冗長なNgram/Basic Elementsを除いた自動要約評価指標

本多 右京[1]　　平尾 努[2]　　永田 昌明[2]
[1] 奈良先端科学技術大学院大学 情報科学研究科
[2]NTT コミュニケーション科学基礎研究所
[1]honda.ukyo.hn6@is.naist.jp, [2]{hirao.tsutomu, nagata.masaaki}@lab.ntt.co.jp

1 はじめに

自動要約評価指標は，自動要約システム研究において重要な役割を果たす．自動生成された要約に人手でスコアを付けるのはコストが大きすぎるため，自動要約システムの良し悪しはもっぱら自動要約評価指標によるスコアで決定されるからである．

自動要約評価指標で最も広く使われているのはROUGE[4] と BE[2] である．ROUGE/BE は，評価対象となる要約とその参照要約との間でいくつのユニット，すなわち ngram/basic elements[1]が一致するかを数える．多くの場合，ngram では unigram か bi-gram，basic elements では (head|modifier|relation) の dependency triple が用いられる．

しかし，この評価方法は次の 2 点で人間の要約評価方法と大きく異なる．1 点目は低情報量のユニットを重く評価すること，2 点目は意味的に重複するユニットに複数回点数を与えることである．

我々は，これら冗長なユニットをそれぞれ次の 2 ステップで取り除いてこの問題を解決する．すなわち，ユニットの頻度カウントを省略すること，意味的に重複するユニットを単語分散表現に基づいて 1 つのクラスタにまとめることである．この 2 ステップを適用した ROUGE/BE を pruned ROUGE/pruned BE (pROUGE/pBE) と呼ぶことにする．

実験の結果，DUC 2003-2007 において pROUGE/pBE がそれぞれ ROUGE/BE より人間の評価結果に対して高い相関を示し，特に basic elements は元々 ngram より冗長なユニットが少ないため，大半のケースで pBE が ROUGE の相関を大きく上回ることが確認された．また，関連研究の ROUGE-WE との比較では，TAC 2011 の自動要約評価指標コンペティションに参加した指標の中で，pBE が最も高い順位相関係数を示した．

[1]名称の曖昧性を避けるため，自動要約評価指標としての Basic Elements は BE，一致を見るユニットとしての Basic Elements は basic element(s) と表記する．

2 関連研究

我々の提案手法と最も関連するのは ROUGE-WE[7] と BEwT-E[11] である．ROUGE-WE は単語分散表現を採用し，ngram 同士の一致は 0/1 ではなく，コサイン類似度にて [0,1] の連続値で表現される．この手法は，評価対象要約と参照要約間で ngram の意味的な一致を考慮することができる点において本研究に類似しているが，その片方の要約内でも意味的に重複する ngram があるかを判断しない点で異なる．また，ROUGE-WE は頻度カウントを省略しない．これらの違いから，ROUGE-WE では本研究で目的とするところの冗長なユニットの除去はなされない．

BEwT-E は basic elements を変形し，これらが表層形の違いを超えて一致できるようにする手法である．ROUGE-WE と同様，表層形以外の要素を考慮して一致を判断させようとする点で本研究に類似するが，この変形は複雑なルールを必要とする．さらにこのルールは英語固有の性質に基づいたものも多く，他の言語に適用することが難しい．提案手法は単語分散表現のみで意味的な重複を判断することができるため，多言語に適用可能である．BEwT-E はまた，頻度カウントの有無による性能への影響も調べている．しかし単にハイパーパラメータの一つとして操作しているだけであり，なぜ頻度カウントが性能に影響するのかは述べられていない．本研究の貢献は，頻度カウント省略の効果は冗長なユニットの除去であり，そのため意味的に重複するユニットの除去と組み合わせるととりわけ大きく効果を発揮すると説明を与える点である．

3 提案手法

この章では ROUGE/BE の概要について述べた後，提案手法の pROUGE/pBE を 2 ステップに分けて説

明する．

3.1 ROUOGE/BE

ROUGE/BE は次の式で定義される：

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

K 個の参照要約 $\mathbf{R} = \{R_1, ..., R_K\}$ と評価対象要約 S，$\mathbf{R_k}$ に現れるユニットのセットを $F_k = \{f_1^k, ..., f_M^k\}$ とおく．$N(f_m^k, R_k)$，$N(f_m^k, S)$ はそれぞれ，f_m^k が何回 R_k，S に出現するかを返す関数である．上式からわかるとおり，ROUGE/BE は評価対象要約 S の中に f が何回出現するかでスコアを定める．

