Modified Key-Sentence Extraction by RICOH at NTCIR-2 TSC

Masayuki KAMEDA
Software Research Center, RICOH COMPANY, LTD
1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, Japan
kameda@src.ricoh.co.jp

Abstract

We participated at NTCIR-2 in the TSC subtask A-1 and A-2 using the output of QJR/KSE, a function of key-sentence extraction. Through examining the evaluation results and the human-extracted sentence data of the dryrun subtask A-1, we tried to make an experimental hybrid system incorporating the lead method into the original. In the formal run, the evaluation results of the hybrid system in the subtask A-1 and A-2 were in the top class among the twelve systems, much better than the results obtained in the dryrun. Especially in the F-measures of the 30% summarization rate, our submitted two systems got the top and second rankings. As a result of our participating in TSC, we have examined some internal scores to study the effectiveness of their results and had a chance to make some improvements and tuned up the hybrid system to increase its performance.
Keywords: *NTCIR, TSC, summarization, important sentence extraction, QJR/KSE.*

1 Introduction

We proposed and developed a Quick Japanese text Reading support system, QJR [2], with functions for supporting the screening, skimming, skip-reading, and analytic-reading of Japanese text by using our portable and Quick Japanese Parser, QJP[1] [1]. QJR has a key-sentence-extraction function for skip-reading support, QJR/KSE [3,4].

The work of TSC, Text Summarization Challenge [5], at NTCIR-2 started in 2000. We decided to participate in subtask A, composed of A-1 for extracting important sentences and A-2 for producing summaries, using QJR/KSE. First, we submitted the results of the original QJR/KSE for the dryrun and then revised, based on those results, the original QJR/KSE and made an experimental hybrid system incorporating the lead method into the QJR/KSE and submitted their results for the formal run.

In this paper, we first introduce the original QJR/KSE. Then, we describe the participating systems, including our hybrid system, the evaluation results, the examination, and some other experiments using human-extracted important sentences for subtask A-1 of the dryrun and the formal run. Lastly, some considerations are given regarding the internal scores of the original QJR/KSE, the hybrid system and the characteristics of documents of the dryrun and formal run.

[1] QJP is a morphological and syntactic-kakariuke analyzer for Japanese sentence, which requires only 50KB of the dictionary and only 250KB of memory.

2 QJR/KSE

QJR/KSE assigns a key-sentence level to every sentence in a text according to not only the ranking-value of the sentence but also the ranking-value of the paragraph; based on preference-ranking sentences within a preference-ranked paragraph. The level is intended to be used to highlight key-sentences of the highest level 0 (and additionally the second highest level 1) on a text browser for supporting skip-reading a text rather than to extract sentences for summary.

The scores for ranking sentences/paragraphs are calculated by the relevance degree between two sentences/paragraphs, determined by counting the component words common to two groups of keyword candidates extracted by QJP. Two types, a referring type and a referred-to type, of relevance degrees are used. The referring type and the referred-to type of relevance degree of a sentence/paragraph to the other is the degree of common words against words in the self and the other, respectively. Each sentence/paragraph is internally given the average of the referring type of relevance degrees of the self to the all other sentences/paragraphs, r1, the average of the referred-to type of relevance degrees of the self to the all others, r2, and the average of r1 and r2, the basic score, r3. We think that r1 and r2 are suitable for primitive indexes for important sentence/paragraph because of the following reasons;

- r1 is likely to be high in case of a short sentence(/paragraph) including one or a few important word(s) common to some other sentences(/paragraphs).
- r2 is likely to be high in case of a rather long sentence/paragraph including many important words, one or a few of which is common to rather short sentences/paragraphs.

Then, relevance degrees, r1, r2, and r3 are respectively used as the importance score of the (sub)title sentence, the paragraph, and the sentence. Some points are added to the basic score in case of a sentence, including clue words or phrases.

