 doi:10.20944/preprints201908.0073.v1

Self-supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling

Prafull Sharma, Yingbo Li*
Naister, France
psharma@naister.com,yli@naister.com

***Abstract*—In this paper we propose a novel self-supervised approach of keywords and keyphrases retrieval and extraction by an end-to-end deep learning approach, which is trained by contextually self-labelled corpus. Our proposed approach is novel to use contextual and semantic features to extract the keywords and has outperformed the state of the art. Through the experiment the proposed approach has been proved to be better in both semantic meaning and quality than the existing popular algorithms of keyword extraction. In addition, we propose to use contextual features from bidirectional transformers to automatically label short-sentence corpus with keywords and keyphrases to build the ground truth. This process avoids the human time to label the keywords and do not need any prior knowledge. To the best of our knowledge, our published dataset in this paper is a fine domain-independent corpus of short sentences with labelled keywords and keyphrases in the NLP community.**

Keywords—Contextual Keyword Extraction, BERT, Word Embedding, LSTM, Transformers, Deep Learning

I. INTRODUCTION

In a digitalisation driven world, we are witnessing a huge growth in unstructured data. Text data such as social media opinions, tweets, digital documents and blogs are growing over the internet very fast.

For instance, Wikipedia [2] has over 5,836,552 articles, while English Wikipedia has more than 27 billion words in 40 million articles in 293 languages. To leverage and reap the benefits of the growing text data, capturing the importance of text and representing them in a succinct way is a popular area of research in Natural language processing. One method of representing a large text in succinct way is representing them by Keywords and Keyphrases. Thus, Keyword and keyphrase extraction is one of the fundamental research topics in NLP domain [1] [3] [4] [5]. Keywords and keyphrases play an important role in getting the idea behind text data quickly without having to read through the whole text. It finds application in content management space such as Search Engine Optimization, advertisement, and recommendation systems for users. For instance, while visiting an advertisement or a web site, the end users get attracted if the keywords are relevant to their needs. Thus, it is important that keywords capture the meaning, and important aspects of a text data.

A Keyword is a word which could succinctly and accurately describe the subject fully or partially in a document [5]. A keyword is a unigram while a keyphrase is N-grams i.e multiple words, for example 'family' is a keyword and 'family vacation' is a keyphrase. In terms of understandability, human beings prefer keyphrase over keyword because keyphrase contains contextually more information and meaning compared to the keyword whose contextual meaning may be variant in different text environment. For example, the word 'bank' could mean a banking organisation, or it could mean river bank. Thus, context is an important aspect. In this paper we leverage contextual features of text corpus through transformer architecture and use them to develop model for keyword extraction.

While it is easy to extract keywords and keyphrases from long corpus, it is a bit difficult task to extract the same from a shorter sentence. There are several proposed algorithms which successfully extract keywords from long sentence corpus, however, their performance is comparatively less for short sentences. We will discuss some of the methods in proceeding in the further sections. In this paper, the proposed approach, SCKKRS (Self-supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling) is suitable for long as well as short sentence corpus to semantically and contextually retrieve keywords and keyphrases. Our proposed method extracts keywords while focussing on contextual features, and thus, outperforms some of the existing methods. Furthermore, the other advantage is that it performs equally well on long, as well as short sentence corpus for keyword extraction task.

Deep learning-based approaches of keyword and keyphrase extraction requires a well labelled training corpus. Labelling a training data manually is extremely time consuming, yet indispensable to model development. Therefore, in the paper we propose a novel self labelling approach in SCKKRS to achieve self-supervised learning. We use a bidirectional LSTM, coupled with proposed self-labelling algorithm to develop an end-to-end solution.

We open source one domain-independent corpus of short sentences with labelled keywords and keyphrases in this paper. According to our knowledge, this would be the first for this kind of dataset in the NLP community, because most existing corpora are domain-specific and are not short sentence level. [8] [9].

*Corresponding author: Yingbo Li, Email: yli@naister.com

 doi:10.20944/preprints201908.0073.v1

This paper is structured as follows: Section II would be the state of the art for the keyword and keyphrase extraction. We describe in detail the proposed approach, SCKKRS in Section III. While we demonstrate the experimental results of SCKKRS from all different aspects in Section IV. This paper is concluded in Section V.

