



**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	3
1 Problem statement	4
1.1 Data to text generation	4
1.2 Fantasy sports	4
1.3 My goal	4
2 Data	5
2.1 General description	5
2.2 WikiBIO dataset	5
2.2.1 Structured Data	5
2.2.2 Summaries	5
2.3 Rotowire dataset	6
2.3.1 Structured Data	6
2.3.2 Summaries	7
2.3.3 Relation of summaries and tables	8
3 Neural Network Architectures	10
3.1 RNN	10
3.2 LSTM	10
3.3 Attention	10
3.4 Copy mechanism	10
3.5 Truncated Backpropagation Through Time	10
3.6 Beam search	10
3.7 Transformers	10
4 Models	11
4.1 Tensorflow	11
4.2 Sequence to Sequence architecture	11
4.3 Encoder	11
4.4 Decoder	11
4.5 Base Model	11
4.6 Joint Copy model	11
5 Experiments	12
5.1 BLEU	12
5.2 Manual evaluation	12
5.3 Other evaluation approaches	12
5.4 Results of the baseline model	12
5.5 Dropout	12
5.6 Scheduled Sampling	12
5.7 Copy methods	12
Conclusion	13
Bibliography	14

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 $r.v = \{r.v_1, \dots, r.v_m\}$. 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

Na to, aby som vedel rozoznať, čím je generovanie popisu športového zápasu náročné, rozhodol som sa pracovať ešte s jednou úlohou a to s generovaním krátkych životopisov na základe infoboxu z wikipédie. Touto úlohou sa zaoberali tvorcovia datasetu WikiBIO [Lebret et al., 2016].

2.2.1 Structured Data

Charakteristikou tohoto datasetu je, že vstupná tabuľka je výrazne rôznorodá. Vyplýva to z toho, že osoby s inou kariérou majú výrazne iné charakteristiky. U niekoho hovoríme o titule, u iného o povolání, u niekoho spomenieme významné práce, u niekoho tímy, v ktorých hraje. Zatiaľ čo zrejme pri každej osobe sa vyskytne napríklad položka "dátum narodenia", pri žijúcich sa zrejme nevyskytne "dátum úmrtia". Dá sa teda povedať, že štruktúrované dáta z tohoto datasetu sú veľmi ťažké, pretože generačný systém nemôže predpokladať štruktúru, ale musí sa naučiť rozoznať ju. Príkladom môže byť infobox pre T.G. Masaryka 2.1.

2.2.2 Summaries

Ako sumár infoboxu slúži prvá veta z článku na wikipedii, ku ktorému prislúcha daná tabuľka. Z pozorovaní vstupných dát vyplýva, že prvá veta väčšinou obsahuje meno, dátum narodenia, prípadne dátum úmrtia a najdôležitejšie činnosti a úspechy osoby, ktorej sa infobox týka. Už z pozorovaní tvorcov datasetu [Lebret et al., 2016] vyplýva, že okolo tretiny tokenov nachádzajúcich sa v prvej vete pochádza z tabuľky. Generačný systém pracujúci s týmto datasetom teda bude

musieť správne určiť, kedy sa spoľahnúť na jazykové zručnosti a kedy kopírovať. Ako príklad možno použiť sumár prislúchajúci k infoboxu 2.1.

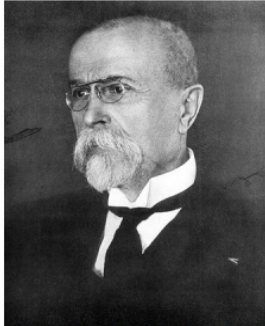
2.3 Rotowire dataset

Pre samotnú úlohu generovania textového popisu zápasu zo štruktúrovaných dát som si vybral dataset RotoWire [Wiseman et al., 2017]. Štruktúrované dáta sú vo forme tabuľky. Tabuľka obsahuje jednak hodnoty popisujúce celkové tímové štatistiky, jednak hodnoty, popisujúce štatistiky hráčov. Ako výstupné dáta slúžia sumáre jednotlivých zápasov z portálu venujúceho sa real-time fantasy sports news <https://www.rotowire.com/basketball/> (see below).

2.3.1 Structured Data

V tejto podsekcii si rozoberieme, aké dáta sa v tabuľke nachádzajú, a čo z nich sa dá využiť pri vytváraní sumáru. Pre prehľadnosť ich rozdelím na tri časti, kontextové údaje, tímové údaje a štatistiky a hráčske údaje a štatistiky. Keďže kontextové údaje sú vo vstupnej tabuľke zastúpené len ako informácia v ktorý deň sa zápas konal a vo väčšine sumárov táto informácia nie je použitá, rozhodol som sa ju tiež úplne vynechať z tréningových dát. Všetky číselné údaje v tabuľkách

tomáš garrigue masaryk , sometimes called thomas masaryk in english (7 march 1850 - 14 september 1937) , was a czechoslovak politician , sociologist and philosopher , who as an eager advocate of czechoslovak independence during world war i became the founder and first president of czechoslovakia .

