Abstract

The ability to accurately judge the similarity between sentences is important for dialog system development in various areas such as utterance verification, context reasoning, utterance clustering. However, standard text similarity measures fail when directly applied to dialog sentences which are usually very short and have many ungrammatical omissions and inversions. This paper presents a method for sentence similarity refining method using discourse similarity of dialog sentences. First, we propose a novel discourse similarity based on the dialog act taxonomy. Given discourse similarity, we then present a novel way of rescoring original sentence score by explicitly adding discourse score to it. Experiments on test data sets demonstrate that the proposed measure significantly outperforms traditional similarity scoring measures.

Index Terms: sentence similarity, discourse similarity, dialog act, dialog system

1. Introduction

In dialog system, dialog sentence similarity measures play an increasingly important role. Various applications in dialog system are based on sentence similarity; for accept/rejecting speech recognition result comparing with expected sentences, for collecting paraphrase sentences from web or other dialog resource, and for clustering semantically similar sentences for rapid dialog act labeling. The widely-used techniques for measuring sentence similarity are based on co-occurred words in two sentences. However, in dialog sentence, word co-occurrence may be rare or even null, because dialog sentences are usually very short. Due to the data sparseness and the lack of context, it is difficult to determine semantically equivalence on short dialog sentences, and thus directly using standard sentence similarity measures may lead to only limited performances for dialog system.

To address these drawbacks, this paper aims to develop a dialog sentence similarity measuring method. Our work consist of the following two main parts: 1) computing discourse similarity and 2) developing a novel way of rescoring sentence similarity. First, we compute discourse similarity by newly deriving simplified dialog act taxonomy based on SWBD-DAMSL [1]. Given discourse similarity, we further apply sentence similarity rescoring methods to original sentence score, in order to finally obtain discourse-equipped sentence similarity.

The paper is organized as follows: The next section reviews some related works briefly. Section 3 presents a new method for measuring dialog sentences. Section 4 presents a number of experiments on dialog sentence similarity. The paper concludes in Section 5 that, based on the test data set, our similarity method outperforms existing methods.

2. Related Works

Traditional sentence similarities fall into roughly five categories: word overlap, edit operation, TF-IDF, sequence alignment and linguistic measures. Word overlap measures are a family of combinatorial similarity measure that compute similarity score based on a number of words shared by two sentences. Jaccard similarity coefficient is a measure that compares the similarity between two feature sets. Edit operation based sentence similarity measures such as Levenshtein distance [2] are usually defined as the minimum number of edits needed to transform one sentence into the other, with the allowable edit operations such as insertion, deletion, or substitution of a single word. The standard TF-IDF based sentence similarity is defined as cosine similarity between vector representation of two sentences. Term weights are computed from TF-IDF score. However, the TF-IDF scores are likely to be noisy in dialog sentences due to the shortness. Needleman-Wunsch [3] and Smith-Waterman [4] algorithms are well known sentence similarities based on word alignment score. Linguistic measures utilize linguistic knowledge such as semantic relations between words and their syntactic compositions, to determine the similarity of sentences [5, 6].

In general, there is extensive literature on measuring the similarity between documents or long texts, but only few works have investigated on short text or sentences [6, 7] and very few works have dealt with the similarity on dialog sentences [8]. Li et al. suggested a semantic-vector approach to measure relatively short sentence similarity [6]. In their work, sentences were transformed into feature vectors consisting of individual words from sentence pair as a feature set. Term weights were then derived from the maximum semantic similarity score between words in the feature vector. They used WordNet for calculating word-word similarities. However, their benchmark test sentences are not dialog sentences, and consist of dictionary definitions of words. Ferri et al. used [6]'s method to match a multimodal dialog sentence and define the multimodal templates similarity [8]. Jung et al. used Needleman-Wunsch algorithms [3] for measuring similarity between simulated spoken utterances in user simulation problem [9].

