EL-BERT at SemEval-2020 Task 10: A Multi-Embedding Ensemble Based Approach for Emphasis Selection in Visual Media

Chandresh S. Kanani
Department of CSE
Indian Institute of Technology
Patna, India
cskanani@gmail.com

Sriparna Saha
Department of CSE
Indian Institute of Technology
Patna, India
sriparna@iitp.ac.in

Pushpak Bhattacharyya
Department of CSE
Indian Institute of Technology
Patna, India
pb@iitp.ac.in

Abstract

In visual media, text emphasis is the strengthening of words in a text to convey the intent of the author. Text emphasis in visual media is generally done by using different colors, backgrounds, or font for the text; it helps in conveying the actual meaning of the message to the readers. Emphasis selection is the task of choosing candidate words for emphasis, it helps in automatically designing posters and other media contents with written text. If we consider only the text and do not know the intent, then there can be multiple valid emphasis selections. We propose the use of ensembles for emphasis selection to improve over single emphasis selection models. We show that the use of multi-embedding helps in enhancing the results for base models. To show the efficacy of proposed approach we have also done a comparison of our results with state-of-the-art models.

1 Introduction

The SemEval-2020 Task 10 (Shirani et al., 2020) challenge focuses on emphasis selection in visual media. Emphasis is the process of giving importance to some parts of communication to convey the message in a better way. It is used to draw the attention of readers to a specific section of the information. It is also used for removing the ambiguity in the message. In vocal communication, the emphasis is generally conveyed by giving stress on the specific word. In visual communications like flyers, posters, advertisements, the emphasis is conveyed by using different fonts, colors, or backgrounds for the text.

Text designing systems like Adobe Spark [1] can automatically provide template-based layouts for text. However, these algorithms generally rely on the visual features of the text like word length and suggest designs based on those features. Sometimes, this type of method does not emphasize on the proper words and might not help in conveying important information or may even convey wrong information. In Fig 1a, we show the automatic design provided by Adobe Spark. Even though Figure 1a is aesthetically pleasing, it is not giving emphasis to essential words and might fail in conveying the message. Instead, Figure 1b uses a different layout and emphasizes on vital words.

Given only the text and not the intent of the message, there can be multiple valid emphases, and different authors can prefer to emphasize different words. Therefore, we cannot use a single label to say whether a word should be emphasized or not. To tackle this, we use learning label distribution (LDL) (Gao et al., 2017). LDL is used for assigning a real number to each word showing the probability of a word being emphasized.

Our main contributions can be summarized as:

- We propose the use of different embeddings for the task of emphasis selection and also show that different combinations of embedding can improve the emphasis selection.
- We show that encoded sentences and words using different embedding (multi-embedding) have introduced new information to models and improved the performance.

[1] https://spark.adobe.com

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1645–1651
Barcelona, Spain (Online), December 12, 2020.

(a)

(b)

Figure 1: Two different designs with different emphasized words. Figure a is automatically generated using Adobe Spark. Figure b is generated using ground truth annotations.

- We propose the use of ensemble models for emphasis selection in visual media.
- We show a comparison of our results with baselines and state-of-the-art models and also give a qualitative comparison of different methods.

2 Related Work

In text data, the majority of works focus on important keyword identification from long texts. There are mainly two methods for keyword extraction: supervised and unsupervised. Supervised methods generally treat the keyword identification as a classification problem and classify words as either keyword or not (Frank et al., 1999; Tang et al., 2004; Medelyan and Witten, 2006). Unsupervised methods usually utilize TF-IDF (Hasan and Ng, 2010) scores or clustering methods (Liu et al., 2009) for keyword identification. Recently, (Zhang et al., 2016) also proposed a model using RNNs for keyword identification.

Different methods are proposed for the emphasis selection in audio. Most works use acoustic features like loudness, pitch to detect emphasized words in audio data (Kochanski et al., 2005; Wang and Narayanan, 2007). Recently, some works are proposed to predict word emphasis in text to improve text-to-speech (TTS) systems (Nakajima et al., 2014; Mass et al., 2018). (Sun, 2002) proposed ensemble-based models for emphasis selection in audio.

(Shirani et al., 2019) proposed the use of label distribution learning (LDL) for emphasis selection in short texts in visual media. Here, we show that the use of ensemble models trained with different embeddings performs better than base models.

3 Approach

3.1 Problem Definition

Given a sentence S with tokens $C = \{w_1, w_2, ..., w_n\}$, where $1 < |S| < n$, emphasis selection is the task to find candidate words in C to emphasize on for conveying the meaning of message in an effective manner.