II. Overview of KeyWord and KeyPhrase Extraction

The keywords and keyphrases retrieval methods can be broadly classified into following: statistical approach, graph-based approach, linguistic approach, machine learning approach and hybrid approach. In this section, we discuss each of the approaches and concepts behind them.

A. Statistical Approach

In a statistical approach of keywords and keyphrases extraction, the frequency measure for statistical features is used to choose top *n* candidates based on linguistic corpus. Most statistical approaches are language independent, so they could be applied to every language if the large corpus is available.

Gerard Salton and Christopher Buckley [3] discussed the importance of an appropriate term weighting system for an effective information retrieval system. Using an external resource such as Wikipedia to ascertain the importance of the candidate phrase [4] is also another possibility.

Additionally, statistical association among candidate keyphrases can be used as a possible proxy of semantic coherence. Rapid Automatic Keyword Extraction (RAKE) [11] by M.W.Berry et al. is a popular keyword extraction algorithm for single document that can be extended to multiple documents. Yutaka Matsuo and Mitsuru Ishizuka [5] presented another statistical algorithm to extract keyword from a single document without relying on a corpus and TF-IDF measurement. In their proposed algorithm, the first frequent terms are determined and then those are clustered based on some similarity measures. The degree of bias of the probability distribution for co-occurrence of any term with those clusters is investigated. If there exists a bias, then it is very likely that the term is a keyword.

We should however notice that most of the statistical approaches are based on frequency metrics of the words in a corpus, and outputs of the algorithms are very much prone to noisy words present in a corpus.

B. Graph Based Approach

Graph based approaches [18] use bag of words with co-occurrence metric and come up with *N*-dimensional vector for each document, where *N* is the number of all possible words in the corpus. The documents may be represented by a cosine similarity matrix from the *N*-dimensional vector. Thus, when we build the graph relation among words and documents, the words in the corpus become the vertices while the edges represent the calculated similarities. Finally, multiple centrality algorithms could be chosen to extract the top nodes as keywords and keyphrases, such as pure degree centrality, Eigenvector centrality [12] and Pagerank [13].

In PageRank [13], the importance of a node is decided by the edges to the neighbouring nodes representing votes for relevance. The ranking score is recursively calculated by considering the weights of these edges and the rank of the neighbouring nodes. While, textrank [14] can be applied to both text summarization and keyword extraction. Textrank uses the concept of prestige in the network and Pagerank to rank the nodes of the graph. The top *n* key words or sentences in the graph are the top ranked nodes. In this way, a list of keywords is extracted from the sentence.

C. Linguistic Approach

Linguistic approach utilizes the linguistic features of the words for keyword detection, so linguistic approach is language dependent. The popular algorithms used in linguistic approach include POS pattern, *n*-gram, NP chunks, etc. Linguistic approach is popularly used in domain dependent corpus [15] [16] [17]. Linguistic approach is popular to use the rule to decide the keyphrase extraction. For example, Adjective+Noun, e.g. linear algebra, and Noun+Noun, e.g. Computer Virus.

D. Machine Learning Approach

Machine learning approaches of keyword extraction are like other machine learning approaches, which are supervised learning methods and need the prior knowledge - training data to learn and output a trained model. The training data for keyword and keyphrase extraction is the corpus and their corresponding labelled keywords and keyphrases in advance. Many successful approaches such as HMM, Naive Bayes, and Support Vector Machine fall into this category.

With the development of deep learning, especially LSTM [6] [7], deep learning has shown its strong capability to processing language and text problems. Zhang Q et al [19] proposes a keyphrase extraction approach for Twitter-like corpus and sites. The authors use deep recurrent neural network (RNN) model to exploit contextual information among keywords to retrieve keyphrases. The rule-based approach is used by the authors to build the training tata from Twitter corpus. Keywords is a mandatory prior knowledge in the approach too.

Rui Meng et al. [20] proposed another deep learning model for keyphrase retrieval. The authors attempt to capture the deep semantic meaning of the content with a deep learning method of generative model for keyphrase prediction with an encoder-decoder framework. The proposed approach is domain specific for scientific publications.

