Paraphrase Extraction from Comparable Corpus based on Document Structure Models

Wangjie Wang[1], **Makoto Miwa**[2], **Yoshimasa Tsuruoka**[1], **Takashi Chikayama**[1]
[1]School of Engineering, The University of Tokyo, Japan
[2]National Centre for Text Mining, the University of Manchester, UK
{wang-wj, tsuruoka, chikayama}@logos.t.u-tokyo.ac.jp
makoto.miwa@manchester.ac.uk

1 Introduction

Paraphrasing techniques, especially data-driven approaches heavily rely on paraphrase resources of different units of language, e.g. words, phrases, or sentences.

A lot of methods have been developed to extract paraphrases from various language resources. In most of the traditional methods, mutual or common patterns in texts such as n-grams occurrences (Barzilay and Lee, 2003), distributional similarity (Marton et al., 2009), or dependencybased features (Wan et al., 2006) are considered to be good indicators. However, these methods use only attributes of the target sentence itself, and some of them need plenty of training data to build a model. A recent study conducted by (Regneri and Wang, 2012) pointed out that the context of the sentence in drama plots, namely discourse information, is an important paraphrase identifier. They proved their simple but critical hypothesis that sentences in the same discourse context are more likely to be paraphrases, even if little similarity in traditional models is found.

Inspired by their work, we tried to solve the problem of paraphrase extraction from general comparable corpora instead of a specific source. This paper proposed a Document Structure (DS) Model for paraphrase extraction. We applied our model to a corpus consisting of semantically comparable essays, and managed to enhance paraphrase extraction systems based on the traditional similarity measure.

2 Paraphrase Extraction Using a Document Structure Model

Given a collection of documents,which we call corpus D. A document is a collection of sentences and their positions. Since for different doncuments in D, they may have different numbers of sentences. We first map the position of a sentence to a normalized form. The Normalized Position (NP) can be defined as:

$$NP(s) = \frac{index\ of\ s}{number\ of\ sentences\ in\ d}$$

$$0 < NP(s) \leq 1,\ \forall s \in D$$

Any sentence s in a document d can be completely represented by its semantic content $SE(s)$ and its position $NP(s)$.

$$s = (SE(s), NP(s)) = (se, np)$$

Consider a probabilistic model of a joint distribution by a pair of sentences $(s_1, s_2) \in d_1 \times d_2$, the probability of them being paraphrase pair in specific positions (np_1, np_2) can be inferred by Bayes's rule:

$$\begin{aligned} & P(se_1 = se_2 \mid np_1, np_2, d_1, d_2) \\ &= P_{d_{1,2}}(se_1 = se_2 \mid np_1, np_2) \\ &= \frac{P_{d_{1,2}}(se_1 = se_2) P_{d_{1,2}}(np_1, np_2 | se_1 = se_2)}{P_{d_{1,2}}(np_1, np_2)} \\ &\propto P_{d_{1,2}}(se_1 = se_2) P_{d_{1,2}}(np_1, np_2 | se_1 = se_2) \end{aligned} \quad (1)$$

In Equation 1, $P_{d_{1,2}}(se_1 = se_2)$ denotes a semantic similarity model. We ignore the relevance of the sentences semantic and the documents, simplifying it to a general sentence-to-sentence model which has been well studied, e.g. (Fernando and Stevenson, 2008). It can be approximated by a semantic similarity measure such as the following.

$$P_{d_{1,2}}(se_1 = se_2) \propto Sim_{sem}(SE(s_1), SE(s_2))$$

Our target is to calculate the other factor $P_{d_{1,2}}(np_1, np_2 | se_1 = se_2)$, which is also a probablistic model, distributed over the positions of candidate sentence pairs in d_1 and d_2.

Obviously, It is difficult to model document structure for arbitrary documents. However, for the corpus consisting of comparable documents, a sentence of particular semantics has a good chance to show up in a certain position of document. It is not uncommon to see a narration consisting of a series of events in a time based order, or an essay led by topic sentences, followed by several evidences, and ended up with conclusions. Thus an assumption can be given that in a comparable corpus, the position pairs subject to the same document structure distribution. If D here is a comparable corpora, we will have a Document Structure model as following:

$$\forall(s_1, s_2) \in \{d_1 \times d_2\} \subset \{D \times D\}$$

$$\begin{aligned} &P(np_1, np_2|se_1 = se_2, d_1, d_2) \\ &= P(np_1, np_2|se_1 = se_2, D) \\ &= P_D(np_1, np_2|se_1 = se_2) \end{aligned}$$

