nagoy Team's Summarization System at the NTCIR-14 QA Lab-PoliInfo

Yasuhiro Ogawa, Michiaki Satou, Takahiro Komamizu, and Katsuhiko Toyama

Nagoya University, Japan yasuhiro@is.nagoya-u.ac.jp

Abstract. The *nagoy* team participated in the NTCIR-14 QA Lab-PoliInfo's summarization subtask. This paper describes our summarization system for assembly member speeches using random forest classifiers. Because we encountered an imbalance in the data, we were unable to achieve good results in this subtask when training on all data. To solve this problem, we developed a new summarization system that apples multiple random forest classifiers training on different-sized data sets step by step. As a result, our system achieved good performance, especially in the evaluation by ROUGE scores. In this paper, we also compare our system with a single random forest classifier using probability.

Team Name. *nagoy*

Subtasks. Summarization

Keywords: NTCIR-14 · summarization · random forest · progressive ensemble random forest

1 Introduction

The NTCIR-14's QA Lab-PoliInfo [4] (Question Answering Lab for Political Information) deals with political information and sets forth three tasks: segmentation, summarization, and classification. Our team participated in the summarization task. We previously developed a summarization system [5] for Japanese statutes, which are also political, that is based on random forest classifiers [1] and achieves better results than other summarization systems. Thus, we expected our system to perform equally well for assembly member speeches. However, we were confronted with a data imbalance problem between summarization for statutes and that for assembly member speeches. To overcome this problem, we introduced a new approach that applies multiple random forest classifiers training on different-sized data sets in a step-by-step manner.

This report describes our summarization system for assembly member speeches and discusses not only the official results, but additional comparison results as well.

2 System

Our summarization system consists of two modules: a sentence extraction module using random forest classifiers and a sentence reduction module. In Section 2.1, we explain how to construct training data for random forest classifiers and, in Section 2.2, show the features we used. We solve the training data imbalance by applying multiple classifiers, as described in Section 2.3. In Section 2.4, we provide the details of the sentence reduction module.

2.1 Training Data

In this task, a summary of an assembly member's speech is provided as a description of Togikai dayori[1]. This summary was not made by sentence extraction methods; that is, a sentence in the summary may not appear in the original speech.

Since our method is based on sentence extraction methods, we need training data that consists of positive and negative sentences, where the "positive" or "negative" sentence means that it is or is not used for making the summary, respectively. Thus, we determined which sentence is used for making a summary as follows.

When we are given a pair consisting of an assembly member's speech and its summary, we find the sentence in the speech that contains the most words in the summary. We consider this sentence to be positive and the others negative. Since this summarization task has a length limit, if the length of the positive sentence is shorter than the length limit, we choose the sentence with the second-most summary words. In order to make the training data more correct, we should consider redundancy; that is, we should account for the overlap of the first positive sentence and the second, but we simply chose the second without considering the degree of overlap.

In the formal run, 596 assembly member speeches consisting of 9,979 sentences[2] were given. Hereafter, we refer to these speeches as the "source documents." Using the above method, we assembled training data that included only 825 (8.3%) positive sentences. This differs considerably from the summarization of Japanese statutes. In our study of statute summarization [5], we used *outlines of Japanese statutes*, which are official summaries of statutes published by the Japanese government, as the gold standard. In this case, the ratio of positive data is over 70% [5]. Because of this difference, we cannot apply the statute summarization methods to assembly member speeches, so we developed another method, described in Section 2.3.

2.2 Random Forest Features

In order to train a random forest classifier, we used the following features: sentence position, sentence length, and presence of a word. Here, we chose words

[1] https://www.gikai.metro.tokyo.jp/newsletter/ (in Japanese)
[2] Two speeches have no sentences.

Table 1. The number of sentences extracted by each classifier

ID	×1	×2	×3	×4	×5	number of sentences
111	1	0	0	0	0	45
106	9	5	2	0	0	11
19	7	3	3	1	0	8
23	3	2	1	0	0	34
92	5	3	1	1	1	13

that are nouns, occur more than once in the summary, and are not within the top 20 of the number of occurrences in all the source documents.

For the formal run data, we used the presence of 992 words as features.

2.3 Progressive Ensemble Random Forest

Since the above training data includes only 8.3% positive data and is imbalanced, using all of the training data results in poor performance. In fact, when we trained a random forest classifier on all training data, the classifier chose no sentences for 135 of the 146 documents in the test data.

