EDU-Based Similarity for Paraphrase Identification

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{bachnx,nguyenml,shimazu}@jaist.ac.jp

Abstract. We propose a new method to compute the similarity between two sentences based on elementary discourse units, EDU-based similarity. Unlike conventional methods, which directly compute similarities based on sentences, our method divides sentences into discourse units and uses them to compute similarities. We also show the relation between paraphrases and discourse units, which plays an important role in paraphrasing. We apply our method to the paraphrase identification task. By using only a single SVM classifier, we achieve 93.1% accuracy on the PAN corpus, a large corpus for detecting paraphrases.

Keywords: Paraphrase Identification, Elementary Discourse Unit, Text Similarity, MT Metrics.

1 Introduction

Paraphrase identification is the task of determining whether two sentences have essentially the same meaning. This task has been shown to play an important role in many natural language applications, including text summarization [4], question answering [15], machine translation [6], and plagiarism detection [34]. For example, detecting paraphrase sentences would help a text summarization system avoid adding redundant information.

Although the paraphrase identification task is defined in the term of semantics, it is usually modeled as a binary classification problem, which can be solved by training a statistical classifier. Many methods have been proposed for identifying paraphrases. These methods usually employ the similarity between two sentences as features, which are computed based on words [10,16,21,25], n-grams [11,21], syntactic parse trees [11,30,33], WordNet [21,25], and MT metrics, the automated metrics for evaluation of translation quality [17,23].

Recently, several studies have shown that discourse structures deliver important information for paraphrase computation. For example, to extract paraphrases, Dolan et al. [14] take the first sentences from comparable documents and consider them as paraphrases. Regneri and Wang [29] introduce a method for collecting paraphrases based on the sequential event order in the discourse.

E. Métais et al. (Eds.): NLDB 2013, LNCS 7934, pp. 65–76, 2013.

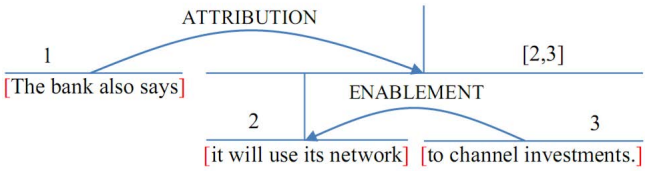

Fig. 1. An example of a discourse tree in RST-DT

However, they only consider some special kinds of data, which the discourse structures can be easily achieved.

Complete discourse structures like in the RST Discourse Treebank (RST-DT) [7] are difficult to achieve though they can be very useful for paraphrase computation [29]. In order to produce such complete discourse structures for a text, we first segment the text into several elementary discourse units (EDUs) (discourse segmentation step). Each EDU may be a simple sentence or a clause in a complex sentence. Consecutive EDUs are then put in relation with each other to create a discourse tree (discourse tree building step) [24]. An example of a discourse tree with three EDUs is shown in Figure 1. Existing full automatic discourse parsing systems are neither robust nor very precise [3,29]. Recently, however, several discourse segmenters with high performance have been introduced [2,19]. The discourse segmenter described in Bach et al. [2] gives 91.0% in the F_1 score on the RST-DT corpus when using Stanford parse trees [20].

In this paper, we present a new method to compute the similarity between two sentences based on elementary discourse units (EDU-based similarity). We first segment two sentences into several EDUs using a discourse segmenter, which is trained on the RST-DT corpus. These EDUs are then employed for computing the similarity between two sentences. The key idea is that for each EDU in one sentence, we try to find the most *similar* EDU in the other sentence and compute the similarity between them. We show how our method can be applied to the paraphrase identification task. Experimental results on the PAN corpus [23] show that our method is effective for the task. To our knowledge, this is the first work that employs discourse units for computing similarity as well as for identifying paraphrases.

The rest of this paper is organized as follows. We first present related work and our contributions in Section 2. Section 3 describes the relation between paraphrases and discourse units. Section 4 presents our method, EDU-based similarity. Experiments on the paraphrase identification task are described in Section 5. Finally, Section 6 concludes the paper.