An intuitive way to estimate this model is directly sampling positions of paraphrase pairs. However, in such multidimensional case, we need to take plenty of samples to lower the risk. Therefore in practice, we need to reduce the dimension of variable. Taking the essays for example, we observed that the distance between a pair of sentences is a representative feature for document structure modeling. We define the relative distance (RD) of two sentences as the variable of document structure model.

$$RD(s_1, s_2) = |np_1 - np_2|$$

$$0 \leq RD(s_1, s_2) < 1 \text{ },\forall s_1, s_2 \in D$$

With all the positions (np_1, np_2) replaced by the relative distance, the target model can be written as:

$$\Downarrow P_D(np_1, np_2|se_1 = se_2)$$

$$P_D(RD(s_1, s_2)|se_1 = se_2) \triangleq \psi(RD)$$

The Kernel density estimation (Parzen, 1962) is applied to approximate the model.

$$\widehat{\psi}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), 0 \leq x < 1$$

An example of the estimated DS density curve is illustrated in Figure 1. In an essay based corpus, which will be explained in Section 3.1, the Gaussian kernel $K(u)$ is used to smooth the curve and the bandwidth h is set to 0.1. We find 78 samples out of 200 annotated pairs of sentences.

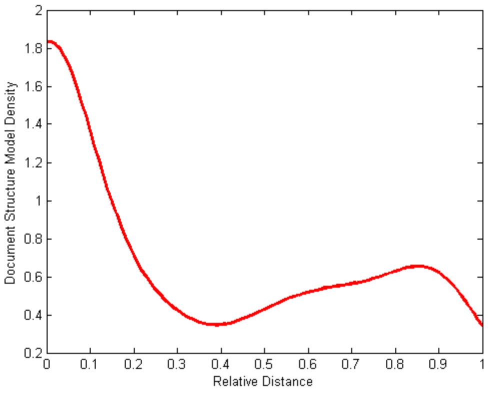

Figure 1: The density curve of an example DS model

Finally, by applying a semantic similarity measure, the overall similarity of (s_1, s_2) in Equation 1 can be calculated as the following:

$$\begin{aligned} &Sim_{overall}(SE(s_1) = SE(s_2))|RD(s_1, s_2)) \\ &= \widehat{\psi}(RD(s_1, s_2))Sim_{sem}(SE(s_1), SE(s_2)) \end{aligned} \tag{2}$$

3 Evaluation

3.1 Dataset Preparation

We use the International Corpus Network of Asian Learners of English[1](ICNALE) to make a dataset for evaluation. The ICNALE is one of the largest corpora of Asian leaners' English. It contains 1.3 million words of controlled essays written by 2,600 college students in 10 Asian regions and 200 English native speakers. The participants are asked to write essays on each of the given two topics. The length of each essay is between 200-300 words.

Since the strictly controlled format, it is easy to collect a dataset of comparable documents. To make full use of ICNALE, we choose the documents written by students from two different regions, the native English speakers (ENS) and Singaporean students (SIN). Among these documents, those of the topic of smoking in public place are selected. Hence we get a monolingual comparable corpus which is also a domain paralleled corpus. In the ENS domain, there are 1,959 sentences out of 200 essays, while in the SIN the number is 2,345, with the same essay number. The extracted paraphrases can also be a useful resourse

[1] http://language.sakura.ne.jp/icnale/index.html

to fill in the gap between the two domains. We evaluate our method on this dataset.

3.2 Evaluation Setting

To show the contribution of the DS model, we create three baseline systems which use the measures of TF-IDF (Jones, 1972), BLEU (Papineni et al., 2002) and matrixJcn (Fernando and Stevenson, 2008) to calculate the similarity of sentences. The first system accumulates TF-IDF weighted word co-occurrence as a score of similarity. The second system establishes the average 1-to-4-gram overlap of two sentences. In the third system, the similarity of words is evaluated by Jcn metric. The Scores are used to update a matrix indexed by the words of the candidate sentences, the similariy of sentences is defined as the maximum matched score of the matrix.

We use a development dataset to determine the threshold of each system, and for the matrixJcn based system we use the same threshold setting as they described in the paper.

3.3 Gold Standard

According to the size of the corpus, there are 4,593,855 candidate pairs in total, and only a small portion of them are expected to be paraphrases. Manually labeling all the sentences pairs is infeasible. In addition, random sampling would end up with few valid paraphrases.

We thus choose samples from both the output of baselines (300) and a random selection (100). The evaluation set consists of 400 sentence pairs without reduplication, 200 for model estimation and development (see Section 2) and 200 for test.

Sentence pairs are labeled to three categories. Follows are the examples of each category.