Based on the ranking of sentences and paragraphs using these importance scores, each sentence is given a key-sentence level, r4. Actually, the ranking-value of sentence is shifted down by the ranking-value of the paragraph including the sentence to become the key-sentence level, we called it paragraph-shift method; the n-th ranked sentence within the N-th ranked paragraph is given the value of (n-1)+(N-1) as the key-sentence level. Additionally a sentence is leveled up in the cases where it is a (sub)title or is highly relevant to the (sub)title sentence at a high level or it has a high basic score. However, a sentence is leveled down in the case of having a quite low basic score.

Though the primary index of QJR/KSE is this key-sentence level, QJR/KSE re-ranks sentences by the key-sentence level as the first ascending sort-key and the basic score as the second descending sort-key. The re-ranked value of a sentence, R5, is used to extract important sentences in the specified sentence number, like the subtask A-1.

Furthermore, QJR/KSE has a sentence-shortening function which eliminates verbose words or phrases in a sentence. This function is to be used in the subtask A-2.

3 Dryrun

3.1 Basic system

For our participation in subtasks A-1 and A-2, we added a pre-processing module that removes TSC-tags of the input TSC text to output plain text as the input of QJR/KSE. We also added a post-processing module that extracts sentences by referring to the QJR/KSE experimental output including some internal scores, r1 to r4, with their ranked values and the re-ranked value, R5.

The additional processes are the noted below.

(1) Subtask A-1 for extracting important sentences

During the dryrun, the post-module extracts sentences based on R5 according to the specified number of a sentence. The subtask A-1 excludes a title sentence to be extracted, then the post-module ignores a title.

(2) Subtask A-2 for producing summaries

Although the aim of A-2 is to produce summaries, we submitted almost the same results as A-1. In the case where a title is extracted, a new line has to be output after the title.

If more sentences are to be extracted using the sentence shortening function, a set of the shortened sentences is selected.

3.2 Evaluation results

We submitted the output of the above system to the dryrun, and got the following evaluation results. All of the evaluations were much worse of the participated systems .

(1) Subtask A-1

The evaluation results of subtask A-1 are given as the F-measure[2] calculated by (2*P*R/(P+R)) using the precision, P, and the recall, R, of the system-extracted sentences against the human-extracted sentences.

The average of the F-measures, FULL, of the 10%, 30%, and 50% summarization rates, whose real rates are 15%, 46% and 75% respectively, was 0.472, which was the second worse of the twelve systems, including the two base-line systems, using the lead method, Lead, and the term-frequency method, TF [Table 1 (including the results of other three revised systems described later)].

＼ Sum. Rate System	10% [15%]	30% [46%]	50% [75%]	FULL
(Top data)	*.428*	*.560*	*.779*	*.554*
Lead2	.456	.513	.754	.574
Lead(top sys)	①.428	⑧.488	⑧.747	①.554
2nd system	⑦.324	①.560	④.759	②.548
QJR/KSE2	.322	.527	.768	.539
TF	⑧.318	⑤.505	⑦.751	⑧.525
QJR/KSE1	.243	.492	.741	.492
Our system	⑩.207	⑩.474	⑨.734	⑩.472

Table 1 [3]: Evaluation results of Dryrun A-1

In the dryrun A-1, the lead system shows the top performance. Main reason was thought that 30 documents of the dryrun (and also the formal run) were news articles selected from Mainichi newspaper [6]. The most important content of news report articles are to be written in a few leading paragraphs.

Our dryrun system had a problem in that the sentence segmentation was different from the segmentation done by TSC-sentence tags. The problem-fixed system, QJR/KSE1, improved the F-measure 0.472 to 0.492, though it did not raise the ranking of the system.

[2] In the subtask A-1, P, R, and the F-measure are actually the same, so only the F-measures are shown in this paper.

[3] The number in a circle "○ " is the ranking in the evaluation.

(2) Subtask A-2

Two types of evaluations were done: content-based and subjective type.

- **Content-based evaluation**

  The cosine distance between the content-word vectors of a human summary based on selected important parts, and a system summary. The average distance of the 20% and 40% rates of ours was 0.557, which was ranked 8th for the 11 systems.