Thus, we used an undersampling technique to solve this problem; however, we questioned how much negative data we should use. We prepared the following five random forest classifiers trained on the same positive data with different sized negative data:

1. classifiers trained on training data consisting of same-sized positive and negative data,
2. classifiers trained on training data where the size of the negative data is two times the positive data,
3. classifiers trained on training data where the size of the negative data is three times the positive data,
4. classifiers trained on training data where the size of the negative data is four times the positive data,
5. classifiers trained on training data where the size of the negative data is five times the positive data.

Table 1 shows how many sentences each random forest classifier extracted from the source documents of the test data. In this table, "ID" indicates the identification number of the target documents and "$\times\ n$" indicates the result of n-th random forest classifier. ID 111 consists of 45 sentences; the first classifier extracted just one sentence, but other classifiers extracted no sentences. On the other hand, ID 106 consists of 11 sentences and the first classifier extracted 9 sentences, which is too many. In this case, the third classifier, which extracted two sentences, seems better. As can be seen from these results, the most suitable classifier varies from document to document.

How should the classifier be chosen? Our solution is to use all the classifiers step by step, which we call "progressive ensemble."

First, we use the fifth classifier. If that classifier does not extract any sentences, then we use the fourth classifier. If the fourth classifier also extracts no sentences, then we use the third one. We repeat this process until we obtain a sentence. Note that we use the next classifier if the length of extracted sentences is ten less than the limit, because we find that such extracted sentences are insufficient for summarization. As a result, the length of extracted sentences may exceed the limit.

In addition, the test data in the formal run consists of "single-topic" and "multi-topic" data. We assume that "multi-topic" data needs multiple sentences for summarization, so we choose at least two sentences for "multi-topic" data.

2.4 Sentence Reduction

Since extracted sentences are redundant and sometimes exceed the length limit, we need to reduce them. Our sentence reduction method is a typical one using a Japanese dependency analyzer. We analyze extracted sentences by CaboCha [6] and choose the important *bunsetsu* segments (hereafter "segment"). We calculate importance scores using segment features, such as dependency depth, case information, and frequency in all summaries, not used in traditional sentence reduction methods. If a segment contains a noun, its frequency in all summaries of the training data is used as a weight. The weights of other features are adjusted by hand.

When we reduce an extracted sentence, we first choose the last segment. Next we choose the segment with the highest importance score, where we also choose the other segments on the path between the segment with the highest importance score and the last segment to avoid creating ungrammatical sentences. We add the next segments unless the sentence length exceeds the limit.

Although this sentence reduction method always chooses the last sentence, it is sometimes redundant. Thus, we introduce a replacement process for preprocessing that simply replaces the end of the sentence, as described in Table 2.

3 Result of the Formal Run

Tables 3 and 4 show our official formal run results. Table 3 shows the human evaluation results and Table 4 shows the evaluation by ROUGE scores. Bolded scores indicate that we achieved the best results among the participants. As shown, our system achieved good performance, especially in the ROUGE scores evaluation. However, the formed score was less than other systems, which indicates that our reduction module created some unnatural sentences.

Figure 1 shows a successful example of our system. In this example, the bolded sentence is extracted and successfully reduced. Figure 2 shows an unsuccessful example of our system. In this example, our system successfully extracted the target sentence, but failed to reduce the sentence. Our reduction module

Table 2. String replacement as preprocessing

target string	replaced string
でございます。	です。
伺います。	。
ております。	ている。
でおります。	でいる。
てまいります。	ていく。
でまいります。	でいく。
であります。	です。
いたします。	する。
と思います。	。
と思っている。	。

Table 3. Quality question scores in the formal run (max is 2)

	all-topic				single-topic				multi-topic			
	content		formed	total	content		formed	total	content		formed	total
	X=0	X=2			X=0	X=2			X=0	X=2		
nagoy	**0.886**	1.104	1.619	0.899	**0.953**	**1.179**	1.642	**1.028**	**0.810**	1.016	1.592	0.750

deleted the object "め り張り", but left the verb "つけ", which made the reduced sentence unnatural. This is because our module tries to leave as many words as possible within the length limit. To solve this problem, we should delete the verb if we delete its object. In addition, we need to further adjust the weights of the features; for example, in this case we should delete "新年度予算編成作業を進めるべきと考えますが" and leave "め り張りをつけ".

