Kernel-based Approach for Automatic Evaluation of Natural Language Generation Technologies: Application to Automatic Summarization

Tsutomu Hirao
NTT Communication Science Labs.
NTT Corp.
hirao@cslab.kecl.ntt.co.jp

Manabu Okumura
Precision and Intelligence Labs.
Tokyo Institute of Technology
oku@pi.titech.ac.jp

Hideki Isozaki
NTT Communication Science Labs.
NTT Corp.
isozaki@cslab.kecl.ntt.co.jp

Abstract

In order to promote the study of automatic summarization and translation, we need an accurate automatic evaluation method that is close to human evaluation. In this paper, we present an evaluation method that is based on convolution kernels that measure the similarities between texts considering their substructures. We conducted an experiment using automatic summarization evaluation data developed for Text Summarization Challenge 3 (TSC-3). A comparison with conventional techniques shows that our method correlates more closely with human evaluations and is more robust.

1 Introduction

Automatic summarization, machine translation, and paraphrasing have attracted much attention recently. These tasks include text-to-text language generation. Evaluation workshops are held in the U.S. and Japan, e.g., the Document Understanding Conference (DUC)[1], NIST Machine Translation Evaluation[2] as part of the TIDES project, the Text Summarization Challenge (TSC)[3] of the NTCIR project, and the International Workshop on Spoken Language Translation (IWSLT)[4].

These evaluation workshops employ human evaluations, which are essential in terms of achieving high quality evaluations results. However, human evaluations require a huge effort and the cost is considerable. Moreover, we cannot automatically evaluate a new system even if we use the corpora built for these workshops, and we cannot conduct re-evaluation experiments.

To cope with this situation, there is a particular need to establish a high quality automatic evaluation method. Once this is done, we can expect great progress to be made on natural language generation.

In this paper, we propose a novel automatic evaluation method for natural language generation technologies. Our method is based on the Extended String Subsequence Kernel (ESK) (Hirao et al., 2004b) which is a kind of convolution kernel (Collins and Duffy, 2001). ESK allows us to calculate the similarities between a pair of texts taking account of word sequences, their word sense sequences and their combinations.

We conducted an experimental evaluation using automatic summarization evaluation data developed for TSC-3 (Hirao et al., 2004a). The results of the comparison with ROUGE-N (Lin and Hovy, 2003; Lin, 2004a; Lin, 2004b), ROUGE-S(U) (Lin, 2004b; Lin and Och, 2004) and ROUGE-L (Lin, 2004a; Lin, 2004b) show that our method correlates more closely with human evaluations and is more robust.

2 Related Work

Automatic evaluation methods for automatic summarization and machine translation are grouped into two classes. One is the longest common subsequence (LCS) based approach (Hori et al., 2003; Lin, 2004a; Lin, 2004b; Lin and Och, 2004). The other is the N-gram based approach (Papineni et al.,

[1] http://duc.nist.gov
[2] http://www.nist.gov/speech/tests/mt/
[3] http://www.lr.titech.ac.jp/tsc
[4] http://www.slt.atr.co.jp/IWSLT2004

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 145–152, Vancouver, October 2005.

Table 1: Components of vectors corresponding to S1 and S2. Bold subsequences are common to S1 and S2.

d	subsequence	S1	S2	d	subsequence	S1	S2	d	subsequence	S1	S2
1	**Becoming**	1	1	2	**Becoming-is**	λ^2	λ^2	2	astronaut-DREAM	0	λ^2
	DREAM	1	1		**Becoming-my**	λ^3	λ^3		astronaut-ambition	0	λ^2
	SPACEMAN	1	1		***SPACEMAN*-*DREAM***	λ^3	λ^2		astronaut-is	0	1
	a	1	0		SPACEMAN-ambition	0	λ^2		astronaut-my	0	λ
	ambition	0	1		SPACEMAN-dream	λ^3	0		cosmonaut-DREAM	λ^3	0
	an	0	1		SPACEMAN-great	λ^2	0		cosmonaut-dream	λ^3	0
	astronaut	0	1		***SPACEMAN*-is**	1	1		cosmonaut-great	λ^2	0
	cosmonaut	1	0		***SPACEMAN*-my**	λ	λ		cosmonaut-is	1	0
	dream	1	0		a-DREAM	λ^4	0		cosmonaut-my	λ	0
	great	1	0		a-SPACEMAN	1	0		great-DREAM	1	0
	is	1	1		a-cosmonaut	1	0		great-dream	1	0
	my	1	1		a-dream	λ^4	0		**is-*DREAM***	λ^2	λ
2	**Becoming-*DREAM***	λ^5	λ^4		a-great	λ^3	0		is-ambition	0	λ
	Becoming-*SPACEMAN*	λ	λ		a-is	λ	0		is-dream	λ^2	0
	Becoming-a	1	0		a-my	λ^2	0		is-great	λ	0
	Becoming-ambition	0	λ^4		an-DREAM	0	λ^3		**is-my**	1	1
	Becoming-an	0	1		an-SPACEMAN	0	1		**my-*DREAM***	λ	1
	Becoming-astronaut	0	λ		an-ambition	0	λ^3		my-ambition	0	1
	Becoming-cosmonaut	λ	0		an-astronaut	0	1		my-dream	λ	0
	Becoming-dream	λ^5	0		an-is	0	λ		my-great	1	0
	Becoming-great	λ^4	0		an-my	0	λ^2				

2002; Lin and Hovy, 2003; Lin, 2004a; Lin, 2004b; Soricut and Brill, 2004).

