Discriminative Dialog Analysis
Using a Massive Collection of BBS comments

Eiji ARAMAKI Takeshi ABEKAWA
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan
eiji.aramaki@gmail.com
abekawa@p.u-tokyo.ac.jp

Yohei MURAKAMI Akiyo NADAMOTO
NICT (National Institute of Information and
Communications Technology)
3-5 Hikari-dai Seika-cho Souraku-gun
Kyoto, Japan
yohei@nict.go.jp
nadamoto@nict.go.jp

ABSTRACT

With the rapid growth of the Internet, the amount of electronic dialog, such as discussions in blogs or bulletin board systems (BBSs), has been increasing day by day. Although dialogs in blogs and BBSs are full of knowledge, they are sometimes split into several segments because multiple topics are discussed simultaneously. This paper addresses the challenge of determining whether two comments in a BBS are related to each other or not. We use two types of indications: a content relevance, which captures the similarity between two comments, and a functional relevance, which captures corresponding phrase pairs such as "*please tell me why...*" and "*It is because ...*". We use a measure proposed in previous studies for the content relevance, while for the functional relevance, we propose a new measure based on a co-occurrence ratio in dialogs. We also propose a method to gather a large collection of BBS comments. Experimental results showed that two types of relevance individually contribute to the accuracy, demonstrating the basic feasibility of our approach.

Categories and Subject Descriptors

H.4.m [**Information Systems and Applications**]: Miscellaneous

General Terms

Algorithms

Keywords

semantic similarity, Web mining

1. INTRODUCTION

With the rapid growth of the Internet, the amount of available text data increases day by day. Especially, texts in blogs and Bulletin Board Systems (BBS) explodes, because everyone easily posts comments or messages. Such huge texts can be promising source of knowledge.

Table 1 shows an example of a BBS dialog. From the comments (1) and (3), we can know that "*N12*"is a "*light and small mp3 player*". The comment (5) also tells us that "*N12*"is now under "stopped producing". Another comment chain (2) and (4) goes in the same way.

Table 1: Examples of Comments in a BBS ("*mp3 player*" Community).

(1) What is the most light or small mp3 player? iPod Shuffle is the best way to do?

(2) please tell me why my nano sometimes stops even battery still remains.

(3) How about iriver N12? extremely light and small.

(4) It is because battery display approaches approx. Even battery runs out, display sometimes shows it is still left.

(5) iriver N series has stopped producing.

In this way, knowledge in a BBS sometimes emerges from the collaboration of many individuals. However, a BBS text produces a problem of gaps between related comments. (i.e., a comment chain, (1)-(3)-(5), has two gaps, (2) and (4)). Figure 1 shows the frequency of the distance between a comment and its response[1]. As shown in the figure, the ratio of successive responses (distance=1) is only 48.8%, while the others have gaps (distance>1).

This paper addresses the problem of determining whether two comments are related or not. This problem is not only a pre-processing for a BBS analysis, it can be a new challenging task for dialog studies as well.

To do this, we assume that three different types of clues are available.

(1) Content Relevance

The first one is the similarity between two comments. For example, in the two sentences (1) and (3) from Figure 1, we can deduce that they are probably related because "*light*" and "*small*" occur in both sentences as follows:

NLPIX2008, April 22, 2008, Beijing, China.
.

[1] This distance is counted in the test-set in the experiment described in Section 5.

(1) What is the most light or small mp3 player?

(3) . . . . extremely light and small.

We call this type of indication a **content relevance**. To calculate the content relevance, we use a co-occurrence-based similarity [4].

(2) Functional Relevance

The second clue is a discourse relationship between two comments. For example, a phrase pair, "*please tell me why...*" in (2) and "*it is because...*" in (4), can provide evidence of their relationship as follows:

(2) please tell me why my nano sometimes . . . .

(4) It is because battery display approaches . . . .

We refer such a phrase pair to a **corresponding pair**, and refer this type of clue to **functional relevance**.

In calculating this relevance, the first key issue is how to define a rich set of corresponding pairs. Our approach is to extract highly co-occurring phrase pairs from a large collection of comment pairs (a comment and its response), and regard them as corresponding pairs.

The above approach produces another issue of how to build the collection of comment pairs. This introduces a chicken-and-egg situation because we need a collection of comment pairs to obtain a comment pair. To circumvent this dilemma, we relied on a very large volume of dialog data. First, we collected 17,300,000 comments from the Web, and then extracted only reliable parts of them using manually designed lexical patterns. Although the reliable parts were only 1.4% of the total volume, this still left us with 121,699 comment pairs.

(3) Context Information

The final clue is information that comes from outside (context) of comments, such as the previous dialog history, the distance between comments, and their time stamps. Although such information could be a strong evidence, this study does not utlize such knowledge from the view point of dialog study. This study relies only on information from two comments.

The point of this study is three-fold:

(1) We challenge a new task (to determine whether two comments correspond to each other or not).

