Pruning Basic Elements for Better Automatic Evaluation of Summaries

Ukyo Honda[1] **Tsutomu Hirao**[2] **Masaaki Nagata**[2]
[1] Nara Institute of Science and Technology [2] NTT Communication Science Laboratories
[1] honda.ukyo.hn6@is.naist.jp
[2] {hirao.tsutomu, nagata.masaaki}@lab.ntt.co.jp

Abstract

We propose a simple but highly effective automatic evaluation measure of summarization, pruned Basic Elements (pBE). Although the BE concept is widely used for the automated evaluation of summaries, its weakness is that it redundantly matches basic elements. To avoid this redundancy, pBE prunes basic elements by (1) disregarding frequency count of basic elements and (2) reducing semantically overlapped basic elements based on word similarity. Even though it is simple, pBE outperforms ROUGE in DUC datasets in most cases and achieves the highest rank correlation coefficient in TAC 2011 AESOP task.

1 Introduction

Automatic evaluation measures have a significant impact on the research on summarization. Since there is no other practical way to quickly evaluate the quality of system summaries, summarization studies work on raising the scores that are given by automatic evaluation measures.

Among the automatic evaluation measures, the most popular ones are ROUGE (Lin, 2004) and BE (Hovy et al., 2006). ROUGE/BE counts the number of ngrams/basic elements[1] that match those in manual reference summaries. ROUGE normally employs unigrams or bigrams while BE uses dependency triples (head|modifier|relation) as their units. It is known that both ROUGE and BE are well correlated with human judgment.

Their evaluation approach, however, is quite different from humans' in two ways: they score low-information units higher and ignore the semantic overlap of units. The first problem is caused by scoring units according to their frequencies. We found that the units that occur multiple times in a summary are highly likely to be function-word bigrams (e.g., "of the") or basic elements that represent only single nouns (e.g., (house|the|det)); such units are less informative than units connected with verbs (e.g., "John went" and (went|John|nsubj)). The second problem is that ROUGE/BE sometimes gives scores twice or more to the units that are semantically overlapped but spelled differently. This is due to the fact that ROUGE/BE only considers the surface level of unit matching, which also yields inaccurate scoring of paraphrased units.

Our method is aimed at solving these problems by cutting back redundant units. We use BE, but with Universal Dependencies (UD) (Nivre et al., 2016), a more ideal form of annotation that is available for multiple languages, and introduce two steps to prune basic elements. The first step is to disregard the frequency count of basic elements, and the other one is to reduce semantically overlapped basic elements using word embeddings. We call this new measure pruned BE (pBE). Our experiments show that pBE outperforms ROUGE in most DUC datasets and achieves the highest rank correlation coefficient in TAC 2011 AESOP task.

2 Related Work

ROUGE-WE (Ng and Abrecht, 2015) and BEwT-E (Tratz and Hovy, 2008) are closely related to our method in that they aim to improve unit matching. ROUGE-WE exploits word embeddings to softly match ngrams based on their cosine similarities. Although this also takes semantic correspondence into consideration, it is different from pBE because it does not judge word similarity within one summary, but only between a target sum-

[1]We use "BE" to represent the evaluation method Basic Elements, "basic element(s)" to represent the fragments of Basic Elements and "unit" as a general term of ngrams and basic elements.

Proceedings of NAACL-HLT 2018, pages 661–666
New Orleans, Louisiana, June 1 - 6, 2018.

mary and its reference summaries. Furthermore, ROUGE-WE does not remove the frequency count of ngrams as pBE does. As a result, ROUGE-WE does not achieve our goal of reducing redundant units.

BEwT-E transforms basic elements to help in matching. However, it requires complex transformation rules, which are difficult to apply to languages other than English. pBE, on the other hand, needs no resources other than word embeddings and UD parsers, and so can be implemented in many other languages. BEwT-E was checked as to whether the frequency count of basic elements affected its performance. The focus, however, was not to prune basic elements and there was no clear explanation as to why disregarding frequency count was effective. Our contribution is that we have identified why disregarding frequency count is effective; it yields the pruning of low-information basic elements, and thus works well in combination with reducing semantic overlaps.