In [22] the authors proposed an approach of keyword extraction of product review based on a bi-directional long short-memory (LSTM) recurrent neural network (RNN). The training data is crawled product review from jd.com.

E. Hybrid Approach

Hybrid approaches combine the advantages of all of the above approaches. The methods using heuristics, such as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2019 doi:10.20944/preprints201908.0073.v1

position, and HTML tags around the words belong to the hybrid approach [23].

The proposed approach in this paper, SCKKRS, is an end to end, and a variant of hybrid approach which could extract both keywords and keyphrases from 'domain-independent' corpus of long paragraphs as well as short sentences. To build the training data of SCKKRS, contextual word features are leveraged. Further, POS pattern, *n*-gram, and NER (Named Entity Recognition) [24] of linguistic approach are used to enhance fine tune the labels for training data. These contextually labelled data are then fed to a model, where we pose the keyword and keyphrase extraction as a problem of word classification by Bidirectional LSTM in deep learning.

III. Proposed Method for Keyword and Keyphrase Labelling and Extraction

The proposed approach focuses on an end-to-end solution for keyword and keyphrases extraction. The "end-to-end" here means that it can perform self labelling of unlabeled corpus, reduction of manual efforts, and contextual keyword extraction. This method provides a generic approach as is it uses domain-independent corpus, as we introduce randomness in picking up data for corpus. Once the self labelling is finished, we then pose the task of keyword extraction as words classification in deep learning by Bidirectional LSTM [6] [7]. So, the contextually labelled training data coupled with bidirectional LSTM methods results in a complete solution.

The proposed approach is explained in following stages as shown in Figure 2: Domain-independent corpus collection, corpus cleaning, corpus self-labelling, keyword extraction model training by bidirectional LSTM. The self labelling stage extracted contextual features from the text by leveraging Bidirectional Transformer Encoders, and out-performs the keyword labels obtained from some of the approaches discussed above, such as RAKE [11] and TextRank [14].

A. *Data collection*

In NLP (Natural Language Processing) community several datasets with corpus especially long paragraphs and their keywords and/or keyphrase are available [8] [9]. However, the existing datasets are not suitable for us because of the following reasons: 1) They are domain specific so cannot be used for generic datasets; 2) They are normally long paragraphs, but not short sentence-length corpus; 3) These dataset are not big enough with respect to the volume; 4) The labelling of keywords and keyphrase are based on frequency based methods, instead of contextual relevance, and are thus, less closer to the ground truth.

For collecting the data, we used Wikipedia as a source. Wikipedia [26] is a popular text corpus source for the research community. Since we need to build a domain-independent corpus, we collect the sentences from Wikipedia web pages randomly in order to ensure the generic nature of collected data, it ensured that the corpus does not belong to a particular domain (for example, sports, politics).

B. *Data cleaning*

Because the sentences of wikipedia article contains special characters and stop words, the dataset from previous step contains a high volume of special characters and stop words. So, we exploit the traditional regular expressions and existing toolkits to pre-process and clean the data.

Raw Data	Cleaned Data
Descent Pass (77°51′S 163°5′E) is a pass leading from Blue Glacier to Ferrar Glacier, in Victoria Land, Antarctica. 77 ° 51.	Descent Pass is a pass leading from Blue Glacier to Ferrar Glacier, in Victoria Land, Antarctica.

Figure 1 - Cleaned sample Data

Figure 1 shows the output after cleaning stage of a sample sentence.

C. *Data self-labelling and Dataset building*

Our novel approach of keywords and keyphrases labelling for sentence-length paragraphs uses a novel self-supervised method of labelling keywords and keyphrases. The approach is to extract keywords based on their contextual relevance with the sentence. Unlike frequency based statistical approaches, which heavily rely on co-occurrences and term frequencies to extract keywords, our proposed approach considers contextual relevance of words to the sentence. Thus, it leverages both the semantic and the contextual features of words and phrases while extracting them.

