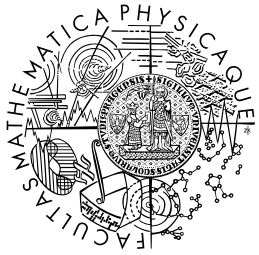

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

František Trebuňa

Generating text from structured data

Institute of Formal and Applied Linguistics (ÚFAL)

Supervisor of the bachelor thesis: Mgr. Rudolf Rosa, Ph.D.

Study programme: Computer Science (B1801)

Study branch: General Computer Science Bc. R9 (NIOI9B)

Prague 2021

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Dedication.

Title: Generating text from structured data

Author: František Trebuňa

Institute: Institute of Formal and Applied Linguistics (ÚFAL)

Supervisor: Mgr. Rudolf Rosa, Ph.D., Institute of Formal and Applied Linguistics (ÚFAL)

Abstract: Abstract.

Keywords: text generation structured data natural language processing neural networks

Contents

Introduction

1. Problem statement

In september 2020 I read a blog by [Karpathy, 2015]. He created a neural network consisting of only one LSTM cell and trained it to predict a character based on all the previous ones. The network was trained on a corpus of all the plays by Shakespeare. During inference the last predicted character was fed as the input to the network and this way it could create a really good looking Shakespeare-like text. Then I began to explore the possibilities of generating a text conditioned on some input parameters. How to construct a network that could be told to generate a sad, happy, or sarcastic sounding text?

1.1 Data to text generation

Known datasets (WIKIBIO, WeatherGov, RoboCup) -¿ short description, the generated summaries are one-two sentences long. Rotowire -¿ really long summaries although only a fraction of the number of unique tokens from the WIKIBIO dataset. Only short description of the dataset, the statistics and observations are in the second chapter.

1.2 Fantasy sports

What it is, where it originates, the relation to basic optimisation problems, why NLGenerated summaries could be added value.

1.3 My goal

Fluent text which captures the important statistics from the table

2. Data

One needs a lot of data if he wants to train his neural network. E.g. [Sennrich et al., 2016] trains the neural machine translation system on 4.2 million English-German sequence pairs. The Data-to-Text generation task má vyššie nároky na kvalitu datasetu. Potrebujeme, aby boli vstupné dáta štandardizované a aby výstupné texty zodpovedali vstupným dátam. Existuje viacero datasetov, ktoré spĺňajú túto podmienku. V tejto kapitole predstavím datasety WikiBIO a Rotowire.

2.1 General description

Obidva tieto datasety používajú notáciu, ktorá bola predstavená v článku od [Liang et al., 2009], preto ju najprv aj tu zadefinujeme.

Ako vstup používame postupnosť záznamov (recordov) $\mathbf{s} = \{r_i\}_{i=1}^{J}$. Každý record r má svoj typ $r.t \in \mathcal{T}$. Množina typov $\mathcal{T}$ je dopredu definovaná. Ďalej má typ, množinu hodnôt $r.v = \{r.v_1, ..., r.f_m\}$. Napríklad v datasete WeatherGOV: $r.t == windSpeed$, $r.v = \{time, min, mean, max, mode\}$. Na základe týchto záznamov následne predpovedáme výstupný text $\mathbf{w} = \{w_i\}_{i=1}^{|\mathbf{w}|}$

2.2 WikiBIO dataset

Štruktúrované dáta vo WikiBIO datasete sú vo forme tabuliek. Encoder-Decoder architektúra však vyžaduje, aby do nej boli dáta vkladané sekvenčne. [potrebujem tu pridať príklad tabuľky, pomocou obrázka]. Preto po vzore [Lebret et al., 2016] je tabuľka sploštená do série recordov.Typy sú anotácie riadkov, napríklad meno, dátum narodenia. Príslušných hodnôt však môže byť variabilne veľa. Preto musia vytvorené recordy reflektovať aj poradie hodnôt, čo dosahujeme tým, že pridáme pozičnú informáciu. Recordy sú teda podávané vo formáte $name_1 = Frantisek, name_2 = Trebuna, birthplace_1 = Kosice, birthplace_2 = Slovensko \ldots$ – pridať informáciu o odlišných dĺžkach tabuliek – lepší príklad, než vlastné meno –

2.2.1 Statistics of WikiBIO dataset

o tomto píšem zajtra

2.2.2 Preprocessing of WikiBIO dataset

o tomto píšem zajtra

2.3 Rotowire dataset

Štruktúrované dáta v RotoWire datasete sú taktiež vo forme tabuľky. Väčšina hodnôt je vo forme čísla, ktoré buď predstavuje informáciu o absolútnom počte nejakej hodnoty (e.g. počet bodov, ktoré hráč strelil) alebo o relatívnom (e.g.

percento úspešných pokusov z poľa). Zvyšné hodnoty sú mená miest, tímov a hráčov. Neskôr uvediem, ako som za pomoci preprocessingu zariadil, aby nebola potrebná pozičná informácia ani v týchto prípadoch. Na rozdiel od WikiBIO datasetu, kde všetky hodnoty pojednávali o jednej entite, v RotoWire potrebujeme ešte špeciálnu informáciu. Teda je potrebné pridať *r.e*

2.3.1 The statistics of the dataset

General statistics of the dataset before preprocessing - number of tokens, unique tokens, player names, city names. Why we want to make the number of tokens lower while keeping the ability to have rich vocabulary.

