Extract Domain-specific Paraphrase from Monolingual Corpus for Automatic Evaluation of Machine Translation

Lilin Zhang, Zhen Weng, Wenyan Xiao, Jianyi Wan, Zhiming Chen, Yiming Tan, Maoxi Li, Mingwen Wang

School of Computer Information Engineering, Jiangxi Normal University

{1006806747, 1091013334, 51852710, 1363955817}@qq.com,

{qqchenzhiming, tt_yymm, mosesli, mwwang}@jxnu.edu.cn

Abstract

Paraphrase can help match synonyms or match phrases with the same or similar meaning, thus it plays an important role in automatic evaluation of machine translation. The traditional approaches extract paraphrase in general domain from bilingual corpus. Because the WMT16 metrics task consists of three subtasks, namely news domain, medical domain, and IT domain, we propose to extract domain-specific paraphrase tables from monolingual corpus to replace the general paraphrase table. We utilize the M-L approach to filter the large scale general monolingual corpus into a domain-specific sub-corpus, and exploit Markov Network model to extract paraphrase tables from the sub-corpus. The experimental results on WMT15 Metrics task show that METEOR metric using the domain-specific paraphrase tables outperforms that using the paraphrase table in general domain extracted from the bilingual corpus.

1 Introduction

Machine translation (MT) automatic evaluation metrics, such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee et al., 2005), TER (Snover et al., 2006), MAXSIM (Chan et al., 2008) etc., evaluate the quality of the MT system output by calculating the similarity between the translation output and the human reference. Accurately matching words or phrases with the same or similar meaning is critical to the performance of the automatic evaluation metrics (Li et al., 2013; Li et al., 2016).

Recently, many works enhanced traditional metrics by adding paraphrase match. For instance, in the latest version of METEOR package (Denkowski and Lavie, 2014), the paraphrase match was added after the standard exact word match, stem match and synonym match. And the latest version of TER package (Bannard et al., 2005) relaxes the condition of word match or chunk shift by adding paraphrase match. Note that the paraphrase tables used in latest METEOR and TER metrics belong to the general domain and they are extracted from bilingual parallel corpus by the Pivot approach (Bannard et al., 2005). However, the WMT16 metrics task consists of sub-tasks on specific domains involving several different languages. Confronted with the changes, we propose a Monolingual Paraphrase Extraction method based on Domain Adaptation (MPEDA), and use the new domain-specific paraphrase table to replace the traditional paraphrase tables in the latest METEOR package.

2 Related Work

In statistical natural language processing, both the scale and the quality of the training data have a direct impact on the performance of statistical learning. Take statistical MT for an example, if the size of training data is larger and the more it covers *n*-gram appeared in the test set, the quality of the MT outputs will be better.

To expand the scale of the existing domain-specific corpus, Moore and Lewis (2010) trained models with general corpus and domain-specific corpus, and computed cross entropy of each sentence in the general corpus to extract a sub-corpus much larger than the existing domain-specific corpus. In this way, a large scale domain-specific training corpus for statistical MT was established. Along this approach, Amittai et al. (2011) proposed a bilingual parallel data selection approach based on cross entropy to improve the MT performance for spoken language translation. And Juri et al. (2015) filtered training data for automatic extraction of paraphrase by using Moore and Lewis' approach to extract paraphrases from the filtered training data via the Pivot approach.

Automatically extracting paraphrases from the large scale corpus is low cost. Barzilay and McKeown (2001) presented an unsupervised learning approach to extract paraphrases of

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 511–517,
Berlin, Germany, August 11-12, 2016.

words and phrases from different English translations of the identical source language sentences. Bannard and Callison-Burch (2005) employed the word alignment technique of statistical MT to extract paraphrases from bilingual parallel corpus. Shinyama et al. (2002) used the named entity recognition features to extract paraphrases from monolingual comparable corpus. Barzilay and Lee (2003) used text strings alignment algorithm to learn paraphrases at sentence level from the unannotated comparable corpus. Yet, there are still great restrictions of the latter two monolingual paraphrase extraction methods. Therefore, we adopt the Markov-based method proposed by Weng et al. (2015) to extract paraphrases in specific domain from monolingual corpus because that it has no restrictions on monolingual corpus in the target language as it can extract paraphrase by constructing the Markov networks of words. Prior to the paraphrase extraction, we first filter large scale monolingual corpus into sub-corpus close to the domain of the human reference. Compared with general training corpus, the filtered sub-corpus is smaller and more related to the target domain, which results in the improvement on the quality of paraphrase table as well as the performance when the paraphrase table is applied in automatic evaluation metric.