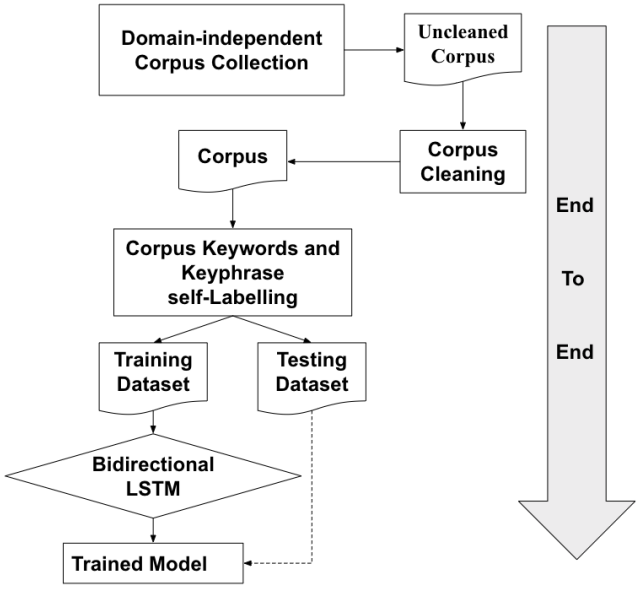

Figure 2 - flowchart of the proposed approach

The contextual features of the words in sentences are extracted using Bidirectional Transformers [10], which are based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality compared to sequence models.

BERT - Bidirectionally Transformers [27] is basically deep bidirectionally trained, similar to OpenAI's GPT model [28] which is trained unidirectionally. BERT has reportedly obtained some very significant results on 11 NLPtasks, including pushing the GLUE benchmark to 80.4% (7.6%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2019 doi:10.20944/preprints201908.0073.v1

absolute improvement), MultiNLI accuracy to 86.7 (5.6% absolute improvement), and the SQuAD v1.1 question-answering test *F1-score* to 93.2 (1.5% absolute improvement), outperforming human performance by 2.0%. So we use the BERT as the feature extraction of words in the corpus.

We feed the sentence to BERT, and obtain the contextual feature vector of each word, as shown in Figure 3. The vectors of words in a sentence is averaged in order to get its sentence embedding vector. Then we choose the words close to sentence embedding vector. The idea is that a keyword should capture meaning of sentence, and thus should be closer to the sentence embedding. The similarity of the embeddings to the sentence embedding is obtained using cosine similarity metric (Equation 1).

$$Sim_i = cos(w_i, W) \quad (1)$$

Sim_i is the cosine similarity between the word embedding vector w_i of a word i, and the sentence embedding vector. Once the candidate keywords are extracted, we obtain our keyphrases through the rule of adjacent keywords.

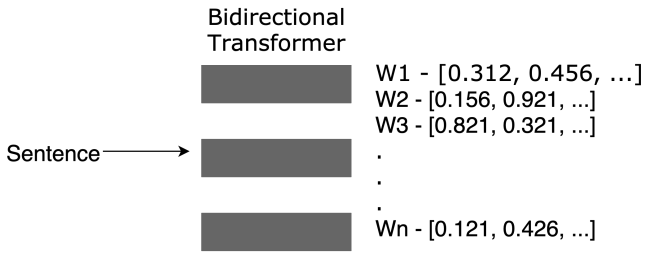

Figure 3 - Embedding vector of the word by BERT

The self-labelling of corpus without human intervention reduces a major dependency on manual construction of well labelled corpus for keyword and keyphrase extraction. Key-word Extraction Model

After the self labelling stage, the labelled corpus was divided into training and validation sets, which was then fed to deep learning-based keyword extraction model. As presented in Figure 2, we pose the problem of keyword extraction as a classification problem i.e. given the contextual features of sentence, which of the words in the sentence can be classified as a candidate for keyword. Thus, posing the problem as a binary classifier.

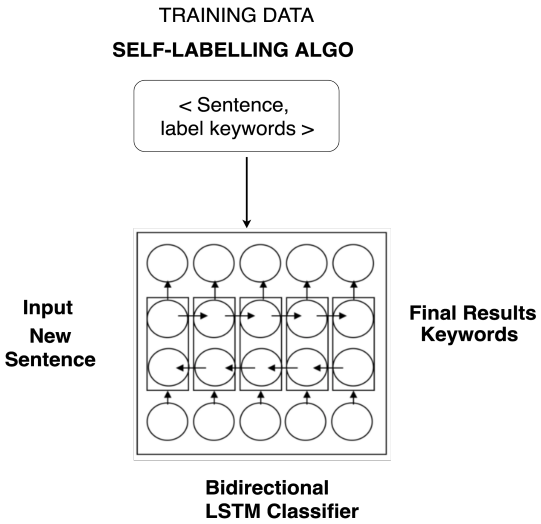

Figure 4 - Keyword extraction training process in bidirectional LSTM

The bidirectional LSTM [6] [7] takes a sentence as a sequence, along with the keywords and keyphrases labels coming from the self labelling. The labels are one hot encoded in following manner -

1 - Word is a keyword.