- **Subjective evaluation**

  The measure is a rank-value of four kinds of summaries: a human summary based on important parts, a human freely-written summary, a summary using the lead method, and a system summary. Two kinds of averages of the rank-value of the 20% and 40% summarization rates were 3.32 for readability and 3.55 for content, which were ranked in the7th and 8th among 10 systems.

4 Improvement and examination using the dryrun human data

The human-extracted data of the 15%, 46%, and 75% real rates was provided to participants after the evaluation. Dr. Sekine of New York University provided the additional human-extracted data of the 10%, 30%, and 50% rates used in the task.

We examined and experimented with extractions using these human-extracted data.

4.1 Small improvement

Reviewing the human-data, we noticed the following related to (sub)titles which tended to be extracted because the level was rather high.

- A line composed of some symbols, such as "◇", or a sentence surrounded by parentheses was mis-judged as a (sub)title to be extracted.

- Most of (sub)titles were extracted by QJR/KSE though they were not extracted in the human data.

We added the processing to avoid having extracting (sub)titles in the pre- and post-modules. The revised system, QJR/KSE2, improved the F-measure 0.492 of QJR/KSE1 to 0.539. The F-measures of the 30% [46%] and 50% [75%] became rather higher, but the F-measure of the 10% [15%] was rather lower [Table1].

Adding the same processing to the lead method, the revised lead, Lead2, also improved the F-measure.

4.2 Performance of the internal scores

The dryrun system used the re-ranked value, R5, to extract important sentences. As previously described, QJR/KSE has some other internal scores for every sentences such as,

- r0: total count of the common words
- r1: average of the referring type of relevance degrees
- r2: average of the referred-to type of relevance degrees
- r3: basic score as the average of r1 and r2
- r4: key-sentence level

We can re-rank the sentences based on each of these scores instead of R5. So, we tried to extract sentences using the ranking-values R0, R1, R2, R3, and R4 based on r0, r1, r2, r3 and r4, respectively, to evaluate the F-measures by comparing them to the human data. The evaluated F-measures shown in Table 2 are for the dryrun human data and those in Table 3 are for the additional dryrun human data provided by Dr. Sekine.

＼ Sum. Rate System	10% [15%]	30% [46%]	50% [75%]	FULL
Lead2	*.456*	*.513*	*.754*	*.574*
Lead	*.428*	*.488*	*.747*	*.554*
R5	.322	.527	.768	.539
R4 (r4)	.313	.533	.765	.537
R3 (r3)	.317	.518	.756	.530
R2 (r2)	.296	.531	.763	.530
R1 (r1)	.264	.492	.761	.506
R0 (r0)	.230	.491	.734	.495

Table 2: Evaluation results of extraction by internal scores compared to dryrun human data.

Sys ＼ Rate	10%	30%	50%	FULL
Lead2	*.442*	*.513*	*.541*	*.476*
Lead	*.397*	*.488*	*.517*	*.452*
R5	.298	.474	.568	.447
R4 (r4)	.315	.497	.575	.462
R3 (r3)	.336	.471	.557	.455
R2 (r2)	.291	.466	.573	.444
R1 (r1)	.204	.389	.532	.375
R0 (r0)	.176	.377	.534	.362

Table 3: Evaluation results of extraction by internal scores compared to the additional dryrun human data provided by Dr. Sekine.

The F-measures by R0 are much worse than the ones by R2, R3, R4, and R5. Considering that r0 of a sentence is total count of the words common to the

other sentences, its performance should be similar to that of the TF method, which was also rather poor.

Of the two types of relevance degrees, the referred-to type's average, r2, is better than the referring type's average, r1. It is also superior to the lead method at a more than 30% summarization rate .

At the less than 15% rates, the basic score, r3, is superior to both r1 and r2, but at the more than 30% rate, r3 was below r2.

The key-sentence level, r4, is below r3 at the less than 15% rates, but is over r3 at all other rates and for the average.

In the dryrun human data the re-ranked value, R5 , is over those by the level, r4, but in the dryrun human data provided by Dr. Sekine it is vice versa.

