DOI: 10.21917/ijsc.2020.0297

DEEPREPLY - AN AUTOMATIC EMAIL REPLY SYSTEM WITH UNSUPERVISED CLOZE TRANSLATION AND DEEP LEARNING

P.V. Rajaraman[1] and M. Prakash[2]

[1]*Department of Computer Science and Engineering, Rajalakshmi Engineering College, India*
[2]*Department of Information Technology, Karpagam College of Engineering, India*

Abstract

Electronic mail (E-mail) has been the primary mode of communication for official purposes and it continues to be the same in all work environments even today. With the growing number of emails and most of them requiring only trivial replies, more tools are needed to generate replies to emails by reusing past replies. Although there are expert systems that can assist us in replying to incoming emails, they produce a generic reply to all. So an intelligent system that can generate replies for an incoming email in a very precise manner and generating the text reply in the user's style is the identified requirement. This work is divided into two portions. First, translating an incoming email into cloze representation and extract the entities from it for generating a context, question and answer triplets. This is used for synthesising the training data for Extractive Question Answering later. The mentioned triplets are generated from a corpus of random emails belonging to different contexts and then the answers are extracted by recognising the named entities and random phrases of nouns from these paragraphs. The second ploy is to find the similarity between an incoming email that requires a reply and an old email that contains the reply to it. As a solution to these challenges, we propose a new deep neural network-based approach that relies on coarse-grained sentence modelling using CNN and a LSTM model. Our experimental results show that the approach outperforms the state-of-the-art approaches that are existing on a cleaner corpus.

Keywords:

Deep Learning, E-mail, Unsupervised, Questioning

1. INTRODUCTION

In the world of social media, Email is still a powerful tool for exchanging messages and files. Facts show that by the year 2015 there were 5.2 billion email accounts. The same year had around 2.5 billion email users [1]. Approximately 246 billion email messages are sent and received on a daily basis [1]. After the increase of smart mobile phones there has been an increase in email users. The statistics detailed below prove the claims mentioned above.

Further analysis reveal that 72% of about email transactions contain received emails, while 28% of them are sent emails [1]. Entailing the above statistical analysis there are 128.8. Billion business email alone sent/received per day.

Most of the emails in a business domain could be narrowed down to some common or previously answered issues. To name some of such work scenarios: Professors receiving common doubts from different students, educational help desks, socially organized events and many more. Research suggests that creating an ontology that is specific to the domain can improve the chances of retrieving the appropriate result [3]. But, this might not be useful for cases that are not domain specific. A system that uses a technique called Case-Based Reasoning [2]. This method, retrieves similar past problems from the knowledge base and then reuses the solutions to answer a new email.

Table.1. Worldwide Email Accounts and User Forecast (M), 2015–2019

% Growth	2015	2016	2017	2018	2019
Worldwide Email Accounts (M)	4,353	4,626 6%	4,920 6%	5,243 7%	5,594 7%
Worldwide Email Accounts (M)	2586	2,672 3%	2,760 3%	2,849 3%	2,643 3%
Avg. Accounts Per User	1.7	1.7	1.8	1.8	1.9

Table.2. Worldwide Daily Email Traffic (B), 2015-2019

% Growth	2015	2016	2017	2018	2019
Total Worldwide Emails Sent/Received per Day (B)	205.6	215.3 5%	225.3 5%	235.6 5%	246.5 5%
Business Emails Sent/Received per Day (B)	112.5	116.4 3%	120.4 3%	124.5 3%	128.8 3%
Consumer Emails Sent/Received per Day (B)	93.1	98.9 6%	104.9 6%	111.1 6%	117.7 6%

The task of answering questions in an email given a context email is Extractive Question Answering (EQA) in the email domain. This is working on an assumption that the answer is contained in the short message. The main characteristic of an intelligent system is to comprehend the input text like how human beings do the same. For the SQuaD dataset, an EQA dataset, the existing models beat human capabilities. For the dataset SQuaD 2.0, based on BERT is now matching the human performance. In the case of the newly introduced Natural Language corpus, human performance will soon be reached. The common factor in all the cases discussed above is the availability of huge amounts of noise free training data. But, in the case of newer domains and other languages, data collection becomes a non-trivial process and a barrier to progress and also would require significant resources.