(2) To solve this task, we formalize relevance using two indicators (content relevance and functional relevance), and the experimental results empirically demonstrate their individual contribution.

(3) To calculate the functional relevance, we propose a method to automatically build a large set of comment pairs from the Web.

2. METHOD

First, we formalize our task as follows:

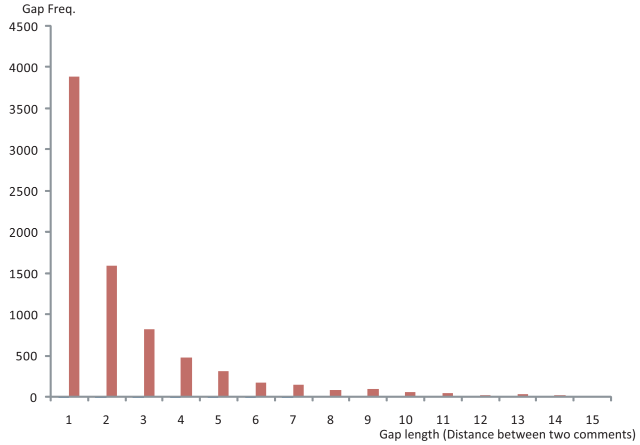


Figure 1: Gap length (between a Comment and its response) and its Occurrence.

Input: two comments, the ith comment and jth comment ($j > i$) from the same BBS.

Output: True, if the jth comment is a response to the ith comment, and otherwise, False.

For simplicity, this paper uses the notation P for the ith comment and Q for the jth comment.

We use two types of indicators: content relevance (Section 2.1) and functional relevance (Section 2.2). Both of relevances are converted into features, and a Support Vector Machine (SVM) classifier[32] learns their relationships.

2.1 Content Relevance

Various metrics have already been proposed to measure the similarity between two sentences, starting from a simple word overlap ratio to a recent statistical similarity (such as *sentence_relevance*[5]). We used a Web-based point-wise mutual information ($WEBPMI$) [4] to calculate content relevance (REL_c), which gave the best performance in our experiments. This value is based on a co-occurrence point-wise mutual information (PMI) of two words in Web pages as follows:

$$REL_c(P, Q) = \sum_{p \in W_P} \max_{q \in W_Q} WEBPMI(p, q), \qquad (1)$$

where W_P is a set of words in P, W_Q is a set of words in Q, and $WEBPMI$ is defined as follows:

$$\begin{aligned}&\mathrm{WEBPMI}(p, q)\\ &= \begin{cases} 0 & \text{if } H(p \cap q) \le c, \\ \log \frac{\frac{H(p \cap q)}{N}}{\frac{H(p)}{N}\frac{H(q)}{N}} & \text{otherwise,} \end{cases}\end{aligned} \qquad (2)$$

where $H(p)$ is the number of retrieved documents from a Web search engine resulting from the query "p," $H(q)$ is the number of retrieved documents resulting from the query "q," and $H(p \cap q)$ is the number of retrieved documents corresponding to the conjunction query "$p + q$." To avoid small number noise, we filter out any query that returns less than a threshold c number of documents[2]. N is the number of documents indexed by the search engine.

We used a search engine "TSUBAKI" developed by Shinzato et al.[28], which provided a precise occurrence number.

2.2 Functional Relevance

To capture a functional relevance, we propose a new measure, Corresponding-PMI ($CPMI$). This measure is similar to $WEBPMI$, but has the following two differences:

(1) While $WEBPMI$ is defined by the co-occurrence ratio in Web pages, $CPMI$ is defined by the co-occurrence ratio in a set of comment pairs.

(2) To capture a corresponding phrase (not single word), $CPMI$ deals with the n-gram co-occurrence ratio ($n = 1..3$).

We describe the method for building a collection of comment pairs in the next section. In this section, we describe how to calculate the functional relevance using the collection.

First, we build three databases using a set of comment pairs (Ps and Qs):

DATABASE-A: a database of n-gram occurrences in Ps.

DATABASE-B: a database of n-gram occurrences in Qs.

DATABASE-C: a database of possible combinations of $n : m$-gram pair (possible n-grams in P: possible m-grams in Q) occurrences ($1 \leq n \leq 3, 1 \leq m \leq 3,$). For example, given a comment pair, P"*How about i-pod*" and Q"*Nice idea*", we get the $n : m$-grams as shown in Figure 2.

We define the functional relevance ($REL_f(P, Q)$) using those databases as follows:

$$REL_f(P, Q) = \sum_{p \in N_P} \max \sum_{q \in N_Q} CPMI(p, q), \qquad (3)$$

where N_P is a set of n-grams in P, N_Q is a set of n-grams in Q, and $CPMI$ is defined as follows:

$$CPMI(p, q) = \begin{cases} 0 & \text{if } H_c(p \cap q) \leq c, \\ \log \frac{\frac{H_c(p \cap q)}{M}}{\frac{H_a(p)}{M} \frac{H_b(q)}{M}} & \text{otherwise}, \end{cases} \qquad (4)$$

where $H_a(p)$ is the number of occurrences of n-gram p in the DATABASE-A, $H_b(q)$ is the number of occurrences of n-gram q in the DATABASE-B, and $H_c(p \cap q)$ is the number of occurrences of the n-gram pair p and q in the DATABASE-C. We filtered out queries that returned less than a threshold c to avoid the noise of small numbers. M is the number of comment pairs.

