ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

A Comparative Study of Learning to Paraphrase for Question Answering

Vinaya A V[1], Naseer C[2]

PG Student, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[1]

Assistant Professor, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[2]

Abstract: Question Answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing, which is used to build systems that automatically answer questions posed by humans in a natural language. The system propose learning to paraphrase for question answering, Here turns to Paraphrases as a means of capturing this knowledge and present a general framework which learns different paraphrases for various QA tasks. Our method is trained different question-answer pairs. A question and its paraphrases input to a neural scoring model which assigns higher weights to each expressions most likely to yield correct answers. Here discuss about different paraphrase generation methods and check the generated paraphrase sentence with a scoring model. The advantages and limitations of these methodologies are also discussed.

Keywords: Question Answering (QA), Natural Language Processing (NLP), Paraphrase generation, Paraphrase Scoring Model Question Answering Model

I. INTRODUCTION

A question Answering (QA) system constructs its answers automatically by querying from a structured database known as a knowledgebase or an unstructured collection of documents and a set of questions. Paraphrase approaches are widely used to solve paraphrasing problems in natural language QA systems. To construct the answers, the QA systems must be able to understand the language resources which have a paraphrasing property. The property makes the QA systems difficult to retrieve an answer to the question, as questions and answers can be expressed in a various ways. To solve the paraphrastic problem in the QA system, some studies have presented with a paraphrasing approach in an effort to understand questions and retrieve answers in a QA system.

A paraphrase is a restatement of the meaning of a text or passage using other words [wikipedia]. The term itself is derived via Latin paraphrasis from Greek word meaning “additional manner of expression”. The act of paraphrasing is also called “paraphrasis”. A paraphrase typically explains or clarifies the text that is being paraphrased.

Question Answering (QA) is challenging due to the many different ways natural language expresses the same information need. For example: QA system must recognize that the questions “who created Microsoft” and “who started Microsoft” have the same meaning and that they both convey the founder relation in order to retrieve the correct answer from data.

The Natural Language Processing (NLP) techniques used many QA systems may range from simple lexical and semantic disambiguation of question stems to complex processing that combines syntactic and semantic features of the questions with pragmatic information derived from the context of candidate answers,[3] Question Answering (QA) is challenging due to the many different ways natural language expresses the same information need. The problem of paraphrases conceals a number of different linguistic problems, which in our opinion need to be treated in separate ways. In fact, paraphrases can happen at various levels in language.

There are various techniques used for Paraphrase generation, such as evaluate our approach on QA over Freebase and text based answer sentence selection. Here discuss different paraphrase models based on the Paraphrase Database [8],

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE
International Journal of Advanced Research in Computer and Communication Engineering
PCON-2019
National Conference and Seminar on Innovations in Engineering & Technology
Govt Polytechnic College, Palakkad
Vol. 8, Special Issue 1, January 2019

neural machine translation [10], and rules mined from the Wiki Answers corpus [1]. Which evaluated on three datasets, shows that our framework constantly improves the performance. it achieves good results on Graph Questions and competitive performance on two additional benchmark datasets using simple QA models.

This paper is organized as follows: Section II Model description of the question answering model of paraphrasing. Section III discusses various approaches of paraphrase question answering system. Table 1 gives comparison of different methods and Section IV gives a brief concluding comment. Section V discusses various future research directions of paraphrase question answering system.

II. MODEL DESCRIPTION OF THE QUESTION ANSWERING MODEL OF PARAPHRASING

The proposed Learning paraphrase question answering system mainly contain two components:

Paraphrase generation and Paraphrase Scoring Model. The architecture and the important steps in building the proposed system is described here. Figure .1 shows an overall work flow or architecture of the proposed paraphrase question answering system for learning.

