Evaluation Measures Considering Sentence Concatenation for Automatic Summarization by Sentence or Word Extraction

Chiori Hori, Tsutomu Hirao and Hideki Isozaki
NTT Communication Science Laboratories
{chiori, hirao, isozaki}@cslab.kecl.ntt.co.jp

Abstract

Automatic summaries of text generated through sentence or word extraction has been evaluated by comparing them with manual summaries generated by humans by using numerical evaluation measures based on *precision* or *accuracy*. Although sentence extraction has previously been evaluated based only on *precision* of a single sentence, ***sentence concatenations*** in the summaries should be evaluated as well. We have evaluated the appropriateness of sentence concatenations in summaries by using evaluation measures used for evaluating word concatenations in summaries through word extraction. We determined that measures considering sentence concatenation much better reflect the human judgment rather than those based only on the precision of a single sentence.

1 Introduction

Summarization Target and Approach

The amount of text is explosively increasing day by day, and it is becoming very difficult to manage information by reading all the text. To manage information easily and find target information quickly, we need technologies for summarizing text. Although research into text summarization started in the 1950's, it is still largely in the research phase (Mani and Maybury, 1999). Several projects on text summarization have been carried out. [1] In these project, text summarization has so far focused on summarizing single documents through sentence extraction. Recently, summarizing multiple documents with the same topic has been made a target. The major approach to extracting sentences that have significant information is statistical, i.e., supervised learning from parallel corpora consisting of original texts and their summarization (Kupiec et al., 1995) (Aone et al., 1998) (Mani and Bloedorn, 1998).

Several summarization techniques for multimedia including image, speech, and text have been researched. Manually transcribed newswire speech (TDT data) and meeting speech (Zechner, 2003) have been set as summarization targets. The need to automatically generate summaries from speech has led to research on summarizing transcription results obtained by automatic speech recognition instead of manually transcribed speech (Hori and Furui, 2000a). This summarization approach is word extraction (sentence compaction) that attempts to extract significant information, exclude acoustically and linguistically unreliable words, and maintain the meanings of the original speech.

The summarization approaches that have been mainly researched so far are extracting sentences or words from original text or transcribed speech. There has also been research on generating an "abstract" like the much higher level summarization composed freely by human experts (Jing, 2002). This approach includes not only extracting sentences but also combining sentences to generate new sentences, replacing words, reconstructing syntactic structure, and so on.

Evaluation Measures for Summarization

Metrics that can be used to accurately evaluate the various appropriateness to summarization are needed.The simplest and probably the ideal way of evaluating automatic summarization is to have human subjects read the summaries and evaluate them in terms of the appropriateness of summarization. However, this type of evaluation is too expensive for comparing the efficiencies of many different approaches precisely and repeatedly. We thus need automatic evaluation metrics to numerically validate the efficiency of various approaches repeatedly and consistently.

Automatic summaries can be evaluated by comparing them with manual summaries generated by humans. The similarities between the targets and

[1] SUMMAC in the Tipster project by DARPA (http://www-nlpir.nist.gov/related_projects/tipster_summac) and DUC in the TIDES project (http://duc.nist.gov/) in the U.S. TSC (http://research.nii.ac.jp/ntcir/) in the NTCIR by NII (The National Institute of Informatica) in Japan.

the automatically processed results provide metrics indicating the extent to which the task was accomplished. The similarity that can better reflect subjective judgments is a better metric.

To create correct answers for automatic summarization, humans generate manual summaries through sentence or word extraction. However, references consisting of manual summaries vary among humans. The problems in validating automatic summaries by comparing them with various references are as follows:

- correct answers for automatic results cannot be unified because of subjective variation,
- the coverage of correct answers in the collected manual summaries is unknown, and
- the reliability of references in the collected manual summaries is not always guaranteed.

When the similarity between automatic results and references is used for the evaluation metrics, the similarity determination function counts overlapping of each component or sequence of components in the automatic results. If concatenations between components in a summary had no meaning, the overlap of a single component between the automatic results and the references can represent the extent of summarization. However, concatenations between sentences or words have meanings, so some concatenations of sentences or words in the automatic summaries sometimes generate meanings different from the original. The evaluation metrics for summarization should thus consider each concatenation between components in the automatic results.