4 Discussion

In the traditional sentence extraction summarization method, all sentences in the source documents are scored and are chosen in the order of their scores until the given length limit is reached. This technique often considers redundancy such as Maximal Marginal Relevance [2].

Although classifiers such as random forest or support vector machine (SVM) basically return a binary or multivalued output, many classifier implementations can return a continuous output. Thus, we can use such output as a score for summarization. For example, a summarization method using an SVM classifier uses the distance from the hyperplane [3].

We used scikit-learn's random forest classifier, which can return the probability of each sample. Thus, we compared the method using the probability with our proposed method. We used the same classifiers in the formal run and added one more classifier trained on all data without undersampling. We modified the classifiers to output the probability and chose the sentences in the order of their

Table 4. ROUGE scores in the formal run (all-topic)

	recall							F-measure						
	N1	N2	N3	N4	L	SU4	W1.2	N1	N2	N3	N4	L	SU4	W1.2
Surface Form	**0.459**	**0.200**	**0.131**	**0.089**	**0.394**	**0.229**	**0.186**	**0.361**	0.151	0.097	0.064	0.305	**0.169**	**0.192**
Stem	**0.479**	**0.217**	**0.145**	**0.101**	**0.412**	**0.247**	**0.197**	**0.377**	**0.165**	0.108	0.074	0.319	**0.184**	**0.205**
Content Word	**0.326**	**0.164**	**0.094**	0.046	**0.315**	**0.168**	**0.201**	**0.249**	**0.123**	0.067	0.036	**0.239**	**0.110**	**0.187**

Source Document	次に、新しい公共について伺います。 新しい公共という考え方は、私たちが国家戦略の柱として、地域主権改革とともに、これからのあるべき社会像として掲げたものです。 日本では、古くから連、結、講、座、あるいは若者組などの住民組織や市井の寺子屋、隠居という名のボランティア的な活動などが活力ある市民社会を担っていました。 新しい公共の考え方は、以前あったこのような社会を現在にふさわしい形で再構築することを目指すものです。 東日本大震災の被災地では、数々のボランティア活動が行われています。 強制ではなくみずからの意思で支援活動をされていた多くの方々の姿は感動的であり、改めて人々のつながりと助け合いの大切さを感じさせられました。 石原都知事は、都の防災対応指針において、自助、共助の徹底について述べられています。 行政依存ではなく、一人一人自立した個が、地域、社会を主体的に働きかけていく協働は、災害時には不可欠なものです。 そこで伺います。**東京都においては、このような新しい公共型社会の実現を目指し、支え合いと活気のある社会を構築していくべきと考えますが、知事の所見を伺います。**
System Output	公共型社会の実現を目指し、支え合いと活気のある社会を構築していくべきと考えますが、知事の所見を。
Gold Standard	支え合いと活気のある社会を構築すべき。知事の所見を。

Fig. 1. Example of successful summarization

probabilities. We chose extra sentences if the length of the chosen sentences was ten less than the limit, as with the proposed method.

We conducted the extraction tests with these classifiers and calculated precision, recall, and F-measure, as shown in Table 5. In these tests, we considered the positive sentences gathered by the method described in Section 2.1 to be the gold standard. In Table 5, “closed” indicates the closed test conducted on the training data in the formal run. Similarly, “open” indicates the open test conducted on the test data in the formal run, where the test data consists of “single-topic” and “multi-topic” data.

Table 5 shows that the first and second classifier scored low because of the low amount of their training data. With precision, although the classifier using all the training data scored the highest in the closed test, we consider this overfitting. The fourth classifier achieved the highest precision score in the open test. With