3.2 Label distribution learning (LDL)

We use "IO" scheme for labels, where "I" represents emphasis and "O" represents non-emphasis. Then label distribution learning is the task of learning probability d_y^w for each word $w \in C$, denoting the degree with which word w belongs to label y. $d_y^w \in [0, 1]$ and $\sum_y d_y^w = 1$.

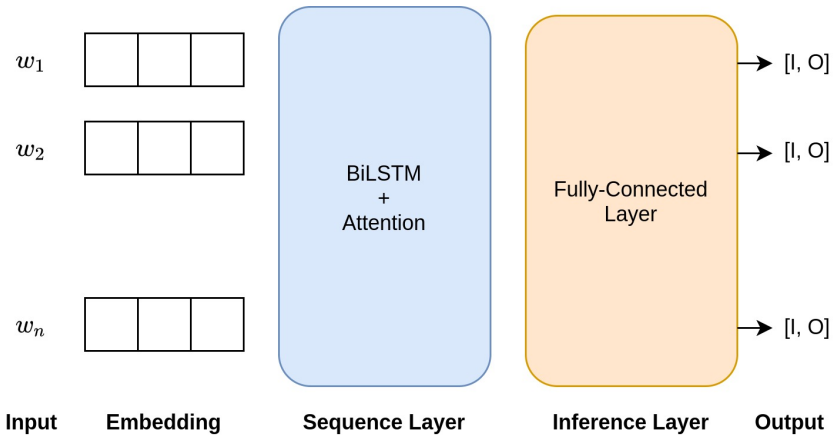

Figure 2: Architecture of DL-BiLSTM model.

3.3 Dataset

The dataset consists of two sub-datasets: 1. Spark dataset, 2. Quotes dataset

Spark dataset: This dataset is collected from Adobe Spark and is a collection of short texts from flyers, posters, or advertisements and includes 1,200 instances.

Quotes dataset: This dataset is a collection of quotes from well-known authors collected from Wisdom Quotes, which contains 2,718 instances.

For further analysis, the dataset is split up into train, test, and development sets with 70%, 20%, and 10% samples, respectively.

4 Model

We use the DL-BiLSTM model proposed by (Shirani et al., 2019) as our base model, to the best of our knowledge this model provides state-of-the-art results for emphasis selection in visual media. Each word of the input sequence is represented with word embedding. Part-of-speech (POS) and sentence embedding are added as additional information. Two bidirectional LSTM (Hochreiter and Schmidhuber, 1997) layers are used to capture the sequence information. The last layer of the model is fully-connected and assigns probability using the hidden state of LSTM. Figure 2 shows the overall architecture of the DL-BiLSTM model.

We have used different sets of embedding to train the models. WordBERT and SentBERT represent word, and sentence embedding generated using pre-trained BERT (Devlin et al., 2019), respectively. POSEmbd represents one-hot-encoded POS tag. In Model-4, Model-5 and Model-6, we use ELMo (Peters et al., 2018) as word embedding.

Below is the list of models and used embeddings:

Model-1: DL-BiLSTM + WordBERT
Model-2: DL-BiLSTM + WordBERT + POSEmbd
Model-3: DL-BiLSTM + WordBERT + POSEmbd + SentBERT
Model-4: DL-BiLSTM + ELMo
Model-5: DL-BiLSTM + ELMo + POSEmbd
Model-6: DL-BiLSTM + ELMo + POSEmbd + SentBERT

4.1 Baseline Models

Here, we discuss the baseline models and their implementation.

SL-BiLSTM: This model has same architecture as DL-BiLSTM, but the distribution is mapped to binary labels. Also, for training the SL-BiLSTM model (Shirani et al., 2019), we use negative log-likelihood loss in place of KL-Divergence loss (Kullback and Leibler, 1951).

CRF (Conditional Random Fields): Similar to (Shirani et al., 2019), this model is trained with handcrafted features like word identity, word suffix, word shape, and POS tag for the current and nearby

words.

5 Experimental Settings

We have used BERT-as-Service framework[2] and 1024 dimension pre-trained cased BERT embedding for encoding the words and sentences. We have used universal POS tagger from NLTK[3] to get the POS tags for sentences, and then have used a bag of word embedding to encode POS tag information. We have also used 2048 dimension ELMo embedding for encoding word information.