Tomáš Garrigue Masaryk


1st President of Czechoslovakia

In office
14 November 1918 – 14 December 1935
Preceded by Position established
Succeeded by [Edvard Beneš](#)

Personal details

Born 7 March 1850
[Hodonín, Austrian Empire](#)
(now [Czech Republic](#))

Died 14 September 1937
[Lány, Czechoslovakia](#)
(now [Czech Republic](#))

Spouse(s) [Charlotte Garrigue](#)
[Alice](#) (1879–1966)
[Herbert](#) (1880–1915)

Children [Jan](#) (1886–1948)
Eleonor (1890–1890)
Olga (1891–1978)

Profession [Philosopher](#)

Signature 

Figure 2.1: Príklad z tréningových dát datasetu WikiBIO

sú tvorcami datasetu upravené na celočíselné hodnoty.

Team Statistics

Medzi tímové údaje rátam jednak mená klubov, mestá, kde kluby sídlia. Štatistiky sú jednak kontextové (počet predchádzajúcich zápasov, ktoré klub vyhral, resp. prehral) jednak zápasové (napríklad počet získaných bodov, percentuálna úspešnosť strelby z poľa atp.). Tímové údaje a štatistiky obsahujú celkovo 15 položiek pre každý tím, tvorcovia datasetu túto časť označujú ako *line score* a ich celý výčet sa dá nájsť na githube tvorcov datasetu ¹ Tabuľka 2.1 z validačného datasetu môže poslužiť ako dobrý príklad.

Name	City	PTS ₁	AST ₂	REB ₃	TOV ₄	Wins	Losses ...
Raptors	Toronto	122	22	42	12	11	6 ...
76ers	Philadelphia	95	27	38	14	4	14 ...

Note: The statistics are accumulated across all the team players

₁ Points; ₂ Assists; ₃ Rebounds; ₄ Turnovers

Table 2.1: Príklad tímových štatistík z datasetu Rotowire

Player statistics

U každého hráča je údaj, za ktorý tím hrá, či je daný tím domáci, alebo či hráč patril do starting roster, alebo nastúpil z lavičky. Štatistiky sú len zápasové a medzi ne patria napríklad počet nahraných bodov, počet minút, ktoré hráč hral atp. Hráčske údaje a štatistiky čítajú 24 položiek, tvorcovia datasetu ich nazývajú *box score* a čitateľa, ktorého zaujíma kompletný výpis položiek opäť odporúčame na github tvorcov datasetu ¹ a uvádzame aj príklad tabuľky z validačného datasetu 2.2.

Name	Team City	S_POS ₁	PTS ₂	STL ₃	BLK ₄ ...
Kyle Lowry	Toronto	G	24	1	0 ...
Terrence Ross	Toronto	N/A	22	0	0 ...
Robert Covington	Philadelphia	G	20	2	0 ...
Jahlil Okafor	Philadelphia	C	15	0	1 ...

Note: N/A means that the statistic couldn't be collected because it is undefined

(e.g. player didn't appear on starting roster therefore his starting position is undefined)

₁ Starting position ; ₂ Points; ₃ Steals; ₄ Blocks

Table 2.2: Príklad hráčskych štatistík z datasetu Rotowire

2.3.2 Summaries

Kladíme vysoké nároky na sumáre. Jednak musia súvisieť s tabuľkou, jednak nechceme neurónovú sieť učiť niečo, čo by sa dalo rýchlo napísať pomocou jednoduchej šablóny. Preto potrebujeme sumáre, ktoré dokážu vytiahnuť z

¹<https://github.com/harvardnlp/boxscore-data>

tabuľky hlbšie štatistiky a správne s nimi pracovať. Ako už vysvetľuje [Wiseman et al., 2017] obyčajné športové spravodajstvo zo stránok <https://www.sbnation.com/nba> nesplňa tieto kritériá, keďže obsahuje priveľa informácií, ktoré sú založené na iných kontextoch, ako sú dáta v zápasovej tabuľke. Práve preto je zaujímavé sledovať fenomén fantasy športov.

