Cleansing, organizing and training: Two guidelines for generating attractive news headlines for social media

ABSTRACT

In this paper, we explore the challenge of automatically generating attractive news headlines for social media, which can be good introductions to make news articles go viral. To this end, we propose a novel method for identifying key sentences that are useful for generating viral news headlines from a given news article. This problem can be formulated as supervised sequence labelling that employs the most popular microblog post mentioning the news article as supervised information, and a recurrent neural network (RNN) can be used for this purpose. However, we show that a naïve implementation of this approach does not work well, due to noises contained in both microblog messages as the ground-truth headlines and news articles, and data cleansing and organizing for news articles and ground-truth headlines are rather critical for improving the performance of sequence labelling. We then propose a method for organizing a dataset for training accurate models for supervised sequence labeling. The experimental results demonstrate that our proposed method greatly improve the accuracy of key sentence identification.

CCS CONCEPTS

I.2.7 [**Artificia Intelligence**]: Natural Language Processing

KEYWORDS

social media, news headline generation, data cleansing, recurrent neural network

1 INTRODUCTION

The spread of social media made it possible for people to read the news easily anytime, anywhere. Not only news articles but also statements written by their friends or acquaintances and corporate advertisements are being distributed on social media. According to a white paper on information and communications[1], 66.5% of Japanese citizen are reported to be social media users, increased from 41.4% in 2012. Especially it's increasingly used by young people and more than 50 % of 20's are reported to use social media such as LINE, Facebook and Twitter, which are now their important information infrastructure for their daily communication. However, the information on social media is a mixture of wheat and chaff, where a lot of inaccurate information such as lies and fake news are included, which is now becoming more and more problematic to the whole society.

During the 2016 US presidential election, fake news with headlines such as "Pope Francis shocks world, endorses Donald Trump for president" and "FBI agent suspected in Hillary email leaks found dead in apparent murder-suicide" were widely shared on social media, which is said to influence the result of the election greatly[2]. As for the 2017 French presidential election, a research shows that 40 % of the related news on Twitter was fake[3]. In Germany where the general election will be held in September 2017, a bill to prevent the spread the fake news was submitted by the government, as such fake news is considered to bring social confusion and to promote cleavages in a diverted society[4]. In Japan as well, fake news spread on social media has been causing the social confusion. For example, at the time of the 2011 Great East Japan Earthquake, a false story about the fire of the Cosmo Oil Refinery in Chiba Prefecture was distributed widely on social media. At the time of the 2016 Kumamoto Earthquakes, a false story of "a lion is running away from the zoo in town" was distributed widely on social media, which resulted in an arrest of a person who posted it online[5]. In this way, the issue of fake news in Japan has been caught as a problem occurring in the disaster such as the earthquake. Therefore, it can be said that the studies related to fake news have been done as the false story diffusion at the disaster.

Previous researches revealed that the inaccurate information such as lies and fake news tend to spread more widely on social media than the factual news [1,14]. However, it is still unclear how factual news and accurate information spread on social media. News organization today are requested to deliver accurate factual news to interested readers effectively. News organizations have been trying to direct the readers to their articles, letting them select their articles among vast amounts of news articles. According to Japan Newspaper Publishers and Editors Association, 43 Japanese newspaper companies among 83 utilize social media [6]. Some are making efforts to write creative introductions for their news articles posted on social media to attract more readers. A social media account manager of Asahi Shimbun, one of the most popular newspaper in Japan, recognizes that traditional headlines are not attractive enough to make the news go viral on social media and novel approaches should be introduced for organizing attractive headlines [7].

[1] http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nc242220.html

[2] http://www.newsweek.com/fake-news-trump-clinton-pizzagate-paedophile-election-521797

[3] http://af.reuters.com/article/idAFKBN17M314?pageNumber=2&virtualBrandChannel=0

[4] http://www.bbc.com/news/technology-39269535

[5] http://www.asahi.com/articles/ASJ7N6HWDJ7NTIPE034.html

[6] http://www.pressnet.or.jp/data/media/media01.html

[7] http://web-tan.forum.impressrd.jp/e/2015/07/14/20270

However, organizing attractive news headlines for social media heavily relies on the experience and institution of the editor.