2 Related Work and Our Contributions

There have been many studies on the paraphrase identification task. Finch et al. [17] use some MT metrics, including BLEU [28], NIST [13], WER [26], and

PER [22] as features for a SVM classifier. Wan et al. [36] combine BLEU features with some others extracted from dependency relations and tree edit-distance. They also take SVMs as the learning method to train a binary classifier. Mihalcea et al. [25] use pointwise mutual information, latent semantic analysis, and WordNet to compute an arbitrary text-to-text similarity metric. Kozareva and Montoyo [21] employ features based on longest common subsequence (LSC), skip n-grams, and WordNet. They use a meta-classifier composed of SVMs, k-nearest neighbor, and maximum entropy models. Rus et al. [30] adapt a graph-based approach (originally developed for recognizing textual entailment) for paraphrase identification. Fernando and Stevenson [16] build a matrix of word similarities between all pairs of words in both sentences. Das and Smith [11] introduce a probabilistic model which incorporates both syntax and lexical semantics using quasi-synchronous dependency grammars for identifying paraphrases. Socher et al. [33] describe a joint model that uses the features extracted from both single words and phrases in the parse trees of the two sentences.

Most recently, Madnani et al. [23] present an investigation of the impact of MT metrics on the paraphrase identification task. They examine 8 different MT metrics, including BLEU [28], NIST [13], TER [31], TERP [32], METEOR [12], SEPIA [18], BADGER [27], and MAXSIM [8], and show that a system using nothing but some MT metrics can achieve state-of-the-art results on this task. In our work, we also employ MT metrics as features of a paraphrase identification system. The method of using them, however, is very different from the method in previous work.

Discourse structures have only marginally been considered for paraphrase computation. Regneri and Wang [29] introduce a method for collecting paraphrases using discourse information on a special type of data, TV show episodes. With such kind of data, they assume that discourse structures can be achieved by taking sentence sequences of recaps. Our work employ the recent advances in discourse segmentation. Hernault et al. [19] present a sequence model for segmenting texts into discourse units using Conditional Random Fields. Bach et al. [2] introduce a reranking model for discourse segmentation using subtree features. Two segmenters achieve 89.0% and 91.0%, respectively, in the F_1 score on RST-DT when using Stanford parse trees.

The aim of our work is to exploit discourse information for computing paraphrases in general texts. Our main contributions can be summarized in the following points:

1. We show the relation between discourse units and paraphrasing, in which discourse units play an important role in paraphrasing.
2. We present EDU-based similarity, a new method for computing the similarity between two sentences based on elementary discourse units.
3. We apply the method to the task of paraphrase identification.
4. We conduct experiments on the PAN corpus [23] to show that EDU-based similarity is effective for the task of identifying paraphrases.

[Or his needful holiday has come,]$_{1A}$ [and he is staying at a friend's house,]$_{1B}$ [or is thrown into new intercourse at some health-resort.]$_{1C}$

[Or need a holiday has come,]$_{2A}$ [and he stayed in the house of a friend]$_{2B}$ [or disposed of in a new relationship to a health resort.]$_{2C}$

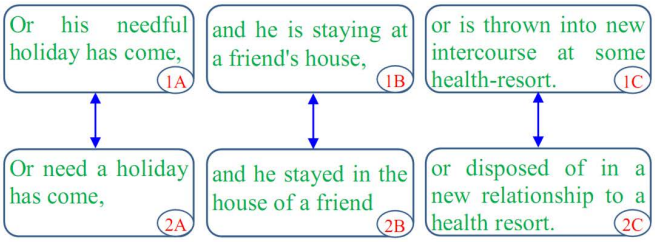

Fig. 2. A paraphrase sentence pair in the PAN corpus [23]

3 Paraphrases and Discourse Units

In this section, we describe the relation between paraphrases and discourse units. We will show that discourse units are blocks which play an important role in paraphrasing.