To evaluate sentence automatically generated with taking consideration word concatenation into by using references varied among humans, various metrics using *n-gram precision* and *word accuracy* have been proposed: **word string precision** (Hori and Furui, 2000b) for summarization through word extraction, **ROUGE** (Lin and Hovy, 2003) for abstracts, and **BLEU** (Papineni et al., 2002) for machine translation. Evaluation metrics based on word accuracy, *summarization accuracy* (**SumACCY**), using a word network made by merging manual summaries has been proposed (Hori and Furui, 2001). In addition, to solve the problems for the coverage of correct answers and the reliability of manual summaries as correct answers, *weighted summarization accuracy* (**WSumACCY**) in which **SumACCY** is weighted by the majority of the humans' selections, has been proposed (Hori and Furui, 2003a).

In contrast, summarization through sentence extraction has been evaluated using only single sentence *precision*. Sentence extraction should also be evaluated using measures that take into account sentence concatenations, the coverage of correct answers, and the reliability of manual summaries.

This paper presents evaluation results of automatic summarization through sentence or word extraction using the above mentioned metrics based on *n-gram precision* and *sentence/word accuracy* and examines how well these measures reflect the judgments of humans as well.

2 Evaluation Metrics for Extraction

In summarization through sentence or word extraction under a specific summarization ratio, the order of the sentences or words and the length of the summaries are restricted by the original documents or sentences. Metrics based on the *accuracy* of the components in the summary is a straight-forward approach to measuring similarities between the target and automatic summaries.

2.1 Accuracy

In the field of speech recognition, automatic recognition results are compared with manual transcription results. The conventional metric for speech recognition is recognition accuracy calculated based on *word accuracy*:

$$\mathbf{ACCY} = \frac{Len - (Sub + Ins + Del)}{Len} \times 100[\%], \quad (1)$$

where Sub, Ins, Del, and Len are the numbers of substitutions, insertions, deletions, and words in the manual transcription, respectively. Although *word accuracy* cannot be used to directly evaluate the meanings of sentences, higher accuracy indicates that more of the original information has been preserved. Since the meaning of the original documents is generated by combining sentences, this metric can be applied to the evaluation for sentence extraction. *Sentence accuracy* defined by eq. (1) with words replaced by sentences represents how much the automatic result is similar to the answer and how well it preserves the original meaning.

Accuracy is the simplest and most efficient metric when the target for the automatic summaries can be set as only one answer. However, there are usually multiple targets for each automatic summary due to the variation in manual summarization among humans. Therefore, it is not easy to use *accuracy* to evaluate automatic summaries. Subjective variation results into two problems:

- how to consider all possible correct answers in the manual summaries, and
- how to measure the similarity between the evaluation sentence and multiple manual summaries.

If we could collect all possible manual summaries, the one most similar to the automatic result could be chosen as the correct answer and used for the evaluation. The sentence or word accuracy compared with the most similar manual summary is denoted as **NrstACCY**. However, in real situations, the number of manual summaries that could be collected is limited. The coverage of correct answers in the collected manual summaries is unknown. When the coverage is low, the summaries are compared with inappropriate targets, and the **NrstACCY** obtained by such comparison does not provide an efficient measure.

2.2 N-gram Precision

One way to cope with the coverage problem is to use local matching of components or component strings with all the manual summaries instead of using a measure comparing a word sequence as a whole sentence, such as **NrstACCY**. The similarity can be measured by counting the *precision*, i.e., the number of sentence or word n-gram overlapping between the automatic result and all the references.

Even if there are multiple targets for an automatic summary, the *precision* of components in each original can be used to evaluate the similarity between the automatic result and the multiple references. *Precision* is an efficient way of evaluating the similarity of component occurrence between automatic results and targets with a different order of components and different lengths.

In the evaluation of summarization through extraction, a component occurring in a different location in the original is considered to be a different component even if it is the same component as one in the result. When an answer for the automatic result can be unified and the lengths of the automatic result and its answer are the same, *accuracy* counts insertion errors and deletion errors and thus has both the *precision* and *recall* characteristics.