Syntactically and semantically richer structures are free from low-information units. In this sense, PEAK (Yang et al., 2016) is related to our method in that it tries to employ predicate-argument structures as primitive units for matching. However, the predicate-argument structures are more difficult to extract than dependency triples. It is reported that PEAK scored only about 0.7 in Pearson coefficient for the DUC 2006 dataset (Yang et al., 2016), whereas ROUGE achieved around 0.83.

3 pruned BE (pBE)[2]

In this section, we describe our implementation of BE and the two steps of pruning basic elements.

3.1 Our Implementation of BE

BE was proposed to compensate some of the shortcomings of ngrams (Hovy et al., 2006). ROUGE usually uses short ngrams such as unigrams and bigrams, but these can be low-information content because they are simply extracted without considering the syntactic relations of the words. For example, the sentence "John went to the store on foot" is decomposed into the bigrams ["John went", "went to", "to the", "the store", "store on", "on foot"]. The function-word pair "to the" bears almost no meaning but is frequently found since function words appear in sentences quite often. On the other hand, a dependency triple holds the syntactic information that the dependency of "to" is not "the" but "store". Although BE requires applying parsers to summaries, syntactic dependencies enable BE to avoid making low-information units[3].

Accordingly, while we use BE, the annotation is UD based, an approach not employed in previous studies. Since UD focuses on the relations between content words, UD triples are able to represent key components of sentences more directly. For example, the sentence above can be decomposed in UD as [(went|John|nsubj), (store|to|case), (store|the|det), (went|store|nmod:to), (foot|on|case), (went|foot|nmod:on)][4], while it is [(went|John|nsubj), (went|to|prep), (store|the|det), (to|store|pobj), (went|on|prep), (on|foot|pobj)] in Stanford Dependencies (de Marneffe et al., 2006). In UD, the predicate-object relation is directly expressed as (went|store|nmod:to), instead of having intermediate triples (went|to|prep) and (to|store|pobj). Moreover, UD has another key advantage, that it is available in many languages. This makes our method available for multiple languages other than English.

We use (head|modifier|relation) triples of UD v1 relations which correspond to narrow-sense dependencies and multiword expression (MWE) dependencies of UD v2[5]. One thing to note here is that we excluded auxpass and mwe relations. It is because the information of these is mostly contained in other relations such as nsubjpass, nmod and advcl. Auxpass is a special relation of aux, which indicates that a verb is passive. Aux indicates a verb's modality or tense, which is not mentioned by nsubj relation alone. Auxpass also indicates an important information of a verb, its voice. However, the information of voice is already contained in the relation of nsubjpass. Mwe is used for multiword expressions with function words that behave like a single function word.

[2]Code will be available at https://github.com/ukyh/prunedBE

[3]It can be pointed out that bigrams of function words can be avoided if we remove function words. However, this is just an ad hoc measure, which leads to another meaningless bigram "store foot".

[4]root relation triple is omitted because we do not include it in our basic elements. See the next footnote.

[5]That is, nsubj, nsubjpass, dobj, iobj, csubj, csubjpass, ccomp, xcomp, nmod, advcl, advmod, neg, vocative, discourse, expl, aux, cop, mark, nummod, appos, acl, amod, det, case, compound, name, foreign and dislocated.

The whole information of mwe, however, is generally contained in nmod or advcl relations in enhanced++ UD representation (Schuster and Manning, 2016) (e.g., (fruits|apple|nmod:such_as) and (bought|fixing|advcl:insted_of)). Counting these relations can lead to redundant unit matching. In fact, the performance was better when we excluded these relations.

3.2 Step 1: Disregard Frequency Count

ROUGE/BE score is defined as follows:

$$\text{ROUGE/BE}(\mathbf{R}, S) = \frac{\sum_{k=1}^{K}\sum_{m=1}^{M_k}\min\{N(f_m^k, \mathbf{R}_k), N(f_m^k, S)\}}{\sum_{k=1}^{K}\sum_{m=1}^{M_k}\{N(f_m^k, \mathbf{R}_k)\}}. \quad (1)$$

Given K reference summaries $\mathbf{R} = \{\mathbf{R}_1, ..., \mathbf{R}_K\}$, target summary S, and the set of units that appear in $\mathbf{R}_k$ as $F_k = \{f_1^k, ..., f_{M_k}^k\}$ ($|F_k| = M_k$), ROUGE/BE counts how many times each f occurs in target summary S. Let $N(f_m^k, \mathbf{R}_k)$ be the frequency of f_m^k in $\mathbf{R}_k$ and $N(f_m^k, S)$ be the frequency of f_m^k in S. Unit f contributes to ROUGE/BE scores according to its frequency[6].