ROUGE と BE の相違点は，ROUGE は ngram，BE は basic elements をユニットにとる点である．Basic elements は ngram の欠点を補う目的で提案された [2]．ROUGE は unigram や bigram のような短い ngram を用いることが多いが，これらはしばしば情報量の低い ngram を生み出す．例えば，“John went to the store on foot” という文は [“John went”, “went to”, “to the”, “the store”, “store on”, “on foot”] という bigram に分解される．ここで見られる “to the” のような，機能語のみから成る bigram はほぼ意味をなさないが，機能語は文中に頻出するためこのような bigram は頻繁に出現する．これに対して basic elements は係り受け情報を持つので，“to” と対になるのは “the” ではなく “store” だということが正しく反映される[2]．

我々は単語の依存構造として，Universal Dependencies (UD) [8] を使用する．UD は内容語を中心に係り受けをとるため，文の要点をより直接的に表す dependency triple がとれるという利点がある．例えば，上記の文は Stanford Dependency[1] では [(went|John|nsubj), (went|to|prep), (store|the|det), (to|store|pobj), (went|on|prep), (on|foot|pobj)] と分解されるのに対し，UD では [(went|John|nsubj), (store|to|case), (store|the|det), (went|store|nmod:to), (foot|on|case), (went|foot|nmod:on)] となる[3]．UD の triple では，述語-目的語の関係が，(went|to|prep) と (to|store|pobj) のような中間的な表現を経ることなく (went|store|nmod:to) と表現されている．また UD のもう一つの利点として，多言語に応用可能であるということが挙げられる．

Basic elements として扱うのは，UD の係り受けラベルのうち，狭義の係り受けラベルと multiword expression (MWE) の係り受けラベル[4]をもつすべての dependency triple (head|modifier|relation) とする．これは，この組み合わせの BE が最も良い性能を示したためである．

3.2 Step 1: 頻度カウントの省略

前節のとおり，ROUGE/BE ではユニットの出現頻度がスコアに大きな影響を持つ[5]．問題は，ROUGE/BE で複数回点数を与えられるユニットは情報量が低いものであることが多いということである．前述のように，bigram では機能語の対が頻出しこれらが他の bigram より大きくスコアを稼ぐ．BE ではこのようなユニットは現れない．しかし BE でも det，compound の triple が不当にスコアを付与されるという問題点が残る．例えば DUC 2003 では，$\min\{N(f_m^k, \mathbf{R}_k), N(f_m^k, S)\}$ で 2 以上を返す basic elements が 301 あるうち，139 が compound，96 が det で，全体の約 78%を占めた．これは，これらの triple が一つの名詞しか表していないに等しいためだと考えられる[6]．これに対して，文の中心である動詞と結びついた，より情報量のあると思われる係り受け関係 nsubj，dobj，iobj などはほぼ頻度による重み付けを受けない．Det と compound の triple は不要ではないが，nsubj らより重く評価する重要性はない．

そこで我々は，単純に頻度カウントを省略することでこの問題に対処する．頻度カウントを省略した ROUGE/BE を次の式で定義する：

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

$O(f_m^k, \mathbf{R}_k)$ と $O(f_m^k, S)$ は，それぞれ f_m^k が $\mathbf{R}_k$，S に入っていれば 1 を返し，そうでなければ 0 を返す．

[2] 単に機能語を取り除くだけでは，“store foot” のような意味のない bigram ができてしまう．

[3] ROOT の triple は，後述するように本稿での basic elements から除外されているためここでも表記していない．

[4] nsubj, dobj, iobj, csubj, ccomp, xcomp, obl, vocative, expl, dislocated, advcl, advmod, discourse, aux, cop, mark, nmod, appos, nummod, acl, amod, det, clf, case, fixed, flat, compound.