3 MPEDA: Monolingual Paraphrase Extraction Based on Domain Adaptation

We extract domain-specific paraphrases from the monolingual corpus which are the most related to the test data. Our approach aims at accurately matching synonyms and phrases with the same or similar meaning in MT outputs and in human references with the help of the domain-specific paraphrase. We first filter a sub-corpus from a large general corpus by the extended M-L method, and then extract paraphrases based on Markov Network model and finally apply the paraphrase table to METEOR metric.

3.1 Extracting paraphrases based on word chunks

According to the Markov Network model, we first use the term co-occurrence in the text set to calculate the correlation among terms and construct a term Markov network where the correlation between two words in the network (edge weight) is computed by the joint conditional probability of two terms in the text set according to Formula (1) - (3), in which conditional probability $P(t_i|t_j)$ and $P(t_j|t_i)$ are not equal.

$$R(t_i, t_j) = \frac{P(t_i \mid t_j) + P(t_j \mid t_i)}{2} \quad (1)$$

$$P(t_i \mid t_j) = \frac{C(t_i, t_j)}{C(t_j)} \quad (2)$$

$$P(t_j \mid t_i) = \frac{C(t_i, t_j)}{C(t_i)} \quad (3)$$

In Formula (1) – (3), t_i and t_j stand for two terms, $C(t_i, t_j)$ is the number of documents that in the whole training data term t_i and term t_j co-occur in the same window, $C(t_i)$ and $C(t_j)$ denote the numbers of documents that term t_i and term t_j occur in the whole training data respectively, $R(t_i, t_j)$ denotes the correlation between term t_i and term t_j. The greater the R value, the higher the correlation between the two terms.

Extracting paraphrases from the constructed term Markov network is built on the following hypothesis: the more word chunks co-occurring between two terms, the more similar their semantic meanings are, and thus the two terms are a paraphrase pair. Therefore, we need to build an n-gram word chunk set for each term and then calculate the ratio between the number of co-occurring word chunks of two terms and the total number of word chunks with one term occurring. The ratio is considered as the possibility of the two terms constructing a paraphrase pair, which can be obtained by formula (4) - (6). Formula (6) is used to calculate the weight of n-gram word chunk.

$$pos(t_i, t_j) = \frac{W_3(t_i, t_j)}{\frac{1}{2}(W_3(t_i) + W_3(t_j))} \quad (4)$$

$$W_3(t_i, t_j) = \sum_{k \neq i \wedge k \neq j \wedge t_k \in clique(t_i, t_j, t_k)} w_3(t_i, t_j, t_k) \quad (5)$$

$$w_n\{t_1, t_2, \ldots t_n\} = \frac{\sum_{1 \leq i, j \leq n} R(t_i, t_j)}{\frac{1}{2} n(n-1)} \quad (6)$$

In the above formulas, $pos(t_i, t_j)$ is the paraphrase probability of term t_i and term t_j, $W_3(t_i, t_j)$ is the sum of weights of all the *3*-gram word chunks containing term t_i and term t_j, $W_3(t_i)$ is the sum of weights of all the *3*-gram word chunks containing term t_i, $W_3(t_j)$ denotes the sum of weights of all the *3*-gram word chunks containing term t_j, n denotes the number of nodes in word chunk, $R(t_i, t_j)$ denotes the correlation between term t_i and term t_j.

We use the terms co-occurrence to construct a term Markov network and extract phrases in the corpus as a node of Markov network. Figure 1 shows an example of *3*-gram word chunk, where t_1 stands for the term "*computer*", t_2 stands for the term "*Internet*", t_3 stands for the term "*calculating machine*", t_4 stands for the term "*electronic*". In this example, the *3*-gram word chunk set for each term is $S(C_3(t_1))$= {{ t_1, t_2, t_3}, {t_1, t_3, t_4}}, $S(C_3(t_2))$={ t_1, t_2, t_3}, $S(C_3(t_3)$= {{ t_1, t_2, t_3}, {t_1, t_3, t_4}}, $S(C_3(t_4))$={ t_1, t_3, t_4}. It can be observed that $S(C_3(t_1))$= $S(C_3(t_3)$= {{ t_1, t_2, t_3}, {t_1, t_3, t_4}}, hence, there is a high correlation between the two terms of t_1 and t_3 . Based on the hypothesis of this paper, we think term t_1, "*computer*", and term t_3, "*calculating machine*", in this example is a paraphrase pair.