This research aims to develop a system to automatically generate attractive news headlines posted on social media, which is expected to explain the key points of the news articles as well as contribute to their social distribution. To this end, we focus on the problem of identifying key sentences in a given news article as material for generating attractive news headlines. This problem can be regarded as supervised sequence labeling that employ the most popular social post as the ground-truth headline, and recurrent neural networks (RNN) are known to be one of the promising models for this purpose [3]. However, we present that naively following the above approach yields unpleasant results since both news articles and social posts that forms the training data contain unfavorable noises. We then proposed a method for organizing a dataset for accurate model training. Experimental results demonstrate that out proposed method greatly improces the accuracy of identifying key sentences in a given news article.

2 RELATED WORK

The goal of our proposed method is to generate headlines for news articles, which has been well explored by the natural language community. Kourogi et al. [2] tried to reveal the source of virality as being news headlines. They interviewed journalists who are writing news headlines on social media as their professions, and revealed the important factors to make news articles go viral on social media by investigating articles of Huffington Post Japan and their official Twitter posts for introducing the articles, and proposed a method to select the most appropriate social posts among a given set of candidates. Zajic et al. [9] proposed an HMM-based method for selecting headline words from the beginning of a given news article. Dorr et al. [10] and Wang et al. [11] focused on the use of lead paragraphs, which are believed to be the most appropriate for summarizing news articles, and proposed methods for shortening a lead paragraph to generate a headline. Headline generation is closely related to and can be regarded as a sub-task of document summarization. Rush et al. [25] first introduced a neural attention model for abstractive sentence summarization, and evaluated the performance for the headline generation task. Takase et al. [26] incorporated structural syntactic and semantic information into a neural attention model. Several news providers have a strong interest in recent progress on automatic headline generation, and they have applied several methods and published them on their platforms [12]. However, almost all previous research has been aimed at summarizing news articles, and virality on social media was not considered.

Several studies have also tried to understand, quantify and predict content virality. Virality can be roughly categorized in terms of two sources, namely content itself and diffusion routes. As regards the contribution of content to virality, Kourogi et al. [2] attempted to reveal the source of virality as being news headlines and the relationships between headlines and news articles. Guerini et al. [15] and Deza et al. [16] proposed methods

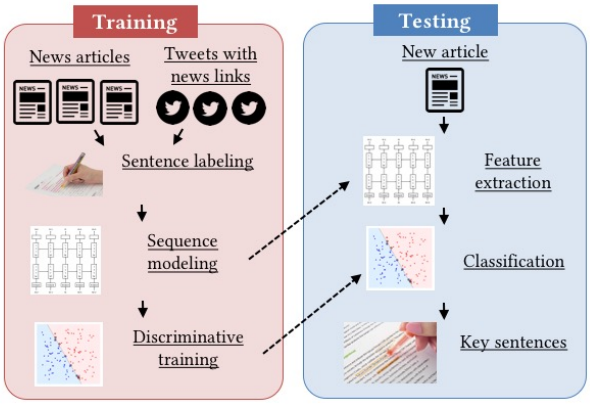

Figure 1: Framework of proposed method

for estimating the virality of image content. In another approach, Park et al. revealed that bad news spreads much faster than other types of information on social media [17]. Jenders et al. analyzed Twitter and user functions [18]. Resnick et al. [19] built a system to allow journalists to distinguish between rumors and their corrections. Gabielkov et al. [20] revealed the way in which many news articles become known by investigating the number of clicks on news articles that are shared on Twitter. However, these previous studies focused solely on understanding content virality, and the generation of content with high virality has yet to be explored.

3 Our model

As described in the introduction, our proposed method for identifying key sentences useful for generating attractive news headlines can be formulated as supervised sequence labeling. In this section, we present out model for supervised sequence labeling that combines an RNN and a support vector machine (SVM).

