663*
ROUGE2	**0.936**	**0.907**	**0.857**	**0.869**	**0.907**	0.544
ROUGESU4	**0.919**	**0.883**	**0.857**	**0.871**	**0.886**	0.543
Responsiveness	**0.917**	**0.878**	0.611	**0.968**	**0.900**	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.nd.bin.aeh	**0.928**	**0.894**	*0.743*	**0.860**	**0.926**	0.521
BEwT-E.nd.rt.aeh	**0.926**	**0.891**	*0.802*	**0.925**	**0.924**	*0.642*
BEwT-E.nd.tot.aeh	**0.926**	**0.890**	*0.826*	**0.953**	**0.920**	*0.707*
BEwT-E.d.bin.aeh	**0.924**	**0.888**	*0.719*	**0.837**	**0.910**	0.621
BEwT-E.d.rt.aeh	**0.926**	**0.891**	*0.778*	**0.897**	**0.901**	*0.710*
BEwT-E.d.tot.aeh	**0.920**	**0.882**	**0.850**	**0.924**	**0.890**	*0.763*
BEwT-E.nd.bin.on	**0.926**	**0.890**	*0.802*	**0.865**	**0.921**	*0.665*
BEwT-E.nd.rt.on	**0.927**	**0.892**	*0.826*	**0.926**	**0.921**	*0.729*
BEwT-E.nd.tot.on	**0.925**	**0.888**	**0.850**	**0.954**	**0.917**	*0.762*
BEwT-E.d.bin.on	**0.927**	**0.892**	**0.886**	**0.843**	**0.904**	*0.725*
BEwT-E.d.rt.on	**0.922**	**0.883**	**0.898**	**0.898**	**0.896**	*0.793*
BEwT-E.d.tot.on	**0.914**	**0.872**	**0.898**	**0.923**	**0.885**	**0.824**
BEwT-E.nd.bin.off	**0.929**	**0.894**	*0.755*	**0.872**	**0.929**	*0.718*
BEwT-E.nd.rt.off	**0.927**	**0.893**	*0.778*	**0.942**	**0.928**	*0.749*
BEwT-E.nd.tot.off	**0.925**	**0.889**	*0.778*	**0.967**	**0.924**	*0.754*
BEwT-E.d.bin.off	**0.923**	**0.885**	*0.826*	**0.856**	**0.914**	*0.776*
BEwT-E.d.rt.off	**0.922**	**0.884**	**0.898**	**0.921**	**0.904**	*0.787*
BEwT-E.d.tot.off	**0.913**	**0.870**	**0.898**	**0.946**	**0.889**	**0.790**
Original BE	**0.917**	**0.877**	*0.683*	**0.905**	**0.912**	0.464
ROUGE2	**0.920**	**0.882**	0.587	**0.882**	**0.909**	0.579
ROUGESU4	**0.927**	**0.893**	**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 of BEwT-E and 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.nd.bin.aeh	**0.971**	**0.958**	0.476	**0.887**	**0.956**	0.439
BEwT-E.nd.rt.aeh	**0.969**	**0.955**	0.595	**0.941**	**0.954**	0.474
BEwT-E.nd.tot.aeh	**0.967**	**0.952**	0.571	**0.962**	**0.951**	0.500
BEwT-E.d.bin.aeh	**0.970**	**0.956**	0.500	**0.868**	**0.943**	0.543
BEwT-E.d.rt.aeh	**0.970**	**0.956**	0.619	**0.919**	**0.935**	0.582
BEwT-E.d.tot.aeh	**0.966**	**0.950**	*0.667*	**0.940**	**0.924**	0.616
BEwT-E.nd.bin.on	**0.967**	**0.952**	0.595	**0.893**	**0.951**	*0.642*
BEwT-E.nd.rt.on	**0.965**	**0.949**	*0.690*	**0.942**	**0.949**	0.614
BEwT-E.nd.tot.on	**0.966**	**0.950**	*0.667*	**0.962**	**0.946**	0.590
BEwT-E.d.bin.on	**0.968**	**0.953**	0.571	**0.874**	**0.936**	*0.652*
BEwT-E.d.rt.on	**0.964**	**0.947**	*0.738*	**0.919**	**0.929**	*0.665*
BEwT-E.d.tot.on	**0.958**	**0.938**	*0.738*	**0.939**	**0.918**	*0.670*
BEwT-E.nd.bin.off	**0.972**	**0.960**	0.381	**0.895**	**0.956**	0.424
BEwT-E.nd.rt.off	**0.969**	**0.955**	0.571	**0.952**	**0.955**	0.450
BEwT-E.nd.tot.off	**0.969**	**0.955**	0.571	**0.970**	**0.951**	0.466
BEwT-E.d.bin.off	**0.970**	**0.956**	*0.690*	**0.884**	**0.944**	*0.650*
BEwT-E.d.rt.off	**0.965**	**0.949**	*0.738*	**0.938**	**0.934**	*0.646*
BEwT-E.d.tot.off	**0.957**	**0.936**	*0.738*	**0.956**	**0.920**	*0.650*
Original BE	**0.957**	**0.938**	0.190	**0.915**	**0.943**	0.054
ROUGE2	**0.959**	**0.942**	-.024	**0.896**	**0.942**	-.014
ROUGESU4	**0.952**	**0.931**	0.357	**0.859**	**0.925**	0.333
Responsiveness	**0.948**	**0.925**	*0.695*	**0.980**	**0.949**	*0.741*

Table 7. Correlation of BEwT-E and modified Pyramid scores on the TAC 2008 update summaries by peer type.