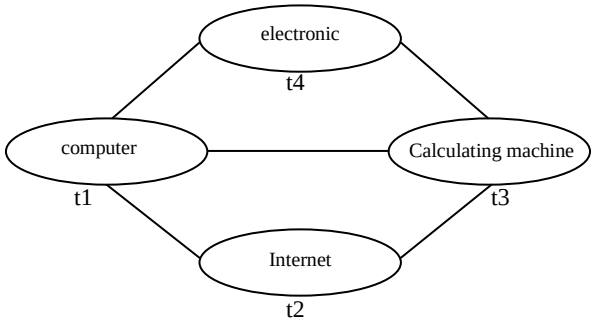

Figure 1: 3-gram word chunk

3.2 Corpus filtering

3.2.1 M-L corpus filtering

The corpus filtering method is built similar to the M-L method proposed by Moore and Lewis (2010). To extract a sub-corpus of target domain from the large general corpus, we first select a domain-specific corpus and a general large scale corpus. To improve the automatic MT metric, we use the human references of each sub-task in the metric tasks as the domain-specific corpus, and train the language model of the two corpora respectively, furthermore, we calculate the cross entropy of the two models. Finally, the similarity between the sentences and the human references is measured by calculating the difference of two cross entropy of the same sentence according to Formula (7). Generally, smaller value means the sentence is closer to the target domain.

$$\delta_{s_i} = H_{ref}(S_i) - H_{train}(S_i) \quad (7)$$

In formula (7), S_i denotes the *i*-th sentence, H_{ref} denotes the cross entropy of the language model trained from the human references, while H_{train} denotes the cross entropy of the language model trained from the training data.

3.2.2 Document sets filtering

The Markov network-based automatic paraphrase extraction approach requires divide a general monolingual corpus into different document sets. Weng et al. (2015) divided the text of a fixed length into a document without considering the correlation among documents. Hence, we form the sentences in the corpus into cluster via *K*-means clustering algorithm, and then use the bag of word model to create a vector for each sentence in the corpus. Thus the distance between two sentences can be obtained by calculating the cosine value of the two vectors. Each cluster is viewed as a document. In the process of clustering, dividing documents via *K*-means algorithm can guarantee that the sentences in a document approximately belong to the same domain.

Then, the M-L method is used to extract the sub-sets of documents which are close to the target domain from the clustered general document sets. This signifies that it is the document not the sentence that is regarded as the smallest filtering unit in the process of corpus filtering. And we want to identify documents which are similar to our target domain by summing up the difference of cross entropy of each sentence in the document. However, when dividing the large-scale corpus into documents via *K*-means algorithm, the number of sentences in the documents varies, thus we calculate the mean after summing up the difference of cross entropy of each sentence to obtain the score of each document δ_{D_i} by Formula (8),

$$\delta_{D_i} = \frac{\sum_{j=1}^{n}(H_{ref}(S_j) - H_{train}(S_j))}{n} \quad (8)$$

where δ_{D_i} is the score of the *i*-th document, $H_{ref}(S_j)$ is the cross entropy of the *j*-th sentence in the document D_i derived from the language model of the references, $H_{train}(S_j)$ is the cross entropy of the *j*-th sentence in the document D_i derived from the language model of the training data, *n* is the number of sentences in the document D_i. Then we sort δ_{D_i} in ascending order. The lower score implies the document is more like the human references.

4 Experiments

To test the quality of the domain-specific paraphrase extracted from monolingual corpus by the proposed approach, we conducted experiments on WMT15 Metrics task.

The METEOR-Universal metric (Denkowski and Lavie, 2014) using the paraphrase tables which were extracted from the bilingual parallel corpus was set as the baseline metric. We used the paraphrase tables in general domain extracted by the Markov Network model, and the domain-specific paraphrase tables extracted by our approach substituted for the original paraphrased tables, respectively. The updated metrics are called as METEOR-Markov and METEOR-MPEDA. We compared the METEOR-MPEDA metric with the METEOR-Markov metric and METEOR-Universal metric to demonstrate the quality of the domain-specific paraphrase table extracted by our approach. Besides, we compared the METEOR-MPEDA with METEOR metric (Banerjee et al., 2005) which only uses the exact word match, stem match and synonym match.

Data	en-cs	en-de	en-fr	en-fi	en-ru	cs-en	de-en	fr-en	fi-en	ru-en
T-corpus	1000k	1920k	2007k	1926k	1074k	2218k	2218k	2218k	2218k	2218k
ref	2656	2169	1500	1370	2818	2656	2169	1500	1370	2818

Table 1. The statistics of the corpus

Data	en-cs	en-de	en-fr	en-fi	en-ru	cs-en	de-en	fr-en	fi-en	ru-en
D-corpus	28230	39684	39763	39921	28643	39684	39684	39684	39684	39684

Table 2. The number of documents in training data

4.1 Corpus

The training data and the human references we used in the experiment are all provided in WMT15 Translation task and Metrics task (Bojar et al., 2015), every training data has its corresponding references. Table 1 shows the number of sentences in the corpora. The row "*T-corpus*" denotes the training data, while the row "*ref*" denotes the references.