3.1 FRAMEWORK

Figure 1 shows the framework of the proposed method. It requires ground-truth data that contain news articles and sentence-wise labels representing the sequences that should be extracted as key sentences from a given news article. One of the main contributions of this paper is to demonstrate that we have to carefully collect, cleanse and organize datasets for key sentence identification, and this is much more important than developing sophisticated machine learning models for this purpose. An initial collection will be presented in Section 4.1, its problems will be discussed in Section 4, and a refined collection will be described in detail in Section 5. We used the obtained ground-truth datasets to train an RNN model for supervised sequence labeling, where the input is a news article and the target is its sentence-wise labels. Although the proposed RNN itself can estimate key sentences from a given news article, we add an SVM classifier to provide a post-processing step for

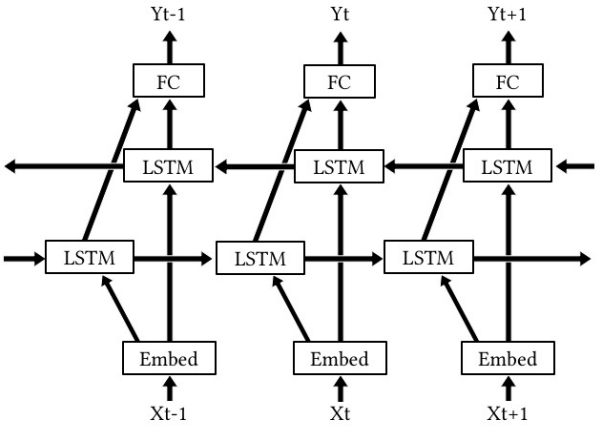

Figure 2: Architecture of proposed model

improving performance. The proposed model will be described in detail in Section 3.2.

3.2 MODEL ARCHITECTURE

We used an RNN to select key sentences from a news article, and trained it with the data constructed in the previous section. Figure 2 shows the architecture, where we follow the bi-directional long-short term memory (BLSTM) model [24], which has been demonstrated to provide promising performance as an acoustic model for speech recognition. This BLSTM model can handle previous and future contexts by introducing bidirectional structures and long-term dependences into a model to achieve key sentence selection in a holistic manner. Since our BLSTM model requires word-wise labels for training, we annotate all the words in a sentence with a sentence label indicating whether or not the sentence should be selected as a key sentence. Word embedding is first pre-trained with Word2Vec [Mikolov2013] with Japanese Wikipedia articles as a corpus, and later the entire network is fine-tined to minimize the cross entropy between the ground-truth labels and soft-max estimates of outputs. Since we are handling Japanese documents, we need a tokenizer to extract words from sentences. We used MeCab [22] for this purpose, and ipadic-NEologd [8] as an additional dictionary for accurately extracting brand-new named entities that frequently appeared in news articles.

Key sentence identification can be achieved simply by providing the trained BLSTM model with a news article. However in this research, in order to achieve more accurate prediction, we introduce another classifier as a post-processing step. More specifically, the output of the 2nd (backward) LSTM layer located just before the output layer is extracted as a feature vector, and an SVM with radial basis function (RBF) kernels is adopted to obtain the final estimated labels. Since each time step of the RNN model corresponds to a word, estimates of the SVM classifier are obtained for each word. On the other hand, the purpose of the proposed method is to select key sentences, and thus we take the average of the estimates for all the words in a sentence.

4 PROBLEMS OF NAÏVE IMPLEMENTATION

In this section, we first show our preliminary experimental results and demonstrate that a naive use of collected data does not produce promising results.

4.1 DATA COLLECTION

To construct the ground-truth datasets, we crawled news articles published on web pages by a major Japanese news publisher and Twitter posts that contained a link to collected news articles, resulting in 6K articles and 800K tweets. For every article, we selected the most viral tweet that contained a link to the article. We here employ the sum of the retweet and favorite counts as a measure of virality. After removing the tweets with low (less than 4) virality scores and the corresponding articles, we finally obtained 4022 pairs of tweets and articles.

Next, for every sentence in every article, we performed a manual annotation indicating whether or not the sentence was a key sentence in terms of generating attractive news headlines. We gave positive labels to the sentences that included the same semantic meaning in part of the paired tweet, and negative labels for all the others. We found that several news articles used unusual characters for displaying the ends of sentences, and comments and speeches were often quoted in the middles of sentences. Therefore, we used all the possible characters as delimiters, including unusual characters and quotations, to make it possible to split articles into sentences automatically.