Hori et. al (2003) proposed an automatic evaluation method for speech summarization based on word recognition accuracy. They reported that their method is superior to BLEU (Papineni et al., 2002) in terms of the correlation between human assessment and automatic evaluation. Lin (2004a; 2004b) and Lin and Och (2004) proposed an LCS-based automatic evaluation measure called ROUGE-L. They applied ROUGE-L to the evaluation of summarization and machine translation. The results showed that the LCS-based measure is comparable to N-gram-based automatic evaluation methods. However, these methods tend to be strongly influenced by word order.

Various N-gram-based methods have been proposed since BLEU, which is now widely used for the evaluation of machine translation. Lin et al. (2003) proposed a recall-oriented measure, ROUGE-N, whereas BLEU is precision-oriented. They reported that ROUGE-N performed well as regards automatic summarization. In particular, ROUGE-1, *i.e.*, unigram matching, provides the best correlation with human evaluation. Soricut et. al (2004) proposed a unified measure. They integrated a precision-oriented measure with a recall-oriented measure by using an extension of the harmonic mean formula. It performs well in evaluations of machine translation, automatic summarization, and question answering. However, N-gram based methods have a critical problem; they cannot consider co-occurrences with gaps, although the LCS-based method can deal with them. Therefore, Lin and Och (2004) introduced skip-bigram statistics for the evaluation of machine translation. However, they did not consider longer skip-n-grams such as skip-trigrams. Moreover, their method does not distinguish between bigrams and skip-bigrams.

3 Kernel-based Automatic Evaluation

The above N-gram-based methods correlated closely with human evaluations. However, we think some skip-n-grams (n$\geq$3) are useful. In this paper, we employ the Extended String Subsequence Kernel (ESK), which considers both n-grams and skip-n-grams. In addition, the ESK allows us to add word senses to each word. The use of word senses enables flexible matching even when paraphrasing is used.

The ESK is a kind of convolution kernel (Collins and Duffy, 2001). Convolution kernels have recently attracted attention as a novel similarity measure in natural language processing.

3.1 ESK

The ESK is an extension of the String Subsequence Kernel (SSK) (Lodhi et al., 2002) and the Word Sequence Kernel (WSK) (Cancedda et al., 2003).

The ESK receives two node sequences as inputs

and maps each of them into a high-dimensional vector space. The kernel's value is simply the inner product of the two vectors in the vector space. In order to discount long-skip-n-grams, the decay parameter λ is introduced.

We explain the computation of the ESK's value whose inputs are the sentences (S1 and S2) shown below. In the example, word senses are shown in braces.

S1 Becoming a cosmonaut:{*SPACEMAN*} is my great dream:{*DREAM*}
S2 Becoming an astronaut:{*SPACEMAN*} is my ambition:{*DREAM*}

In this case, "cosmonaut" and "astronaut" share the same sense {*SPACEMAN*} and "ambition" and "dream" also share the same sense {*DREAM*}. We can use WordNet for English and *Goitaikei* (Ikehara et al., 1997) for Japanese.

Table 1 shows the subsequences derived from S1 and S2 and its weights. Note that the subsequence length is two or less. From the table, there are fifteen subsequences[5] that are common to S1 and S2. Therefore, $\mathrm{ESK}^{d=2}(\mathrm{S1},\mathrm{S2}) = 7 + \lambda + 2\lambda^2 + \lambda^3 + \lambda^4 + \lambda^5 + \lambda^6 + \lambda^9$. For reference, there are three unigrams, one bigram, zero trigrams and three skip-bigrams common to S1 and S2.

Formally, the ESK is defined as follows. T and U are node sequences.

$$\mathrm{ESK}^d(T,U)=\sum_{m=1}^{d}\sum_{t_i\in T}\sum_{u_j\in U} K_m(t_i,u_j) \tag{1}$$

$$K_m(t_i,u_j)=\begin{cases} val(t_i,u_j) & \text{if } m=1 \\ K'_{m-1}(t_i,u_j)\cdot val(t_i,u_j) & \text{otherwise}\end{cases} \tag{2}$$

Here, d is the upper bound of the subsequence length and $K'_m(t_i,u_j)$ is defined as follows. t_i is the i-th node of T. u_j is the j-th node of U. The function $val(s,t)$ returns the number of attributes common to given nodes s and t.

$$K'_m(t_i,u_j)=\begin{cases} 0 & \text{if } j=1 \\ \lambda K'_m(t_i,u_{j-1}) + K''_m(t_i,u_{j-1}) & \text{otherwise}\end{cases} \tag{3}$$

$K''_m(t_i,u_j)$ is defined as follows:

$$K''_m(t_i,u_j)=\begin{cases} 0 & \text{if } i=1 \\ \lambda K''_m(t_{i-1},u_j) + K_m(t_{i-1},u_j).\end{cases} \tag{4}$$

[5]Bold subsequences in Table 1.