[5] BE については頻度カウントの有無が既にオプションとして用意されているが [11]，我々はなぜ頻度カウントを無視するオプションが有効なのか説明する．

[6] 例えば複合名詞 “ Donald Trump” は，“Donald Trump won the election.” や “Donald Trump will visit China next week.” など複数のトピックにわたって何度も用いられる可能性があるが，主語-述語関係を表した “Trump won” などは Trump が何かに勝ったという特定のトピックでしか現れず，またそのようなトピックが一つの要約で複数回出てくることは考えにくい．

3.3 Step 2: 意味的に重複する単語を同じクラスタに割り当てる

人間は意味的な同一性を判断できる．もし我々が，要約の中にある要点が含まれているか否かを判断するように言われたら，要点と意味的に一致する記述があるかどうかを見るはずである．しかし ROUGE/BE は表層形の一致しか判断できない．このことから，表層形が異なっていても意味的には同じであるようなユニットについて，ROUGE/BE では重複してスコアを与えたり，あるいは不当に与えなかったりすることがある．例えば，ある評価対象要約に “John killed” と “John murdered” の記述があり，その各記述がそれぞれ別の参照要約に含まれていたとする．この場合，評価対象要約は同じ内容に対して二重にスコアを与えられることになる．また，今度は評価対象要約に “John killed”，参照要約に “John murdered” が含まれていたとすると，正しい内容を含んでいるのにもかかわらず表層形の相違から正しくスコアが与えられない．

この問題に対処するため，意味的に同一と思われる単語を同じクラスタに割り当てる．K 個の参照要約 $\mathbf{R} = \{R_1, ..., R_K\}$ と，評価対象要約 S, $\mathbf{R}$ と S に含まれる unigram のセット $U = \{u_1, ..., u_P\}$, U 内の unigram Q 個に対する単語分散表現のセット $V = \{v_1, ..., v_Q\}$ $(Q \leq P)$ が与えられているとする．V に基づいて階層クラスタリングを行い，U をクラスタ ID のセット $C = \{c_1, ..., c_N\}$ に割り当てる．クラスタ数 N はハイパーパラメータである．次に，$\mathbf{R}$ と S 中の unigram を該当するクラスタ ID c に置き換える．クラスタ ID を付されていない unigram は置き換えず残す．置き換え操作後の参照要約と評価対象要約をそれぞれ $\mathbf{R}'$，S' とおき，$\mathbf{R}'_k$ 中のユニットのセットを $F'_k = \{f'^k_1, ..., f'^k_M\}$ とおく．これを step 1 と組み合わせると，冗長なユニットを最大限除去した提案手法は以下のように定義される：

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

4 実験設定

提案手法の性能を調べるにあたって，人手の評価スコアとの相関係数を比較した．使用したのは DUC 2003-2007，TAC 2011 の複数文書要約データセットである．詳細は表 1 を参照．相関係数は，参照要約を除く全システム要約について計算した．

表 1: データセット詳細．Ref と System は，それぞれトピックごとの参照要約とシステム要約の数である．

	人手評価法	制限語数	トピック	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	250	44	4	51

まず，ROUGE/BE と比較して pROUGE/pBE の性能がどの程度向上したかを確かめるため DUC のデータセットで実験をした．Basic elements には dependency triple を使うので，ngram で最もこれに近いのは bigram であると考え，ROUGE-2[7] と比較した．次に，ROUGE-WE が高い性能を示した，TAC 2011 AESOP task のデータセットで比較を行った．先行研究では ROUGE-WE-1 が最も良い数字を出していたので [7]，ROUGE には pROUGE-1 を用いた．

その他詳細な実験設定は以下に記す．

Parser： Stanford CoreNLP[5] の neural-network dependency parser を使用した．係り受けアノテーションは enhanced++ Universal Dependencies [10] に設定した．

クラスタリング： 最長距離法の階層クラスタリングを用いた．クラスタ数 N は，pROUGE では $0.95 * Q$，pBE では $0.975 * Q$ とした．

単語分散表現： word2vec[6] を使って Google-News で学習された，300 次元 300 万単語のものを使用した[8]．

5 結果と考察

実験結果を表 2 と表 3 に示す．表 2 から，DUC 2003 以外ではすべて ROUGE より pROUGE の相関係数が高くなり，すべてのケースで BE と比べて pBE の相関係数が高くなっていることがわかる．特に pBE の相関係数の高さは顕著であり，7 つのデータセットのうち 4 つで最高値を出している．これは，元々冗長なユニットの少ない BE が，提案手法によってさらにこれを取り除かれたためだと考えられる．

表 2 では，+cls が相関係数を下げてしまうことが多いのにもかかわらず，-cnt+cls の組み合わせが他の指標より高い相関係数をとっていることも見てとれる．こ

[7] 本稿の実験において，ROUGE-1, 2 と BE はすべて著者らの実装である．我々の実装では，できるだけ BE と ROUGE の条件を近づけるため，tokenization はどちらも Stanford CoreNLP の tokenizer を使用した．(p)ROUGE-2 は stemming なしで stopword 除去なし，pROUGE-1 は stemming なしで stopword 除去の設定とした．