Figure 2 shows an example of a paraphrase sentence pair. In this example, the first sentence can be divided into three elementary discourse units (EDUs), 1A, 1B, and 1C, and the second sentence can also be segmented into three EDUs, 2A, 2B, and 2C. Comparing these six EDUs, we can see that they make three aligned pairs of paraphrases: 1A with 2A, 1B with 2B, and 1C with 2C. Therefore, if we consider the first sentence is the original sentence, the second sentence can be created by paraphrasing each discourse unit in the original sentence.

Figure 3 shows a more complex case. The first sentence consists of four EDUs, 3A, 3B, 3C, and 3D; and the second sentence includes four EDUs, 4A, 4B, 4C, and 4D. In this case, if we consider the first sentence is the original one, we have some remarks:

- The discourse unit 4A is a paraphrase of the discourse unit 3B,
- The unit 4B is a paraphrase of the combination of two units, 3A and 3C, and
- The combination of two units 4C and 4D is a paraphrase of the unit 3D.

By analyzing paraphrase sentences, we found that discourse units are very important to paraphrasing. In many cases, a paraphrase sentence can be created by applying the following operations to the original sentence:

1. Reordering two discourse units,
2. Combining two discourse units into one unit,
3. Dividing one discourse unit into two units, and
4. Paraphrasing a discourse unit.

[Age of consent legislation,]$_{3A}$ [as applied to the question of social vice,]$_{3B}$ [is one thing,]$_{3C}$ [and consent as applied to the question of slavery , quite another thing.]$_{3D}$

[When applied to social vices,]$_{4A}$ [age of consent legislation is one thing,]$_{4B}$ [when the legislation is applied to slavery,]$_{4C}$ [a totally different and epidemic problem exists.]$_{4D}$

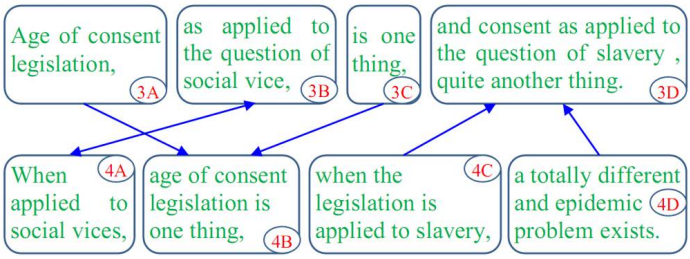

Fig. 3. Another paraphrase sentence pair in the PAN corpus

An example of Operation 1 and Operation 2 is the case of units 3A, 3B, and 3C in Figure 3 (reordering 3A and 3B, and then combining 3A and 3C). Unit 3D illustrates an example for Operation 3. The last operation is the most important operation, which is applied to almost all of discourse units.

4 EDU-Based Similarity

Motivated from the analysis of the relation between paraphrases and discourse units, we propose a method to compute the similarity between two sentences. Our method considers each sentence as a sequence of EDUs.

First, we present the notion of *ordered similarity functions*. Given two arbitrary texts t_1 and t_2, an ordered similarity function $Sim_{ordered}(t_1, t_2)$ will return a real score, which measures how t_1 is similar to t_2. Note that in this function, the roles of t_1 and t_2 are different, in which t_2 can be seen as a *gold* standard and we want to evaluate t_1 based on t_2. Examples of ordered similarity functions are MT metrics, which evaluate how a hypothesis text (t_1) is similar to a reference text (t_2).