Finally, we define the similarity measure between T and U by normalizing ESK. This similarity can be regarded as an extension of the cosine measure.

$$\mathrm{Sim}^d_{\mathrm{esk}}(T,U)=\frac{\mathrm{ESK}^d(T,U)}{\sqrt{\mathrm{ESK}^d(T,T)\mathrm{ESK}^d(U,U)}}. \tag{5}$$

3.2 Automatic Evaluation based on ESK

Suppose, $\mathcal{C}$ is a system output, which consists of ℓ sentences, and $\mathcal{R}$ is a human written reference, which consists of m sentences. c_i is a sentence in $\mathcal{C}$, and r_j is a sentence in $\mathcal{R}$. We define two scoring functions for automatic evaluation. First, we define a precision-oriented measure as follows:

$$P^d_{\mathrm{esk}}(\mathcal{C},\mathcal{R})=\frac{1}{\ell}\sum_{i=1}^{\ell}\max_{1\le j\le m}\mathrm{Sim}^d_{\mathrm{esk}}(c_i,r_j) \tag{6}$$

Symmetrically, we define a recall-oriented measure as follows:

$$R^d_{\mathrm{esk}}(\mathcal{C},\mathcal{R})=\frac{1}{m}\sum_{j=1}^{m}\max_{1\le i\le \ell}\mathrm{Sim}^d_{\mathrm{esk}}(c_i,r_j) \tag{7}$$

Finally, we define a unified measure, *i.e.*, F-measure, as follows:

$$F^d_{\mathrm{esk}}(\mathcal{C},\mathcal{R})=\frac{(1+\beta^2)\times R_{\mathrm{esk}}(\mathcal{C},\mathcal{R})\times P_{\mathrm{esk}}(\mathcal{C},\mathcal{R})}{R_{\mathrm{esk}}(\mathcal{C},\mathcal{R})+\beta^2\times P_{\mathrm{esk}}(\mathcal{C},\mathcal{R})} \tag{8}$$

β is a cost parameter for R_{esk} and P_{esk}. β's value is selected depending on the evaluation task. Since summary should not miss important information given in the human reference, recall is more important than precision. Therefore,a large β will yield good results.

3.3 Extension for Multiple References

When multiple human references (correct answers) are available, we define a simple function for multiple references as follows:

$$F^{\mathrm{mean}}_{\mathrm{esk}}(\mathcal{C},R)=\frac{1}{n}\sum_{i=1}^{n}F_{\mathrm{esk}}(\mathcal{C},\mathcal{R}_i), \tag{9}$$

Here, equation (9) gives the average score. R indicates a set of references; $R=\{\mathcal{R}_1,\cdots,\mathcal{R}_n\}$.

4 Experimental Evaluation

To confirm and discuss the effectiveness of our method, we conducted an experimental evaluation using TSC-3 multiple document summarization

evaluation data and our additional data.

4.1 Task and Evaluation Metrics in TSC-3

The task of TSC-3 is multiple document summarization. Participants were given a set of documents about a certain event and required to generate two different length summaries for the entire document set. The lengths were about 5% and 10% of the total number of characters in the document set, respectively. Thirty document sets were provided for the official run evaluation. There were ten participant systems; one provided by the TSC organizers as a baseline system.

The evaluation metric follows DUC's SEE evaluation scheme (Harman and Over, 2004). For each document set, one human subject makes a reference summary and uses it as a basis for evaluating ten system outputs. This human evaluation procedure consists of the following steps:

Step 1 For each reference sentence $r_j (\in \mathcal{R})$, repeat Steps 2 and 3.

Step 2 For r_j, the human assessor finds the most relevant sentence set S from the system output.

Step 3 The assessor assigns a score, $e(r_j, S)$, $0, 0.1, \cdots, 1.0$. 1.0 means perfect. in terms of how much of the content of r_j can be reproduced by using only sentences in S.

Step 4 Finally, the evaluation score of output $\mathcal{C}$ for reference $\mathcal{R}$ is defined $H(\mathcal{R}, \mathcal{C}) = \sum_j e(r_j, S)/|\mathcal{R}|$.

The final score of a system is calculated by applying the above procedure and normalized by the number of topics, *i.e.*, $\sum_{t=1}^{30} H(\mathcal{R}_t, \mathcal{C}_t)/30$. When multiple references $R(=\{\mathcal{R}1, \cdots, \mathcal{R}_n\})$ are available, the scores are given as follows: $H^{\mathrm{mean}}(R, \mathcal{C}) = \sum_k H(\mathcal{R}_k, \mathcal{C})/|R|$.

4.2 Variation of Human Assessors

In TSC-3's official run evaluation, system outputs were compared with one human written reference summary for each topic. There were five topic sets and five human assessors (A-E in Table 2) for each topic set.

Before we use the one human written reference summary as the gold-standard-reference, to examine variations among human assessors, we prepared two additional human summaries for each topic sets.