The problem is that the units found multiple times tend to be low-information units. ROUGE-2 often finds function-word bigrams, which leads to their overweighting. While BE is free from function-word bigrams, it still contains improperly weighted basic elements: compound and det. For example, in DUC 2003, 302 basic elements are returned more than 1 in $\min\{N(f_m^k, \mathbf{R}_k), N(f_m^k, S)\}$ of which 139 were compound and 96 were det; together they occupy about 78% of the total. This is because these relations represent only single nouns. Since they are not associated with verbs, which are key components of sentences, they appear in many sentences even within one summary[7]. It is not that compound and det are meaningless units, but that they should not be weighted more than other relations such as nsubj, dobj and iobj, which are associated with verbs.

[6]In BE, it is optional to consider or disregard this frequency count (Tratz and Hovy, 2008). We describe why dispensing with the frequency count affects the results below.

[7]"Donald Trump" can be used in various sentences like "Donald Trump won the election." and "Donald Trump will visit China next week." But "Trump won" can only occur in the specific situation where Trump won something, which is unlikely to be described in a summary more than once.

Therefore, we simply get rid of the frequency count. We define our scoring function as follows:

$$\text{pBE}_{-\text{cnt}}(\mathbf{R}, S) = \frac{\sum_{k=1}^{K}\sum_{m=1}^{M_k}\{O(f_m^k, S)\}}{\sum_{k=1}^{K}\sum_{m=1}^{M_k}\{O(f_m^k, \mathbf{R}_k)\}}. \quad (2)$$

Here $O(f_m^k, \mathbf{R}_k)$ and $O(f_m^k, S)$ are functions that return 1 if f_m^k is in $\mathbf{R}_k$ and S respectively, and otherwise return 0. This way, we can simplify equation (1) and avoid undue weighting.

3.3 Step 2: Cluster Basic Elements Using Word Embeddings

We are able to detect semantic correspondence. If we are given key points to be included in the summary, we can judge whether the key points are in the summary or not on the semantic level. ROGUE/BE, however, judges the correspondence of key points only on the surface level. Since the same content can be expressed in various surface forms, ROUGE/BE sometimes scores semantically overlapped units multiple times or does not score units that semantically correspond to each other but are significantly different on the surface level[8].

To deal with this problem, we put semantically identical words into one cluster based on word similarity. Our method only requires word embeddings trained with word2vec (Mikolov et al., 2013), and so offers multilingual capability.

Given K reference summaries $\mathbf{R} = \{\mathbf{R}_1, ..., \mathbf{R}_K\}$, target summary S, a set of all unigrams in $\mathbf{R}$ and S as $U = \{u_1, ..., u_P\}$, and a set of Q word embeddings for the unigrams as $V = \{v_1, ..., v_Q\}$ ($Q \leq P$), we put U into the set of cluster IDs $C = \{c_1, ..., c_N\}$ by hierarchical clustering using word similarities. The number of clusters, N, is a hyperparameter. Next, we convert the unigrams of $\mathbf{R}$ and S into the cluster ID c. If unigram u_i has no word embeddings, we leave it in its surface form. Let the converted reference summaries and target summary be $\mathbf{R}'$ and S', respectively. We define the set of basic elements in $\mathbf{R}'_k$ as $F'_k = \{f_1'^k, ..., f_{M_k}'^k\}$ ($|F'_k| = M_k$).

[8]Suppose the phrases "John killed" and "John murdered" are in a target summary and each reference summaries. Here, the target summary gets double scores for the semantically same units. On the other hand, if "John killed" is only in the target summary and "John murdered" is only in the reference summaries, the target summary gets no score for the semantic correspondence.