Given an ordered similarity function $Sim_{ordered}$, we can define the similarity between two arbitrary texts t_1 and t_2 as follows:

$$Sim(t_1, t_2) = \frac{Sim_{ordered}(t_1, t_2) + Sim_{ordered}(t_2, t_1)}{2}. \quad (1)$$

Let (s_1, s_2) be a sentence pair, then s_1 and s_2 can be represented as sequences of elementary discourse units: $s_1 = (e_1, e_2, \ldots, e_m)$ and $s_2 = (f_1, f_2, \ldots, f_n)$, where m and n are the numbers of discourse units in s_1 and s_2, respectively. We define an ordered similarity function between s_1 and s_2 as follows:

$$Sim_{ordered}(s_1, s_2) = \sum_{i=1}^{m} Imp(e_i, s_1) * Sim_{ordered}(e_i, s_2) \quad (2)$$

where $Imp(e_i, s_1)$ is the importance of the discourse unit e_i in the sentence s_1, and $Sim_{ordered}(e_i, s_2)$ is the ordered similarity between the discourse unit e_i and the sentence s_2.

In this work, we simply consider that all words contribute equally to the meaning of the sentence. Therefore, the importance function can be computed as follows:

$$Imp(e_i, s_1) = \frac{|e_i|}{|s_1|} \tag{3}$$

where $|e_i|$ and $|s_1|$ are the lengths (in words) of the discourse unit e_i and the sentence s_1, respectively.

The ordered similarity $Sim_{ordered}(e_i, s_2)$ is computed based on the discourse unit f_j in the sentence s_2, which is the most similar to e_i:

$$Sim_{ordered}(e_i, s_2) = Max_{j=1}^{n} Sim_{ordered}(e_i, f_j). \tag{4}$$

Substituting (3) and (4) into (2) we have:

$$Sim_{ordered}(s_1, s_2) = \sum_{i=1}^{m} \frac{|e_i|}{|s_1|} Max_{j=1}^{n} Sim_{ordered}(e_i, f_j). \tag{5}$$

Finally, from (5) and (1) we have the formula for computing the similarity between two sentences based on their discourse units (EDU-based similarity), where the ordered similarity between two units is computed directly using the definition of the ordered similarity function, as follows:

$$\begin{aligned} Sim(s_1, s_2) &= \frac{Sim_{ordered}(s_1, s_2) + Sim_{ordered}(s_2, s_1)}{2} \\ &= \frac{1}{2} * \sum_{i=1}^{m} \frac{|e_i|}{|s_1|} * Max_{j=1}^{n} Sim_{ordered}(e_i, f_j) \\ &+ \frac{1}{2} * \sum_{j=1}^{n} \frac{|f_j|}{|s_2|} * Max_{i=1}^{m} Sim_{ordered}(f_j, e_i). \end{aligned} \tag{6}$$

We now present an example of computing the EDU-based similarity between two sentences in Figure 2 using the BLEU score. Table 1 shows the basic information of the calculation step by step. Line 1 and line 2 present two tokenized sentences and their lengths in words. Lines 3 through 5 compute the similarity between two sentences directly based on sentences. By using this method, the similarity is 0.5332. Elementary discourse units of two sentences are shown in lines 6 through 11. The computation of EDU-based similarity is described in lines 12 through 20. By using this method, the similarity is 0.5369, which is slightly higher than the similarity computed directly using sentences.

5 Experiments

This section describes our experiments on the paraphrase identification task using EDU-based similarities as features for an SVM classifier [35]. Like the