Roughly speaking, this equation searches the highest co-occurring m-gram in Q for each n-gram in P, and sums up their PMI values.

2.3 SVM Classifier

Features

We obtain two values from the two types of relevance, the content relevance and the functional relevance. We regard them as SVM features, as with no normalization. In addition, we directly use lexicons in P and Q as features.

Training-set

The SVM training requires two types of data: (1) positive examples and (2) negative examples.

For positive examples, we used a comment pair ($P : Q$) described in the next section (Section 3).

For negative examples, we randomly replaced a response comment (Q) in a positive example by another previous comment (Q') from the same BBS. This gave us the same amount of positive data (P, Q) and negative data (P, Q').

3. AUTOMATIC DIALOG CORPUS BUILDING

3.1 Pattern-based Extraction

This section describes how to extract dialog pairs from the Web. First, we crawled 130,000 Japanese BBS sites to extract 17,300,000 comments.

Although we generally could not capture their relationships because of the gaps mentioned in Section 1, we could readily identify response targets in the following comments:

(6) Hi, John! Maybe you should ...
(7) John> Maybe you should ...
(8) 119 > *Maybe you should ...*

The comment (6) and (7) are responses to the latest comment by "*John*". In comment (8), a number "119" indicates a comment-ID. In this case, we also guess its response target as well.

To capture such indications, we manually designed the lexical patterns shown in Figure 3. By using these patterns, we extracted 890,000 comment pairs (10.2% of all comments).

Although these patterns are language-dependent, we believe that such patterns are available in most of languages (such as "*Hi, <person-name>*," in English.).

3.2 Long Comment Filtering

Figure 5 indicates the length (the number of characters) and the frequencies of the comment pairs. Because a long comment sometimes includes complex phenomena such as a response for two or more comments, or a long quotation from other comments, we focused only on short comments of less than 100 characters[3].

Figure 4 shows the distribution of comment pair length. Figure 4 LEFT is full scale; RIGHT is our target part (0-100 characters). This left us with 121,699 comment pairs (1.4% of the total volume).

3.3 Observation on Comment Pairs

We randomly extracted 140 pairs and manually classified them into nine categories. Previous studies have proposed a fine grained category set such as 40 categories for utterances

[2] Based on the work of Bollegala et al.[4], we set $c = 5$ in our experiments.

[3] A sequence of 100 Japanese characters approximately equals to 30–40 English words.

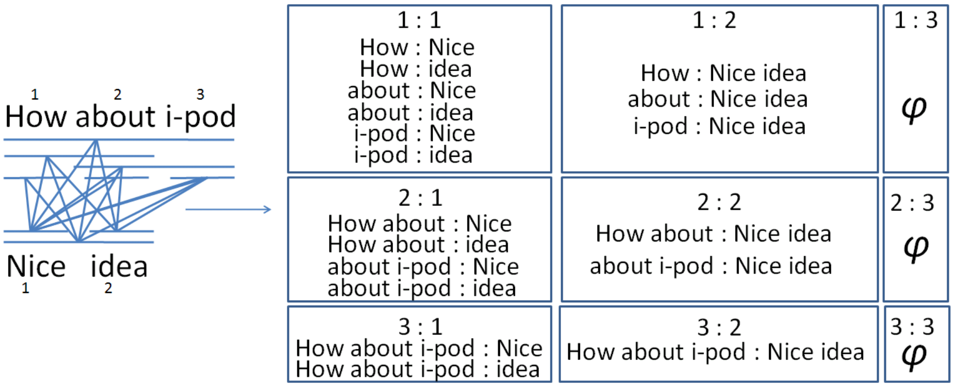


Figure 2: $n : m$-gram Examples.

