Employing Bert Embeddings for Customer Segmentation and Translation Matching

Tim vor der Brück

Lucerne School of Computer Science and Information Technology
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
E-mail: tim.vorderbrueck@hslu.ch

Abstract—**In this work, we investigate the performance of Bert (Bidirectional Encoder Representations from Transformers) embeddings for two Natural Language Processing (NLP) scenarios based on semantic similarity and conduct a comparison with ordinary Word2Vec embeddings. The Bert embeddings are pretrained on a multi-lingual dataset from Google consisting of several Wikipedias. The semantic similarity between two input texts is estimated in the usual way of applying the cosine measure on the two embeddings centroids. In case of Bert, these centroids are determined by two different approaches. In the first approach, we just average the embeddings of all the word vectors of the associated sentence. In the second approach, we only average the embeddings of a special sentence start token that contains the whole sentence representation. Surprisingly, the performance of ordinary Word2Vec embeddings turned out to be considerably superior in both scenarios and both calculation methods.**

***Keywords*–Bert embeddings; Targeted Marketing; Translation Matching.**

I. Introduction

Word2Vec Word Embeddings [1] enjoy high popularity due to their ease of use and good performance for estimating semantic similarity between words, sentences, or entire texts. However, they do lack one important property: they cannot directly convey phenomena like homography or polysemy. Thus, the same word used in a completely different meaning (like *space* as universe and *space* as location) would still be assigned the same word vector. Thus, Bert and ELMo (Embeddings from Language Models) embeddings [2][3] were introduced to overcome this issue. These embeddings are completely context dependent and can therefore no longer be expressed by global lookup tables as it is the case for Word2Vec Embeddings. Instead, they are generated by a deep neural network applied to a given text segment. In this work, we compare the performance of Bert embeddings with ordinary Word2Vec embeddings on two different NLP application scenarios.

II. Scenario 1 - Customer Segmentation

Our business partner operates a commercial youth platform for the Swiss market, where registered members get access to third-party offers such as discounts and special events like concerts or castings [4]. Actually, several hundred online contests per year are launched over this platform sponsored by other firms, an increasing number of them require the members to write short free-text snippets. Depending on these text snippets, the members should be automatically mapped to the best fitting marketing target group (called youth milieus) to allow for more customer-focused and precise marketing campaigns. The 6 employed youth milieus are:

- progressive postmodern youth: people primarily interested in culture and arts
- young performers: people striving for a high salary with a strong affinity for luxury goods
- freestyle action sportsmen
- hedonists: rather poorly educated people who enjoy partying and disco music
- conservative youth: traditional people with a strong concern for security
- special groups: comprises all those who cannot be assigned to one of the upper five milieus.

In total, our business partner conducted three online contests, where the participants should

1) elaborate on a perfect holiday at a destination of their choice in case of a contest sponsored by a travel agency (Contest 1),
2) fantasize what they could do with a pair of new sneakers (Contest 2) and
3) how they would use one of several possible prizes (Contest 3).

To accomplish this matching, all marketing target groups are described by a set of keywords that conveys their typical characteristics. We then generate word embedding centroids for both the snippet and the keyword list. Afterward, we select that marketing target group for a certain text snippet, for which the cosine measure between both embedding centroids is maximal (see Figure 1). These selections are then compared with a gold standard annotation conducted independently by three different marketers.

III. Scenario 2 - Translation Matching

In this scenario, we investigate two independent translations of the same novel (*The purloined letter by Edgar Allen Poe*) into German. In particular, we aim to match each sentence of the first translation to the associated sentence of the second translation. The matching procedure is analogous to scenario 1, which means that we generate embedding vectors for all sentences and determine the sentence pairs with maximal cosine similarity between the associated embedding centroids.