	DUC03	DUC04	DUC05	DUC06	DUC06 pyr	DUC07	DUC07 pyr
ROUGE-2	.906/.821/.617	.909/.838/.691	**.932/.931/.792**	.836/.767/.584	**.905/.884/.740**	.880/.873/.715	.979/.989/.949
ROUGE-S4	.851/.791/.617	.876/.816/.647	.915/.889/.727	.829/.759/.574	.888/.880/.732	.850/.836/.646	.971/.956/.872
ROUGE-SU4	.782/.774/.600	.854/.772/.559	.925/.893/.731	.849/.790/.601	.885/.850/.706	.835/.832/.650	.961/.973/.897
BE	.928/.862/.700	.936/.868/.721	.897/.863/.706	.831/.757/.587	.881/.848/.688	.890/.890/.732	.982/.978/.923
pBE$_{-\text{cnt}}$	.930/**.871/.717**	.938/.873/.735	.904/.882/.723	.854/.793/.628	.894/.848/.714	**.902/.906/.760**	**.985**/.978/.923
pBE$_{+\text{cls}}$	.929/**.871/.717**	.940/.877/.735	.897/.862/.702	.834/.768/.601	.886/.849/.697	.890/.894/.736	.980/.967/.897
pBE$_{-\text{cnt}+\text{cls}}$	**.932/.871/.717**	**.943/.885/.765**	.905/.877/.718	**.859/.801/.631**	.898/.849/.714	**.902/.906**/.756	**.985/.995/.974**

Table 1: Correlation coefficients of pBE and ROUGE. The coefficients are written in the order of "Pearson/Spearman/Kendall".

	Pearson	Spearman	Kendall
ROUGE-SU4	**.981**	.894	.737
C_S_IIITH3	.965	.903	.758
ROUGE-WE-1	.949	.914	.753
pBE$_{-\text{cnt}+\text{cls}}$	.947	**.915**	**.774**

Table 2: Correlation coefficients of pBE and other participants with manual pyramid scores in TAC 2011. ROUGE-SU4/ROUGE-WE-1/C_S_IIITH3 (Kumar et al., 2011) achieved the highest correlation coefficient in Pearson/Spearman/Kendall correlation among the past results.

	Evaluation	Limit	Topic	Ref	System
DUC 2003	coverage	100	30	4	16
DUC 2004	coverage	100	50	4	17
DUC 2005	responsiveness	250	50	4 or 9	32
DUC 2006	responsiveness	250	50	4	35
	pyramid		20		22
DUC 2007	responsiveness	250	45	4	32
	pyramid		23		13
TAC 2011	pyramid	100	44	4	51

Table 3: The details of the datasets. "Evaluation" represents manual evaluation methods and "Limit" represents word limits of summarization.

Combined with step 1, fully pruned BE is defined as follows:

$$\text{pBE}_{-\text{cnt}+\text{cls}}(\mathbf{R}, S) = \frac{\sum_{k=1}^{K}\sum_{m=1}^{M_k}\{O(f'^k_m, S')\}}{\sum_{k=1}^{K}\sum_{m=1}^{M_k}\{O(f'^k_m, \mathbf{R}'_k)\}}. \quad (3)$$

4 Experimental Setup

To assess the effectiveness of pBE, we computed the correlation coefficient between pBE scores and human judgments, as well as between the scores of other automatic evaluation measures and manual scores for comparison. We used multi-document summarization datasets DUC 2003 - 2007 and TAC 2011. The correlation was computed between all system summaries, excluding reference summaries.

Our first experiment compared the performance of pBE and ROUGE on DUC datasets. Since a dependency triple is a type of bigram/skip-bigram, we chose ROUGE-2 and ROUGE-S4 for comparison. We also examined ROUGE-SU4[9] because it is known as a strong baseline that outperforms most of other measures in TAC 2011 AESOP task (Owczarzak and Dang, 2011).

The second experiment was designed to see how well pBE worked compared with our related method ROUGE-WE. We chose the latest AESOP dataset, TAC 2011, for which ROUGE-WE achieved the highest Spearman coefficient (Ng and Abrecht, 2015).

The details of our experimental setup are given in Table 3 and below.

Parser: We used the neural-network dependency parser of Stanford CoreNLP (Manning et al., 2014). Dependencies were set to enhanced++ Universal Dependencies (Schuster and Manning, 2016).