Category 1. Paraphrase

- Therefore, it is right to ban smoking, not only in restaurants, but in general on health and moral grounds.

- Therefore, it is right to ban smoking, not only in restaurants, but in general on health and moral grounds.

Category 2: Related

- Some may argue that *smokers have their right to do what they want* but if this right infringes on others who have the right to have fresh air, the smokers should just give in.

- However, these kinds of places are for adults, and *adults should be able to do what they want to do*

Category 3: Unrelated

- This way, the smokes would not affect the non-smoking customers.

- I don't know about Japan, but in Australia the government is always telling us what to do and we think that it would be nice if they asked us what we wanted sometimes.

Note that the category of related indicates that only a fragment of a candidate sentence can be matched to a paraphrase in the other sentence. Such annotation policy may necessarily result in additional unmatched content, or even the opposite semantics, but we think such related sentences are useful.

Firstly, the fragments can be extracted, which becomes paraphrase pairs at phrase level. Secondly, there are too few paraphrases perfectly matched at the sentence level. If we discard all the related pairs, we can hardly collect enough paraphrases to build a sentence aligned language resource.

Thus, in the evaluation, we prepared two test sets based on different partition. In the related paraphrase test set (T1), both paraphrases and related paraphrases are accepted, while in the paraphrase identification test set (T2), only paraphrases are accepted. The development set are used to determine the thresholds for baseline systems.

The inter-annotator agreement is measured by Cohen's Kappa coefficient (Cohen and others, 1960). In this annotation, the overall coefficient for the annotation of three categories is $k = 0.49$, which indicates a moderate agreement. For a conflict in annotation, two annotators codetermine it again to reach an agreement. Among all the gold standard sentence pairs, we found 53 paraphrases, 104 partial paraphrase cases, and 243 unrelated.

3.4 Results

Figure 2 lists the results of three systems on two test sets (T1 and T2), evaluated by F1score and accuracy. For each system, the red column shows the result cooperated with the DS model. We can observe from the result that all three systems have been improved by our model on both T1 and T2.

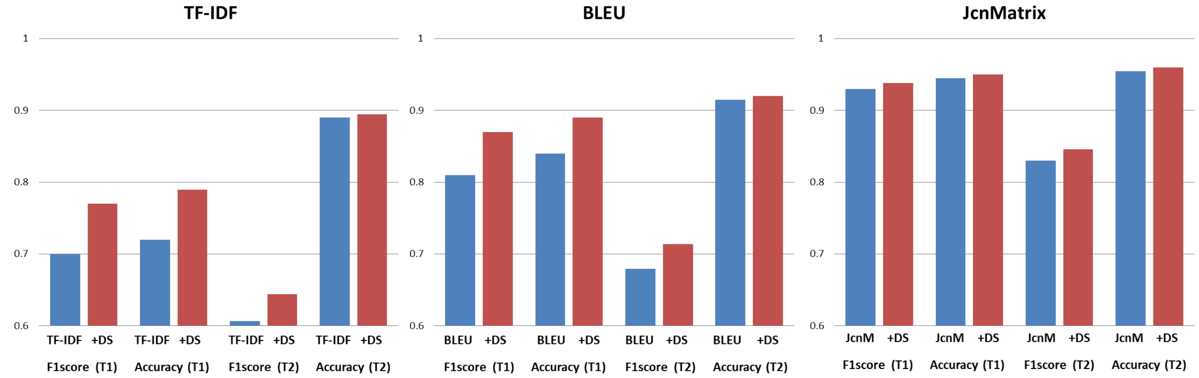

Figure 2: The improved performance of three systems by using the Ducument Structure model

However, when the performance of a system is already good, e.g. the Jcn based system, the improvement is not significant. This may due to the errors of the DS model generated in modeling, one-dimension simplification and estimation, which led to a limited prediction.

The results are not compared to other published results since the test set is collected from the system outputs, which results in a biased distribution of the samples.

4 Conclusion

In this paper, we addressed the problem of extracting paraphrases from comparable corpus. We proposed a model leveraging the document structure information. The empirical results showed that our model's capability of improving systems based on some existing similarity measures. The DS_Jcn based system extracted 51,680 related sentences pairs for the future research.

References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 16–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Cohen et al. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. *Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloqium.*

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 381–390, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Michaela Regneri and Rui Wang. 2012. Using discourse information for paraphrase extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 916–927, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen Wan, Mark Dras, Robert Dale, and Cecile Paris. 2006. Using dependency-based features to take the 'para-farce' out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 131–138, Sydney, Australia, November.