0 - Word is not a keyword.

This <sentence, label> pair is then passed to the model for training. Regularisation methods, such as dropout were adopted to avoid high variance and low bias.

It should be noted that the labels are extracted based on contextual features. We use Figure 4 to illustrate this training process in bidirectional LSTM.

IV. Experimental Results

In the experimental results, we will discuss the experimental results of self-labelling and its closeness with the ground truth. The closer the results are to the ground truth, the better the performance is. Further, we discuss the comparative analysis of results of our proposed method to the results of other existing solutions such as RAKE, SG-Rank, TextRank. It can be observed that, our proposed solution, comes out to be closer to the ground truth on a contextual scale.

A. Contribution of Open source corpus of short sentences

In the community, majority of the open-source and public corpus are domain specific, in addition to that, the corpus with the labelled keywords and keyphrases are even more rare. Furthermore, most of them are corpus with longer sentences, sometimes as long as a paragraph, so these difficulties make them unsuitable for building deep learning model.

Therefore, we opensource partial of our sentence-length corpus to the community, which is available here: https://github.com/naister/Keyword-OpenSource-Data. To the best of our knowledge, it is the first public sentence-length corpus with labelled keywords and keyphrases in the community.

B. Self-labelling of keywords and ground truth keyphrases

Manual labelling of text corpus for keywords poses following problems -

- Time consuming.
- Requires human effort
- Requires domain expertise.
- Infeasible for the huge unlabeled corpus.

Our proposed self-labelling approach resolves these issues by making the process automatic. For analysing the performance of our self-labelling approach, we visually show, the contextual closeness of self-labelled keywords compared to the ground truth. Then we compare the contextual closeness of keyworother methods - RAKE and TextRank, to the ground truth.

For performance evaluation we use the corpus of well-known INSPEC dataset [29] with, and DUC Dataset [30].

Figure 5 shows keywords/keyphrases from our proposed approach and the keywords of ground truth. We could see that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 August 2019 doi:10.20944/preprints201908.0073.v1

our approach has retrieved all the keywords/keyphrases, and even given more useful keywords and keyphrases compared to ground truth.

Long Sentence
accelerated simulation of the steady state availability of non markovian. A general accelerated simulation method for evaluation of the steady state availability of non markovian systems is proposed .it is applied to the investigation of ad class of systems with repair. numerical examples are given.

Keyword/Keyphrases Proposed approach	Gold Dataset Keywords/Keyphrases
accelerated simulation	accelerated simulation
steady state availability	steady state availability
general accelerated simulation	general accelerated simulation
Non markovian	non markovian systems
Numerical examples	numerical examples
Markovian method	
Numerical	
system	
evaluation	

Figure 5 - Self-labelling vs Ground-truth vs Other-keyword extraction methods

Furthermore, we compute the cosine similarity with BERT features between corpus sentences for Inspec dataset and DUC dataset keywords/ keyphrases from approaches of ground truth, RAKE, TextRank, and our proposed approach. The results are shown in Figure 6. In Figure 6, *g* denotes ground truth keywords, *r* denotes RAKE generated keywords, *t* denotes TextRank generated keywords and *p* denotes proposed self-labelled keywords.

