WIKIFICATION

A PREPRINT

Anil Ozdemir
Department of Computer Science
Sabancı University
Istanbul, TR
aozdemir@sabanciuniv.edu

Furkan Coşkun
Department of Computer Science
Sabancı University
Istanbul, TR
furkancoskun@sabanciuniv.edu

June 22, 2020

1 Introduction

Information retrieval has played a huge role in the representation of human knowledge. With the amount of textual data are exponentially increasing, creating these representations successfully became an important goal [1].However, the problem of understanding the context and background of information became a harder task for the user. Thus, the main focus of the task is to envision new information that enhances the learning experience of the user by filling the gap in the context. In this project, we focus on the Wikification which is the task of identifying concepts and entities in text-based datasets and disambiguating them into their corresponding Wikipedia page. In other words, given a document D containing set of concepts and entities mentions M, the ultimate goal of wikification is the finding the most accurate mapping from mentions to Wikipedia links. This kind of mapping requires an understanding of the text as well as background knowledge that is often needed to determine the most appropriate title [2].

The motivation of the Wikification and derives from the need of linking the mentions of concepts, artifacts and entities to external or internal sources with in-depth information that will equip the reader with a better understanding of the articles. In the literature, the authors explored this idea under the title of entity linking.Entity linking a task in Information Extraction that links the entity mentions in a text collection with their corresponding knowledge-base (KB) entries. [3]. Thanks to the EL, we may take advantage of a large amount of information available in publicly available knowledge bases, such as Wikipedia, DBpedia, Wikidata to access real world entities. Thus, Wikification has become a more complex version of the entity linking that requires systems to map mentions to the Wikipedia pages describing the entities mentioned.

The work in this area seems to have gained momentum in 2012 and 2013 and among its first examples is the "PARMA" which is shortening for the A Predicate Argument Aligner [4]. Parma is a cross-document, semantic predicate and argument alignment system that combines several linguistic resources familiar to researchers in the fields such as recognizing textual entailment and question answering. There are significant previous works that use the neural context and entity encoders in entity linking systems. In the [5] authors proposed a novel embedding method for named entity disambiguation that jointly maps words and entities into the same continuous vector space.

In the [6], authors found a solution to an entity linking problem where both the entity mentions and the target entities are within the same social media platform. For this aim, the authors constructed a dataset called Yelp-EL and linked the business mentions in Yelp reviews to their corresponding businesses on the platform. Another study named as End-to-End Neural Entity linking proposed in 2018 [7].The designed system jointly discovers and links entities in a text document that consider all possible spans as potential mentions and learn contextual similarity scores over their entity candidates. Authors proposed the first neural end-to-end entity linking model by jointly optimizing the entity recognition and linking.

In this work:

(i) We made experiment on the task of Wikification(see 2.1) which focuses on enhancing the learning experience.

(ii) We performed our experiments on the Wikidump and TREC Washington Post Collection data (see 3).

2 Task Description

2.1 Wikification

The goal of the Wikification task is to help users contextualize news articles while reading. This will be achieved by linking entities, artifacts, mentions of concepts, etc to another news article or Wikipedia articles which will help the user understand the article better. Briefly, Wikification is the task of hyperlinking of concepts, entities, or references to a different resource that helps to gather more information.

For implementing such a task, first we need to have a knowledge base that contains the target entity. After that for each token in the sentence, we are trying to estimate a probability distribution of knowledge base and called them as a candidates.If the probability output of these candidates greater then a certain confidence level which is a hyper-parameter , then we assign the corresponding entity to the token. To estimate probabilities , we used a Bi-directional LSTM model and perform many-to-many sequence prediction.

3 Dataset

The data which will be used for this task is the Wikipedia Dumps and TREC Washington Post Collection.

Wikipedia Dumps - Wikipedia offers all available content to interested users via Wikipedia dumps.These databases may be used for mirroring, personal use, informal backups or database queries and in the form of wikitext source and metadata embedded in public XML files. In this project, we used XML dumps that have large volume of text data that contains articles and corresponding Wikipedia links [1] in the 2020.