Since meanings are basically conveyed by word strings rather than single words, ***word string precision*** (Hori and Furui, 2000b) can be used to evaluate linguistic precision and the maintenance of the original meanings of an utterance. In this method, word strings of various lengths, that is n-grams, are used as components for measuring precision. The extraction ratio, p_n, of each word string consisting of n words in a summarized sentence, $V = v_1, v_2, \ldots, v_M$, is given by

$$p_n = \frac{\sum_{m=n}^{M} \delta(v_{m-n+1}, \ldots, v_{m-1}, v_m)}{M - n + 1} \quad , \qquad (2)$$

where

$$\delta(u_n) = \begin{cases} 1 & \text{if } u_n \in U_n \\ 0 & \text{if } u_n \notin U_n \end{cases} \quad , \qquad (3)$$

u_n: each word string consisting of n words
U_n: a set of word strings consisting of n words in all manual summarizations.

When n is 1, p_n corresponds to the precision of each word, and when n is the same length as a summarized sentence ($n = M$), p_n indicates the precision of the summarized sentence itself.

2.3 Summarization Accuracy: SumACCY

Summarization accuracy (**SumACCY**) was proposed to cope with the problem of correct answer coverage and various references among humans (Hori and Furui, 2001). To cover all possible correct answers for summarization using a limited number of manual summaries, all the manual summaries are merged into a word network. In this evaluation method, the word sequence in the network closest to the evaluation word sequence is considered to be the target answer. The *word accuracy* of the automatic result is calculated in comparison with the target answer extracted from the network.

Since summarization is processed by extracting words from an original; the words cannot be replaced by other words, and the order of words cannot be changed. Multiple manual summaries can be combined into a network that represents the variations. Each set of words that could be extracted from the network consists of words and word strings occurring at least once in all the manual summaries. The network made by the manual summaries can be considered to represent all possible variations of correct summaries.

SUB	The beautiful cherry blossoms in Japan bloom in spring
A	The cherry blossoms in Japan
B	cherry blossoms in Japan bloom
C	beautiful cherry bloom in spring
D	beautiful cherry blossoms in spring
E	The beautiful cherry blossoms bloom

Table 1: Example of manual summarization by sentence compaction

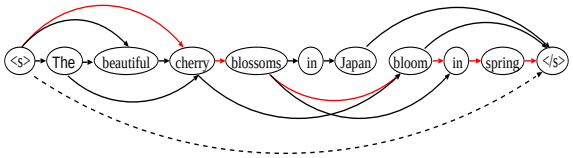

Figure 1: Word network made by merging manual summaries

The sentence "*The beautiful cherry blossoms in Japan bloom in spring.*" is assumed to be manually summarized as shown in Table 1. In this example, five words are extracted from the nine words. Therefore, the summarization ratio is 56%. The variations of manual summaries are merged into a word network, as shown in Fig. 1. We use <s> and </s> as the beginning and ending symbols of a sentence. Although "*Cherry blossoms bloom in spring*" is not among the manual answers in Table 1, this sentence, which could be extracted from the network, is considered a correct answer.

When references consisting of manual summaries cannot cover all possible answers and lack the appropriate answer for an automatic summary, **SumACCY** calculated using such a network is better than **NrstACCY** for evaluating the automatic result. This evaluation method gives a penalty for each word concatenation in the automatic results that is excluded in the network, so it can be used to evaluate the sentence-level appropriateness more precisely than matching each word in all the references.

2.4 Weighted SumACCY: WSumACCY

In **SumACCY**, all possible sets of words extracted from the network of manually summarized sentences are equally used as target answers. However, the set of words containing word strings selected by many humans would presumably be better and give more reliable answers. To obtain reliability that reflects the majority of selections by humans, the summarization accuracy is weighted by a posterior probability based on the manual summarization network. The reliability of a sentence extracted from the network is defined as the product of the ratios of the number of subjects who selected each word to the total number of subjects. The weighted summarization accuracy is given by

$$\mathbf{WSumACCY} = \frac{\tilde{P}(v_1 \dots v_M|R) \times \mathbf{SumACCY}}{\tilde{P}(\hat{v}_1 \dots \hat{v}_{\hat{M}}|R)}, \quad (4)$$

where $\tilde{P}(v_1 \dots v_M|R)$ is the *reliability score* of a set of words $v_1 \dots v_M$ in the manual summarization network, R, and M represents the total number of words in the target answer. The set of words $\hat{v}_1 \dots \hat{v}_{\hat{M}}$ represents the word sequence that maximizes the *reliability score*, $\tilde{P}(\cdot|R)$, given by

$$\tilde{P}(v_1 \dots v_M|R) = \left(\prod_{m=2}^{M} \frac{C(v_{m-1}, v_m|R)}{H_R} \right)^{\frac{1}{M-1}}, \quad (5)$$

where v_m is the m-th word in the sentence extracted from the network as the target answer, and $C(x, y|R)$ indicates the number of subjects who selected the word connection of x and y. Here, "word connection" means an arc in the manual summarization network. H_R is the number of subjects.