2.3.2 Cleaning

Lorem Ipsum as one of the summaries, the paper doesn't mention if it's used as augmentation of the dataset or if it's a bug - removed. Player initials "C.J. McCollum" in table "CJ McCollum" in text.

2.3.3 Transformations of player names, city names and team names

The motivation - making data denser. In a lot of summaries the player is firstly introduced "It was up to LeBron James to take over for Cleveland" and then referenced only by his surname "James finished with 27 points, 14 assists and 8 rebounds..." Many transformations were ommitted although possible. The text looks more natural if the network learns that Philladelphia is sometimes mentioned as Philly. Those nuances are left in the text.

2.3.4 Other transformations made to the dataset

Here I'll mention lowercasing, number transformations.

2.3.5 Byte pair encoding

What it is, who introduced it. How it decreases the number of tokens.

2.3.6 The statistics of the transformed dataset

What is achieved with the use of all the mentioned transformations. The number of unique tokens decreased almost than 4 times (11 300 to 2 900). The fraction of tokens mentioned more than 5 times increased from 47% to 89,5%. Highlight the differences to processing of

3. Neural Network Architectures

Why only neural network approach is used - mention many authors and papers which approach the problem of NLG by making use of deep neural networks.

3.1 RNN

What it is, creating the representation of sequence, etc. etc. Gradient vanishing and gradient explosion problems.

3.2 LSTM

What it is, how it solves the mentioned problems, cite the paper Massive Exploration of Neural Machine Translation Architectures which experiments with LSTMs, GRUs and vanilla RNNs and shows that LSTMs are the most promising option for the sequence to sequence tasks.

3.3 Attention

What it is, cite Bahdanau, Luong, possible subsection about the input feeding approach. It should select the most relevant entry from the database of match statistics.

3.4 Copy mechanism

What it is, right now I've implemented only Joint Copy mechanism, possibly add Conditional Copy

3.5 Truncated Backpropagation Through Time

Why it is infeasible to generate sequences of average 350 tokens with full backpropagation. Which types of truncated BPTT exist and which I've chosen.

3.6 Beam search

Why greedy search isn't enough, what is beam search, when is it used.

3.7 Transformers

Right now I don't think I'll get this far in my exploration and implementation of DNN architectures.

4. Models

The main goal is to explore the architectures. Therefore each model is manually implemented making use of the tensorflow library

4.1 Tensorflow

What is it, mention the paper introducing it, mention other frameworks and motivation why this is the one which is used.

4.2 Sequence to Sequence architecture

Encoder, decoder. Encoder creates the representation of the input in some meta language, decoder creates the output from the representation.

4.3 Encoder

Mention the embedding and MLP encoding (the main approach used in the rotowire paper) - MLP is used instead of LSTM in the encoding process, then the 2 initial decoder states are obtained by mean pooling over the MLP encodings of the embedded source records.

4.4 Decoder

2 layer LSTM, embeddings, dimensionality.

4.5 Base Model

Seq2Seq architecture with attention, both Luong style Dot Attention and Bahdanau style Concat Attention are used, input feeding approach. Maybe some pictures.

4.6 Joint Copy model

Uses 2 attention mechanisms. Definitely some pictures.

5. Experiments

This chapter should present the observations about the generated data and the steps taken to improve the generations.

5.1 BLEU

What it is, why do I use such a metric for evaluating the data.

5.2 Manual evaluation

How the summaries for manual evaluation are chosen, how many people do evaluate the predicted summaries.

5.3 Other evaluation approaches

Which approaches are presented in the read papers, which improvements should be made.

5.4 Results of the baseline model

Learned which teams play, what are the greatest stars of each team, although the summaries diverge, only first few sentences from the generated summaries are relevant.

5.5 Dropout

What it is, where I apply the dropout - on the decoder LSTM cells.

5.6 Scheduled Sampling

What it is, how it solves the divergence of the summaries.

5.7 Copy methods

How do they help the model to choose more relevant data from the table, how do they fare in the concurrence of the baseline model.

Conclusion

Bibliography

Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks, 2015. URL http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

Remi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain, 2016.

Percy Liang, Michael Jordan, and Dan Klein. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P09-1011.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

List of Tables