In a comparison to the lead method, the F-measure for each of r2, r3, r4, and R5 is better than it at the more than 30% rates, but not at the average.

4.3 The hybrid system with the lead method

As discussed in 4.2, the QJR/KSE's internal scores could not exceed the lead method at the less than 15% rates. Based on this, we tried to incorporate the lead method into QJR/KSE for the small summarization rates to make a hybrid system.

To incorporate the lead method, we introduced a framework for preferring sentences located in some leading range in a text against the QJR/KSE ranking. Two primitive conditions were used;

- X(p): sentences included within the first p paragraphs.
- Y(n,m) : sentences from the first sentence to the (N/n+m)-th, where N is the number of total sentences in the text.

An additional condition below is used to exclude less important sentences of the lead sentences;

- Z(k): sentences whose ranked value of R5 is less than N*k where $0 \leqq k \leqq 1$.

The two patterns below were tried as a combination of one of the ranks, R1, R2, R3, R4, and R5, by QJR/KSE internal scores and the lead method ;

- A(Rn; p,n,m,k) : QJR/KSE(Rn) and {{X(p) or Y(n,m)} and Z(k)}.
- B(Rn; p,n,m,k) : QJR/KSE(Rn) and {{X(p) and Y(n,m)} and Z(k)}.

Through a lot of trial and error, we found much better parameter sets for the dryrun human data and Dr. Sekine's data.

For the original dryrun data,

- A1: A(5; 2,5.5,1,0.9)
- B1: B(5; 2,3.0,1,1.0).

For the additional dryrun data by Dr.Sekine,

- A2: A(5; 2,4,1,0.8)
- B2: B(5; 2,4,3,0.8).

The evaluation results of these combinations for the hybrid system are shown in Table 4 and Table 5 (including A3 and B3 that will be described later).

The F-measures obtained by each of the four hybrid systems, A1, B1, A2, and B2 were better than those obtained by R5 and the other Rn. And they were also better than those by the lead method at almost rates partially including the smaller 10% and 15% rates [Table 4].

The best performance was obtained by A1 for the dryrun data, and by B2 for the additional dryrun data provided by Dr. Sekine [Table 5].

＼ Sum. Rate System	10% [15%]	30% [46%]	50% [75%]	FULL
A1	.461	.569	.778	.603
B1	.457	.559	.775	.597
A2	.446	.549	.769	.588
B2	.447	.562	.769	.593
A3	.346	.529	.771	.549
B3	.372	.540	.767	.559
Lead2	*.456*	*.513*	*.754*	*.574*
Lead	*.428*	*.488*	*.747*	*.554*
R5	*.322*	*.527*	*.768*	*.539*
R4 (r4)	*.313*	*.533*	*.765*	*.537*
R2 (r2)	*.296*	*.531*	*.763*	*.530*

Table 4: Evaluation results of extraction by the hybrid system compared to the dryrun human data

Sys ＼ Rate	10%	30%	50%	FULL
A1	.425	.513	.606	.515
B1	.436	.515	.594	.515
A2	.442	.515	.584	.513
B2	.443	.542	.599	.528
A3	.352	.491	.570	.471
B3	.423	.498	.582	.501
Lead2	*.442*	*.446*	*.541*	*.476*
Lead	*.397*	*.441*	*.517*	*.452*
R5	*.298*	*.474*	*.568*	*.447*
R4 (r4)	*.315*	*.497*	*.575*	*.462*
R2 (r2)	*.291*	*.466*	*.573*	*.444*

Table 5: Evaluation results of extraction by the hybrid system compared to dryrun human data provided by Dr. Sekine

5 Formal run

5.1 Formal run system

We participated in the formal run using the B2 hybrid system because it had the best performance at the 10%, 30%, and 50% rates used in the formal run. At the formal run's subtask A-1, two systems results could be submitted, thus, we submitted the result obtained from the QJR/KSE2, R5, as well.