In this process, the worst possibility could be the lack of availability of training data itself. This issue is addressed in the first part of this work. We start to explore the possibility of using an unsupervised EQA for which there is no question, answer and context. The idea is that, if we have a method which does not require QA supervision to generate correct questions for a given context paragraph, we could train a Question Answering system using only the generated questions. This methodology allows us have progress in QA, just as in the case of pretrained routines and architectures based on models. The merits of the proposed method are that it is both flexible and scalable. This method can also be used to generate additional data for the purpose of training in semi-supervised experiments.

The method proposed by us as shown in Fig.1 comprises the following three steps to generate EQA:

1. We sample an email in a domain, Enron email dataset in our case;
2. The candidate answers are sampled from the context as shown in the schematic diagram. We do this with the help of pretrained components like Named Entity Recognizer and Noun Checkers to identify the entities. These do not require aligned question-answer or question-context pairs. When a context and candidate answer is given, the masked cloze questions are extracted and
3. Finally, these cloze questions are converted into natural questions using a cloze to natural language unsupervised question translator.

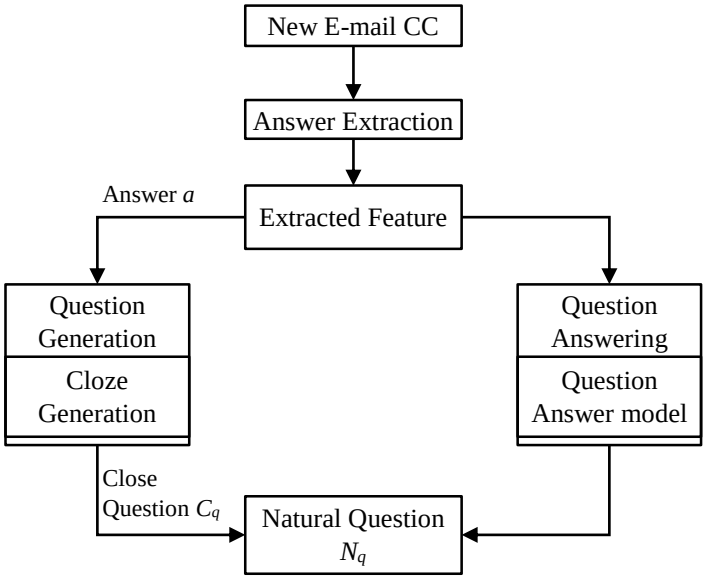

Fig.1. Schematic of Approach

The most challenging part of the steps listed above was the conversion of cloze representation to natural language representations. There are a lot of rule -based methods available that transform statements to questions, but their performance for QA was found to be very weak. Other supervised methods require annotated data that is unavailable for this task. This challenge is answered by using the recent development in unsupervised MT.

We used a seq2seq model to match natural and cloze questions from a large corpus that contains natural questions and their cloze representations. Other techniques like back-translation and de-noising auto encoding were also used in combination with seq2seq mode.

2. UNSUPERVISED EQA

When considering the EQA we are given a question q and a context mail m and are required to provide an answer $a=(s,t)$ starting with s and end t character indices in m. The right wing of Fig.1 shows this. It is proposed to address the unsupervised QA in two stages. In the first a generative model $g(q,a,m)$ with no supervision and then train a more discriminative model $g_r(a|q,m)$ using g_{as} training data generator. The generator $g(q,a,m) = g(m)g(a|m)p(q|a,m)$ will be used to generate data in the reverse direction. First the context is sampled using $p(m)$, second the answer contained in the email via $p(a|m)$ and third a question for the answer and email via $p(q|a,m)$.

2.1 GENERATING CONTEXT AND ANSWER

Given the Enron email dataset containing 0.5 million emails our context generator $p(m)$ uniformly samples an email m of appropriate length from any email content, the generation step creates answer spans a for m via $p(a|m)$. This incorporates prior probability beliefs about what makes a good answer. There are two simple variants of $p(a|m)$:

Named Entities: Here, we extract all the named entities using an NER system and sample from them. This step is very useful in restricting the variety of questions that e_i can answer.

Noun Phrases: Here, all the noun phrases from e_i are extracted. This gives the exhaustive list of answers that is contained in e_i.

2.2 GENERATING QUESTIONS

As already mentioned, the challenging part of the whole work happens to be modelling the relation between the question and the answer. We capture this using $p(q|a,m)$ that will produce questions from a given answer in the email. This is split into two steps: generating cloze representation $q' = cloze(a,m)$ and translation, $g(q|q')$.