4.2 PRELIMINARY EXPERIMENTS

We undertook experiments to evaluate the performance of the proposed model with the collected data presented in Section 3.1. We used LIBSVM [27] and Chainer [28] to implement SVM classifiers and neural network models, respectively. The dataset was divided into 10 parts, and all the hyper-parameters of the neural network model and the SVM classifier were determined with 9-fold cross validation, meaning that 8 out of 10 parts were used to train a specific combination of hyper-parameters and 1 part was used to select hyperparameters. We used a receiver operating characteristic (ROC) curve to evaluate performance. We compared the proposed method with two simple baseline methods where one randomly extracts sentences from a given news article as a key sentence and the other extracts sentences from the beginning of the article.

Figure 3 summarizes our experimental results. We observed that our method outperformed a random guess (dashed line indicating true positive rates = false positive rates) with the area under the ROC curve (AUC) = 0.740, however, it could not beat the other baseline method which employed order-based (AUC=0.903).

[8] https://github.com/neologd/mecab-ipadic-neologd

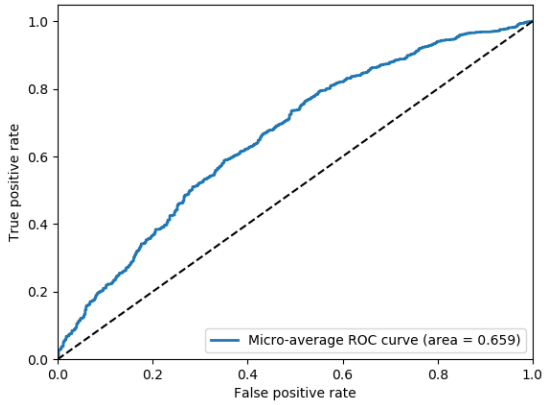

Figure 3: Preliminary experimental results

4.3 IN-DEPTH ANALYSIS OF RESULTS

We investigated the reasons why the proposed method did not work well even with a state-of-the-art neural network model. More specifically, we manually compared the sentences selected by our proposed method and the ground truths, and found three possible factors influencing the performance, namely large variations in the microblog messages used as supervised information and difficulties in determining sentence boundaries.

First, we found that microblog posts used as supervised information for key sentence identification have large variations in terms of writing style. As described in Section 3.1, we selected the most viral microblog post containing the URL of the target news article as the ground-truth headline. This means that we did not control who posted the headline candidate. Every social media user has his/her own style when writing microblog posts. A naive use of the most viral microblog post leads to large undesired variations in writing style, which make it difficult to train machine learning-based models. Some posts do not properly summarize the content of news articles and instead express their own impressions or feelings, or present opinions that are totally unrelated to the content of the news articles. Figure 4 shows an example of the social post (upper) and the corresponding news article (lower), where underlined phrases contain parts of the news article.

We also found that our procedure for separating a news article into sentences has a serious problem. As described in Section 3.1, we used all the possible characters as delimiters, and these include unusual characters and quotations. However, this rule for automatic separation yielded over-splitting, which means that a single sentence was often separated into multiple parts. Figures 5 shows examples.

In summary, we found that the following two factors influence the performance of key sentence identification. (1) A naive use of the most viral microblog posts leads to larg variations in terms of writing style. (2) The automatic separation of articles into sentences often fails.

そういえば、皆さんこの記事って見られましたか？ぐんまちゃんのルーツを知る、貴重なインタビューなので是非ご覧下さい mainichi.jp/artic・・・

Have you checked this article? Please take a look at this, a great interview that can learn the root of Gunma-Chan. mainichi.jp/artic・・・

ゆるキャラグランプリ（ＧＰ）２０１４で初優勝を果たした「ぐんまちゃん」は、群馬県職員の中嶋史子さん（４７）＝前橋市＝によって今から２０年前にデザインされた。「この子と…

"Gunma-Chan" who won his first victory at Yuru Character Grand Prix (GP) 2014 was designed by Fumiko Nakajima (47), a staff member of Gunma prefecture twenty years ago from now...

Figure 4: Example of microblog messages posted by general users and corresponding news article

AKB 島崎遥香：こじはるが「おしゃれ認定」も塩対応ファッションに無関心？ - 毎日新聞 #ぱるる #島崎遥香 mainichi.jp/mantan/news/2015・・・

AKB Haruka Shimazaki : Kojiharu had unfriendly attitude for Fashionable Accreditation Indifferent to fashion – Mainichi shimbun #Paruru #HarukaShimazaki mainichi.jp/mainichi/2015/...