Table 1. An example of computing sentence-based and EDU-based similarities

Line	Computation		
1	s_1: Or his needful holiday has come , and he is staying at a friend 's house , or is thrown into new intercourse at some health-resort .		Length=27
2	s_2: Or need a holiday has come , and he stayed in the house of a friend , or disposed of in a new relationship to a health resort .		Length=29
	Sentence-based Similarity		
3	BLEU$(s_1, s_2) =$ **0.5333**		
4	BLEU$(s_2, s_1) =$ **0.5330**		
5	$Sim(s_1, s_2) = \frac{BLEU(s_1,s_2)+BLEU(s_2,s_1)}{2} =$ **0.5332**		
	Discourse Units		
6	e_1: Or his needful holiday has come ,		Length=7
7	e_2: and he is staying at a friend 's house ,		Length=10
8	e_3: or is thrown into new intercourse at some health-resort .		Length=10
9	f_1: Or need a holiday has come ,		Length=7
10	f_2: and he stayed in the house of a friend ,		Length= 10
11	f_3: or disposed of in a new relationship to a health resort .		Length=12
	EDU-based Similarity		
12	BLEU$(e_1, f_1) =$ **0.7143**	BLEU$(e_1, f_2) = 0.0931$	BLEU$(e_1, f_3) = 0.0699$
13	BLEU$(e_2, f_1) = 0.1818$	BLEU$(e_2, f_2) =$ **0.5455**	BLEU$(e_2, f_3) = 0.0830$
14	BLEU$(e_3, f_1) = 0.0833$	BLEU$(e_3, f_2) = 0$	BLEU$(e_3, f_3) =$ **0.4167**
15	EDU_BLEU$(s_1, s_2) = \frac{7}{27} * 0.7143 + \frac{10}{27} * 0.5455 + \frac{10}{27} * 0.4167 =$ **0.5416**		
16	BLEU$(f_1, e_1) =$ **0.7143**	BLEU$(f_1, e_2) = 0.1613$	BLEU$(f_1, e_3) = 0.0699$
17	BLEU$(f_2, e_1) = 0.1000$	BLEU$(f_2, e_2) =$ **0.5429**	BLEU$(f_2, e_3) = 0$
18	BLEU$(f_3, e_1) = 0.0833$	BLEU$(f_3, e_2) = 0.0833$	BLEU$(f_3, e_3) =$ **0.4167**
19	EDU_BLEU$(s_2, s_1) = \frac{7}{29} * 0.7143 + \frac{10}{29} * 0.5429 + \frac{12}{29} * 0.4167 =$ **0.5321**		
20	EDU_Sim$(s_1, s_2) = \frac{EDU_BLEU(s_1,s_2)+EDU_BLEU(s_2,s_1)}{2} =$ **0.5369**		

work of Madnani et al. [23], we employed MT metrics as the ordered similarity functions. However, we computed MT metrics based on EDUs in addition to MT metrics based on sentences. To segment sentences, we implemented the discourse segmenter described in Bach et al. [2]. In all experiments, parse trees were obtained by using the Stanford parser [20].

5.1 Data and Evaluation Method

We conducted experiments on the PAN corpus, a corpus for paraphrase identification task created from a plagiarism detection corpus [23]. Table 2 shows statistics on the corpus. The corpus includes a training set of $10,000$ sentence pairs and a test set of $3,000$ sentence pairs. On average, each sentence contains

Table 2. PAN corpus for paraphrase identification

	Training Set	**Test Set**
Number of sentence pairs	10,000	3,000
Number of EDUs per sentence	4.31	4.33
Number of words per sentence	40.07	41.12

about 4.3 discourse units, and about 40.1 words in the training set, 41.1 words in the test set. We chose this corpus for these reasons. First, it is a large corpus for detecting paraphrases. Second, it contains many long sentences. Our method computes similarities based on discourse units. It is suitable for long sentences with several EDUs. Last, according to Madnani et al. [23], the PAN corpus contains many realistic examples of paraphrases.

We evaluated the performance of our paraphrase identification system by accuracy and the F_1 score. The accuracy was the percentage of correct predictions over all the test set, while the F_1 score was computed only based on the paraphrase sentence pairs[1].

5.2 MT Metrics

We investigated our method with six different MT metrics (six types of ordered similarity functions). These metrics have been shown to be effective for the task of paraphrase identification [23].

1. BLEU [28] is the most commonly used MT metric. It computes the amount of n-gram overlap between a hypothesis text (the output of a translation system) and a reference text.
2. NIST [13] is a variant of BLEU using the arithmetic mean of n-gram overlaps. Both BLEU and NIST use exact matching. They have no concept of synonymy or paraphrasing.
3. TER [31] computes the number of edits needed to "fix" the hypothesis text so that it matches the reference text.
4. TERP [32] or TER-Plus is an extension of TER, that utilizes phrasal substitutions, stemming, synonyms, and other improvements.
5. METEOR [12] is based on the harmonic mean of unigram precision and recall. It also incorporates stemming, synonymy, and paraphrase.
6. BADGER [27], a language independent metric, computes a compression distance between two sentences using the Burrows Wheeler Transformation (BWT).