The training data was processed by text clustering. We used *K*-means clustering algorithm to gather the corpus sentences in different clusters, and then adopted the bag of word to create a vector for each sentence. By computing the cosine value of the two vectors, we obtained the distance between two sentences. Each cluster was viewed as a document. The *i*-th document in training data was named D_i, and the number of sentences in each document was different. Table 2 is the number of documents after training data clustering. The row "*D-corpus*" is the number of document used in the training data.

4.2 Experiments Settings

After dividing the training data into documents, we processed the corpus by the following procedure: tokenize the training data and the references; delete the punctuations; transform the capitalized letters of words into lower case. Then, we employed *4*-gram language model with *Kneser-Ney* discounting to train corresponding language models for training data and the references. The difference of cross entropy of each sentence in the training data language model was calculated. Then we summed up and normalized the difference of the cross entropy of the documents' sentences. Thus every document in the training data received a score. The smaller the value is, the closer the document is to the reference. Later, we arranged the values in an ascending order, meanwhile, a threshold value was set, and the corpus beyond the threshold was abandoned. In this way, we obtained a smaller sub-corpus with the approximately same domain with the training data. Finally, we gave different threshold value to the different sub-tasks, in other words, we selected the top *n* documents after ordering.

We used the Markov network to build a term Markov network model in the sub-corpus, then we calculated the relation among words according to words co-occurrence, next, we extracted the word chunks in the Markov network, and computed the likelihood that two words are a paraphrase pair by comparing the two chunks' similarity. In this work, we extracted ten paraphrase tables for ten sub-tasks in six languages on WMT15.

4.3 Results

The Pearson Coefficient is used to compute the system-level correlation between automatic evaluation and human judgments as follows:

$$r = \frac{\sum_{i=1}^{n}(H_i - \bar{H})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \bar{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \bar{M})^2}} \quad (9)$$

where H_i and M_i are the i-th system scores of human judgment and that of the automatic evaluation metrics, respectively.

The system-level correlation for the three metrics is given in Table 3 and Table 4, from the tables, we found that the system-level correlation of METEOR-MPEDA metric is better than METEOR, METEOR-Universal and METEOR-Markov on average.

Furthermore, Kendall's τ coefficient was used to compute the correlation between automatic evaluation metrics and human judgments at segment -level as follows:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (10)$$

where *Concordant* denotes the set where the human judgment and the automatic evaluation metrics' score are concordant, while *Discordant* denotes the set where they are discordant.

The segment-level correlation is given in Table 5 and 6. It can be observed that the segment-level correlation of METEOR-MPEDA metric on evaluation translation into English tasks is better than METEOR, METEOR-Universal metric and METEOR-Markov metric on average. However, when evaluating translation out of English tasks, the performance of the METEOR-MPEDA metric is slightly lower than METEOR-Universal metric. It can be explained that when we have a large amount of bilingual parallel training data, the paraphrase table extracted from the bilingual corpus is better than that from monolingual corpus for automatic evaluation of MT.

Metrics	de-en	cs-en	fr-en	fi-en	ru-en	Average
METEOR	0.926	0.973	**0.979**	0.929	0.959	0.953
METEOR-Universal	0.953	**0.974**	**0.979**	0.934	0.964	0.961
METEOR-Markov	0.950	**0.974**	0.978	0.929	**0.965**	0.959
METEOR-MPEDA	**0.959**	**0.974**	**0.979**	**0.939**	0.963	**0.963**

Table 3. The system-level correlation of metrics on evaluation translation into English on WMT15 Metrics task

Metrics	en-de	en-cs	en-fr	en-fi	en-ru	Average
METEOR	0.680	0.957	0.951	0.713	0.864	0.833
METEOR-Universal	0.722	0.940	0.952	**0.724**	0.845	0.837
METEOR-Markov	0.705	**0.954**	0.949	0.712	0.845	0.833
METEOR-MPEDA	**0.735**	0.938	**0.955**	0.714	**0.851**	**0.839**

Table 4. The system-level correlation of metrics on evaluation translation out of English on WMT15 Metrics task