2.5 Evaluation Experiments

Newspaper articles and broadcast news speech were automatically summarized through sentence extraction and word extraction respectively under the given summarization ratio, which is the ratio of the numbers of sentences or words in the summary to that in the original.

The automatic summarization results were subjectively evaluated by ten human subjects. The subjects read these summaries and rated each one from 1 (incorrect) to 5 (perfect). The automatic summaries were also evaluated by using the numerical metrics **SumACCY**, **WSumACCY**, **NrstACCY**, and **n-gram precision** ($1 \leq n \leq 5$) in comparison with reference summaries generated by humans. The precisions of 1-gram, ..., 5-gram are denoted PREC1, ..., PREC5. The numerical evaluation results were averaged over the number of automatic summaries.

Note that the subjects who judged the automatic summaries did not include anyone who generated the references. To examine the similarity of the human judgments and that of the manual summaries, the kappa statistics, κ, was calculated using eq. (A-1) in the Appendix.

Finally, to examine how much the evaluation measures reflected the human judgment, the correlation coefficients between the human judgments and the numerical evaluation results were calculated.

Sentence extraction

Sixty articles in Japanese newspaper published in 94, 95, and 98 were automatically summarized with a 30% summarization ratio. Half the articles were general news report (NEWS), and other half were columns (EDIT).

The automatic summarization was performed using a Support Vector Machine (SVM) (Hirao et al., 2003), random extraction (RDM), the lead method

(LEAD) extracting sentences from the head of articles. In comparison with these automatic summaries, manual summaries (TSC) was also evaluated.

These 4 types of summaries, SVM, RDM, LEAD, and TSC were read and rated 1 to 5 by 10 humans. The summaries were evaluated in terms of extraction of significance information (SIG), coherence of sentences (COH), maintenance of original meanings (SEM), and appropriateness of summary as a whole (WHOLE).

To numerically evaluate the results using the objective metrics, 20 other human subjects generated manual summaries through sentence extraction. These manual summaries were set as the target set for the automatic summaries.

Word extraction

Japanese TV news broadcasts aired in 1996 were automatically recognized and summarized sentence by sentence (Hori and Furui, 2003b). They consisted of 50 utterances by a female announcer. The out-of-vocabulary (OOV) rate for the 20k word vocabulary was 2.5%, and the test-set perplexity was 54.5. Fifty utterances with word recognition accuracy above 90%, which was the average rate over the 50 utterances, were selected and used for the evaluation. The summarization ratio was set to 40%.

Nine automatic summaries with various *summarization accuracies* from 40% to 70% and a manual summary (SUB) were selected as a test set. These ten summaries for each utterance were judged in terms of the appropriateness of the summary as a whole (WHOLE).

To numerically evaluate the results using the objective metrics, 25 humans generated manual summaries through word extraction. These manual summaries were set as a target set for the automatic summaries, and merged into a network. Note that a set of 24 manual summaries made by other subjects was used as the target for SUB.

2.6 Evaluation Results

Figures 2 and 3 show the correlation coefficients between the judgments of the subjects and the numerical evaluation results for EDIT and NEWS. They show that the measures based on accuracy much better reflected human judgments than those of the n-gram precisions for evaluating SIG and WHOLE for both EDIT and NEWS. On the other hand, PREC2 better reflected the human judgments for evaluating COH and SEM. These results show that measures taking into account sentence concatenations better reflected human judgments than *single component precision*. The precisions of longer