Among six MT metrics, TER and TERP compute a translation error rate between a hypothesis text and a reference text. Therefore, the smaller they are, the more similar the two texts are. When using these metrics in computing EDU-based similarities, we replaced the *max* function in Equation (6) by a *min* function.

[1] If we consider each sentence pair as an instance with label +1 for *paraphrase* and label -1 for *non-paraphrase*, the reported F_1 score was the F_1 score on label +1.

Table 3. Experimental results on each individual MT metric

	Sentence-based similarities		**+ EDU-based similarities**	
MT Metric	**Accuracy**(%)	$\mathbf{F}_1$(%)	**Accuracy**(%)	$\mathbf{F}_1$(%)
BLEU(1-4)	89.0	88.4	**89.6(+0.6)**	89.1(+0.7)
NIST(1-5)	84.6	82.7	**87.6(+3.0)**	86.8(+4.1)
TER	88.2	87.3	**88.5(+0.3)**	87.7(+0.4)
TERP	91.0	90.6	91.1(+0.1)	90.8(+0.2)
METEOR	90.0	89.6	89.8(-0.2)	89.4(-0.2)
BADGER	88.1	87.8	88.2(+0.1)	87.8(-)

5.3 Experimental Results

In all experiments, we chose SVMs [35] as the learning method to train a binary classifier[2].

First, we investigated each individual MT metric. To see the contributions of EDU-based similarities, we conducted experiments in two settings. In the first setting, we directly applied the MT metric to pairs of sentences to get the similarities (sentence-based similarities). In the second one, we computed EDU-based similarities in addition to the sentence-based similarities. Like Madnani et al. [23], in our experiments, we used BLEU1 through BLEU4 as 4 different features and NIST1 through NIST5 as 5 different features[3]. Table 3 shows experimental results in two settings on the PAN corpus. We can see that, adding EDU-based similarities improved the performance of the paraphrase identification system with most of the MT metrics, especially with NIST(3.0%), BLEU (0.6%), and TER (0.3%).

Table 4 shows experimental results with multiple MT metrics on the PAN corpus. With each MT metric, we computed the similarities in both methods, based directly on sentences and based on discourse units. We gradually added MT metrics one by one to the system. After adding the TERP metric, we achieved 93.1% accuracy and 93.0% in the F_1 score. Adding more two metrics METEOR and BADGER, the performance was not improved.

Two last rows of Table 4 shows the results of Madnani et al. [23] when using 4 MT metrics, including BLEU, NIST, TER, and TERP (Madnani-4) and when using all 6 MT metrics (Madnani-6)[4]. Compared with the best previous results, our method improves 0.8% accuracy and 0.9% in the F_1 score. It yields a 10.4% error rate reduction. Also note that, the previous work employs a meta-classifier with three constituent classifiers, Logistic regression, SVMs, and instance-based learning, while we use only a single classifier with SVMs.

We also investigated our method on long and short sentences. We divided sentence pairs in the test set into two subsets: Subset1 (long sentences) contains

[2] We conducted experiments on LIBSVM tool [9] with the RBF kernel.

[3] BLEUn and NISTn use n-grams.

[4] Madnani et al. [23] show that adding more MT metrics does not improve the performance of the paraphrase identification system.