2.2.1 Cloze Generation:

Cloze representations are statements with hidden answer. First the scope of the email has to be mapped to the same level of detail of the original questions in EQA. In the context and answer as shown in Fig.1, consider the following message:

"George, Saturday night sounds great, Darlene is in _____ for the next 2 weeks so I'm ready. Call me at home 713-588-5176 or at work 713-853-3917. Golf also sound good, another guy from Calgary who now works down here Chris Dorland, I think you've met him before, wants to golf as well. Look forward to seeing you".

2.2.2 Cloze Translation:

After close equation q' is generated it has to be translated into a form that is closer to what looks for more questions that are used in QA tasks. They are explored in the following sections.

Rule-Based Approach: Converting an answer contained in a statement into a (q,a) pair can be considered as a transformation with respect to the syntax with wh- occurrences and a type-independent choice of wh-word. There are a lot of software available for English to serve this purpose. Among them is used in this work, which has a set of rules that generates a lot of candidate questions. It also has a ranking system to find the best among the generated questions.

Seq2seq Modelling: This approach does not require any prior knowledge on any rule-based systems. We present an unsupervised training of a seq2seq model that helps in translating the cloze to questions in natural language. The details of this can be found in the results section.

3. PARAPHRASE DETECTION

Having received a new message, we need to classify whether it requires a reply or not. If it requires a reply, we then trigger a response to it. We analyse the content and detect an old email that already contains the answer in it. This is where we face a challenge called Paraphrase Detection (PD)

PD is viewed as a classification problem in NLP. When there are two sentences given, the system will then determine the similarity between the two sentences based on the semantics. If in case they convey the same meaning, it is then labelled as paraphrase or otherwise as a non-paraphrase. The existing systems that handle this task do a fairly great job when a clean text corpus like the Microsoft Paraphrase Corpus (MSRP) is given. But, trying to achieve the same amount of accuracy with other noisy datasets will be more challenging due to the spelling mistakes, structure and style and acronyms.

Adding to the above points, when we measure the semantic similarity between two messages, it becomes very difficult because of the lack of lexical features [4]. There has been some attention given to PD in noisy short-texts till now, and has been reported on the SemVal 2015 twitter dataset [3] [4] [5]. One of the main goals of this work is to build a paraphrase detection model based on deep learning techniques. The model is expected to accurately detect paraphrasing in both noisy and clean texts. To be more specific, we are proposing a hybrid deep neural modelled architecture that is comprised of both Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) model, adding a word-pair similarity module.

Our proposed model consists of two components 1) Sentence modelling, and 2) pair-wise word similarity matching. We first have to build an effective text representation system that is required for sentence modelling. For this we build a hybrid CNN and RNN architecture wherein the features extracted by the CNN are fed as input to the RNN. We do word embeddings and its output is fed to the CNN. After the convolutional and pooling operations, the features that are encoded and mapped are given as input to the RNN. The final state learned by RNN model is the sentence level representation. The main idea behind using both CNN and RNN is that, CNN is capable of learning the features as n-grams of the texts and RNN takes words in a sequential order and will be learning the long-term dependencies of the texts rather than the local features.

As our second task, a pair-wise matching similarity model is built to extract the similar information between pairs of email passage sentences. A pair-wise similarity matrix is built by computing the similarity of each and every word in a given sentence to all the words in another email message.

The purpose of this paper is to reveal how this model proposed for PD can produce better results if an extra set of features is statistically extracted from the input text. Our main contributions in this paper are:

- We have proposed a novel deep neural architecture that leverages coarse-gained features at the sentence level and fine-grained features at the word-level for paraphrase detection on emails from Enron dataset. This model combines the information gained both at the sentence level and word-level in a way that it captures the features in all the levels. The word level similarity model provides with useful information when the text contains any grammatical incorrectness. This is how both model work by complimenting each other and finally shows an efficient performance.
- It is shown how a pair-wise similarity model is useful in extracting word-level semantics, and in the PD task.
- We propose combining statistical textual features and the features that were learned from the deep neural network architecture.

