

**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук
Департамент программной инженерии

СОГЛАСОВАНО

Доцент факультета гуманитарных наук Школы
Лингвистики Высшей Школы Экономики

_____/Ф. Фишер/
« ____ » _____ 2019 г.

УТВЕРЖДАЮ

Академический руководитель образовательной
программы «Программная инженерия» профессор де-
партамент программной инженерии, канд. техн. наук

_____/В.В. Шилов/
« ____ » _____ 2019 г.

**Аннотация
к курсовой работе**

на тему «An Approach to the Generation and Evaluation of Scripts for Television Series»

по направлению подготовки бакалавров 09.03.04 «Программная инженерия»

Выполнил студент группы БПИ173
образовательной программы 09.03.04
«Программная инженерия»

_____/Д.М. Торилов/
« ____ » _____ 2019 г.

Москва 2019

Keywords

Dialogue system; Evaluation metrics; Deep Learning; Seq2seq; Attention; Language models.

Abstract

Today, the dialogue systems are in great demand, as they allow to formulate requests fuzzy and still get valid answers during human-like conversation. These systems can be used in chat-bots and voice assistants for different purposes, for example they can solve the problem of responding to customers' questions about products and services.

We investigate the application of dialogue systems to the TV Series scripts generation. This task will be solved using seq2seq models, which are widely used in machine translation and in dialogue systems too.

In recent past we've got a lot of good-working algorithms for word embedding generation (word2vec [GL14], GloVe [PSM14], FastText [BGJM16]). Such embeddings carry partial semantics information about the encoded words. There is also an another approach to the embeddings representation where tokens are not whole words, but their parts. These approaches are Byte Pair Encoding (BPE), WordPiece, etc. These embeddings can be used as inputs for dialogue models.

In recent years, one main breakthrough has been made in the field of seq2seq algorithms — the attention mechanism invention and transformer especially. Using attention mechanism made neural machine translation (NMT) more effective than using statistical machine translation (SMT) [VSP⁺17]. Bare LSTM and GRU could not beat SMT in quality before this breakthrough.

An important aspect of dialogue response generation is how to evaluate the quality of the generated answers. Recent works in dialogue systems use adopted metrics from machine translation. There is shown [LLS⁺16] that these metrics correlate very weakly with human judgements. These metrics do encourage the appearance of words of the source sequence in the response, while normal dialogue does not require this. The human dialogue can be very allegorical and full of allusions. For example, the following dialogue is a valid human-like dialogue, but it will have low metrics score:

– *What about going to the cinema with me tonight?*
– *I've got lots and lots of work.*

We have the dataset of «Friends» TV Series transcripts, seasons 1–10. That is approximately 900k

words. The transcripts are better data than just subtitles because they contain scene directions.

The aim of this work is to build a dialogue system for TV series generation which will get high evaluation using old metrics like BLEU [PRWZ02] and METEOR [BL05] and then to find the new metric which will satisfy some popular requirements [LLS⁺16] and have higher correlation with human judgements than old metrics do.

References

- [BGJM16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2016.
- [BL05] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [GL14] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, 2014.
- [LLS⁺16] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, 2016.
- [PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vec-

- tors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Исполнитель студент группы БПИ173
образовательной программы 09.03.04
«Программная инженерия»

_____ /Д.М. Торилов/
«___» _____ 2019 г.

СОГЛАСОВАНО

Доцент факультета гуманитарных наук
Школы Лингвистики Высшей Школы
Экономики

_____ /Ф. Фишер/
«___» _____ 2019 г.