Table 4. Experimental results on combined MT metrics

MT Metrics	Accuracy(%)	F_1(%)
BLEU	89.6	89.1
BLEU+NIST	91.2	90.9
BLEU+NIST+TER	91.8	91.6
BLEU+NIST+TER+TERP	**93.1**	**93.0**
Madnani-4	91.5	91.2
Madnani-6	92.3	92.1

Table 5. Experimental results on long and short sentences

Subset	#sent pairs	#EDUs/sent	#words/sent	Acc.(%)	F_1(%)
Subset1	1317	6.5	56.6	**96.6**	**94.8**
Subset2	1683	2.6	27.2	90.4	92.3

sentence pairs that both sentences have at least 4 discourse units[5], and Subset2 (short sentences) contains the other sentence pairs. Table 5 shows the information and experimental results on two subsets. Subset1 consists of 1317 sentence pairs (on average, 6.5 EDUs and 56.6 words per sentence), while Subset2 consists of 1683 sentence pairs (on average, 2.6 EDUs and 27.2 words per sentence). We can see that, our method was effective for the long sentences, which we achieved 96.6% accuracy and 94.8% in the F_1 score compared with 90.4% accuracy and 92.3% in the F_1 score of the short sentences.

6 Conclusion

In this paper, we proposed a new method to compute the similarity between two sentences based on elementary discourse units, EDU-based similarity. This method was motivated from the analysis of the relation between paraphrases and discourse units. By analyzing examples of paraphrases, we found that discourse units play an important role in paraphrasing. We applied EDU-based similarity to the task of paraphrase identification. Experimental results on the PAN corpus showed the effectiveness of the proposed method. To the best of our knowledge, this is the first work to employ discourse units for computing similarity as well as for identifying paraphrases. Although our method is proposed for computing the similarity between two sentences, it can be also used to compute the similarity between two arbitrary texts.

In the future, we would like to apply our method to other datasets for the paraphrase identification task as well as to other related tasks such as recognizing textual entailment [5] and semantic textual similarity [1]. Another direction is to improve the method of computing similarity, especially how to evaluate the

[5] Number 4 was chosen because on average each sentence contains about 4 EDUs (see Table 2).

importance of a discourse unit in a sentence. In this work, we simply consider that discourse units are independent and all words contribute equally to the meaning of the sentence. Therefore, the importance of discourse units is only calculated based on their lengths (in words). Exploiting the relations between discourse units for computing similarity may be an interesting direction.

Acknowledgements. This work was partly supported by the JAIST's Grant for Fundamental Research.

References

1. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In: Proceedings of SemEval, pp. 385–393 (2012)
2. Bach, N.X., Minh, N.L., Shimazu, A.: A Reranking Model for Discourse Segmentation using Subtree Features. In: Proceedings of SIGDIAL, pp. 160–168 (2012)
3. Bach, N.X., Le Minh, N., Shimazu, A.: UDRST: A Novel System for Unlabeled Discourse Parsing in the RST Framework. In: Isahara, H., Kanzaki, K. (eds.) JapTAL 2012. LNCS (LNAI), vol. 7614, pp. 250–261. Springer, Heidelberg (2012)
4. Barzilay, R., McKeown, K.R., Elhadad, M.: Information Fusion in the Context of Multi-Document Summarization. In: Proceedings of ACL, pp. 550–557 (1999)
5. Bentivogli, L., Dagan, I., Dang, H.T., Giampiccolo, D., Magnini, B.: The fifth Pascal Recognizing Textual Entailment Challenge. In: Proceedings of TAC (2009)
6. Callison-Burch, C., Koehn, P., Osborne, M.: Improved Statistical Machine Translation Using Paraphrases. In: Proceedings of NAACL, pp. 17–24 (2006)
7. Carlson, L., Marcu, D., Okurowski, M.E.: RST Discourse Treebank. Linguistic Data Consortium (LDC) (2002)
8. Chan, Y.S., Ng, H.T.: MAXSIM: A Maximum Similarity Metric for Machine Translation Evaluation. In: Proceedings of ACL-HLT, pp. 55–62 (2008)
9. Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27:1-27:27 (2011)
10. Corley, C., Mihalcea, R.: Measuring the Semantic Similarity of Texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp. 13–18 (2005)
11. Das, D., Smith, N.A.: Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition. In: Proceedings of ACL-IJCNLP, pp. 468–476 (2009)
12. Denkowski, M., Lavie, M.: Extending the METEOR Machine Translation Metric to the Phrase Level. In: Proceedings of NAACL, pp. 250–253 (2010)
13. Doddington, G.: Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In: Proceedings of the 2nd International Conference on Human Language Technology Research, pp. 138–145 (2002)
14. Dolan, B., Quirk, C., Brockett, C.: Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In: Proceedings of COLING, pp. 350–356 (2004)
15. Duboue, P.A., Chu-Carroll, J.: Answering the Question You Wish They had Asked: The Impact of Paraphrasing for Question Answering. In: Proceedings of NAACL, pp. 33–36 (2006)
16. Fernando, S., Stevenson, M.: A Semantic Similarity Approach to Paraphrase Detection. In: Proceedings of CLUK (2008)