Table 2: The relationship between topics and reference summary creators, *i.e.*, human assessors. $S(A)$ indicates a subject A's evaluation score for all systems for corresponding topics.

topic-ID	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_{avg}$
1 - 6	S(A)	S(E)	S(C)	mean(S(A),S(E),S(C))
7 - 12	S(B)	S(A)	S(D)	mean(S(B),S(A),S(D))
13 - 18	S(C)	S(B)	S(E)	mean(S(C),S(B),S(E))
19 - 24	S(D)	S(C)	S(A)	mean(S(D),S(C),S(A))
25 - 30	S(E)	S(D)	S(B)	mean(S(E),S(D),S(B))

Table 3: Correlations between human judgments.

	correlation coefficient (r)				rank correlation coefficient (ρ)			
short	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_{\mathrm{avg}}$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_{\mathrm{avg}}$
$\mathcal{D}_1$	1.00	.968	.902	.988	1.00	.976	.697	.988
$\mathcal{D}_2$	–	1.00	.910	.996	–	1.00	.733	.988
$\mathcal{D}_3$	–	–	1.00	.914	–	–	1.00	.758
$\mathcal{D}_{\mathrm{avg}}$	–	–	–	1.00	–	–	–	1.00
long	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_{\mathrm{avg}}$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_{\mathrm{avg}}$
$\mathcal{D}_1$	1.00	.908	.822	.964	1.00	.964	.939	.964
$\mathcal{D}_2$	–	1.00	.963	.987	–	1.00	.952	1.00
$\mathcal{D}_3$	–	–	1.00	.931	–	–	1.00	.932
$\mathcal{D}_{\mathrm{avg}}$	–	–	–	1.00	–	–	–	1.00

Therefore, we obtained three reference summaries and evaluation results for each topic sets (Table 2).

Moreover, we prepared unified evaluation results of three human judgment as $\mathcal{D}_{\mathrm{avg}}$, which is calculated as the average of three human scores.

The relationship between topics and human assessors is shown in Table 2. For example, subject B generates summaries and evaluates all systems for topics 7-12, 13-18 and 25-30 on $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ respectively. Note that each human subject, A to E, was a retired professional journalist; that is, they shared a common background.

Table 3 shows the Pearson's correlation coefficient (r) and Spearman's rank correlation coefficient ρ for the human subjects. The results show that every pair has a high correlation. Therefore, changing the human subject has little influence as regards creating references and evaluating system summaries. The evaluation by human subjects is stable. This result agrees with DUC's additional evaluation results (Harman and Over, 2004). However, the behavior of the correlations between humans with different backgrounds is uncertain. The correlation might be fragile if we introduce a human subject whose background is different from the others.

4.3 Compared Automatic Evaluation Methods

We compared our method with ROUGE-N and ROUGE-L described below. We used only content words to calculate the ROUGE scores because the correlation coefficient decreased if we did not remove functional words.

WSK-based method

We use WSK instead of ESK in equation (6)-(8).

ROUGE-N

ROUGE-N is an N-gram-based evaluation measure defined as follows (Lin, 2004b):

$$\text{ROUGE-N}(\mathcal{C},\mathcal{R})=\frac{\sum_{\mathcal{S}\in\mathcal{R}}\sum_{\text{gram}_\text{N}\in\mathcal{S}}\text{Count}_{\text{match}}(\text{gram}_\text{N})}{\sum_{\mathcal{S}\in\mathcal{R}}\sum_{\text{gram}_\text{N}\in\mathcal{S}}\text{Count}(\text{gram}_\text{N})} \quad (10)$$

Here, $\text{Count}(\text{gram}_\text{N})$ is the number of an N-gram and $\text{Count}_{\text{match}}(\text{gram}_\text{N})$ denotes the number of n-gram co-occurrences in a system output and the reference.

ROUGE-S

ROUGE-S is an extension of ROUGE-2 defined as follows (Lin, 2004b):

$$\text{ROUGE-S}(\mathcal{C},\mathcal{R})=\frac{(1+\beta^2)\times R_{\text{skip2}}(\mathcal{C},\mathcal{R})\times P_{\text{skip2}}(\mathcal{C},\mathcal{R})}{R_{\text{skip2}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{skip2}}(\mathcal{C},\mathcal{R})} \quad (11)$$

Where R_{skip2} and P_{skip2} are defined as follows:

$$\text{R}_{\text{skip2}}(\mathcal{C},\mathcal{R})=\frac{\text{Skip2}(\mathcal{C},\mathcal{R})}{\text{\# of skip bigram}\in\mathcal{R}} \quad (12)$$

$$\text{P}_{\text{skip2}}(\mathcal{C},\mathcal{R})=\frac{\text{Skip2}(\mathcal{C},\mathcal{R})}{\text{\# of skip bigram}\in\mathcal{C}} \quad (13)$$

Here, function Skip2 returns the number of skip-bi-grams that are common to $\mathcal{R}$ and $\mathcal{C}$.