Washington Post - It first released in 2018. The data contains 608,180 news articles and blog posts between the years 2012 - 2017. The articles are split into content paragraphs. Also, the media sources are referenced by URL to the articles. The exact and near-duplicate articles have been removed from the data.

4 Pre-processing

Throughout the project, since the raw data used is in the form of meta-data, a couple of pre-processing steps are required before performing such a task. For the initial step, we adopted the document cleaning process of previous successful participant [8] to remove unnecessary articles. The approach is documents were processed according to the Lexicographic order of the titles of the documents. During the processing time, meta-data structure that contains author name, publishing date and document title was extracted and stored. Then, we discarded the duplicate meta-data structure from the meta-data set. Also, we checked the document types and removed the ones that belong "Opinion", "The Post-view" and "Letters to the editor". These specific types are declared as "None-background" according to the guideline. Also, we stored the id, title, timestamp and the actual text of each paragraph and paragraph titles were stored. For pre-processing part of dumps, we simply parse the entire input meta-data and extract the lines of texts and corresponding wiki-links of entities. Using that, we create knowledge bases that contain all of the entities. All pre-processing steps are the following:

- Filtering and cleaning corrupt texts and duplications from raw data.
- Accessing data stored in metadata using web scraping tools.
- Removing unnecessary tags and extract the remaining plain text and text with associated Wikipedia link.
- Eliminating stopwords and punctuations.
- Removing corrupt hyperlinks.
- Create dictionary and knowledge base that contains wiki-entities.

5 Experimental Setup and Methods

5.1 Many-to-Many Sequence Classification

Many-to-many sequence classification is mostly applied on the tasks part-of-speech(POS) tagging and Named Entity Recognition(NER). In POS tagging and NER the goal is to assign a label to each of token sequentially. The labels

[1] https://dumps.wikimedia.org/enwiki/

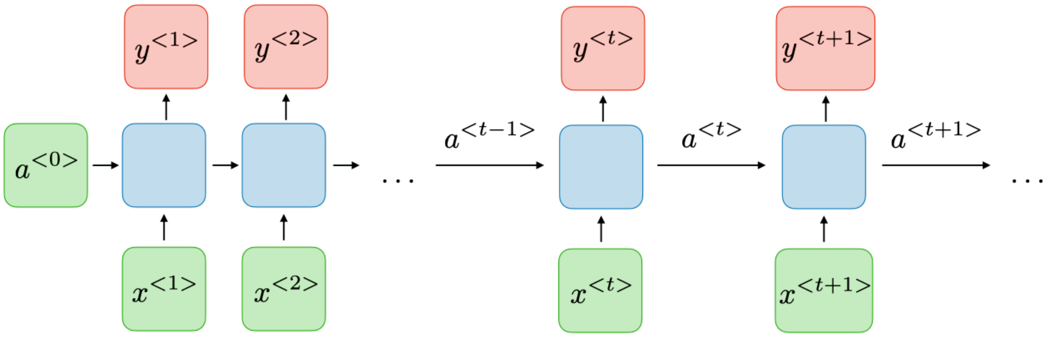


Figure 1: RNN sequence classification image taken from stanford.edu

are defined by the corpus specifications. For example, in the context of POS tagging, the assigned labels are noun, verb, article, adjective, etc. In this work, we adjusted our method to represent the same architecture as in tasks POS tagging and NER. Therefore, the goal is to assign a wiki-link to each token in sequence. Recurrent Neural Networks (RNN) are capable to handle sequential data. The sequence consisting tokens are fed into the RNN units sequentially and each hidden state produces an output as demonstrated in figure 1. The outputs are the vectors with the size of the knowledge base containing Wikipedia links. The RNN architecture allows us to understand the information contained in the sentence sequentially at each time step with hidden state outputs are being fed recursively. However, recurrent neural networks fail to capture long-term dependencies. Therefore, Long-Short Term Memory (LSTM) networks that are specifically proposed to obtain long-term dependencies are preferred in this work [9].