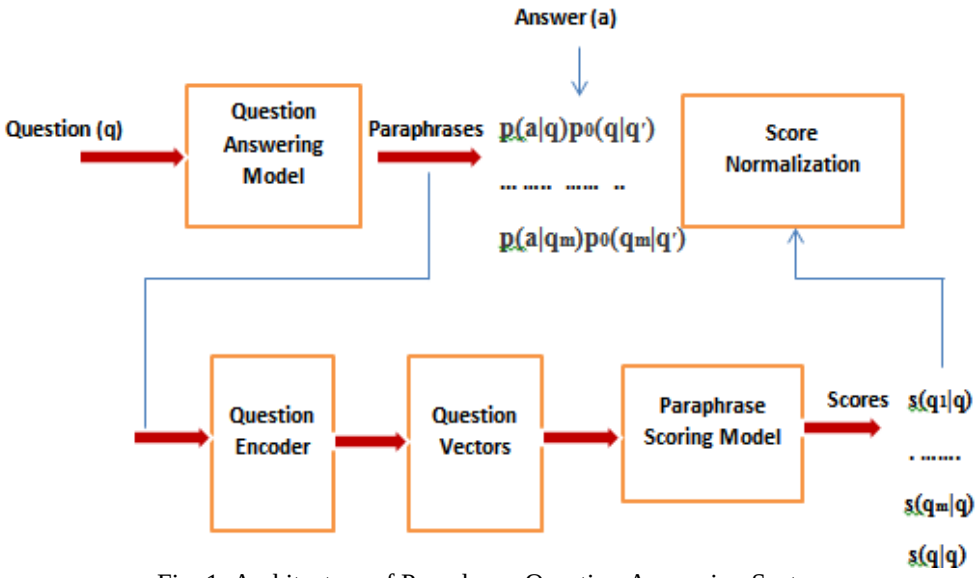

Fig. 1 Architecture of Paraphrase Question Answering System

A. Paraphrase Generation

Here mainly 3 methods are used to generate paraphrase Questions.

1) PPDB-based Generation: Ganitkevitch et.al.[8]here released a new paraphrase database, PPDB. Its English portion, PPDB: Eng, contains over 220 million paraphrase pairs, consisting of 73 million phrasal and 8 million lexical paraphrases, as well as 140 million paraphrase patterns, which capture many meaning preserving syntactic transformations. The most used approaches to paraphrasing Bilingual pivoting, bit uses bilingual parallel corpora to learn paraphrases based on techniques from phrase based Statistical Machine Translation(SMT) The method is that two English strings that translate to the same foreign string that convey the same meaning. The method first extracts a bilingual phrase table and then obtains English paraphrases by pivoting through foreign language phrases. In paraphrase database contain over a billion of paraphrase pairs in 24 different languages. Using bidirectional entailing rule from PPDB. Here focus on single and multiword rules which here use to paraphrase questions by replacing words and phrases.

2) NMT-based Generation: Here Bahdanau et.al.[5] revisit the bilingual pivoting in the context of Neural Machine Translation (NMT) and present a paraphrasing model based on neural networks. NMT is trained end-to-end to maximize the conditional probability of a correct translation given a source sentence, using a bilingual corpus.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

Paraphrases can be obtained by translating an English string into a foreign language and then back-translating it into English. NMT based pivoting models offer advantages over conventional methods such as the ability to learn continuous representations and to consider wider context while paraphrasing. Here select German as our pivot, and pretrain two NMT systems English-to-German (EN-DE) and German-to-English (DE-EN).

3) Rule-Based Generation: Next paraphrase generation approach uses rules mined from the Wiki Answers corpus which contains more than 30 million question clusters labeled as paraphrases by Wiki Answers users. This corpus is a large resource but is relatively noisy due to its collaborative nature 45 percentage of question pairs are merely related rather than genuine paraphrases. First extracted question templates that appear in at least ten clusters. Any two templates co-occurring in the same cluster and with the same arguments were deemed paraphrases.

B. Paraphrase Scoring

In paraphrase scoring q is the natural language question and a is its answer. Estimate p (a |q), the conditional probability of candidate answers given the question.[1] decompose p(a | q) as:

$$p(a|q) = \sum_{q' \in H_q \cup \{q\}} \underbrace{p_\psi(a|q')}_{\text{QA Model}} \underbrace{p_\theta(q'|q)}_{\text{Paraphrase Model}}$$

Hq is the set of paraphrases for question q, Ψ are the parameters of a QA model, and θ are the parameters of a paraphrase scoring model. q = q1……q|n| denote an input question. Every word is initially mapped to a d-dimensional vector. GloVe used for vector conversion. Use a bi-directional recurrent neural network with long short term memory units as the question encoder, shared by the input questions and their paraphrases. Encoder recursively processes tokens one by one, and uses the encoded vectors to represent questions. After obtaining vector representations for q and q', compute the score s (q'|q).