ROUGE-SU

ROUGE-SU is an extension of ROUGE-S, which includes unigrams as a feature defined as follows (Lin, 2004b):

$$\text{ROUGE-SU}(\mathcal{C},\mathcal{R})=\frac{(1+\beta^2)\times R_{\text{su}}(\mathcal{C},\mathcal{R})\times P_{\text{su}}(\mathcal{C},\mathcal{R})}{R_{\text{su}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{su}}(\mathcal{C},\mathcal{R})} \quad (14)$$

Where R_{su} and P_{su} are defined as follows:

$$\text{R}_{\text{su}}(\mathcal{C},\mathcal{R})=\frac{\text{SU}(\mathcal{C},\mathcal{R})}{\text{(\# of skip bigrams + \# of unigrams)}\in\mathcal{R}} \quad (15)$$

$$\text{P}_{\text{su}}(\mathcal{C},\mathcal{R})=\frac{\text{SU}(\mathcal{C},\mathcal{R})}{\text{(\# of skip bigrams + \# of unigrams)}\in\mathcal{C}} \quad (16)$$

Here, function SU returns the number of skip-bigrams and unigrams that are common to $\mathcal{R}$ and $\mathcal{C}$.

ROUGE-L

ROUGE-L is an LCS-based evaluation measure defined as follows (Lin, 2004b):

$$\text{ROUGE-L}(\mathcal{C},\mathcal{R})=\frac{(1+\beta^2)\times R_{\text{lcs}}(\mathcal{C},\mathcal{R})\times P_{\text{lcs}}(\mathcal{C},\mathcal{R})}{R_{\text{lcs}}(\mathcal{C},\mathcal{R})+\beta^2 P_{\text{lcs}}(\mathcal{C},\mathcal{R})} \quad (17)$$

where R_{lcs} and P_{lcs} are defined as follows:

$$R_{\text{lcs}}(\mathcal{C},\mathcal{R})=\frac{1}{u}\sum_{r_i\in\mathcal{R}}\text{LCS}_\cup(r_i,\mathcal{C}) \quad (18)$$

$$P_{\text{lcs}}(\mathcal{C},\mathcal{R})=\frac{1}{v}\sum_{r_i\in\mathcal{R}}\text{LCS}_\cup(r_i,\mathcal{C}) \quad (19)$$

Here, $\text{LCS}_\cup(r_i,\mathcal{C})$ is the LCS score of the union longest common subsequence between reference sentences r_i and $\mathcal{C}$. u and v are the number of words contained in $\mathcal{R}$, and $\mathcal{C}$, respectively.

The multiple reference version of ROUGE-N S, SU or L, RN^{mean}, RS^{mean}, RSU^{mean}, RL^{mean} can be defined in accordance with equation (9).

4.4 Evaluation Measures

We evaluate automatic evaluation methods by using Pearson's correlation coefficient (r) and Spearman's rank correlation coefficient (ρ). Since we have ten systems, we make a vector $\mathbf{x}=(x_1,x_2,\cdots,x_i,\cdots,x_{10})$ from the results of an automatic evaluation. Here, $x_i=1/30\sum_{t=1}^{30} f(\mathcal{R}_t,\mathcal{C}_{i,t})$. $\mathcal{R}_t$ indicates a reference for the t-th topic. f indicates an automatic evaluation function such as F_{esk}, F_{wsk}, ROUGE-N, ROUGE-S, ROUGE-SU and ROUGE-L. Next, we make another vector $\mathbf{y}=(y_1,y_2,\cdots,y_i,\cdots,y_{10})$ from the human evaluation results. Here, $y_i=1/30\sum_{t=1}^{30} H(\mathcal{R}_t,\mathcal{C}_{i,t})$. Finally, we compute r and ρ between $\mathbf{x}$ and $\mathbf{y}$[6].

4.5 Evaluation Results and Discussions

Table 4 shows the evaluation results obtained by using Pearson's correlation coefficient r. Table 5 shows the evaluation results obtained with Spearman's rank correlation coefficient ρ. The ta-

[6]When using multiple references, functions f and H for making vectors $\mathbf{x}$ and $\mathbf{y}$ are substituted for f^{mean} and H^{mean}, respectively.

Table 4: Results obtained with Pearson's correlation coefficient."stop" indicates with stop word exclusion, "case" indicates w/o stop word exclusion.