5.2 The evaluation results

(1) Subtask A-1

The average of F-measures of the B2 hybrid system and QJR/KSE2 at the 10%, 30%, and 50% summarization rates, were 0.454 and 0.434, whose rankings were 3rd and 6th among the 12 systems, including the two base-line system [Table 6]. For the lead method, the hybrid system was superior and QJR/KSE2 was in the same rank as the lead. At the 30% rate, QJR/KSE2 got the top F-measure and the hybrid system got the second.

Sys ＼ Rate	10%	30%	50%	FULL
(Top data)	*.363*	*.483*	*.612*	*.467*
Top system	③.337	③.451	①.612	①.467
2nd system	①.363	⑤.435	②.589	②.462
Our system A	④.305	②.473	⑤.585	③.454
Lead	⑥.284	⑦.432	④.586	⑥.434
Our system B	⑪.241	①.483	⑤.578	⑥.434
TF	⑧.276	⑫.367	⑫.530	⑫.391

Table 6: Evaluation results of Formal run A-1

(2) Subtask A-2

- **Content-based evaluation**

  The summary by the hybrid system was compared with two kinds of human summaries.

  The average distances of the 20% and 40% rates were 0.553 for the human freely-written summary and 0.567 for the human summary based on important parts, both of which were ranked 2nd among the 11 systems, including the two base-line systems.

- **Subjective evaluation**

  The evaluation measure is a rank-value in four kinds of summaries: two human summaries, a summary using the lead method and a system summary. The two kinds of averages of the rank-value of the 20% and 40% were 2.65 for readability and 3.10 for content, respectively, which were ranked 2nd or 3rd among the 10 systems. As the dryrun results were 3.32 and 3.55, the readability was much improved.

6 Examination using the formal run's human data

We re-examined the performance of six kinds of QJR/KSE's internal scores, four kinds of the hybrid systems and two lead methods[4] comparing them to the human-extracted sentence data of the formal run.

And we tuned up the two types, A and B, of the hybrid systems to this human data to get following each best set of parameters.

- A3 : A(5; 2,4,1,0.45)
- B3 : B(2; 1,4,1,0.65)

Table 7 shows the F-measures by all of the above.

Sys ＼ Rate	10%	30%	50%	FULL
(top sys)	*.337*	*.451*	*.612*	*.467*
A1	.283	.423	.564	.423
B1	.283	.452	.581	.439
A2	.311	.431	.576	.439
B2(our sys A)	.305	.473	.585	.454
A3	.325	.508	.578	.470
B3	.378	.491	.609	.493
Lead (admin)	*.287*	*.432*	*.586*	*.434*
Lead2	*.283*	*.377*	*.542*	*.401*
Lead	*.276*	*.367*	*.530*	*.391*
R5(our sys B)	.241	.483	.578	.434
R4 (r4)	.249	.511	.579	.447
R3 (r3)	.251	.450	.597	.432
R2 (r2)	.270	.467	.603	.446
R1 (r1)	.109	.423	.586	.372
R0 (r0)	.141	.404	.577	.374

Table 7: Evaluation results of extraction compared to formal run human data

(1) QJR/KSE internal scores

Of six Rn systems, the following were confirmed;

- The F-measures by r0 are much worse than the ones by the others.
- The F-measures by average of referred-to type of relevance degrees, r2, are the best at the 10 % and 50% summarization rates.
- The F-measure by the key-sentence level, r4, is the best at the 30% rate.
- The average F-measures obtained by r2 and r4 exceed those by Lead(admin) and those by R5, our formal run submitted system (sys B).

[4] The F-mesures by simulated lead method, Lead, didn't agree with the one of the lead by the TSC admin which is shown as Lead(admin) in Table 7.

The total count of the words common to the other sentences, r0, is not good for extracting important sentences same as the TF method as described in 4.2. QJR/KSE does not use r0 for extracting key-sentences.