Metrics	de-en	cs-en	fr-en	fi-en	ru-en	Average
METEOR	0.389	0.406	0.375	0.385	0.358	0.378
METEOR-Universal	0.431	**0.437**	**0.386**	0.388	0.379	0.404
METEOR-Markov	0.421	0.429	**0.386**	0.393	0.367	0.400
METEOR-MPEDA	**0.431**	0.434	0.376	**0.404**	**0.383**	**0.406**

Table 5. The segment-level correlation of metrics on evaluation translation into English on WMT15 Metrics task

Metrics	en-de	en-cs	en-fr	en-fi	en-ru	Average
METEOR	0.319	0.389	0.335	0.251	0.373	0.333
METEOR-Universal	0.339	0.388	**0.342**	**0.274**	0.380	**0.345**
METEOR-Markov	0.332	**0.389**	0.339	0.251	**0.381**	0.338
METEOR-MPEDA	**0.342**	0.385	0.341	0.251	**0.381**	0.340

Table 6. The segment-level correlation of metrics on evaluation translation out of English on WMT15 Metrics task

5 Conclusion

In this paper, we describe the submissions of our metric for WMT16 Metrics task in detail. We propose an approach to extract domain-specific paraphrase table from monolingual corpus for automatic evaluation of MT, and use it to replace the original paraphrase table in METEOR metric to improve the correlation between human judgment and automatic evaluation metrics. The proposed approach is tested on the newswire domain. In future work, we will systematically apply it to different specific domains such as the medical domain, IT domain, etc.

Acknowledgments

This research has been funded by the Natural Science Foundation of China under Grant No.61203313, 61462044, 61462045, and 61562042, and supported by the Natural Science Foundation of Jiangxi Provincial Department of Science and Technology of China under Grant No 20151BAB207025, and also supported by the Natural Science Foundation of Jiangxi Educational Committee of China under Grant No. GJJ150352.

References

Amittai Axelrod, Xiaodong He and Jianfeng Gao, 2011. Domain Adaptation via Pseudo In-Domain Data Selection. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 355-362, Edinburgh, Scotland, UK.

Satanjeev Banerjee and Alon Lavie, 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65-72, Ann Arbor.

Colin Bannard and Chris Callison-Burch, 2005. Paraphrasing with Bilingual Parallel Corpora. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pages 597-604, Ann Arbor, Michigan.

Regina Barzilay and Kathleen R. McKeown, 2001. Extracting Paraphrases from a Parallel Corpus. Proceedings of 39th Annual Meeting of the Association for Computational Linguistics, pages 50-57, Toulouse, France.

Regina Barzilay and Lillian Lee, 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 16-23.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia and Marco Turchi, 2015. Findings of the 2015 Workshop on Statistical Machine Translation. Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 1-46, Lisbon, Portugal.

Yee Seng Chan and Hwee Tou Ng, 2008. MAXSIM: A Maximum Similarity Metric for Machine Translation Evaluation. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pages 55-62, Columbus, Ohio.

Michael Denkowski and Alon Lavie, 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 376-380.

George Doddington, 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. Proceedings of the second international conference on Human Language Technology Research, pages 138-145, San Diego, California, CA, USA.

Maoxi Li, Aiwen Jiang and Mingwen Wang, 2013. Listwise Approach to Learning to Rank for Automatic Evaluation of Machine Translation. Proceedings of Machine Translation Summit XIV, pages 51-59, Nice, France.

Maoxi Li, Mingwen Wang, Hanxi Li, Fan Xu, 2016. Modeling Monolingual Character Alignment for Automatic Evaluation of Chinese Translation. ACM Transactions on Asian and Low-Resource Language Information Processing, 15(3), pages 1-16.

Robert C. Moore and William Lewis, 2010. Intelligent Selection of Language Model Training Data. Proceedings of the ACL 2010 Conference (Short Papers), pages 220-224, Uppsala, Sweden.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311-318, Philadelphia, Pennsylvania.

Ellie Pavlick, Juri Ganitkevitch, Tsz Ping Chan, Xuchen Yao, Benjamin Van Durme and Chris Callison-Burch, 2015. Domain-Specific Paraphrase Extraction. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 57–62, Beijing, China.

Yusuke Shinyama, Satoshi Sekine and Kiyoshi Sudo, 2002. Automatic Paraphrase Acquisition from News Articles. Proceedings of the second international conference on Human Language Technology Research, pages 313-318.

Matthew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciulla and Ralph Ma-

khoul, 2006. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of Association for Machine Translation in the Americas, pages 223-231, Cambridge.

Zhen Weng, Maoxi Li, Mingwen Wang, 2015. Enhance Automatic Evaluation of Machine Translation by Markov Network Based Paraphrases (in Chinese). Journal of Chinese Information Processing, 29(6), pages 136-142.