4.3 Effect of Transformations

In addition to studying the overall effect of the transformations, we examined the impact at the topic level. The base and update halves of the TAC data are treated separately, giving us a total of 241 topics including 96 from TAC and 145 from DUC. Tables 8 and 9 display these results.

One Transform On	All		Auto		Human	
	+	-	+	-	+	-
Swap IS-A Nouns	50	32	46	25	10	8
Add/Drop Periods	65	57	66	51	18	16
Prenoun Noun ↔ PP	88	95	91	86	22	15
Role	7	8	8	5	1	3
Nominalization	48	46	41	42	12	11
Denominalization	48	39	46	35	10	7
Adverb to Adjective	7	4	3	4	3	1
Adjective to Adverb	1	0	1	0	0	0
Pronouns	90	82	84	84	28	16
Names	94	98	105	87	41	30
Abbreviations	33	48	35	44	10	4
De/lemmatize	**139**	102	**138**	103	77	80
Synonyms	**134**	107	127	114	57	67
Hyper/Hyponyms	103	**138**	96	**145**	81	74
Pertainyms	119	111	115	113	42	44
Mero/Holonyms	74	72	68	78	23	15
Prepositions	50	57	42	52	9	16

Table 8. Total number of topics across DUC05–07 and TAC08 whose responsiveness scores correlated better (+) or worse (-) after enabling exactly one transformation. Numbers in bold are statistically significant at .05.

One Transform Off	All		Auto		Human	
	+	-	+	-	+	-
Swap IS-A Nouns	65	72	57	66	14	10
Add/Drop Periods	49	72	44	70	19	13
Prenoun Noun ↔ PP	114	122	112	123	44	47
Role	80	64	76	63	14	11
Nominalization	96	95	80	95	16	38
Denominalization	115	116	112	112	54	42
Adverb to Adjective	7	11	6	6	1	3

Adjective to Adverb	11	11	8	7	3	3
Pronouns	110	109	116	101	34	42
Names	103	95	89	107	30	38
Abbreviations	51	38	47	41	8	9
De/lemmatize	121	120	114	127	77	79
Synonyms	113	128	115	126	70	68
Hyper/Hyponyms	**140**	101	**153**	88	71	86
Pertainyms	118	113	113	117	41	41
Mero/Holonyms	70	76	69	76	19	22
Prepositions	85	128	98	105	26	52

Table 9. Total number of topics across DUC05-07 and TAC08 whose responsiveness scores correlated better or worse after exactly one transformation is disabled. Numbers in bold are statistically significant at .01.

While the transformations do not provide statistically significant improvements in the overall correlation scores for the DUC05–DUC07 and TAC08 datasets after the individual topic scores have been averaged, the transformations do have a generally positive impact at the individual topic level. Tables 10 and 11 show the effect of enabling all transformations or all transformations except Hyper/Hyponyms on a per topic basis for the DUC05–07 and TAC08 datasets.

	All		Auto		Human	
	+	-	+	-	+	-
DUC07	29	16	24	21	15	13
DUC06	24	26	20	30	15	15
DUC05	29	21	33	17	24	13
TAC08 Base	22	26	22	26	16	20
TAC08 Update	19	29	17	31	11	15
Total	123	118	116	125	81	76

Table 10. Number of topics whose responsiveness score correlated better or worse to BEwT-E when all transformations are turned on.

	All		Auto		Human	
	+	-	+	-	+	-
DUC07	26	19	31	14	11	17
DUC06	30	20	29	21	14	16
DUC05	38	12	35	15	18	19
TAC08 Base	25	23	24	24	13	23
TAC08 Update	27	21	23	25	11	15
Total	**146**	95	**142**	99	67	90

Table 11. Number of topics whose responsiveness score correlated better or worse when all transformations are turned on except Hyper/Hyponyms. Numbers in bold are significant at .01.

5 Conclusions and Future Work

While we are pleased with the overall results of the BEwT-E, we are curious as to why the transformations did not help more and would like to examine what mistakes they made and how to improve their effectiveness. We are also surprised that the *root* tallying strategy tends to be more consistent than *total*, which is the tallying method that corresponds to the popularity score currently used in the Pyramid Method. As expected, duplicate BEs provide no help and generally have a somewhat negative impact on the correlation scores.

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 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, for example to include additional transformations, a top-of-the-line NER system, and an anaphora resolution capability. Other parsers, including dependency parsers, may produce significantly different results.

BEwT-E will be made available to the public in the near future through http://www.isi.edu/.

6 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 Sum-

mary 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 conference*.

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

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*.