4. RELATED WORK

There have been extensive studies on Natural Language Processing (NLP) with deep learning in the recently bygone years. The majority of the work done on PD have been focused on some selective features like overlapping n-gram features [6], syntactic and machine translation-based features [7], linguistic features [8] [9], semantic networks based on Wikipedia [10], knowledge graphs [11]. The literature suggests that the researchers have shifted their attention towards the semantic representations [12] [13] [14]. It is seen that CNN was used to learn how to interpret the text, which enhanced the results in the classification of sentences [15]. Recurrent neural networks (RNNs) are capable to learn the long-term dependencies in continuous data and are also used to represent the text in the literature [16]. Kim et al. [17] proposed a design using CNN, an over-character highway network from which output is sent to the RNN network. They combine CNN-RNN model which provides better results. Wang et al. [18] proposed to combine both convolution and recurrent neural network to learn the representation of sentences for the sentiment analysis task. For sentence similarity estimation, a variety of deep neural network-based architectures have been proposed, which is a strategy that we also focus in this paper.

5. DEEPMES ARCHITECTURE

In this work a deep learning- based approach is proposed for detecting paraphrase sentences for Emails, with the architecture as shown in Fig.2. Every sentence in a pair is converted into a semantic representative vector by using CNN and RNN. Next, a semantic pair-level vector is generated by using the difference between each vector in the sentence representations. The difference produces the resultant discriminative vector of the pair of sentences that will be used as a feature vector for learning the similarities between two sentences. To add to this, we also extract more fine-grained information using a similarity matrix that contains word-to-word quantification of similarity. We add more convolutional layers on the pair-wise similarity matrix to gain the similarity patterns between the words in the two sentences. At least, a set of features is extracted by using statistical analysis of the text, and is added to the rest of the features learned. The first set of layers are activated using the ReLU function, while we are using sigmoid activation function for transferring the representations into a binary classification rule. Finally, the model is trained for optimizing binary cross-entropy.

5.1 SENTENCE MODELLING WITH CNN & RNN

In this module, every sentence is presented using a combined representation of CNN and RNN architecture. The CNN learns the common features from words to phrases from the text and RNN learns the long-term dependencies form the text. The word embedding is taken as input to the CNN. The features that are obtained are then taken as input to RNN network. This becomes the semantic sentence representations.