All exiting works on similarity methods did not actively and explicitly used discourse information such as dialog acts (statement, question, appreciation, dialog opening/closing and so on). In this work, we propose a new discourse score calculating method based on dialog act ontology, and then present the integration method which intermediates base sentence score and discourse score to improve dialog sentence similarity accuracy.
3. Dialog Sentence Similarity

In this work, we propose a novel approach to calculate pairwise sentence similarity by interpolating traditional sentence similarities and discourse similarity of the sentence. Our general scoring approach is as follows:

\[
Score = (1 - w) \cdot BaseScore + w \cdot DiscourseScore
\]

where \(w\) is a discourse weight for interpolation, and \(BaseScore\) is the score from any traditional sentence similarity methods such as Levenshtein distance. To calculate the discourse score between two sentences, we used dialog act information of the sentences. Dialog act represents the meaning of an utterance at the level of illocutionary force. Thus, dialog act can be considered as a tag set that classifies utterances according to a combination of pragmatic, semantic, and syntactic criteria [1].

3.1. Dialog act taxonomy

There are many dialog act coding schemes around developed for different purposes. In this work, we used SWBD-DAMSL dialog acts [1] which is one of the most widely used dialog act tag set. DAMSL defines a set of primitive communicative actions in four layers: communicative status, information level, forward looking function and backward looking function. SWBD-DAMSL is a variant of DAMSL coding scheme. The tag set distinguishes 42 mutually exclusive utterances types organized as a flat coding scheme without multi-functionality (see [1] for more information). The distribution of the tags are very skewed. For example, only 0.04% of utterances are tagged as Apology. For the simplicity, we simplified SWBD-DAMSL dialog act tags in following rules:

- Exclude communicative-status dialog acts
  - Uninterpretable, Non-verbal, Abandoned, Self and 3rd-party talk
- Exclude a back-channel related dialog act
  - Acknowledge(Backchannel) (e.g. “Uh-huh.”)
- Exclude dialog acts with very low frequency (< 0.05%)
  - exception: Conventional-Opening and Reject

Convention-Opening and Reject tags are included exception-ally into tag sets even though their frequencies are lower than 0.05% since their pair dialog acts such as Conventional-Closing and Agree/Accept are included. In summary, only 14 tags are used for this work. The simplified dialog act scheme is shown in Figure 1.

3.2. Dialog act classification

The Switchboard Dialog Act corpus(SwDA)\(^1\) is used for dialog act classification training corpus. SwDA has more than 200,000 utterances with 42 dialog acts and speaker information. We removed ill-formed sentences and used the utterances which are tagged with the 14 simplified dialog acts. The summary of the training corpus is shown in Table 1.

As a dialog act classifier, we used maximum entropy classifier [10] with simple lexical features. Complex word sense features are not used for the training and testing. That is based on the following consideration: First, we wanted our sentence similarity method to be as simple as possible and not too demanding in terms of computing resources. Second, part of speech tagging and parsing results are not satisfactory on dialog sentences which have many ungrammatical omissions and inversions. The features used in dialog act classification are as follows:

- Bag of words(1-gram, 2-gram and 3-gram)
- Sentence beginning and ending words(1-gram, 2-gram and 3-gram)

The performance of the dialog act classification is 74.50% accuracy in 10-fold cross validation.

3.3. Discourse similarity

Given two sentences, \(s_1\) and \(s_2\), we need to find the discourse similarity \(S_d(s_1, s_2)\). We can do this by measuring the similarity of the dialog acts of two sentences on the dialog act taxonomy. One direct method for the similarity calculation is to find the minimum length of path connecting two dialog acts. For example, the length of shortest path between “Do you have to have any special training? (Yes-No-Question)” and “Yes (Yes_Answers)” is 6. Using the length of path, we set discourse score \(S_d\) as:

\[
S_d(s_1, s_2) = S_d(d_1, d_2) = e^{-\alpha l(1)}
\]

\(^1\)http://www.stanford.edu/~jurafsky/swb1_dialogact_annot.tar.gz

Figure 1: Simplified SWBD-DAMSL dialog act tag sets. Leaf nodes are dialog act tags.
Table 1: The 14 dialog act labels of 93,070 sentences