III. DIFFERENT METHODOLOGIES

Many different approaches have been applied in different QA systems.

This paper Ying Xu et.al.[3] introduces a lexical replacement approach as an initial study of Korean paraphrasing. Knowledge Based Lexical Replacement: This approach generates paraphrase sentences using knowledge based lexical replacement. This method can generate paraphrases using only a morpheme analyzer and thesaurus resources without a large corpus. The method can use fewer computational resources when implemented therefore it is suitable for small modules. In lexical replacement, synonym knowledge is the most important resource. Therefore, collect lexical resources to find the synonyms of lexemes. Here try to get refined synonym resources for the lexical replacement from various Korean lexical resources. However, the system generates simple results, and the WSD problem occurs with homonyms and in cases involving polysemy.

Feature Based Word Sense Disambiguation (WSD): The extraction from integrated lexical knowledge, the Word-Sense Disambiguation (WSD) problem occurs with homonyms and in cases involving polysemy. It selects appropriate synonyms according to the context of a sentence. The meanings of sentences change after this type of replacement, generating unsuitable sentences due to homophones and multi sense words to solve the WSD problem, and extract additional features of phases.

Paraphrase Generation: The preprocessing module extracted complete simple sentences from a complex sentence through a tokenizing and trimming process. The module extracted full morphemes from an input sentence without named entities. We found that the quality of paraphrasing depended on the type of morpheme. Therefore, we selected types of morphemes which were classified as full morphemes. The analyzer trained classifiers of features from a large corpus on the web. Feature extraction module extracted several features of words and phrases. At next step, the WSD module selected appropriate words from the extracted synonyms for lexical replacement. Obtain the candidate synonyms of the words and phrases. Using the synonyms, the generation module performed lexical replacement on

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

morpheme level. After the replacement process, we modified the postpositional particles in the sentence for natural paraphrasing. In Korean, the quality of the postposition determined the naturalness of the paraphrase.

Duboue et.al.[13]use a template based method to heuristically produce standard content descriptions for candidate logical forms, and then compute paraphrase scores between the generated texts and input questions in order to rank the logical forms.

Another proposed method uses paraphrases in the context of neural question answering models (Dong et.al.2015) introduce Multi Column Convolutional Neural Networks (MCCNNs) to understand questions from three different aspects and learn their distributed representations. Also jointly learn low dimensional embedding's of entities and relations in the knowledgebase. Question-answer pairs are used to train the model to rank candidate answers and leverage question paraphrases to train the column networks in a multi task learning manner.

Fabio Rinaldi et.al.[14] The system implements a simple and efficient logic representation of questions and answers that maps paraphrases to the same underlying semantic representation. Further, paraphrases of technical terminology are deal with by a separate process that detects surface variants.

Most of the paraphrase generation techniques need a corpus. For collecting paraphrases as a corpus, Dolan et.al.[13]introduced and experimented with an unsupervised approach to extract a parallel corpus from news articles. Except for extracting sentences with very low edit distances, they assumed that the first two sentences in news articles are usually the summarization of the whole news article, and similar news articles might be reporting the same event. Thus, the summarization sentences of similar news articles, which can be aligned with low cost, are potentially paraphrases. With this hypothesis, they extracted and filtered the first two sentences from news articles which reported the same event, as paraphrases.

When it comes to QA, sometimes the question queries in QA are just noun phrases. For example, advantage of iPhone" and do people like to use iPhone" almost mean the same in QA.Zhao et.al.[6] introduced their approach for extending incomplete, short, phrasal queries to multiple complete questions, with community based QA systems and previous user query logs. It associates user's input queries with the templates extracted from the finally selected questions, and generates questions from an incomplete query by applying the query into the associated templates.

General framework for learning paraphrases for question answering tasks [1] given a natural language question, the model estimates a probability distribution over candidate answers. First generate paraphrases for the question, which can be obtained by one or several paraphrasing systems. A paraphrase scoring model predicts the quality of the generated paraphrases, while learning to assign higher weights to those which are more likely to yield correct answers. The paraphrases and the original question are fed into a QA model that predicts a distribution over answers given the question.