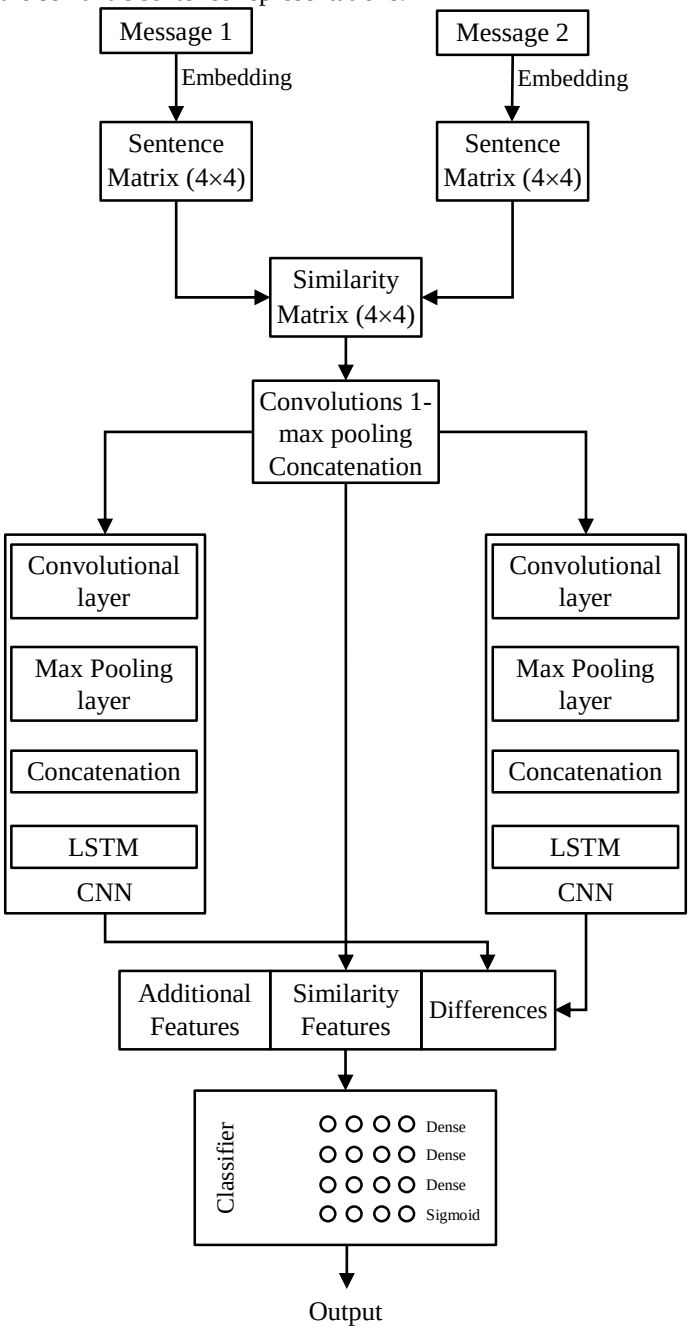

Fig.2. Proposed Deep Learning Architecture

5.2 PAIRWISE WORD SIMILARITY MATCHING

This phase takes two sentences as input, the semantic relativeness between the words in those sentences is taken for determining the similarity in-between those two sentences, and then the pairwise similarity between those two sentences. The n-grams are learnt by applying CNN over the text, then we obtain the word-word similarity pairs from the similarity matrix. This matrix is used as a feature for classification for the paraphrase detection problem.

6. RESULTS AND DISCUSSION ON TWITTER CORPUS

The system is tested with the testing dataset of 840 test entries. The model is trained using this dataset. To remind, there were two models in our proposed work i) Modelling a sentence using CNN and LSTM ii) Matching similarity pair-wise. The reason for using both the models is coarse-grained and fine-grained information in the word level is important for the task of paraphrase detection. First, we used only sentence level modelling for developing the paraphrase detection task. We name this SentenceModArch for detecting paragraphs. The results shown in the Table.3 and Table.4 proves that our models perform very well by giving an F1 score of 0.692. In the next step, pair-wise similarity modelling is used to extract the similarity in the word-level. These features are used only for training the paraphrasing model and it gets a 0.702 as F2 score.

Table.3. Results of the proposed approach compared with state-of-the-art results on SemEval 2015 Twitter dataset

Model	Precision	Recall	F1-Score
Sentence-Mod.	0.724	0.663	0.691
DeepReply Arch.	78.5	0.731	0.742

Table.4. Experimental results for paraphrase detection on the SemEval 2015 dataset

Model	Precision	Recall	F1-Score
Random	0.208	0.500	0.294
Guo and Diab [19]	0.583	0.525	0.655
Zarella et al. [20]	0.569	0.806	0.667
Zhao and Lan [21]	0.767	0.583	0.662
Vo et al. [22]	0.685	0.634	0.659
DeepReply	78.5	0.731	0.742

7. CONCLUSION

In this work, we have introduced a generic approach for paraphrase detection on a deep neural network architecture, which performs well on both user-generated short texts such as tweets. Further, this work figures out the effectiveness of this model's application on emails for answering questions in incoming emails.

Table.5. Example sentence pairs from Enron Email Corpus

Mail 1 (Msg)	Mail 2 (Msg)	Human - Annotation	Prediction	Status
Solomon Burns has made the sale and ended it	The deal was closed by Solomon burns and there no more further talks	Paraphrase	Paraphrase	Correct
1400 customers have invested	Compared to last year 1400 have turned in	Non-paraphrase	Non-paraphrase	Correct
The tech-heavy Nasdaq composite index shot up 5.7% for the week	The Nasdaq composite index advanced 20.59, or 1.3%, to 1,616.50, after gaining 5.7% last week	Non-paraphrase	Paraphrase	Incorrect
Mr. Sheldon said he was incredulous that he would endanger their marriage and family	He hadn't believed he would jeopardize their marriage and family	Paraphrase	Non-Paraphrase	Incorrect

In the process, we have built a pair-wise word similarity finder, which can detect fine-grained semantic equivalent data amid each pair of words in agreed sentences. We have also built a hybrid deep neural network that abstracts coarse-grained data by developing finest semantic illustration of the assumed emails based on CNN and LSTM. The model that we have built contains both sentence modelling and pair-wise word resemblance similarity finding model.