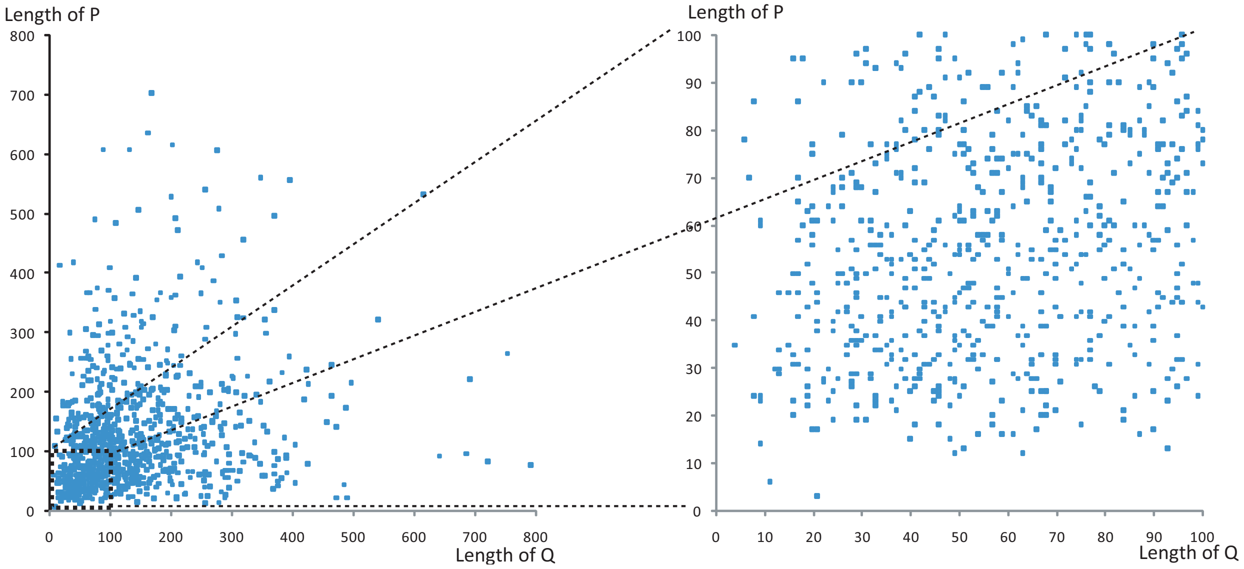


Figure 4: Distribution of comment-pair Length (Length of a comment (P) and its response (Q)).

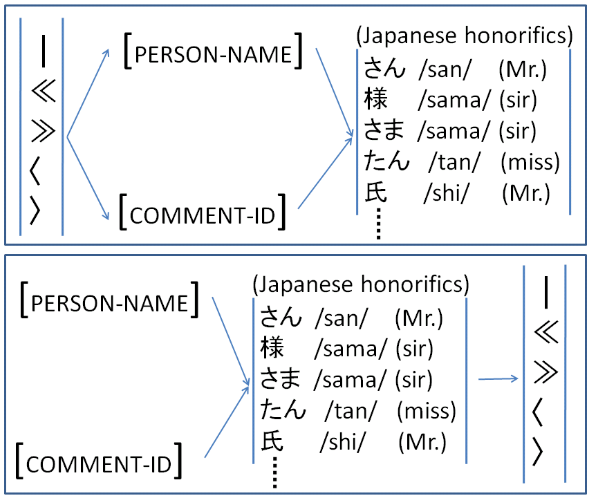


Figure 3: Extracting Patterns.

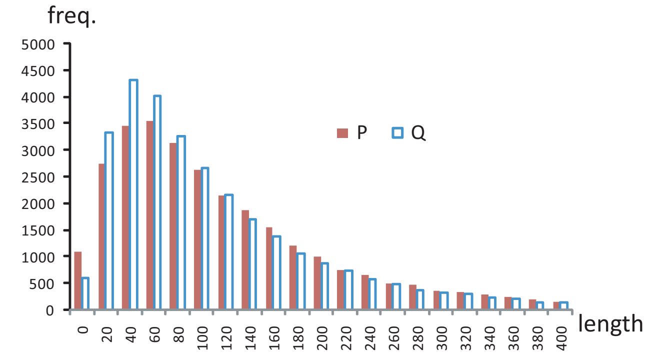


Figure 5: Length and Frequency of comment-pairs (a comment (P) and its response (Q)).

Table 2: Classification of Comment-pairs.

Classification	Freq.	Ratio
ANSWER	39	(27.8%)
THANKING	35	(25.0%)
AGREEMENT	26	(18.5%)
TOPIC-SHIFT	13	(9.2%)
QUESTION	6	(4.2%)
DISAGREEMENT	6	(4.2%)
OTHER	8	(5.7%)
ERROR	4	(2.8%)
APOLOGY	3	(2.1%)
SUM	140	

[30], and more than 20 relationships in Rhetorical Structure Theory Discourse Treebank (RST-DT)[8]). We used only nine categories because our target unit was larger than previous studies (our unit is a comment).

The occurrence of each category is shown in Table 2, and several examples are shown in Table 3. Although the phenomena that we can see in the examples are interesting, those remain subjects for future study.

4. RELATED WORKS

This study relates mainly to four fields: linguistic studies, dialog/discourse studies, topic detection and tracking, and text alignment.

4.1 Linguistic Studies (Pragmatics)

From Grice to recent neo- or post-Gricean, various linguists proposed pragmatic theories for Conversation Analysis (CA).

Grice proposed four conversational maxims (i.e., maxim of quantity, quality, relevance, and manner)[15]. Sperber and Wilson proposed a relevance theory[29], which sums up those maxims into one measure. Levinson proposes Generalized Conversational Implicature (GCI)[20], which is based on Grice's maxims, but covers wider phenomena, such as a usage of quantifier, modality, and anaphora.