IV. CONCLUSION

Here studied various approaches for paraphrase question answering. The framework trained different question answer pairs, paraphrasing is important because it shows you understand the source well enough to write it in your own words. It also gives you a powerful alternative to using direct quotes, which should be used infrequently. This survey compares different methodologies for learning paraphrase for question answering. Paraphrase scoring and QA models are trained on different question answer pairs, which results in learning paraphrases with a purpose. The framework is not for specific paraphrase generator or QA system. Also it allows incorporating several paraphrasing modules, and can serve as a model for testing and exploring their coverage and rewriting capabilities.

V. FUTURE RESEARCH DIRECTIONS

In paper[1]Gives the best method for paraphrase generation and it can be used for other natural language processing tasks which are sensitive to the variation of input (e.g., textual entailment or summarization)

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

Table 1 Comparison of Various Methodologies

Paper	Author	Overview	Remarks
Learning to Paraphrase for Question Answering	Li Dong.et.al (2017)	QA model and Paraphrase scoring model	Perform more complex question answering is difficult.
Learning Paraphrases to Improve a Question-Answering System	Florence Duclaye et.al. (2016)	Expectation Maximization (EM) algorithm.	Unsupervised learning methodology
Paraphrase for Open Question Answering: New Dataset and Methods	Ying Xu et.al(2016)	QA framework over a Knowledge base (KB)	Data is that both arguments are NE, It unable to answer questions with answers that are common nouns or numbers.
Paraphrase Generation Based on Lexical Knowledge and Features for a Natural Language Question Answering System	Kyo-Joong Oh et.al.(2015)	Lexical Knowledge and Features	syntactic structures considered

REFERENCES

[1]. Li Dong, Jonathan Mallinson, Siva Reddy and Mirella Lapata, 2017, Learning to Paraphrase for Question Answering In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 875886 Copenhagen, Denmark, September 711, 2017. c 2017 Association for Computational Linguistics.
[2]. Florence Duclaye, Francois Yvon and Olivier Collin, 2016 Learning Paraphrases to Improve a Question-Answering System In:Proceedings of Association for Computational Linguistics.
[3]. Ying Xu, Pascual Martnez-Gomez, Yusuke Miyao and Randy Goebel Paraphrase for Open Question Answering: New Dataset and Methods Proceedings of 2016 NAACL Human-Computer Question Answering Workshop, pages 5361,San Diego, California, June 12-17, 2016. c 2016 Association for Computational Linguistics
[4]. Kyo-Joong Oh, Ho-Jin Choi, Gahgene Gweon, Jeong Heo, and Pum-Mo Ryu Paraphrase Generation Based on Lexical Knowledge and Features for a Natural Language Question Answering System 978-1-4799-7303- 3/15/ 2015 IEEE.
[5]. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations.
[6]. Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05), pages 597604, Ann Arbor, Michigan. Association for Computational Linguistics.
[7]. Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase driven learning for open question answering. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16081618, Sofia, Bulgaria. Association for Computational Linguistics.
[8]. Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: the paraphrase database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 758764, Atlanta, Georgia. Association for Computational Linguistics.
[9]. Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In Association for Computational Linguistics (ACL).
[10]. Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2016. Paraphrasing revisited with neural machine translation. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.
[11]. N.Madnani and B. J. Dorr, Generating phrasal and sentential paraphrases: A survey of data-driven methods Computational Linguistics, vol. 36, No. 3, pp. 341-387, 2010.lencia, Spain.
[12]. J. Ganitkevitch, C. Callison-Burch, C. Napoles, and B. Van Durme, Learning sentential paraphrases from bilingual parallel corpora for text to text generation Empirical Methods in Natural Language Processing, pp. 1168-1179, 2011
[13]. Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of the 20th International Conference on Computational Linguistics, page 350. Association for Computational Linguistics, 2004.
[14]. Anthony Fader, Luke S Zettlemoyer, and Oren Etzioni. ParaphraseDriven Learning for Open Question Answering. In ACL (1), pages 1608-1618. Citeseer, 2013.