	short								long							
	$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_{\rm avg}$		$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_{\rm avg}$	
	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case
ROUGE-1	**.965**	.884	.931	.888	.937	.879	.956	.906	.906	.876	.919	.916	.897	.891	.918	.948
ROUGE-2	.943	.960	.836	.880	.861	.906	.904	.937	.886	.930	.788	.941	.834	.616	.856	.929
ROUGE-3	.906	.936	.759	.814	.786	.846	.862	.900	.873	.909	.717	.849	.826	.431	.844	.885
ROUGE-4	.878	.914	.725	.752	.729	.794	.837	.871	.850	.890	.651	.787	.836	.292	.836	.865
ROUGE-L	.919	.777	.789	.683	.875	.867	.898	.852	.917	.840	.861	.812	.847	.829	.910	.848
ROUGE-S(∞)	.934	.914	.805	.888	.872	**.938**	.867	.917	.812	.863	.744	.954	.709	.547	.757	.900
ROUGE-S(9)	.929	.935	.783	.899	.808	.917	.856	.939	.840	.903	.735	.951	.730	.617	.787	.927
ROUGE-S(4)	.936	.943	.802	.891	.839	.917	.877	.940	.876	.920	.778	.945	.814	.663	.840	.932
ROUGE-SU(∞)	.934	.914	.805	.887	.872	.937	.867	.917	.811	.864	.743	.954	.707	.547	.756	.900
ROUGE-SU(9)	.926	.938	.765	.890	.789	.906	.845	.936	.829	.904	.705	.948	.701	.586	.766	.925
ROUGE-SU(4)	.930	.945	.772	.865	.810	.889	.861	.927	.868	.921	.730	.928	.785	.620	.818	.925
$F_{\rm esk}^{d=2}(\beta{=}2)$	.942		.927		.921		.957		**.941**		.957		**.967**		**.969**	
$F_{\rm esk}^{d=2}(\beta{=}3)$	.929		**.943**		.928		**.965**		.939		**.962**		.959		.967	
$F_{\rm esk}^{d=3}(\beta{=}2)$	.939		.923		.919		.962		.926		.954		.953		.966	
$F_{\rm esk}^{d=3}(\beta{=}3)$	.927		.933		.920		.964		.920		.947		.904		.949	
$F_{\rm esk}^{d=4}(\beta{=}2)$	.921		.900		.897		.955		.900		.932		.890		.946	
$F_{\rm esk}^{d=4}(\beta{=}3)$	.909		.900		.888		.950		.892		.921		.819		.922	
$F_{\rm wsk}^{d=2}(\beta{=}2)$	.939		.900		.897		.942		.931		.923		.936		.939	
$F_{\rm wsk}^{d=2}(\beta{=}3)$	.928		.921		.909		.958		.932		.939		.950		.950	
$F_{\rm wsk}^{d=3}(\beta{=}2)$	.938		.902		.886		.947		.924		.921		.934		.944	
$F_{\rm wsk}^{d=3}(\beta{=}3)$	.928		.922		.895		.960		.920		.929		.919		.942	
$F_{\rm wsk}^{d=4}(\beta{=}2)$	.929		.896		.873		.947		.910		.913		.908		.938	
$F_{\rm wsk}^{d=4}(\beta{=}3)$	.918		.915		.879		.956		.903		.913		.865		.925	

bles show results obtained with and without stop word exclusion for the entire ROUGE family. For ROUGE-S and ROUGE-SU, we use three variations following (Lin, 2004b): the maximum skip distances are 4, 9 and infinity [7]. In addition, we examine $\beta = 2$and3 for the ESK-based and WSK-based methods. The decay parameter λ for $F_{\rm esk}$ and $F_{\rm wsk}$ is set at 0.5. We will discuss these parameter values in Section 4.6.

From the tables, ROUGE-N's r and ρ decrease monotonically with N when we exclude stop words. In most cases, the performance is improved by including stop words for N ($\geq$2). There is a large difference between ROUGE-1 and ROUGE-4. The ROUGE-S family is comparable to the ROUGE-SU family and their performance is close to ROUGE-1 without stop words and ROUGE-2 with stop words. ROUGE-L is better than both ROUGE-3 and ROUGE-4 but worse than ROUGE-1 or ROUGE-2.

On the other hand, $F_{\rm esk}$'s correlation coefficients (r) do not change very much with respect to d. Even if d is set at 4, we can obtain good correlations. The behavior of rank correlation coefficients (ρ) is similar to the above. The difference between the ROUGE family and our method is particularly large for long summaries. By setting d=2, our method gives the good results. The optimal β is varied in the data sets. However, the difference between β= 2 and β=3 is small.

For ρ, our method outperforms the ROUGE family except for $\mathcal{D}_1$. By contrast, we can see d=3 or d=4 provided the best results. The differences between our method and the ROUGE family are larger than for r.

For both r and ρ, when multiple references are available, our method outperforms the ROUGE family.

Although ROUGE-1 sometimes provides better results than our method for short summaries, it has a critical problem; ROUGE-1 disregards word sequences making it easy to cheat. For instance, we can easily obtain a high ROUGE-1 score by using a sequence of high Inverse Document Frequency (IDF) words. Such a summary is incomprehensible and meaningless but we obtain a good ROUGE-1 score comparable to those of the top TSC-3 systems. By contrast, it is difficult to cheat other members of the ROUGE family or our method.

Our evaluation results imply that $F_{\rm esk}$ is robust

[7] We use β=1,2, and 3. However there are little difference among correlation coefficient regardless of β because the number of the words in reference and the number of the words in system output are almost the same.