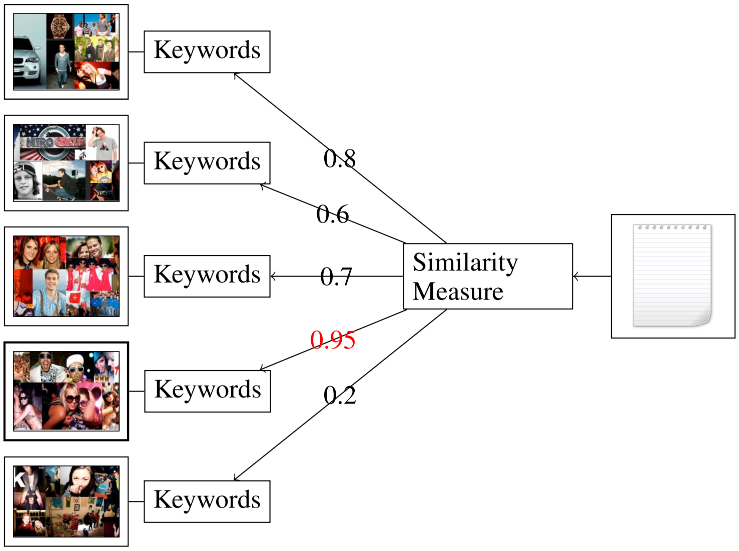

Figure 1. Procedure for mapping a text snippet to the best fitting target group.

TABLE I. Corpus sizes measured by number of words.

Corpus	# Words
German Wikipedia	651 880 623
Frankfurter Rundschau	34 325 073
News journal *20 Minutes*	8 629 955

TABLE II. Obtained accuracy of embedding-based similarity estimation on three online contests.

Method		Accuracy		
	Contest 1	Contest 2	Contest 3	Total
Word2Vec	0.347	0.328	0.227	0.330
Bert (AW)	0.046	0.223	0.061	0.118
Bert (ST)	0.109	0.149	0.136	0.07

TABLE III. Evaluation on translation matching.

Method	Accuracy
W2VC	0.726
Bert (ST)	0.423
Bert (AW)	0.279
Random	0.010

IV. Results

The Bert embeddings were trained on the multilingual data set comprising of several Wikipedias. The centroids of a text snippet were determined using two different approaches:

- average over All Words (AW)
- average only over the Start Tokens (ST) that represent the beginning of a sentence

For the first approach (AW), we used Gluon [5], an NLP library based on MXNet [6], while approach (ST) was based on a PyTorch implementation provided by *Hugging Face* [7].

Word2Vec was trained on the German Wikipedia, the German newspaper *Frankfurter Rundschau* and on the *20 minutes* journal (cf. [4]), which is freely available at various Swiss train stations. The sizes of the three corpora are given in Table I.

The obtained accuracy for the customer segmentation / translation matching is given in Table II / Table III.

V. Discussion

Bert embeddings turned out to be rather unusable for the first task of target group matching. A possible reason is that all text snippets are compared with keyword lists, for which the word order is rather arbitrary and depends on the personal preferences of the marketers. The obtained accuracy values of Bert Embeddings for the second scenario of translation matching were indeed higher, however still considerably lagging behind the use of ordinary Word2Vec word embeddings. Furthermore, calculating the centroids from the Bert embeddings of the Start Token (ST) seems the superior approach to just averaging the individual word embeddings (AVG). A further reason for the rather poor performance of Bert embeddings in both scenarios is the fact that the data set used for training is multi-lingual. We expect the results to be superior in case of a monolingual model, since such a model reduces the number of different tokens possibly occurring in a given word context and, therefore, also the noise in the data.

VI. Conclusion

We applied Bert embeddings to two different German NLP tasks, which are customer segmentation and translation matching. In both scenarios, we obtained a rather poor performance compared to ordinary Word2Vec embeddings. Possible future work comprises the use of monolingual training data for Bert as well as ELMo embeddings and other embedding methods.

Acknowledgment

Hereby, I want to thank all people that supported me with this work. Special thanks go to Marc Pouly who provided me with this dataset.

References

[1] T. Mikolov, I. Sutskever, C. Ilya, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, 2013, pp. 3111–3119.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL, 2019, pp. 4171–4186.

[3] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in Proceedings of NAACL, 2018, pp. 2227–2237.

[4] T. vor der Brück and M. Pouly, "Text similarity estimation based on word embeddings and matrix norms for targeted marketing," in Proceedings of NAACL, 2019, pp. 1827–1836.

[5] "GluonNLP," 2020, https://gluon-nlp.mxnet.io.

[6] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," 2015.

[7] "Huggingface," 2020, https://huggingface.co.