REFERENCES

[1] B. Agarwal, H. Ramampiaro, H. Langseth and M. Ruocco, "A Deep Network Model for Paraphrase Detection in Short Text Messages", *Information Processing and Management*, Vol. 54, No. 6, pp. 922-937, 2018.

[2] K. Amin, "Answering with Cases: A CBR Approach to Deep Learning", *Proceedings of International Conference on Case-Based Reasoning*, pp. 1-12, 2018.

[3] W. Xu, C.C. Burch, W.B. Dolan and Y. Ji, "Extracting Lexically Divergent Paraphrases from Twitter", *Proceedings of International Conference on Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 435-448, 2014.

[4] W. Xu, C.C. Burch and W.B. Dolan, "SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)", *Proceedings of 9th International Workshop on Semantic Evaluation*, pp. 1-7, 2015.

[5] K. Dey, S. Ritvik and K. Saroj, "A Paraphrase and Semantic Similarity Detection System for User Generated Short-Text Content on Microblogs", *Proceedings of International Conference on Computational Linguistics: Technical Papers*, pp. 1-7, 2016.

[6] N. Madnani, T. Joel and C. Martin, "Re-Examining Machine Translation Metrics for Paraphrase Identification", *Proceedings of Conference on North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1-8, 2012.

[7] D. Das and N.A. Smith, "Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition", *Proceedings of the Joint Conference of 47th Annual Meeting of Computational Linguistics*, 2009.

[8] M. Sahi and V. Gupta, "A Novel Technique for Detecting Plagiarism in Documents Exploiting Information Sources", *Cognitive Computation*, Vol. 9, No. 6, pp. 852-867, 2017.

[9] K. Vani and G. Deepa, "Unmasking Text Plagiarism using Syntactic-Semantic based Natural Language Processing Techniques: Comparisons, Analysis and Challenges", *Information Processing and Management*, Vol. 54, No. 3, pp. 408-432, 2018.

[10] Y. Jiang, "Wikipedia-Based Information Content and Semantic Similarity Computation", *Information Processing and Management*, Vol. 53, No. 1, 2017.

[11] Franco-Salvador, Paolo Rosso, and Manuel Montes Y. Gomez. "A Systematic Study of Knowledge Graph Analysis for Cross-Language Plagiarism Detection", *Information Processing and Management*, Vol. 52, No. 4, pp. 550-570, 2016.

[12] S. Arora, Y. Liang and T, Ma. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings", *Proceedings of 5th International Conference on Learning Representations*, pp. 1-12, 2016.

[13] P. Bojanowski, E. Grave and A. Joulin, "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135-146, 2017.

[14] M. Pagliardini, P. Gupta and M. Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional N-Gram Features", *Proceedings of North American Conference on Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 528-540, 2017.

[15] Y. Kim, "Convolutional Neural Networks for Sentence Classification", *Proceedings of Conference on Empirical Methods in Natural Language* Processing, pp. 1746-1751, 2014.

[16] R. Kiros, Y. Zhu, R.S. Zemel and S. Fidler, "Skip-Thought Vectors", *Proceedings of International Conference on Advances in Neural Information Processing Systems*, pp. 3294-3302, 2015.

[17] Y. Kim, "Character-Aware Neural Language Models", *Proceedings of 13th AAAI Conference on Artificial Intelligence*, pp. 1111-1119, 2016.

[18] X. Wang, J. Weijie and L. Zhiyong, "Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts", *Proceedings of 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1-9, 2016.

[19] W. Guo and M. Diab, "Modeling Sentences in the Latent Space", *Proceedings of 50th Annual Meeting of the*

Association for Computational Linguistics: Long Papers, pp. 864-872, 2012.

[20] G. Zarrella, J.C. Henderson, E.M. Merkhofer and L. Strickhart, "MITRE: Seven Systems for Semantic Similarity in Tweets", *Proceedings of 9th International Workshop on Semantic Evaluation*, pp. 12-17, 2015.

[21] J. Zhao and M. Lan, "ECNU: Leveraging Word Embeddings to Boost Performance for Paraphrase in Twitter", *Proceedings of 9th International Workshop on Semantic Evaluation*, pp. 34-39, 2015.

[22] N.P.A. Vo, S. Magnolini and O. Popescu, "Paraphrase Identification and Semantic Similarity in Twitter with Simple Features", *Proceedings of International Conference on Association for Computational Linguistics*, pp. 10-19, 2015.