Table 5: Results obtained with Spearman's correlation coefficient. "stop" indicates with stop word exclusion, "case" indicates w/o stop word exclusion.

	short								long							
	$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_{\text{avg}}$		$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_{\text{avg}}$	
	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case	stop	case
ROUGE-1	**.988**	.964	.842	.891	.842	.855	.927	.903	.818	.830	.903	.806	.867	.855	.842	**.915**
ROUGE-2	.927	.976	.770	.794	.855	.842	.879	.903	.721	**.891**	.721	.855	.794	.648	.818	.903
ROUGE-3	.879	.927	.588	.697	.818	.818	.867	.927	.758	.842	.636	.745	.806	.564	.709	.855
ROUGE-4	.818	.879	.721	.697	.745	.745	.867	.867	.685	.794	.564	.612	.830	.455	.709	.758
ROUGE-L	.927	.830	.661	.600	.806	.818	.879	.806	.842	.770	.576	.612	.636	.709	.879	.697
ROUGE-S(∞)	.939	.939	.673	.818	.794	.818	.818	.927	.770	.879	.636	.818	.697	.527	.709	.867
ROUGE-S(9)	.879	.952	.600	.745	.721	.794	.733	.939	.758	.806	.576	.806	.673	.564	.745	.855
ROUGE-S(4)	.891	.964	.600	.794	.794	.794	.794	.939	.709	.842	.576	.770	.770	.733	.758	.842
ROUGE-SU(∞)	.939	.939	.673	.818	.794	.818	.818	.927	.770	.879	.636	.818	.697	.553	.709	.867
ROUGE-SU(9)	.879	.964	.600	.745	.721	.794	.745	.939	.745	.806	.576	.758	.612	.564	.745	.903
ROUGE-SU(4)	.879	**.988**	.600	.745	.721	.770	.794	.903	.758	.855	.576	.794	.709	.612	.794	.842
$F_{\text{esk}}^{d=2}(\beta{=}2)$	.952		.879		.855		.939		.842		.927		**.903**		.903	
$F_{\text{esk}}^{d=3}(\beta{=}3)$	.952		**.915**		.891		.939		.855		.903		**.903**		.903	
$F_{\text{esk}}^{d=3}(\beta{=}2)$	.964		.867		.867		.976		.818		**.927**		.879		.879	
$F_{\text{esk}}^{d=3}(\beta{=}3)$	.964		.891		**.915**		.976		.758		.903		.709		.891	
$F_{\text{esk}}^{d=4}(\beta{=}2)$	.927		.830		.867		.952		.661		.903		.733		**.915**	
$F_{\text{esk}}^{d=4}(\beta{=}3)$	.927		.842		.842		**.988**		.588		.903		.673		.891	
$F_{\text{wsk}}^{d=2}(\beta{=}2)$	.976		.794		.830		.952		.818		.867		.806		.891	
$F_{\text{wsk}}^{d=2}(\beta{=}3)$	.952		.842		.830		.952		.818		.867		.794		.903	
$F_{\text{wsk}}^{d=3}(\beta{=}2)$	.976		.794		.818		.939		.806		.855		.733		.879	
$F_{\text{wsk}}^{d=3}(\beta{=}3)$	.976		.879		.855		.952		.806		.818		.794		**.915**	
$F_{\text{wsk}}^{d=4}(\beta{=}2)$	.964		.794		.818		.939		.806		.855		.697		**.915**	
$F_{\text{wsk}}^{d=4}(\beta{=}3)$	.964		.867		.855		.976		.745		.855		.770		**.915**	

Table 6: Best scores for each data set.

Pearson's Correlation Coefficient

Length	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_{\text{avg}}$
short	.945	.946	.933	.967
(d, λ, β)	(2,0.7,2)	(2,0.7,4)	(2,0.1,3)	(2,0.7,3)
long	.941	.962	.971	.972
(d, λ, β)	(2,0.6,2)	(2,0.6,3)	(2,0.7,2)	(2,0.8,2)

Spearman's Rank Correlation Coefficient

Length	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_{\text{avg}}$
short	.964	.915	.915	.988
(d, λ, β)	(3,0.9,4)	(2,0.3,4)	(3,0.5,3)	(4,0.7,4)
long	.855	.927	.915	.939
(d, λ, β)	(2,0.8,4)	(3,0.5,2)	(2,0.5,4)	(2,0.8,3)

for d and length of summary and correlates closely with human evaluation results. Moreover, it includes no trivial way of obtaining a good score. These are significant advantages over ROUGE family. In addition, our method outperformed the WSK-based method in most cases. This result confirms the effectiveness of semantic information and the significant advantage of the ESK.

4.6 Effects of Parameters

Our method has three parameters, d, λ, and β. In this section, we discuss the effects of these parameters. Figure 1 shows r and ρ for various λ and β values with respect to $\mathcal{D}_{\text{avg}}$. Note that we set d at 2 in the figure because the tendency is similar when we use other values, namely $d(=3 \text{ or } 4)$. From Fig. 1, we can see that $\beta{=}1$ is not good. With automatic summarization, 'precision' is not necessarily a good evaluation measure because highly redundant summaries may obtain a very high precision. On the other hand, 'recall' is not good when a system's output is redundant. Therefore, equal treatment of 'precision' and 'recall' does not give a good evaluation measure. The figure shows that $\beta{=}2, 3$ and 5 are good for r and $\beta{=}3, 4, 5$ and infinity are good for ρ.

Moreover, we can see a significant differences between $\lambda = 1$ and others from the figure. This implies an advantage of our method compared to ROUGE-S and ROUGE-SU, which cannot handle decay factor for skip-n-grams.

From Fig. 1, we can see that ρ is more sensitive to β than r. Here, $\beta{=}3, 4, 5$ and infinity obtained the best results. $\beta{=}1$ was again the worst. This result indicates that we have to determine the parameter value properly for different tasks. λ does not greatly affect the correlation for $d{=}3, 4, 5$ and infinity as regards the middle range.