Although their theories all provide explanations for various conversation phenomena, we can not utilize them as is because they are not mathematically-modeled theories. This situation motivates our approach, which is based on statistically formalized relevance.

4.2 Dialog and Discourse

In the NLP field, dialogs were mainly studied based on carefully annotated transcription data, such as the dialog act markup in several layers (DAMSL) [11], graph-based dialog annotation [3]. Also, discourse studies are at a similar stage, and various annotation schemes have been proposed, such as RST-DT[8], discourse graph-bank[35] and so on.

In contrast to those previous works, our corpus (a collection of comment pairs) is short by several information:

(1) **Granularity**: While previous studies dealt with an utterance or a phrase, our minimum unit is the comment.

(2) **Classification**: While previous studies prepared a rich set of utterance types or discourse relationships, we

Table 3: Each Category Example.

ANSWER	
P	アドマイヤムーンは秋天で引退ですか？ (Admire-moon[a] plans to retire after the next race?) [a]* Admire-moon is a name of a racehorse.
Q	♠ ＞正確にはドバイへ移籍です (♠ > Precisely, she will move to Dubai.)
THANKING	
P	15 Jul 2006 Sunrise 05:58 Sunset 19:16 だそうです (i (I heard 1993 15 Jul 2006 Sunrise 05:58 Sunset 19:16)
Q	♠! 早速、良い時間帯に予約しようと思います☆ どうもありがとうございました☆ (♠! Thank you very much. I will make a early reservation.)
AGREEMENT	
P	意外なセンで ···· 金の i-pod(笑) ダメかしらん ···· (How about golden i-pod?)
Q	おお，案外ヒットするかも！ (Nice! maybe a surprise hit...)
TOPIC-SHIFT	
P	年金問題が解決する前に、リア・ディゾンが日本語を完璧にマスターする方が早いと思うよ。 (Before resolving the pension problem, Leah Dizon* will master Japanese perfectly.)
Q	アグネス・チャンの方は一体いつになったら日本語が流暢になるのでしょうか？ (Must be. BTW, when Agnes Chan* will master Japanese?) * Leah Dizon and Agnes Chan are TV personalities in Japan.
QUESTION	
P	＞ ♣ さんわかった〜ちなみにうちらクラス一緒だったね (>Mr. ♣! I know you. You and I were in the same class.)
Q	〉 ♠ さんあれ？そうなの〜？誰だろ〜？何かちょっとしたヒント的な… (> Mr. ♠// Oh! Really? Who are you? Please more hint.)
DISAGREEMENT	
P	未読の表示設定を RSS も表示するようすれば、未読 RSS を簡単に表示できますよ (You should change display settings to read un-read RSSs.)
Q	♠ > それではメールと RSS が混ざってしまいます（ ^^; ）ゞ (♠ > If do so, you might confuse your mails with RSSs.)
OTHER	
P	わはー♪新入りです。よろしくおねがいします、 (It's my first comment! Nice to meet you.)
Q	>♠ お、100 レスおめでとうございますヽ(´▽｀) ノわはー (>♠ CONGRATS! You got 100th ID :))
APOLOGY	
P	♣> ラストベガはベガの子供ではありませんよ！父がアドマイヤベガなのでその名前なのだと… (Last-vega[a] is not Vega's child! I think his name comes from his father, Admire-vega.) [a]Last-vega and Admire-vega are names of racehorses.
Q	♠ すみません ^^; 勘違いしてました ┏○" I'm sorry :(I was confused.

* ♠ indicates the P's commenter name. ♣ indicates the Q's commenter name. ♡ indicates another person's name.

use no categories and only one relationship, that of a response or non-response.

In spite of these differences, our approach has two strong points:

(1) **Data Size**: Our data were more numerous than used in previous studies, and this enabled us to use a statistical approach (a PMI-based functional relevance).

(2) **Automatic**: Our large corpus was automatically built.

(3) **Application**: Our task could directly be a practical application, such as an application that indicates comment relationships.

We believe that our corpus can be a promising response for future dialog/discourse studies.

4.3 Topic Detection and Tracking (TDT)

Because a topic in a comment is a strong indicator, Topic Detection and Tracking (TDT)[1] is also a related field. One of the goals of TDT tasks is to detect the segmentation of single topics. Clustering methods are popular for this, including implemental clustering[33], hierarchical clustering[27], and clustering self-organizing neural networks[26].

The difference between those works and ours is the size of the target unit. Because a topic in a BBS sometimes changes comment by comment, the other approaches we have described can hardly capture segmentations. This limitation supports our approach of using functional clues.

4.4 Alignment Studies

The proposed relevance captures corresponding phrases in two texts. This is similar to the bilingual alignment task in Machine Translation (MT). Because alignment is a core technology for MT, tons of alignment methods have been proposed year after year, such as sentence alignment[6, 9, 14, 16, 17, 18, 31, 36], word/phrase alignment [7, 10, 21, 22, 23, 24, 34] and so on. Although those methods succeeded in MT, we cannot use them as is because most alignment studies focus on a parallel corpus, which is exactly corresponding to each other.