Clustering: We employed hierarchical clustering, maximum distance method. The number of clusters, N, was set to $0.975 * Q$.

Word Embeddings: A set of pre-trained Google-News word embeddings[10]. It contains 3 million words, each of which has a word embedding of 300 dimensions.

5 Results and Discussion

Table 1 and 2 show the evaluation results on DUC and TAC data set, respectively.

Regardless of the diversity of datasets, pBE outperformed ROUGE in most cases (table 1). Interestingly, although step 2 itself sometimes did not work well, the combination of both steps gener-

[9] All three ROUGE here were run with stemming but with no removal of stopwords.

[10] https://code.google.com/archive/p/word2vec/

	Relation	BE	$\text{BE}_{+\text{cls}}$	Increased
DUC 2003	compound & det	235	281	46
	subj & obj	1	1	0
DUC 2004	compound & det	426	446	20
	subj & obj	10	10	0
DUC 2005	compound & det	2570	2750	180
	subj & obj	32	39	7
DUC 2006	compound & det	2969	3083	114
	subj & obj	26	49	23
DUC 2007	compound & det	3508	3622	114
	subj & obj	48	57	9

Table 4: The number of basic elements which returned more than 1 in $\min\{N(f_m^k, \mathbf{R}_k), N(f_m^k, S)\}$, before clustering (BE) and after clustering ($\text{BE}_{+\text{cls}}$), and the difference of the numbers, $\text{BE}_{+\text{cls}} - \text{BE}$ (Increased). The relation "subj & obj" includes nsubj, nsubjpass, csubj, csubjpass, iobj and dobj.

ally achieved the best performance. This is because clustering enhanced not only the matching of informative basic elements but also that of low-information basic elements. Table 4 shows how the number of compound and det triples increased, compared with that of subj (nsubj, nsubjpass, csubj and csubjpass) and obj (iobj and dobj) triples. In all datasets, the number of compound and det triples that returned more than 1 in $\min\{N(f_m^k, \mathbf{R}_k), N(f_m^k, S)\}$ increased much more than that of subj and obj, after converting unigrams into cluster IDs. Although clustering reduced semantic mismatches, it worsened the problem of redundant counting. Nonetheless, this problem can be easily solved by applying step 1. This is why the combination of step 1 and 2 was so synergistic.

Another problem with step 2 is that it sometimes makes inappropriate clusters. For example, numbers tend to be put in the same clusters since our word embeddings place them close to each other. In summaries, however, confusing quantitative information such as "two apples" and "five apples" must be avoided . It will be our future work to specify where clustering fails to work and to get rid of inappropriate clusters.

Table 2 shows that pBE achieved the best rank correlation among the other competitors in TAC 2011 and ROUGE-WE. Although its score was lower in Pearson coefficient, it should be noted that the Pearson correlation is based on some strict assumptions: Samples are normally distributed and are linearly related to each other. Since Spearman/Kendall correlation is free from these assumptions, the best rank correlation is a good evidence of pBE's performance.

6 Conclusion

We proposed an automatic evaluation measure of summarization, pBE. It is designed to prune redundant basic elements in two steps: (1) disregarding frequency count of basic elements and (2) using word similarity to reduce semantically overlapped basic elements. Our experiments show that pBE outperforms ROUGE in most cases and achieves the highest rank correlation coefficient in TAC 2011 AESOP task.

References

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Eduard H. Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 06)*.

Niraj Kumar, Kannan Srinathan, and Vasudeva Varma. 2011. Using unsupervised system with least linguistic features for tac aesop task. In *Proceed- ings of the Text Analysis Conference (TAC)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. pages 74–81.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS)*. pages 3111–3119.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1925–1930.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo,

Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*.

Karolina Owczarzak and Hoa Trang Dang. 2011. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *Proceedings of the Text Analysis Conference (TAC)*.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Stephen Tratz and Eduard H. Hovy. 2008. Summarization evaluation using transformed basic elements. In *Proceedings of Text Analytics Conference (TAC)*.

Qian Yang, Rebecca J. Passonneau, and Gerard de Melo. 2016. PEAK: Pyramid evaluation via automated knowledge extraction. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*. AAAI Press.