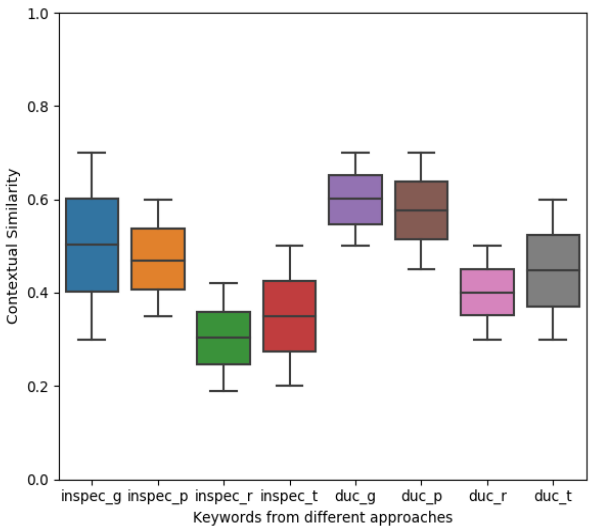

Figure 6- The objective comparison of self-labelling

C. *Keywords and keyphrases extractor*

We use, own corpus having short sentences and the gold standard/ground truth datasets having long paragraphs (such as INSPEC dataset [29] & DUC dataset [30]) to conclude that the proposed model trained by our corpus (with self labelled keywords and keyphrases) outperforms RAKE, SG Rank and TextRank algorithms on both, short sentence-length corpus and long-paragraph corpus.

We feed the training data from our self-labelling corpus into the model training and get model performance i.e. precision, recall, *F1*-score, and support as shown in Table 1. In Table 1, ***1*** means that the word is predicted as the keyword, while ***0*** means the word is predicted as non-keywords.

In order to demonstrate the quality of both extracted keywords for both long-paragraph corpus and sentence-length corpus, we show the sample results in Figure 7 and Figure 8. In subjective view we can conclude that the proposed approach outperforms other existing keywords retrieval algorithms.

Furthermore, we use the gold standard domain-specific dataset INSPEC with human labelled keywords and our testing dataset to validate the performance of the proposed model on a large scale. The results are shown in Figure 9. We could see that in both datasets the proposed approaches have achieved results very close to the ground truth, and better statistical data than other approaches when we consider Figure 6 too. Our result similarity is also better than ground truth of INSPEC and DUC, which proves the better semantic and contextual keyword extraction than gold standard in case of INSPEC and DUC.

In Figure 9, *g* is ground truth**,** *m* represents keywords or keyphrases from our trained model, and *s* is self-labelling keywords or keyphrases from our approach.

Table 1 - Model performance

	Precision	***Recall***	***F1***
0	0.90	0.86	0.88
1	0.76	0.82	0.79

Lisa did the best she could to draw a map on the small piece of paper.
Our keywords: lisa, map, small piece, paper
Rake: draw, could, best, lisa, small piece, paper, map
TextRank: best, piece, map, small

The Holiday destination was so much fun with kids
Our keywords: holiday destination, fun, kids
Rake: much fun, holiday destination, kids
TextRank: destination, fun, Holiday, kids

Figure 7 - Sample results for sentence-length corpus

V. CONCLUSION

In this paper, we have published the first open-source and generic sentence-level corpus with the labelled keywords and keyphrases in the community. The sentence corpus is from Wikipedia, with random articles, to make it generic in nature. The data is then labelled by our novel self-labelling approach based on contextual word features. As can be seen in results, keywords and keyphrases extracted from the proposed self-labelling approach is very close to human-labelling (ground truth). By using our self-labelled corpus we have trained the bidirectional LSTM as keywords and keyphrases extractor. The trained model outperforms the existing approaches of

 doi:10.20944/preprints201908.0073.v1

keywords and keyphrases retrieval. We believe, in future, there is still a high probability to improve our keywords and keyphrases extractor by fine tuning the deep learning model based on bidirectional LSTM.

Not to be confused with the businessman of the same name, chairman of Bwin.Party Digital Entertainment, Simon Duffy (born 13 February 1965) was formerly the chief executive of the social enterprise company In Control.