On the other hand, this study deals with comment pairs, which have at most only several corresponding words.

Note that several alignment studies handle non-parallel (or comparable) corpus[2, 12, 13, 25, 37]. In the future, we incorporate their methods for parallel fragment extraction with this task.

5. EXPERIMENTS

We conducted comprehensive evaluations using two test sets:

SMALL-SET:

140 comment pairs, to compare the performance of humans and several comparable methods.

LARGE-SET:

8400 comment pairs, to investigate relationships between the size of comment pairs and a functional relevance performance.

5.1 Test Set Construction

To construct the test sets, we randomly extracted 140 comment pairs for the SMALL-SET and 8400 for the LARGE-SET from the set of comment pairs described in Section 3.

Then, a half of their responses (Q) are randomly replaced by the previous comment (Q') in the same BBS. This produced a set of TRUE comment pairs (P, Q) and FALSE comment pairs (P, Q').

For the open-test setting, the test-set data were removed from both the SVM training-data and the functional relevance databases.

5.2 Comparable Methods

We used the following methods:

human-**A, B, and C**:

Three humans performed the judgments.

Overlap:

This method was based on a simple word overlap ratio. If this ratio was more than a threshold, the output was TRUE; otherwise it was FALSE.

Content:

This method used only a content relevance. If the content relevance value (REL_c) was more than a threshold, the output was TRUE; otherwise it was FALSE.

Functional:

This method used only a functional relevance. If the functional relevance value (REL_f) was more than a threshold, the output was TRUE; otherwise it was FALSE.

SVM:

This was a SVM-based method using both REL_c and REL_f as described in Section 3.3.

For the LARGE-SET, we could use neither *Content* nor *SVM* because they were based on $WEBPMI$, which requires a large number of Web search queries ($|P| \times |Q| \times \#$ of test set).

For SVM learning, we used TinySVM[4] with a linear kernel.

For detection of Japanese word boundaries, we used JUMAN[19].

5.3 SMALL-SET Results

Table 4 shows the results for the SMALL-SET. Because the performance of the similarity based methods (*Overlap*, *Content* and *Functional*) depended on the threshold, we examined the performance at the highest accuracy point.

Table 5 shows an agreement matrix for the various methods.

Human Upper Bounds

Human accuracy was only 70–79%, demonstrating the extreme difficulty of this task. This is due to two reasons:

False Positive:

Several short responses, such as "*I think so.*" or "*Thank you.*," are universal responses for various comments, leading to false positives.

[4] `http://chasen.org/t̃aku/software/TinySVM/`

Table 5: Agreement Ratio and Kappa Value Matrix.

	Human-A	Human-B	Human-C	*Overlap*	*Content*	*Functional*
Human-A	-	0.78 (0.56)$\oplus$	0.74 (0.49)$\oplus$	0.52 (0.08)$\ominus$	0.60 (0.20)	0.65 (0.28)
Human-B		-	0.73 (0.47)$\oplus$	0.54 (0.09)$\ominus$	0.60 (0.21)	0.62 (0.25)
Human-C			-	0.59 (0.15)$\ominus$	0.52 (0.05)$\ominus$	0.62 (0.25)
Overlap				-	0.63 (0.21)	0.45 (0.13)$\ominus$
Content					-	0.56 (0.16)$\ominus$

* The numbers in brackets indicate κ values. $\ominus$ is a "slight" correlation κ value. $\oplus$ κ is a "moderate" correlation κ value.

Table 4: Results in SMALL-SET.

	Accuracy (%)	Precision (%)	Recall (%)	$F_{\beta=1}$ ×100
human-A	79.28	83.33	75.34	79.13
human-B	75.71	78.26	73.97	76.05
human-C	70.71	71.62	72.6	72.10
Overlap	61.42	58.71	87.67	70.32
Content	61.42	72.09	42.46	53.44
Functional	65.71	66.23	69.86	67.99
SVM	63.28	64.44	79.45	72.10

False Negative:

Some conversation is too specialized or jargon-related for general human judgment, leading to false negatives. The following is one such example:

P: *Does anyone know the name of that song?*

Q: *I think Three Oranges.*

In this example, "*Three Oranges*" is a song title. Without this inside knowledge, a human cannot capture the relationship between the comments.

Due to the above human limitations, the human accuracy is not so high. However, the agreement shown in Table 5 indicates high κ values (moderate agreements), demonstrating that such limitations are almost equally shared. From these results, we can say that our task is difficult, but reasonable.

Independence between Two Relevances (*Content* versus *Functional*)

As shown in Table 4, *Functional* achieved higher accuracy than *Content* and *Overlap*.