人気アイドルグループ「AKB48」が 29 日、国立代々日本最大級の・・・「ガールズアワード 2015 SPRING／SUMMER」に登場し、ライブパフォーマンスを行った。・・・を聞かれた小嶋陽菜さんは「ぱるる(島崎遥香さん)」と回答。一方、島崎さんは・・・“塩対応”だった。・・・

On 29th, a popular idol group “AKB48” appeared in the Japan's largest... “Girls Award 2015 SPRING / SUMMER” and performed live performance..... Ms. Haruna Kojima who asked ... answered “Paruru (Ms. Haruka Shimazaki)”. On the other hand, Ms. Shimazaki had unfriendly attitude.

Figure 5: Example of news articles with many special characters

5 DATA CLEANSING

We refined the process of data collection and dataset organization based on the investigations presented in the previous section. We crawled news articles published on the top web page of Asahi Shimbun, which is recognized as a flagship daily newspaper provider in Japan, and collected Twitter messages posted by the official account of the news provider as the ground-truth headlines, instead of social messages posted by general users. Those messages had been edited by professional

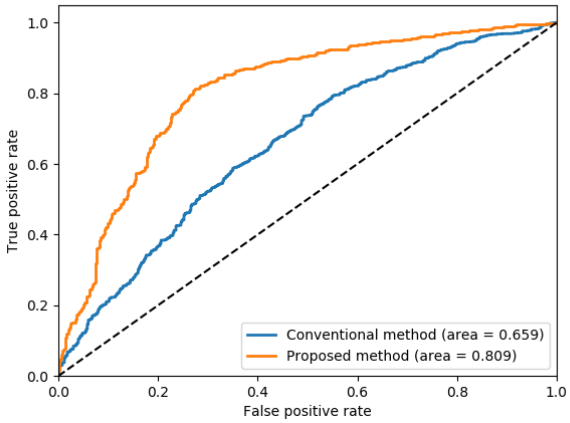

Figure 6: Experimental results

social editors taking virality in social media into consideration, and thus they often differed from the newspaper headlines. We also expect the news articles published on the top page to have a few columns and advertisements that make it difficult to automatically generate news headlines. We recognized through the above investigations that separating articles into sentences automatically and precisely is too difficult, so we separated every article into sentences by hand, and performed a manual annotation to create ground-truth labels for key sentence identification in the same way as that presented in Section 3.1. We used the same models (see Section 3.2) for identifying key sentences. We finally obtained 911 pairs of news articles and corresponding social media posts.

6 EXPERIMENTS

We have undertaken an experiment to evaluate the performance of our proposed method. All the experimental conditions are the same as the one described in Section 4.1. Figure 6 summarizes the results, which indicates that our proposed method (orange) outperformed random guesses (black dashed) and the previous method (blue) with AUC=0.894, and is comparable to the method the utilizes order-based selection.

7 CONCLUTION

In this paper, we dealt with the challenging problem of automatically generating attractive news headlines that have the potential to go viral on social media, and proposed a novel method for extracting key sentences from a given news article as material for generating attractive news headlines. Our main claim in this paper is that we have to carefully collect, cleanse and organize datasets to achieve promising results even with state-of-the-art techniques. We confirmed the effectiveness of the proposed method. In this paper, we focused solely on the problem of identifying key sentences, however, we understand that several post-processing steps will be required to generate news headlines, such as sentence compression [25,29]. There is also the potential to improve the model for key sentence identification. Promising approaches include deep bidirectional LSTM [30] and connectionist temporal classification (CTC) [31] that have already proven effective in the area of speech recognition.