17. Finch, A., Hwang, Y.S., Sumita, E.: Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence. In: Proceedings of the 3rd International Workshop on Paraphrasing, pp. 17–24 (2005)
18. Habash, N., Kholy, A.E.: SEPIA: Surface Span Extension to Syntactic Dependency Precision-based MT Evaluation. In: Proceedings of the Workshop on Metrics for Machine Translation at AMTA (2008)
19. Hernault, H., Bollegala, D., Ishizuka, M.: A Sequential Model for Discourse Segmentation. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 315–326. Springer, Heidelberg (2010)
20. Klein, D., Manning, C.: Accurate Unlexicalized Parsing. In: Proceedings of ACL, pp. 423–430 (2003)
21. Kozareva, Z., Montoyo, A.: Paraphrase Identification on the Basis of Supervised Machine Learning Techniques. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) FinTAL 2006. LNCS (LNAI), vol. 4139, pp. 524–533. Springer, Heidelberg (2006)
22. Leusch, G., Ueffing, N., Ney, H.: A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. In: Proceedings of MT Summit IX (2003)
23. Madnani, N., Tetreault, J., Chodorow, M.: Re-examining Machine Translation Metrics for Paraphrase Identification. In: Proceedings of NAACL-HLT, pp. 182–190 (2012)
24. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory. Toward a Functional Theory of Text Organization. Text 8, 243–281 (1988)
25. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In: Proceedings of AAAI, pp. 775–780 (2006)
26. Niessen, S., Och, F.J., Leusch, G., Ney., H.: An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: Proceedings of LREC (2000)
27. Parker, S.: BADGER: A New Machine Translation Metric. In: Proceedings of the Workshop on Metrics for Machine Translation at AMTA (2008)
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of ACL, pp. 311–318 (2002)
29. Regneri, M., Wang, R.: Using Discourse Information for Paraphrase Extraction. In: Proceedings of EMNLP-CONLL, pp. 916–927 (2012)
30. Rus, V., McCarthy, P.M., Lintean, M.C., McNamara, D.S., Graesser, A.C.: Paraphrase Identification with Lexico-Syntactic Graph Subsumption. In: Proceedings of FLAIRS Conference, pp. 201–206 (2008)
31. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the Conference of the Association for Machine Translation in the Americas, AMTA (2006)
32. Snover, M., Madnani, N., Dorr, B., Schwartz, R.: TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. Machine Translation 23(23), 117–127 (2009)
33. Socher, R., Huang, E.H., Pennington, J., Ng, A.Y., Manning, C.D.: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In: Advances in Neural Information Processing Systems 24 (NIPS), pp. 801–809 (2011)
34. Uzuner, O., Katz, B., Nahnsen, T.: Using Syntactic Information to Identify Plagiarism. In: Proceedings of the 2nd Workshop on Building Educational Applications using Natural Language Processing, pp. 37–44 (2005)
35. Vapnik, V.N.: Statistical Learning Theory. Wiley Interscience (1998)
36. Wan, S., Dras, R., Dale, M., Paris, C.: Using Dependency-Based Features to Take the "Para-farce" out of Paraphrase. In: Proceedings of the 2006 Australasian Language Technology Workshop, pp. 131–138 (2006)