<table>
<thead>
<tr>
<th>Dialog Act</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement-non-opinion</td>
<td>47,478</td>
<td>51.01</td>
</tr>
<tr>
<td>Statement-opinion</td>
<td>15,198</td>
<td>16.33</td>
</tr>
<tr>
<td>Agree/Accept</td>
<td>10,520</td>
<td>11.30</td>
</tr>
<tr>
<td>Appreciation</td>
<td>4,452</td>
<td>4.78</td>
</tr>
<tr>
<td>Yes-No-Question</td>
<td>3,555</td>
<td>3.82</td>
</tr>
<tr>
<td>Yes_answers</td>
<td>2,914</td>
<td>3.13</td>
</tr>
<tr>
<td>Conventional-closing</td>
<td>2,378</td>
<td>2.56</td>
</tr>
<tr>
<td>Wh-Question</td>
<td>1,514</td>
<td>1.63</td>
</tr>
<tr>
<td>No_answers</td>
<td>1,320</td>
<td>1.42</td>
</tr>
<tr>
<td>Response_Acknowledgment</td>
<td>1,277</td>
<td>1.37</td>
</tr>
<tr>
<td>Hedge</td>
<td>1,020</td>
<td>1.09</td>
</tr>
<tr>
<td>Declarative_Yes-No-Question</td>
<td>911</td>
<td>0.98</td>
</tr>
<tr>
<td>Reject</td>
<td>313</td>
<td>0.34</td>
</tr>
<tr>
<td>Conventional-opening</td>
<td>220</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Table 2: Summary of test data set used in the experiment.

<table>
<thead>
<tr>
<th>Summary</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of sentences</td>
<td>1,243</td>
</tr>
<tr>
<td>Average number of words in sentence</td>
<td>8.06</td>
</tr>
<tr>
<td>Average number of characters in sentence</td>
<td>33.83</td>
</tr>
<tr>
<td>Number of unique words</td>
<td>524</td>
</tr>
<tr>
<td>Percentage of unique words covered by WordNet</td>
<td>85%</td>
</tr>
<tr>
<td>Average number of similar sentences for a query</td>
<td>17.75</td>
</tr>
<tr>
<td>Total number of paraphrase set</td>
<td>70</td>
</tr>
</tbody>
</table>

where $\alpha \in [0, 1]$ is a length factor, and $l$ is the length of the shortest path between dialog act $d_1$ of sentence $s_1$ and $d_2$ of sentence $s_2$. In this work, we use $\alpha = 0.6$ as it gave the best performance in the training data set.

### 3.4. Scoring methods

The sentence similarity can be extended by considering the confidence of dialog act classification in the process of interpolation. The discourse factor $\sigma$ can be adjusted by the confidence of dialog act classification as follows:

$$\sigma_c = \sigma \cdot P(d_1|s_1) \cdot P(d_2|s_2)$$

where $\sigma \in [0, 1]$ is the initial discourse factor. In other words, the discourse factor can vary according to the confidence of dialog act classification. If the confidence is high (close to 1.0), then the new discourse factor $\sigma_c$ is the same as initial discourse weight $\sigma$. If the confidence is very low (close to 0.0), then the discourse similarity is not considered in sentence similarity scoring. In summary, we can think two different discourse scoring methods as follows:

$$S_1 = (1 - \sigma)S_h(s_1, s_2) + \sigma S_d(s_1, s_2)$$

$$S_2 = (1 - \sigma_c)S_h(s_1, s_2) + \sigma_c S_d(s_1, s_2)$$

where $S_h$ is the similarity between sentence $s_1$ and $s_2$ which are computed by using a traditional similarity method such as Levenshtein distance.

### 4. Experimental Evaluation

#### 4.1. Test data set

Although a few related studies have been published, there are currently no suitable benchmark data sets for the evaluation of dialog sentence similarity methods. In order to evaluate our similarity measure, a data set of paraphrase sentence sets is constructed. 1,243 unique sentences are collected from human to human real dialogs in traveling guide domain (immigration, hotel reservation, hotel check-in and city tour guide). Sentences are grouped into 70 different paraphrase set. Each paraphrase set consists of semantically equivalent, but lexically different sentences. Table 2 shows the summary of the data set.