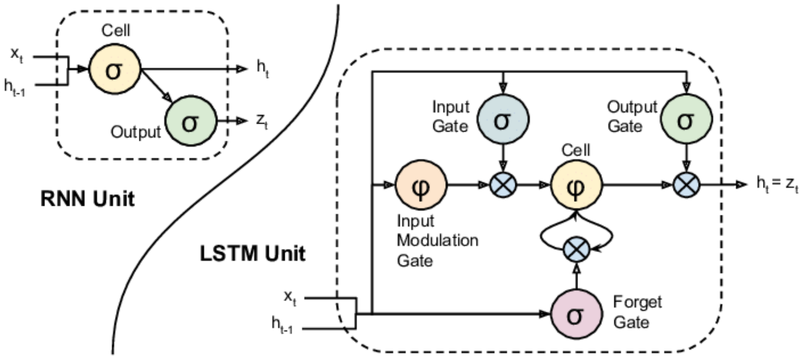


Figure 2: RNN unit(on left) and LSTM unit(on right) [9]

The lack of obtaining long-term dependencies in RNN units is caused by vanishing gradient which gradient is vanished in first states during backpropagation in long sequences. On the other hand, LSTM units consist of input, forget and output gates that update the cell state and hidden state during the sequence (see Figure 2). This provides the capability of learning long-term dependencies [9]. Thus, LSTM units are used in our model.

Even though the LSTM architecture captures the long-distance dependency information, in theory, the architecture can fail in practice. To reduce this issue, even more, bidirectional architectures are leveraged. The bidirectional LSTM is applied with two LSTMs run over the sequence forward and backward to capture both previous and future information for each token. The forward and backward LSTM outputs are concatenated at each time step. The concatenated output is passed to the activation function to produce the final output. This approach provides more information to the token at that time step about the surrounding tokens. The architecture is demonstrated in Figure 3. The figure also demonstrates the experimented model architecture except for the activation function. The activation function used in the experimented model is softmax instead of sigmoid as shown in the figure.

To sum up, the task of predicting a Wikipedia link within the constructed knowledge base, a bidirectional LSTM architecture is leveraged. The tokenized sentence is embedded into vector space and fed into the forward and backward LSTMs. Softmax operation is applied to the concatenated output to produce a confidence-like output sentence.

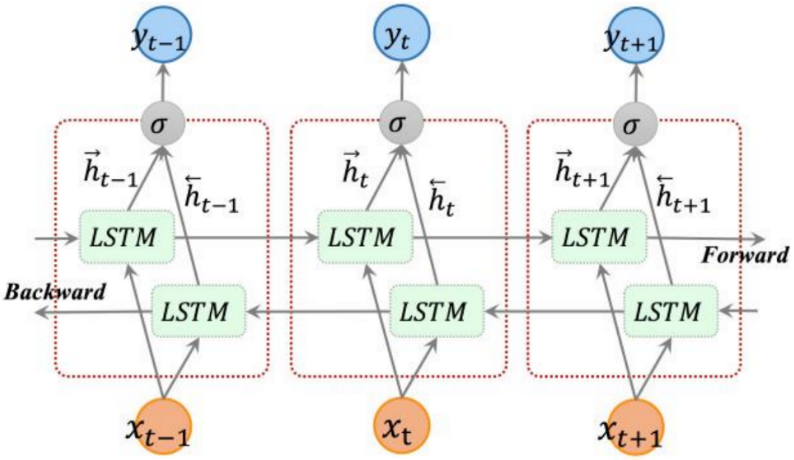


Figure 3: Bidirectional LSTM for sequence labeling [10]