Our Approach	Rake Keywords	SG Rank	TextRank
chief executive, social enterprise, entertainment, simon duffy, party, bwin, chairman, digital, enterprise company,	control, name, formerly, confused, entertainment, simon duffy, chairman, bwin, businessman, party digital social enterprise company	control, company, enterprise, executive, social, february, simon, duffy, chief, entertainme nt, digital	chairman, company, bwin, enterprise, party, social, digital, executive, entertainmentsi mon, chief

Figure 8 - Sample results for long-paragraph corpus

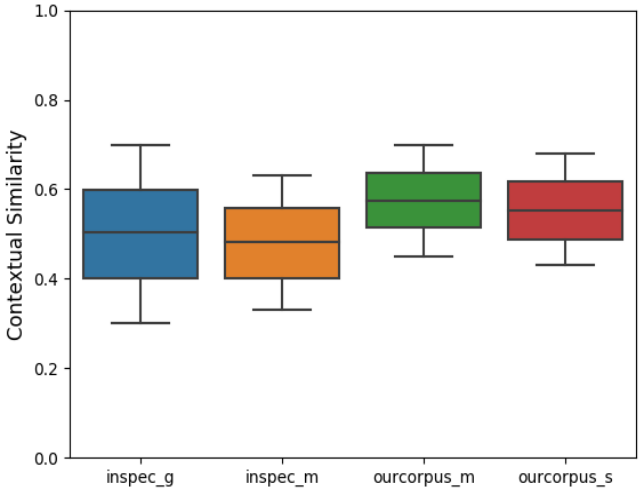

Figure 9 - Comparison of trained model to ground truth

References

[1] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010b. SemEval-2010 Task 5: Automatic key phrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 21–265.

[2] https://en.wikipedia.org/wiki/Wikipedia:Statistics

[3] Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2014.

[4] Bharti, Santosh Kumar, and Korra Sathya Babu. "Automatic keyword extraction for text summarization: A survey." arXiv preprint arXiv:1704.03242 (2017).

[5] Siddiqi S, Sharan A. Keyword and keyphrase extraction techniques: a literature review[J]. International Journal of Computer Applications, 2015, 109(2).

[6] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[7] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM[C]//2013 IEEE workshop on automatic speech recognition and understanding. IEEE, 2013.

[8] https://github.com/zelandiya/keyword-extraction-datasets

[9] http://semeval2.fbk.eu/semeval2.php?location=tasks#T6

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, - 'Attention Is All You Need', Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[11] Rose S, Engel D, Cramer N, et al. Automatic keyword extraction from individual documents[J]. Text mining: applications and theory, 2010: 1-20.

[12] Lahiri S, Choudhury S R, Caragea C. Keyword and keyphrase extraction using centrality measures on collocation networks[J]. arXiv preprint arXiv:1401.6571, 2014.

[13] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.

[14] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[15] R. Barzilay, M. Elhadad, "Using lexical chains for text summarization," Advances in automatic text summarization, 1999, pp. 111-121.

[16] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in: Proceedings of the 2003 conference on Empirical methods in natural language processing, ACL, 2003, pp. 216-223.

[17] G. Salton, A. Singhal, M. Mitra, C. Buckley, "Automatic text structuring and summarization," Information Processing & Management, vol. 33 (2), 1997, pp. 193-207

[18] Beliga S, Meštrović A, Martinčić-Ipšić S. An overview of graph-based keyword extraction methods and approaches[J]. Journal of information and organizational sciences, 2015, 39(1): 1-20.

[19] Zhang Q, Wang Y, Gong Y, et al. Keyphrase extraction using deep recurrent neural networks on Twitter[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 836-845.

[20] Meng R, Zhao S, Han S, et al. Deep keyphrase generation[J]. arXiv preprint arXiv:1704.06879, 2017.

[21] Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, Chris Dyer, "Problems With Evaluation of Word Embeddings Using Word Similarity Tasks", Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP.

[22] Wang Y, Zhang J. Keyword extraction from online product reviews based on bi-directional LSTM recurrent neural network[C]//2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, 2017: 2241-2245.

[23] Humphreys J B K. PhraseRate: An HTML Keyphrase Extractor[J]. Technical report, 2002.

[24] Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study[C]//Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011: 1524-1534.

[25] Loper E, Bird S. NLTK: the natural language toolkit[J]. arXiv preprint cs/0205028, 2002.

[26] http://www.wikipedia.com

[27] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[28] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1: 8.

[29] Debanjan Mahanta, John Kuriakose, Rajiv Ratn Shah, Roger Zimmermann: Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings.

[30] X. Wan and J. Xiao. Single Document Keyphrase Extraction Using Neighborhoofd Knowledge. In Proceedings of AAAI, pages 855–860, 2008