REFERENCES

[1] S. Craig. 2015. Lies, Damn Lies and Viral Content. Tow Center for Digital Journalism.
[2] S. Kourogi, A. Kimura, H. Fujishiro and H. Nishikawa. 2015. Identifying Attractive News Headlines for Social Media. In Proc. CIKM2015.
[3] K. Nagayama, A. Kimura and H. Fujishiro. 2016. Make it go viral – Generating attractive headlines for distributing news on social media. In Proc. Computation + Journalism Symposium2016.
[4] E. Filatova and V. Hatzivassiloglou. 2004. A Formal Model for Information Selection In Multi-Sentence Text Extraction. In Proc. COLING2004.
[5] H. Takamura and M. Okumura. 2009. Text Summarization Model based on Maximum Coverage Problem and its Variant. In Proc. EACL2009.
[6] H. Takamura and M. Okumura. 2009. Text Summarization Model based on the Budgeted Median Problem. In Proc. CIKM2009.
[7] E. Alfonseca, D. Pighin and G. Garrido. 2013. HEADY: News headline abstraction through event pattern clustering. In ACL2013.
[8] S. Xu, S. Yang and F. Lau. 2010. Keyword Extraction and Headline Generation Using Novel Word Features. In AAAI2010.
[9] D. Zajic, B. Dorr and R. Schwartz. 2002. Automatic headline generation for newspaper stories. In ACL Workshop on Text Summarization2002.
[10] B. Dorr, D. Zajic and R. Schwartz. 2003. Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In HLT-NAACL2003 on Text Summarization Workshop2003.
[11] R. Wang, J. Dunnion and J. Carthy. 2005. Machine Learing Approach To Augmenting News Headline Generation. In IJCNLP2005.
[12] S. Wang, E. Han and A. Rush. 2016. Headliner: An integrated headline suggestion system. In Computation + Journalism Symposium2016.
[13] S. Nishiguchi, M. Imono, S. Tsutiya and H. Watabe. 2013. Organization in a table format from news articles according to user's request. In IPSJ SIG Technical Report no.9, pp. 1-6.
[14] S. Takase, R. Sasano , H. Takamura and M. Okumura. 2016. Youyakucyou, buncyou, bunsuseigentsuki news youyaku. In The Association for Natural Language Processing. pp.342-345.
[15] M. Guerini, J. Staiano and D. Albanese. 2013. Exploring image virality in Google Plus. In SocialCom2013.
[16] A. Deza and D. Parikh. 2015. Understanding image virality. In CVPR2015.
[17] J. Park, M. Cha, H. Kim and J. Jeong. 2012. Managing Bad News in social media: A Case Study on Domino's Pizza Crisis. In ICWSM2012.
[18] M. Jenders, G. Kasneci and F. Naumann. 2013. Analyzing and Predicting Viral Tweets. In WWW2013.
[19] P. Resnick, S. Carton, S. Park, Y. Shen and N. Zeffer. 2014. RumorLens: A System for Analyzing the Impact of Rumors and Corrections in Social Media. In Computation + Journalism Symposium2014.
[20] M. Gabielkov, A. Ramachandran, A. Chaintreau and A. Legout. 2016. Social Clicks: What and Who Gets Read on Twitter? In ACM SIGMENTRICS2016.
[21] Y. Yasuda. 2013. Information dissemination in social media: hubs and demagogues. In Kansai University Syakaigakubu yoko vol.45, no.1, pp.33-46.
[22] T. Kudo, K. Yamamoto and Y. Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.230-237,
[23] https://github.com/neologd/mecab-ipadic-neologd/
[24] A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. In Neural Networks vol.18, no.5-6, pp.602-610.
[25] A. Rush, S. Chopra and J. Weston. 2015. Proc. EMNLP. A Neural Attention Model for Abstractive Sentence Summarization. In Proc. EMNLP2015.
[26] S. Takase, J. Suzuki, N. Okazaki, T. Hirao and M. Nagata. 2016. Neural Headline Generation on Abstract Meaning Representation. In Proc. EMNLP2016.
[27] C. Chang and C. Lin. 2013. LIBSVM: A Library for Support Vector Machines. In ACM Trans. IST, 2013.
[28] S. Tokui, K. Oono, S. Hido and J. Clayton. 2015. Chainer: a Next-Generation Open Source Framework for Deep Learning. In Proc. NIPS Workshop on Machine Learning Systems, 2015.
[29] K. Filippova, E. Alfonseca, C. Colmenares, L. Kaiser and O. Vinyals. 2015. Sentence Compression by Deletion with LSTMs. In Proc. EMNLP2015.
[30] A. Graves, A. Mohamed and G. Hinton. 2013. SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS. In Proc. ICASSP2013.
[31] A. Graves, S. Fernandez, F. Gomez and J. Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proc. ICML2006.