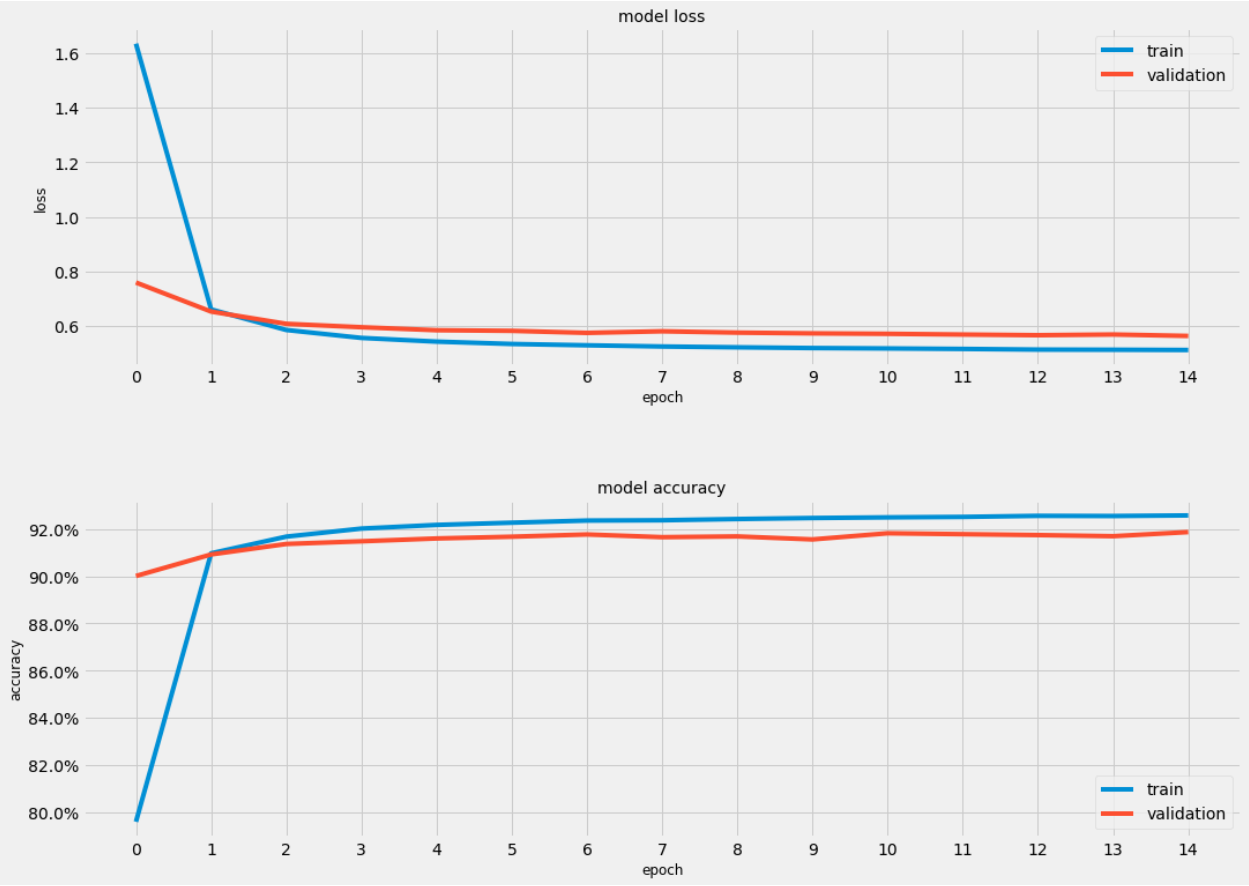
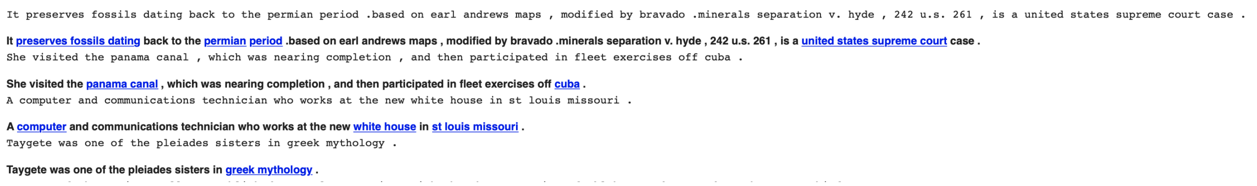