As for the average of referring/referred-to type of relevance degrees, r1 and r2, we treated both of r1 and r2 equally as importance indexes of sentences to make the basic score of the average of the two, r3. Here we have noticed that r2 is much better index for important sentence than r1 through examination using the formal run data and also the dryrun data. If the basic score is given in a form of linear combination, the weight of r2 had better be larger than the weight of r1.

We also confirmed the effectiveness of the key-sentence level, r4, which is the QJR/KSE's primary index. The level is determined by not only sentence's importance but also paragraph's one as described in 2. Introducing paragraph's importance was effective for determining the key-sentence level of sentence.

The F-measure by r4 at 30% was quite higher than the one by R5 which corresponds to the top rank in the formal run subtask A-1. This is thought to come indirectly from r2. Anyway we can say that QJR/KSE is quite good at the 30% rate extraction. But it also poor at 10% rate.

(2) The hybrid systems

The following were confirmed;

- Of the four systems, A1, B1, A2, and B2, tuned by the dryrun data, the B2 system, our formal run submitted system (sys A), have the best performance.

- These systems' performances are worse than A3 and B3 tuned by the formal run data.

- The F-measures by the B3 system are much better. The average F-measure 0.493 is much larger than the 0.467 of the top system in the formal run. The F-measures at 30% and 50% rates by A3 are the best in the formal run.

The hybrid system, incorporating the lead method to compensate the weakness at 10%, improved the F-measures not only at 10% rate but also at other rates. As the best parameter set for the hybrid system tuned up to the dryrun data was some different from the best set for the formal run, and vice versa. If we adopt the hybrid system, we should consider to combine the lead method adaptively according to each of particular document sets.

Another consideration through the hybrid system is on the characteristics of the document set. That the B3's performance is better demonstrates the characteristics of the formal run data. We say that the characteristics of the formal run data are described by R2 and X(1), whereas the characteristics of the dryrun data are described by R5 and X(2), where p of X(p) is the number of preferred leading paragraphs. That is;

- Most of documents in the dryrun were news report articles where lead paragraphs are important, whereas half of documents in the formal run were editorial columns where lead paragraphs are not always important. That difference seemed to appear in X(2) and X(1).

- R2 of the formal run may show that many documents like editorial columns include such rather long and important sentences, because r2 is likely to be high in case of a rather long important sentence as described in 2.

7 Conclusion

We participated in subtask A-1 and A-2 of NTCIR-2 TSC using QJR/KSE. For the dryrun and the formal run human-extracted data, the original QJR/KSE performance was poor at the 10% summarization rate(s), but was much better than that of the other systems at the 30% rate. We confirmed that the average of referred-to type of relevance degree, one of the internal scores, and the key-sentence level , the primary index, were good indexes for determining a sentence's importance. We also obtained good performances at all rates by using the hybrid system, having incorporated the lead method into QJR/KSE, and by tuning up the parameters in accord with the human data used to achieve the high performance reported.

We plan to reconsider the internal scores, especially the average of referred-to type of relevance degree, and to combine them with the lead method adaptively according to each of particular document sets.

References

[1] Masayuki KAMEDA, A Portable & Quick Japanese Parser: QJP. COLING'96, pp. 616-621, 1996.

[2] Masayuki KAMEDA, Support functions for Reading Japanese Text (in Japanese). SIG Notes NL-110, pp. 57-64. Information Processing Society of Japan, 1995.

[3] Masayuki KAMEDA, Extraction of keywords and key-sentences by keyword-candidates correlation methods (in Japanese). Proc. of 2nd Annual Meeting, pp. 97-100, The Association for Natural Language Processing, 1995.

[4] Masayuki KAMEDA, Key-sentences Extraction based on paragraph-shift method (in Japanese). SIG Notes NL-121, pp. 119-126. Information Processing Society of Japan, 1995.

[5] Manabu Okumura and Takahiro Fukushima, Text Summarization Challenge. http://galaga.jaist.ac.jp:8000/tsc/, 2000.

[6] Mainichi shinbunsha, the Mainichi newspaper CD-ROMs 1994, 1995, 1998 versions (in Japanese). 1994,1995, 1998.