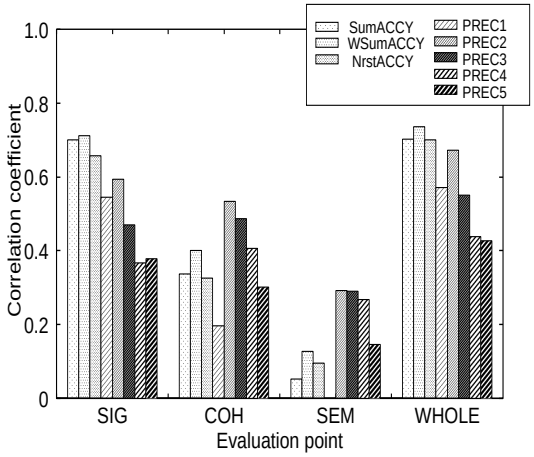

Figure 2: Correlation coefficients between human judgment and numerical evaluation results for EDIT

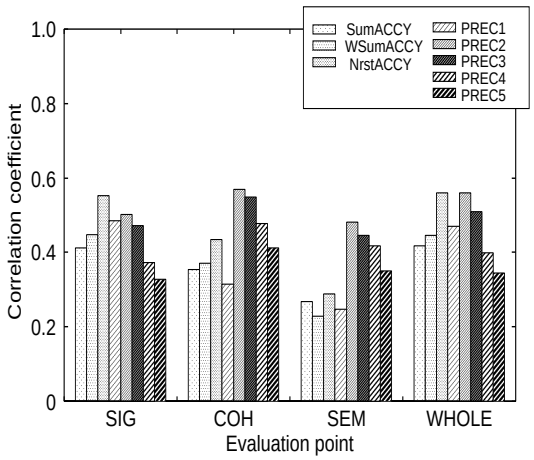

Figure 3: Correlation coefficients between human judgment and numerical evaluation results for NEWS

sentence strings (PREC3 to PREC5) didn't reflect the human judgments for all the conditions. These results show that meanings of the original article can maintain by the concatenations of only a few sentences in summarization through sentence extraction.

Table 2 lists the kappa statistics for the manual summaries and the human judgments for EDIT and NEWS. The manual results varied among humans

DATA	SUMMARIES	κ
EDIT	manual summaries	0.35
NEWS	manual summaries	0.39

Table 2: Kappa statistics for manual summaries and human judgments for sentence extraction.

and the similarity among humans was low. The kappa statistics for NEWS is slightly higher than that for EDIT. The difference of similarities among

manual summaries is due to the difference in structures of information in each article. Although the articles in EDIT had a discourse structure, NEWS had isolated and stereotyped information scattered throughout the articles.

While the human judgments for NEWS were similar, those for EDIT varied.The difficulty in evaluating COH and SEM in EDIT is due to the variation in both manual summaries and human judgment.

Figure 4 shows the correlation coefficients between the judgments of the subjects and the numerical evaluation results for summaries of broadcast news speech through word extraction. Table 3 lists

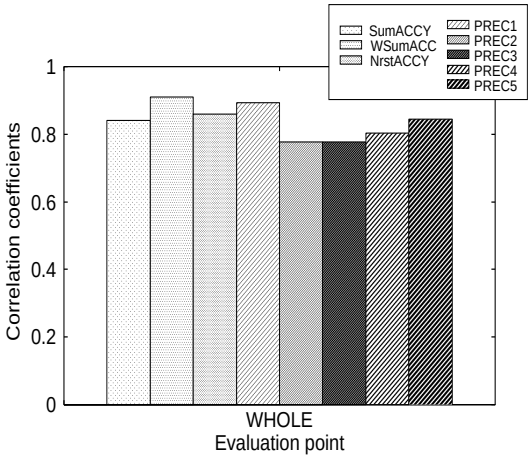

Figure 4: Correlation coefficients between human judgment and numerical evaluation results for summaries through word extraction

the kappa statistics for the manual summaries and the human judgments for summaries through word extraction. In word extraction, the human judgments and the manual summaries were very similar among the subjects.

DATA	SUMMARIES	κ
Broadcast news	manual summaries	0.47

Table 3: Kappa statistics for manual summaries and human judgments for word extraction

As shown in figure 4, **WSumACCY** yielded the best correlation to the human judgments. This means that the correctness as a sentence and the weight (that is how many subjects support the extracted phrases in summarized sentences) are important in summarization through word extraction. In comparison with the results of sentence extraction in Figures 2 and 3, PREC1 effectively reflected the human judgments for word extraction. Since in the manual summarized sentences through word extraction under the low summarization ratio, the sentences were summarized based on significance word extraction rather than syntactic structure maintenance to generate grammatically correct sentences.