Figure 4: Illustration of Loss and Accuracy on Training and Validation set

6 Results and Conclusion

The evaluation part is one of the most difficult parts of the project because there is no study that we can show as a baseline since there is no study using this data to solve the Wikification problem.

The evaluation of the models is done with respect to the quantitative measurement using accuracy and loss. As can be seen in figure 6, we achieved 93 % success on training data and 92% score on validation data. This accuracy is the ratio of the model to accurately predict entities on sentences. We used 20 % of all data for validation and 20 % for testing and the rest for training and we achieved this accuracy by comparing predicted results with labeled data. For this step, this score sufficient for the Bi-LSTM model and qualitative measurements allows us to see the performance of the model more clearly.



```
It preserves fossils dating back to the permian period .based on earl andrews maps , modified by bravado .minerals separation v. hyde , 242 u.s. 261 , is a united states supreme court case .

It preserves fossils dating back to the permian period .based on earl andrews maps , modified by bravado .minerals separation v. hyde , 242 u.s. 261 , is a united states supreme court case .
She visited the panama canal , which was nearing completion , and then participated in fleet exercises off cuba .

She visited the panama canal , which was nearing completion , and then participated in fleet exercises off cuba .
A computer and communications technician who works at the new white house in st louis missouri .

A computer and communications technician who works at the new white house in st louis missouri .
Taygete was one of the pleiades sisters in greek mythology .

Taygete was one of the pleiades sisters in greek mythology .
```

Figure 5: Example outputs for the qualitative analysis of the experimented model

For further investigation of the sequence to the sequence prediction model, the qualitative analysis is demonstrated in Figure 5. The results show a reasonable result in terms of capturing entities and wikification. However, the major drawback is model cannot perform wikification for the entities that did not appear in the training set. The initial reason for this type of problem is the lack of a training set and our equipment is not capable of handling more data. To overcome this problem, a huge amount of data and enough equipment is required.

To conclude, we solved the problem of entity linking and wikification using the Bidirectional Lstm model on the data we have. While solving the problem, we could not compare the model we built with the results in the literature because this data was not studied before and the subject of wikification was the subject of competition for the first time this year. Despite these drawbacks, we achieved reasonable results that can be seen visibly on the problem.

References

[1] Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. Learning cross-context entity representations from text, 2020.

[2] Xiao Cheng and Dan Roth. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, 2013.

[3] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. What should entity linking link? 2018.

[4] Travis Wolfe, Benjamin Van Durme, Mark Dredze, Nicholas Andrews, Charley Beller, Chris Callison-Burch, Jay DeYoung, Justin Snyder, Jonathan Weese, Tan Xu, et al. Parma: A predicate argument aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 63–68, 2013.

[5] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. *arXiv preprint arXiv:1601.01343*, 2016.

[6] Hongliang Dai, Yangqiu Song, Liwei Qiu, and Rijia Liu. Entity linking within a social media platform: A case study on yelp. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2032, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[7] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*, 2018.

[8] Agra Bimantara, Michelle Blau, Kevin Engelhardt, Johannes Gerwert, Tobias Gottschalk, Philipp Lukosz, Shenna Piri, Nima Saken Shaft, and Klaus Berberich. htw saar@ trec 2018 news track. In *TREC*, 2018.

[9] Jeff Donahue, Lisa Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *Arxiv*, PP, 11 2014.

[10] Zhiyong Cui, Ruimin Ke, and Yinhai Wang. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *CoRR*, abs/1801.02143, 2018.