Table 6 show the best results when we examined all parameter combinations. In the brackets, we show the best settings of these parameter combinations. For r, $d{=}\ 2$ provides the best result and middle range λ and $\beta{=}2$ or 3 are good in most cases. On the other hand, the best settings for ρ vary with

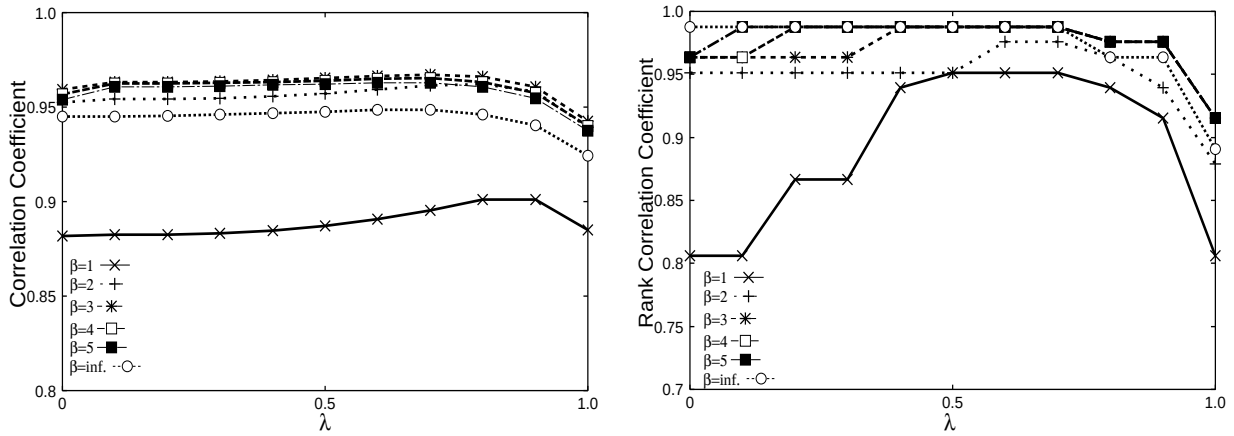

Figure 1: Correlation coefficients for various values of β and λ on $\mathcal{D}_{\mathrm{avg}}$.

the data set. d=2 is not always good for ρ.

In short, we can see that the decay parameter for skips is significant and long skip-n-grams are effective especially ρ.

These results show that our method has an advantage over the ROUGE family. In addition, our method is robust and sufficiently good even if close attention is not paid to the parameters.

5 Conclusion

In this paper, we described an automatic evaluation method based on the ESK, which is a method for measuring the similarities between texts based on sequences of words and word senses. Our experiments showed that our method is comparable to ROUGE family for short summaries and outperforms it for long summaries. In order to prove that our method is language independent, we will conduct an experimental evaluation by using DUC's evaluation data. We believe that our method will also be useful for other natural language generation tasks. We are now planning to apply our method to an evaluation of machine translation.

References

N. Cancedda, E. Gaussier, C. Goutte, and J-M. Renders. 2003. Word Sequence Kernels. *Journal of Machine Learning Research*, 3(Feb):1059–1082.

M. Collins and N. Duffy. 2001. Convolution Kernels for Natural Language. In *Proc. of Neural Information Processing Systems (NIPS'2001)*.

D. Harman and P. Over. 2004. The Effects of Human Variation in DUC Summarization Evaluation. In *Proc. of Workshop on Text Summarization Branches Out*, pages 10–17.

T. Hirao, T. Fukusima, M. Okumura, C. Nobata, and H. Nanba. 2004a. Corpus and Evaluation Measures for Multiple Document Summarization with Multiple Sources. In *Proc. of the COLING*, pages 535–541.

T. Hirao, J. Suzuki, H. Isozaki, and E. Maeda. 2004b. Dependency-based Sentence Alignment for Multiple Document Summarization. In *Proc. of the COLING*, pages 446–452.

C. Hori, T. Hori, and S. Furui. 2003. Evaluation Methods for Automatic Speech Summarization. In *Proc. of the Eurospeech2003*, pages 2825–2828.

S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. 1997. *Goi-Taikei – A Japanese Lexicon (in Japanese)*. Iwanami Shoten.

C-Y. Lin and E. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proc. of the NAACL/HLT*, pages 150–157.

C-Y. Lin and F.J. Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proc. of the ACL*, pages 606–613.

C-Y. Lin. 2004a. Looking for a Good Metrics: ROUGE and its Evaluation. In *Proc. of the NTCIR Workshops*, pages 1–8.

C-Y. Lin. 2004b. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proc. of Workshop on Text Summarization Branches Out*, pages 74–81.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. 2002. Text Classification using String Kernel. *Journal of Machine Learning Research*, 2(Feb):419–444.

K. Papineni, S. Roukos, T. Ward, and Zhu W-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the ACL*, pages 311–318.

R. Soricut and E. Brill. 2004. A Unified Framework for Automatic Evaluation using N-gram Co-occurrence Statistics. In *Proc. of the ACL*, pages 614–621.