The size of the BiLSTM layer in the model depends on the set of embeddings used, Table 1 lists the sizes of BiLSTM layers for different sets of embedding. All the proposed models are trained for 10 epochs with Adam (Kingma and Ba, 2015) as the optimizer and learning rate of 0.001. The baseline models are trained for 160 epochs. To prevent models from over-fitting, we have also used two dropout layers with a dropout-rate of 0.5 in the sequence and inference layers.

For training the DL-BiLSTM model, we have used KL-Divergence loss, and for training SL-BiLSTM model, we have used negative log-likelihood loss.

We have also used different ensembles of the above models, the final score of an ensemble is the average of scores provided by all the models.

Embeddings	BiLSTM Layer Size
WordBERT	1024
WordBERT + POSEmbd	1036
WordBERT + POSEmbd + SentBERT	2060
ELMo	2048
ELMo + POSEmbd	2060
ELMo + POSEmbd + SentBERT	3084

Table 1: Sizes of BiLSTM layers for different sets of embeddings used for models.

6 Results

As proposed by (Shirani et al., 2019), we have used Match_m score for evaluation of our models. Table 2 lists the results of the base, baseline, state-of-the-art and top five ensemble models.

6.1 Analysis

Figure 3 shows a qualitative comparison of ensembles and base models. It can be seen from Figure 3a and 3b that both the single model and the ensemble model produce similar results for most cases. In some cases (examples 1 and 4), results of ensemble model is more similar to target scores shown in Figure 3c.

From Table 2, we have noticed that the addition of POSEmbd improves scores for models with ELMo word embedding (Model-4, Model-5, Model-6) but does not improve results for models with BERT word embedding (Model-1, Model-2, Model-3). This shows that the addition of POSEmbd does not add any additional information to BERT embedding; we can infer that BERT embedding is inherently able to identify POS information.

We have also noticed that the use of BERT sentence embedding (SentBERT) improves results for models with ELMo word embedding (Model-4, Model-5, Model-6) but does not improve results for models with BERT word embedding (Model-1, Model-2, Model-3). This shows the ability of BiLSTM model to encode the sequence property correctly. The improvement in the case of ELMo based models shows that the addition of information from different embedding space can add additional information and helps in improving the performance of models.

We have also noticed that ensemble models perform better than single base models and, in almost all cases, provide better results than base models.

[2]https://github.com/hanxiao/bert-as-service

[3]https://www.nltk.org/book/ch05.html

Models	Dev Set Match$_m$				Test Set Match$_m$			
	m=1	m=2	m=3	m=4	m=1	m=2	m=3	m=4
Base Models								
Model-1	56.1	72.5	78.2	83.0	56.9	70.3	79.1	83.3
Model-2	57.9	71.3	76.4	82.5	54.4	69.9	78.2	83.1
Model-3	57.9	70.1	77.7	81.6	57.5	68.8	77.9	82.8
Model-4	55.6	71.5	79.2	82.2	54.5	71.7	79.1	84.0
Model-5	58.8	74.6	78.6	82.7	**59.1**	73.2	79.7	84.4
Model-6	55.2	72.5	79.2	83.3	55.2	73.1	80.2	84.2
Top 5 Ensembles								
Model-1 + Model-6	59.7	74.6	80.3	84.8	58.1	73.2	**81.0**	85.1
Model-1 + Model-4 + Model-5	59.2	**74.9**	79.8	84.4	58.3	72.7	**81.0**	**85.5**
Model-1 + Model-4 + Model-6	57.9	72.5	**80.8**	84.5	57.1	73.5	81.2	**85.5**
Model-1 + Model-2 + Model-4 + Model-6	**61.5**	73.0	78.9	84.8	57.7	73.0	80.6	85.0
Model-1 + Model-2 + Model-4 + Model-5 + Model-6	59.3	73.0	79.6	**85.1**	58.8	**73.6**	80.8	85.3
SOTA Models								
DL-BiLSTM + ELMo + Attention	-	-	-	-	59.6	72.2	78.8	84.6
DL-BiLSTM + GloVe	-	-	-	-	54.6	69.2	76.5	81.9
DL-BiLSTM + GloVe + Attention	-	-	-	-	57.5	69.7	76.7	80.7
Baseline Models								
SL-BiLSTM + GloVe	-	-	-	-	51.7	66.7	75.0	81.1
SL-BiLSTM + GloVe + Attention	-	-	-	-	52.9	66.5	73.6	80.0
SL-BiLSTM + ELMo	-	-	-	-	54.2	69.0	77.9	83.0
SL-BiLSTM + ELMo + Attention	-	-	-	-	54.2	70.0	78.5	82.8
CRF	-	-	-	-	45.4	66.0	72.8	80.2

Table 2: Table shows Match$_m$ scores obtained for Dev and Test set for various base models and their combinations. We also compare our results with other baseline models and state-of-the-art models.