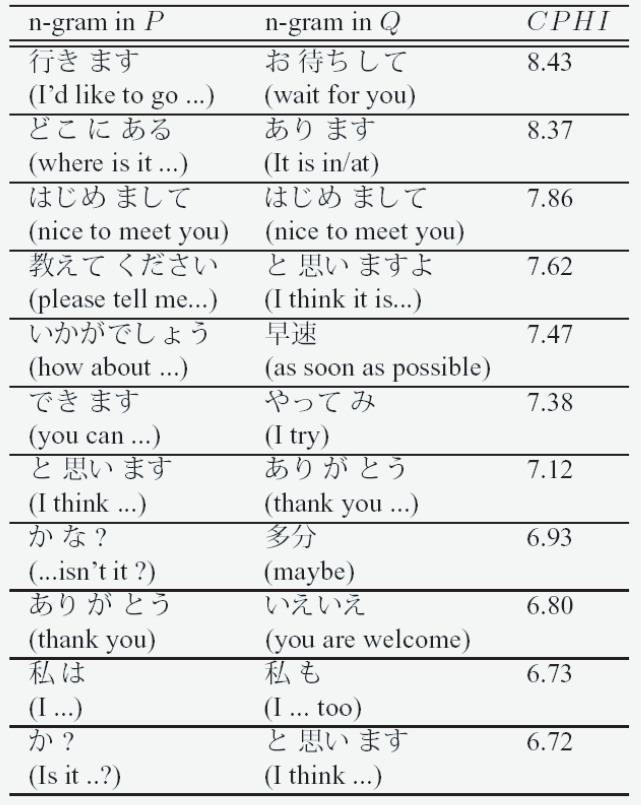
More importantly, although *Overlap* and *Content* had a fair agreement ($0.2 < \kappa < 0.4$) in Table 4, both of them had a slight agreement with *Functional* ($\kappa < 0.2$). From these results, we can conclude that *Content* (or *Overlap*) and *Functional* independent of each other, supporting our apparent that decomposes relevance into two measures.

Table 6 shows several examples of corresponding pairs that have high $CPMI$ values[5].

SVM-based Approach

The SVM-based approach showed an intermediate accuracy between *Content* and *Functional*. This result indicates that the SVM classifier could not integrate both relevance ,measures. We have yet to design suitable features that combine the advantages of both measures.

Table 6: Examples of Corresponding-pairs with high $CPMI$s.



n-gram in P	n-gram in Q	$CPHI$
行き ます (I'd like to go ...)	お 待ち して (wait for you)	8.43
どこ に ある (where is it ...)	あり ます (It is in/at)	8.37
はじめ まして (nice to meet you)	はじめ まして (nice to meet you)	7.86
教えて ください (please tell me...)	と 思い ますよ (I think it is...)	7.62
いかがでしょう (how about ...)	早速 (as soon as possible)	7.47
でき ます (you can ...)	やって み (I try)	7.38
と 思い ます (I think ...)	あり が とう (thank you ...)	7.12
か な ? (...isn't it ?)	多分 (maybe)	6.93
あり が とう (thank you)	いえいえ (you are welcome)	6.80
私 は (I ...)	私 も (I ... too)	6.73
か ? (Is it ..?)	と 思い ます (I think ...)	6.72

* The bracket indicates English translations.

5.4 LARGE-SET Results

Figure 6 shows the relationship between the training set size from 10 to 100% and the *Functional* performance. It is apparent that the accuracy is not saturated at the 100% point, indicating that the current data size was still not large enough, and that further studies are needed.

6. CONCLUSION

In this paper, we proposed a SVM-based classifier that determines if two comments make up a pair, a comment and

[5] The full list of the corresponding pairs will be available at `http://lab0.com/CPMI`.

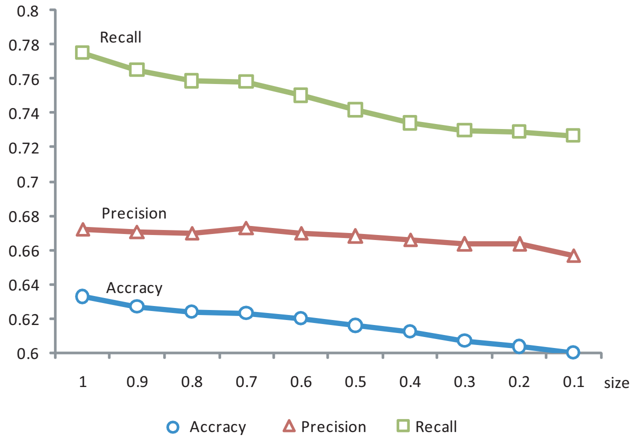


Figure 6: Training-set size (# of comment-pairs) and *Functional* Performance.

its response. We assumed that two consistencies were available: a relevance consistency and a discourse consistency. Experimental results empirically showed the individual contribution of each consistency, demonstrating the feasibility of the proposed approach. We believe that in the future, the proposed technique will contribute to analyzing an even greater volume of electronic dialogs.