#### 4.2. Evaluation Criteria

To evaluate the performance of sentence similarity measures, we cast our task as a retrieval/ranking problem, where given a sentence, the task is to propose a ranked list of sentences, sorted by sentence similarity. That is, each sentence is queried to find most similar(relevant) sentences from 1,243 sentences like traditional information retrieval task. We used the following ranked evaluation criteria for evaluating sentence similarity measurement. Precision-at-k is a precision of top k relevant sentences. R-precision is a precision at R-th position in the ranking of results for a query(given sentence) that has R relevant sentences(paraphrase). Mean average precision(MAP) is a mean of the average precision scores for each query. The mean reciprocal rank(MRR) is the average of the reciprocal ranks of results for a sample of queries.

#### 4.3. Similar sentence ranked retrieval results

To find out the optimal discourse factor $\sigma$, we analyzed the similarity performances with various $\sigma$ settings. Figure 2 shows that R-precisions are best around $\sigma = 0.1$. In this way, we empirically found 0.1 for discourse factor weight.

Table 3 shows overall contribution of discourse similarity on various different sentence similarity measures($Jc=$Jaccard coefficient, $JA=$Jaro distance, $JW=$Jaro-Winkler, $Le=$Levenshtein, $NW=$Needleman-Wunsch, $SW=$Smith-Waterman and $Li =$ the method of Li et al. [6]) . In the sense of R-precision and MAP, there are clear
improvements against the baseline in all similarity measuring methods. It is worth giving some consideration to the small performance improvement of $S_2$ against $S_1$. $S_1$ simply interpolates the given similarity and discourse similarity with fixed discourse factor ($\sigma = 0.1$) while $S_2$ uses variable discourse factor ($\sigma = P(d_1|s_1) \cdot P(d_2|s_2)$) for each sentence pair. It means that the confidence considered similarity - $S_2$ - is relatively more reliable measure against dialog act classification errors. In the sense of MRR, there are little or no improvements from the baseline. It is because that the most similar sentence (1st rank) for the query sentence has very similar lexical forms of query sentence, so any baseline similarity measures can retrieve it as first one.

Figure 3 shows precision and recall curves of baseline and confidence considered method $S_2$ in different base similarity measures. It shows that the proposed discourse-based similarity method outperforms the baseline measures in all range of recall and precision.

Table 4 shows the example of similar sentence retrieved results by Needleman-Wunsch sentence alignment algorithms. It is worth to notice the difference between baseline results and discourse score included $S_2$ results. The baseline algorithm gives high scores to the sentences which have overlapped part "I would like to...". On the other hand, $S_2$ gives high scores to the sentences which have same discourse role "Appreciation".

### 5. Conclusions

This paper presented a method for sentence similarity adaptation using discourse similarity between dialog sentences. First, we simplified SWBD-DAMSL dialog act tag sets for sensing discourse role of sentence, and then proposed a novel discourse similarity based on the dialog act taxonomy using the length of the shortest path between two dialog act tags. Next, we derived a novel sentence similarity by explicitly adding discourse-level score to existing word-based sentence similarity methods. In addition, noting that the classification of dialog act has numerous errors, we further present two sentence similarity scores $S_1$ and $S_2$, where the latter is reflecting a confidence of classification while the former is not. For evaluation, we cast the problem of finding similar sentences as a retrieval task on dialog sentences. The experiment results show that the proposed discourse scores actually contribute to improve sentence similarity performance, outperforming all the baseline similarity scores. Several further works are possible, which include the development of more sophisticated discourse similarity based on dialog act taxonomy, and semantic similarity calculation on phrase level, and so on.

### 6. References

2. V. Lvenshtin, “Binary coors capable or @correcting deletions, insertions, and reversals,” in *Soviet Physics-Doklady*, vol. 10, no. 8, 1966.