Source Document	我が国の経済にあっては、欧州の債務危機や歴史的な円高などが、回復の兆しが見えた景気に冷や水を浴びせています。 企業収益の動向は不透明さを増しており、今後の都税収入への影響は避けられません。 こうした中、都には、少子高齢化や中小企業対策など、山積する課題に対して効果的な手だてを講じ、現下の閉塞感を打ち破り、東京に活力を呼び戻していくことが求められています。 とりわけ、震災への対応は喫緊の課題です。 我が党が立ち上げた東日本大震災復旧・復興対策推進本部で議論を重ね、先月、防災力強化に向けての提言を行いました。 提言内容も含め、高度防災都市の実現に向けた取り組みを加速する上では、法人事業税の暫定措置の撤廃は不可欠であり、約束どおり撤廃するよう国に強く求めるものであります。 この間、国が公共事業を見識ある考えもなく削減し続けたのとは対照的に、都は七年連続で投資的経費を伸ばしてきました。 都税収の回復が当面期待できない今だからこそ、中小企業の受注機会をふやすなど、景気を刺激し、防災力強化にも資する投資的経費に財源を振り向けることが重要であります。 **これまで以上にめり張りをつけ、都民に安心と希望をもたらす予算とするべく、新年度予算編成作業を進めるべきと考えますが、所見を伺います。**
System Output	つけ、都民に安心と希望をもたらす予算とするべく、新年度予算編成作業を進めるべきと考えますが、所見を。
Gold Standard	メリハリをつけて、都民に安心と希望をもたらす予算にするべき。所見は。

Fig. 2. Example of unsuccessful summarization

recall, our proposed method scored the highest in all tests. Since the recall score is more important in summarization, our method is more suitable and thus our system achieved good performance in the formal run.

Another advantage of our method is that it does not need to tune the balance between positive and negative data. Although Table 5 shows that the fourth classifier achieved the highest precision and F-measure scores, this does not always hold. The third classifier might achieve higher performance when we use other training data. However, our method does not need to consider which ratio of positive and negative data is best, and uses multiple training data with different ratios.

5 Conclusions

This paper described our summarization system at the NTCIR-14 QA Lab-PoliInfo. We proposed a progressive ensemble random forest method, which applies multiple random forest classifiers training on different-sized data sets step by step in order to deal with imbalanced data. Although we achieved good performance, especially in the evaluation by ROUGE scores, our sentence reduction module sometimes caused our system to create unnatural sentences.

Thus, our future work is to improve the sentence reduction module. We would also like to investigate the relationship between our progressive ensemble random forest classifiers and the probability they estimated.

Table 5. Extraction results

			proposed	× 1	× 2	× 3	× 4	× 5	all
Precision	closed		0.963	0.860	0.967	0.973	0.983	0.987	**1.00**
	open	all	0.446	0.465	0.471	0.520	**0.526**	0.511	0.523
		single	0.481	0.482	0.464	0.560	**0.588**	0.553	0.571
		multi	0.417	0.450	0.477	**0.483**	0.466	0.473	0.477
Recall	closed		**0.967**	0.785	0.893	0.875	0.886	0.886	0.896
	open	all	**0.523**	0.437	0.406	0.452	0.462	0.457	0.457
		single	**0.526**	0.432	0.411	0.495	**0.526**	0.495	0.505
		multi	**0.520**	0.441	0.402	0.412	0.402	0.422	0.412
F-measure	closed		**0.965**	0.821	0.929	0.921	0.932	0.933	0.945
	open	all	0.481	0.450	0.436	0.484	**0.492**	0.483	0.488
		single	0.503	0.456	0.436	0.525	**0.556**	0.522	0.536
		multi	**0.463**	0.446	0.436	0.444	0.432	0.446	0.442

References

1. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
2. Carbonell, J., Goldstein, J.: The use of mmr and diversity-based reranking for reodering documents and producing summaries. In: Proceedings of ACM-SIGIR '98, pp. 335–336 (1998)
3. Hirao, T., Isozaki, H., Maeda, E., Matsumoto, Y.: Extracting important sentences with support vector machines. In: Proceedings of the 19th International Conference on Computational Linguistics. Vol. 1, pp. 1–7. Association for Computational Linguistics (2002)
4. Kimura, Y., Shibuki, H., Ototake, H., Uchida, Y., Takamaru, K., Sakamoto, K., Ishioroshi, M., Mitamura, T., Kando, N., Mori, T., Yuasa, H., Sekine, S., Inui, K.: Overview of the ntcir-14 qa lab-poliinfo task. In: Proceedings of the 14th NTCIR Conference (2019)
5. Ogawa, Y., Satou, M., Komamizu, T., Toyama, K.: Extracting important sentences with random forest for statute summarization. In: Proceedings of the Annual Conference of JSAI 2019. The Japanese Society for Artificial Intelligence (to appear)
6. Kudo, T. Matsumoto, Y.: Japanese dependency analysis using cascaded chunking. In: CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops). pp. 63–69 (2002)