3 Conclusion

We have presented the results of evaluating the appropriateness of the sentence concatenations in summaries generated using **SumACCY**, **WSumACCY**, **NrstACCY** and **n-gram precision**. We found that the measures taking into account sentence concatenation much better reflected the judgments of humans than did the single sentence precision, so the concatenation of sentences in summaries should be evaluated.

Although the human judgments and the manual summaries for word extraction did not vary much among the subjects, those for sentence extraction for single article summarization greatly varied among the subjects. As a result, it is very difficult to set correct answers for single article summarization through sentence extraction.

Future works involves experiments to examine the efficiency of each numerical measures in response to the coverage of correct answers.

4 Acknowledgments

We thank NHK (Japan Broadcasting Corporation) for providing the broadcast news database. We also thank Prof. Sadaoki Furui at Tokyo Institute of Technology for providing the summaries of the broadcast news speech.

References

C. Aone, M. Okurowski, and J. Gorlinsky. 1998. Trainable scalable summarization using robust NLP and machine learning. In *Proceedings ACL*, pages 62–66.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

T. Hirao, K. Takeuchi, H. Isozaki, Y. Sasaki, and E. Maeda. 2003. SVM-based multi-document summarization integrating sentence extraction with bunsetsu elimination. *IEICE Trans. Inf. & Syst.*, E86-D(9):1702–1709.

C. Hori and S. Furui. 2000a. Automatic speech summarization based on word significance and linguistic likelihood. In *Proceedings ICASSP*, volume 3, pages 1579–1582.

C. Hori and S. Furui. 2000b. Improvements in automatic speech summarization and evaluation methods. In *Proceedings ICSLP*, volume 4, pages 326–329.

C. Hori and S. Furui. 2001. Advances in automatic speech summarization. In *Proceedings Eurospeech*, volume 3, pages 1771–1774.

C. Hori and S. Furui. 2003a. Evaluation methods for automatic speech summarization. In *Proceedings Eurospeech*, pages 2825–2828.

C. Hori and S. Furui. 2003b. A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, 3:368–378.

H. Jing. 2002. Using hidden markov modeling to decompose human-written summaries. *Computational Linguistics*, 28(4):527–543.

J. Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR*, pages 68–73.

Chin-Yew Lin and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings HLT-NAACL*.

I. Mani and E. Bloedorn. 1998. Machine learning of general and user-focused summarization. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 821–826.

I. Mani and M. Maybury. 1999. *Advances in Automatic Text Summarization*. The MIT Press.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings ACL*.

K. Takeuchi and Y. Matsumoto. 2001. Relation between text structure and linguistic clues: An investigation on text structure of newspaper articles. *Mathematical Linguistics*, 22(8).

K. Zechner. 2003. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Appendix

κ is given by

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}, \tag{A-1}$$

where $P(A)$ and $P(E)$ are the probabilities of human agreement and chance agreement, respectively, so κ is adjusted by the possibility of chance agreement. This measure was used to assess agreement of human selections for discourse segmentation (Carletta, 1996).

In this study, $kappa$ was calculated using a table of objects and categories (Takeuchi and Matsumoto, 2001). $P(A)$ was calculated using

$$P(A) = \frac{1}{N}\sum_{i=1}^{N} S_i, \tag{A-2}$$

where N is the number of trials to select one class among all classes, and S_i is the probability that two humans at least agree at the i-th selection:

$$S_i = \frac{\sum_{j=1}^{m} {}_{n_{ij}}C_2}{{}_{k}C_2}, \tag{A-3}$$

where k and m are the number of subjects and classes, respectively. When the task is sentence or word extraction, the number of classes is two, i.e., *extract*/*not extract*. The numerator of eq. (A-3) shows the sum of the combinations that two humans at least agree for each class; n_{ij} is the number of humans who select the j-th class at the i-th selection.

$P(E)$ is the probability of chance agreement by at least two humans:

$$P(E) = \sum_{j=1}^{m} {p_j}^2, \tag{A-4}$$

where p_j is the probability of selecting the j-th class given by

$$P_j = \frac{\sum_{i=1}^{N} n_{ij}}{Nk}, \tag{A-5}$$

where the total number of humans who select the j-th class for each trial is divided by the total number of trials performed by all humans.