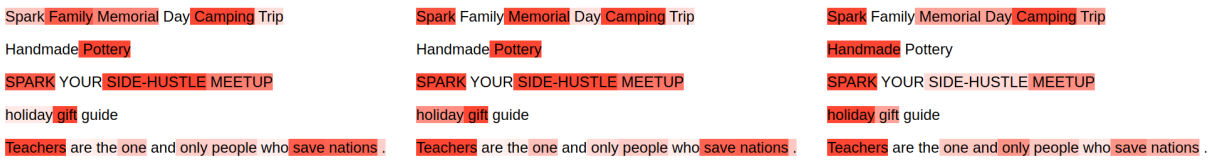

Figure 3: Figure a, b, illustrate emphasis given to input sentences by best single model (Model-5) and best ensemble model (Model-1 + Model-3 + Model-4 + Model-5 + Model-6), respectively. Figure c shows target emphasis weights.

7 Conclusion & Future Work

We show that the use of multiple embeddings for encoding different information (i.e., word and sentences) can help in improving model performance. We also show that ensemble models perform better for emphasis selection then single base models. In the future, we will work on ensembles with different model architectures and methods for emphasis selection. We will also work on generating task-specific word embedding for emphasis selection.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 668–673. Morgan Kaufmann.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Trans. Image Process.*, 26(6):2825–2838.

Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 365–373. Chinese Information Processing Society of China.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. 2005. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2):1038–1054.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 257–266. ACL.

Yosi Mass, Slava Shechtman, Moran Mordechay, Ron Hoory, Oren Sar Shalom, Guy Lev, and David Konopnicki. 2018. Word emphasis prediction for expressive text to speech. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2868–2872. ISCA.

Olena Medelyan and Ian H. Witten. 2006. Thesaurus based automatic keyphrase indexing. In Gary Marchionini, Michael L. Nelson, and Catherine C. Marshall, editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006, Chapel Hill, NC, USA, June 11-15, 2006, Proceedings*, pages 296–297. ACM.

Hideharu Nakajima, Hideyuki Mizuno, and Sumitaka Sakauchi. 2014. Emphasized accent phrase prediction from text for advertisement text-to-speech synthesis. In Wirote Aroonmanakun, Prachya Boonkwan, and Thepchai Supnithi, editors, *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, PACLIC 28, Cape Panwa Hotel, Phuket, Thailand, December 12-14, 2014*, pages 170–177. The PACLIC 28 Organizing Committee and PACLIC Steering Committee / ACL / Department of Linguistics, Faculty of Arts, Chulalongkorn University.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Amirreza Shirani, Franck Dernoncourt, Paul Asente, Nedim Lipka, Seokhwan Kim, Jose Echevarria, and Thamar Solorio. 2019. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1167–1172, Florence, Italy, July. Association for Computational Linguistics.

Amirreza Shirani, Franck Dernoncourt, Nedim Lipka, Paul Asente, Jose Echevarria, and Thamar Solorio. 2020. Semeval-2020 task 10: Emphasis selection for written text in visual media. In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Xuejing Sun. 2002. Pitch accent prediction using ensemble machine learning. In John H. L. Hansen and Bryan L. Pellom, editors, *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*. ISCA.

Jie Tang, Juan-Zi Li, Kehong Wang, and Yue-Ru Cai. 2004. Loss minimization based keyword distillation. In Jeffrey Xu Yu, Xuemin Lin, Hongjun Lu, and Yanchun Zhang, editors, *Advanced Web Technologies and Applications, 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China, April 14-17, 2004, Proceedings*, volume 3007 of *Lecture Notes in Computer Science*, pages 572–577. Springer.

Dagen Wang and Shrikanth S. Narayanan. 2007. An acoustic measure for word prominence in spontaneous speech. *IEEE Trans. Audio, Speech & Language Processing*, 15(2):690–701.

Qi Zhang, Yang Wang, Yeyun Gong, and Xuanjing Huang. 2016. Keyphrase extraction using deep recurrent neural networks on twitter. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 836–845. The Association for Computational Linguistics.