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

表 2:　pROUGE と pBE の相関係数．“Pearson/Spearman/Kendall” の順に記載．

	DUC03	DUC04	DUC05	DUC06	DUC06 pyr	DUC07	DUC07 pyr
ROUGE-2	.872/.832/.667	.928/.814/.662	.912/.889/.718	.828/.736/.557	.882/.874/.714	.885/.882/.728	.982/.978/.923
pROUGE-$2_{-\text{cnt}}$	.881/.832/.667	.934/.841/.691	**.925/.905/.751**	.854/.777/.597	.913/.888/.766	.893/.888/.740	.986/.984/.949
pROUGE-$2_{+\text{cls}}$	.872/.826/.650	.930/.826/.676	.912/.878/.702	.834/.732/.557	.887/.886/.740	.885/.880/.724	.984/.989/.949
pROUGE-$2_{-\text{cnt+cls}}$	.880/.824/.650	.937/.853/.721	.924/.898/.735	.859/.785/.604	**.919/.893/.775**	.892/.886/.736	**.989/.995/.974**
BE	.927/.862/.617	.936/.868/.721	.897/.867/.714	.834/.757/.584	.883/.837/.680	.891/.890/.732	.982/.973/.897
$\text{pBE}_{-\text{cnt}}$	.929/.871/.717	.938/.873/.735	.904/.879/.718	.857/.783/.607	.896/.850/.723	**.902/.906/.760**	.985/.978/.923
$\text{pBE}_{+\text{cls}}$	.929/.871/.717	.941/.877/.735	.897/.862/.702	.837/.779/.607	.888/.844/.688	.890/.893/.732	.981/.967/.897
$\text{pBE}_{-\text{cnt+cls}}$	**.932/.879/.750**	**.944/.885/.765**	.905/.876/.714	**.861/.805/.635**	.901/.841/.688	**.902/.906**/.756	.985/.989/.949

表 3:　pROUGE/pBE，ROUGE-WE とその他 TAC 2011 AESOP task に参加した自動評価指標の相関係数．ROUGE-SU4/ROUGE-WE-1/C_S_IIITH3[3] はそれぞれ Pearson/Spearman/Kendall 相関係数で最高値をとっていた指標 [9]．

	Pearson	Spearman	Kendall
ROUGE-SU4	**.981**	.894	.737
C_S_IIITH3	.965	.903	.758
ROUGE-WE-1	.949	.914	.753
pROUGE-$1_{-\text{cnt+cls}}$	.974	.902	.760
$\text{pBE}_{-\text{cnt+cls}}$	.946	**.915**	**.772**

れは，クラスタリングによって，一つの要約に複数回出現する低情報なユニットが増えたからである．例えば DUC 2003 においては，$\min\{N(f_m^k, \mathbf{R}_k), N(f_m^k, S)\}$ で 2 以上を返す basic elements が 301 あったが，クラスタリング後はこれが 352 にまで増えた．内訳は det が 106，compound が 175 で，全体の約 80%を占めるまでに増加した．このように，クラスタリングは意味的な重複を取り除くと同時に低情報量のユニットの重みを上げてしまうが，この重みは-cnt を適用すれば解消される．-cnt+cls はこの相乗効果の結果，高い相関係数を出すものと思われる．

表 3 では，$\text{pROUGE}_{-\text{cnt+cls}}$ が ROUGE-WE1 に対して Pearson/Kendall 相関係数で上回り，$\text{pBE}_{-\text{cnt+cls}}$ は Spearman/Kendall 相関係数で全評価指標中最高値を出しており，提案手法の有効性を示している．

6　おわりに

本稿では，冗長なユニットを取り除くために次の 2 ステップを適用することを提案した．すなわち，(1) 頻度カウントを省略することと，(2) 意味的に重複する単語を同じクラスタに割り振ることである．以上を ROUGE/BE に適用した結果，使用したほぼすべてのデータセットで性能が向上した．特に BE に 2 ステップを適用したものについては，DUC のデータセット 7 つのうち 5 つにおいて人間の評価結果との相関が ROUGE を大きく上回り，TAC 2011 では順位相関係数 2 つで全指標中最も高い値をとった．

参考文献

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

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

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

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

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

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 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, 2013.

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

[8] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic ˇ , Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*, 2016.

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

[10] Sebastian Schuster and Christopher D. Manning. 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)*, 2016.

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