7. ACKNOWLEDGMENTS

Part of this research is supported by Grant-in-Aid for Information Explosion Project Number: 19024072 (A01-0008)

8. REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[2] E. Aramaki, S. Kurohashi, H. Kashioka, and H. Tanaka. Word selection for ebmt based on monolingual similarity and translation confidence. In *Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003) Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 57–64, 2003.

[3] S. Bird and M. Liberman. Annotation graphs as a framework for multidimensional linguistic data analysis. In *Proceedings of Association for Computational Linguistics (ACL1999) Workshop on Towards Standards and Tools for Discourse Tagging*, pages 1–10, 1999.

[4] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of 16th International World Wide Web Conference (WWW 2007)*, pages 757–766, 2007.

[5] M. D. Boni and S. Manandhar. An analysis of clarification dialogue for question answering. In *Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2003)*, pages 48–55, 2003.

[6] P. F. Brown, J. C. Lai, and R. L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL1991)*, pages 169–176, 1991.

[7] P. F. Brown, S. A. D. Pietra, V. cent J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 1993.

[8] Carlson, D. Marcu, and M. E. Okurowski. Rst discourse treebank, 2002.

[9] S. F. Chen. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL1993)*, pages 9–16., 1993.

[10] C. Cherry and D. Lin. A probability model to improve word alignment. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2003)*, pages 88–95, 2003.

[11] M. G. Core and J. F. Allen. Coding dialogues with the DAMSL annotation scheme. In *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35. American Association for Artificial Intelligence, 1997.

[12] P. Fung and B. Chen. Biframenet: Bilingual frame semantics resource construction by cross-lingual induction. In *Proceedings of The International Conference on Computational Linguistics (COLING2004)*, pages 931–937, 2004.

[13] P. Fung and P. Cheung. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In D. Lin and D. Wu, editors, *Proceedings of EMNLP 2004*, pages 57–63. Association for Computational Linguistics, 2004.

[14] W. A. Gale and K. W. Church. A program for aligning sentences in billingual corpora. *Computational Linguistics*, 19(1), 1993.

[15] H. P. Grice. *Logic and conversation*. In Cole, P. and Morgan, J. (eds.) Syntax and semantics, vol 3. New York: Academic Press, 1975.

[16] M. Haruno and T. Yamazaki. High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL1996)*, pages 131–138, 1996.

[17] M. Kay and M. Roscheisen. Text-translation alignment. *Computational Linguistics*, 19(1), 1993.

[18] T.-L. Kueng and K.-Y. Su. A robust cross-style bilingual sentences alignment mmodel. In *Proceedings of the International Conference on Computational Linguistics (COLING2002)*, pages 502–508, 2002.

[19] S. Kurohashi, T. Nakamura, Y. Matsumoto, and M. Nagao. Improvements of japanese morphological analyzer juman. pages 22–28, 1994.

[20] S. C. Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press, 2000.

[21] K. Matsumoto and H. Tanaka. Autonatic alignment of Japanese and English newspaper articles using an MT system and a bilingual company name dictionary. In *Proceedings of the international conference on Language Resources and Evaluation (LREC2002)*, pages 480–484, 2002.

[22] A. Menezes and S. D. Richardson. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2001) Workshop on Data-Driven Methods in Machine Translation*, pages 39–46, 2001.

[23] F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP1999)*, pages 20–28, 1999.

[24] F. J. Och, N. Ueffing, and H. Ney. A comparison of alignment models for statistical machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING2000)*, pages 1086–1090, 2000.

[25] C. Quirk, R. U. U., and A. Menezes. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of MT Summit XI*.

[26] K. Rajaraman and A. Tan. Topic detection, tracking and trend analysis using self-organizing neural networks. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2001)*, pages 102–107, 2001.

[27] J. M. Schultz and M. Liberman. Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of DARPA Broadcast News Workshop*, pages 189–192, 1999.

[28] K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP2008)*, pages 189–196, 2008.

[29] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Cambridge:Harverd University Press, 1986.

[30] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. V. Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Lingustics*, 26(3):340–373, 2000.

[31] T. Utsuro, H. Ikeda, M. Yamane, Y. Matsumoto, and M. Nagao. Bilingual text matching using bilingual dictionary and statistics. In *Proceedings of the International Conference on Computational Linguistics (COLING1994)*, pages 1076–1082, 1994.

[32] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

[33] F. Walls, H. Jin, S. Sista, and R. Schwartz. Topic detection in broadcast news. In *Proceedings of DARPA Broadcast News Workshop*, pages 193–198, 1999.

[34] H. Watanabe, S. Kurohashi, and E. Aramaki. Finding structural correspondences from bilingual parsed corpus for corpus-based translation. In *Proceedings of the International Conference on Computational Linguistics (COLING2000)*, pages 906–912, 2000.

[35] F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.

[36] D. Wu. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL1994)*, pages 80–87, 1994.

[37] B. Zhao and S. Vogel. Adaptive parallel sentences mining from web bilingual news collection. *icdm*, 00:745, 2002.