Fantasy sports

Podľa [Tozzi, 1999] môžeme datovať počiatky fenoménu fantasy športu do šesťdesiatych rokov dvadsiateho storočia. Podľa stránky RotoWire² je základom fantasy športu vytvorenie tímov reálnych hráčov z ligy a získavanie bodov na základe ich výkonov v skutočných zápasoch. Bodovanie berie do úvahy, či ide o defenzívneho, alebo ofenzívneho hráča a je nutné vybrať hráča na každú pozíciu v hre. (čím sa zaručuje, že nemôžu existovať fantasy tímy, ktoré majú len point guardov). Disponujete obmedzenými zdrojmi na nákup hráčov a hráči, u ktorých je pravdepodobnejšie, že skórujú viac bodov, resp. získajú viac lôpt a tak podobne, stoja viac. Skutočné pravidlá sú jemne zložitejšie a tí, ktorých by fantasy ligy zaujímali odporúčame na dané stránky²

Fantasy sports news

Na to, aby hráč mohol uspieť vo fantasy lige, potrebuje mať dokonalý prehľad o štatistikách hráčov, o trendoch, o zraneniach, o tímoch, ktorým sa darí aj o tých, ktorým sa možno začne dariť neskôr. Práve preto už od začiatku organizovania fantasy lig existujú formy spravodajstva, ktoré sa špecializujú práve na hráčov fantasy športov. Podľa legendy popísanej [Tozzi, 1999] sa hráči jednej z prvých fantasy lig schádzali v reštaurácii La Rotisserie Francaise, a podľa tejto reštaurácie je pomenovaná aj stránka špecializujúca sa na spracovávanie štatistík pre fanúšikov fantasy lig, <https://www.rotowire.com/>. Články písané o zápasoch oveľa viac berú do úvahy, čo sa v zápase udialo a zároveň poskytujú aj hlbší náhľad do kontextu zápasu. Preto ako píše [Wiseman et al., 2017] je pre generálne systémy ideálnejšie učiť sa generovať práve články podobné tým z RotoWire.

2.3.3 Relation of summaries and tables

V tejto kapitole na jednoduchom príklade ukážem, ako sú štatistiky z tabuľky previazané so sumárom z RotoWire. Vo figure 2.2 je krásne vidieť, že väčšina údajov zo sumáru pochádza z tabuľky. Napriek tomu však text obsahuje niektoré údaje (**zvýraznené žltou farbou**), ktoré v tabuľke vôbec nie sú a niektoré údaje (**zvýraznené modrou farbou**), ktoré sú v tabuľke len implicitne a je ich potrebné odvodiť. Zatiaľ čo fakt, že keďže Terrence Ross má nedefinovanú štartovaciu pozíciu, tak musel nastúpiť do zápasu z lavičky je celkom zrejmý, informácia o tom, že Joel Embiid nehraje je dôležitá len vtedy, pokiaľ je Joel Embiid natolko dôležitý hráč pre tím, že stojí za zmienku ho spomenúť. To je však kontext, ktorý

²<https://www.rotowire.com/basketball/advice/>
².

v tabuľke spomenutý nie je a navyše informácia o tom, prečo nehrá nemôže byť jasná ani z kontextu celého korpusu dát.

TEAM	WIN	LOSS	PTS ₁	FG_PCT ₂	REB ₃	AST ₄ ...
Raptors	11	6	122	55	42	22
76ers	4	14	95	42	38	27

PLAYER	City	PTS ₁	AST ₄	REB ₃	FG ₅	FGA ₆	S.POS ₇ ...
Kyle Lowry	Toronto	24	8	4	7	9	G
Terrence Ross	Toronto	22	0	3	8	11	N/A
Robert Covington	Philadelphia	20	2	5	7	11	G
Jahlil Okafor	Philadelphia	15	0	5	7	14	C
DeMar DeRozan	Toronto	14	5	5	4	13	G
Jonas Valanciunas	Toronto	12	0	11	6	12	C
Ersan Ilyasova	Philadelphia	11	3	6	4	8	F
Sergio Rodriguez	Philadelphia	11	7	3	4	7	G
Richaun Holmes	Philadelphia	11	1	9	4	10	N/A
Nik Stauskas	Philadelphia	11	2	0	4	9	N/A
Joel Embiid	Philadelphia	N/A	N/A	N/A	N/A	N/A	N/A
...							

The host Toronto Raptors defeated the Philadelphia 76ers , 122 - 95 , **at Air Canada Center on Monday** . **The Raptors came into this game as a monster favorite** and they did n't leave any doubt with this result . Toronto just continuously piled it on , as they won each quarter by at least four points . The Raptors were lights - out shooting , as they went 55 percent from the field and 68 percent from three - point range . They also held the Sixers to just 42 percent from the field and dominated the defensive rebounding , 34 - 26 . Fastbreak points was a huge difference as well , with Toronto winning that battle , 21 - 6 . **Philadelphia (4 - 14) had to play this game without Joel Embiid (rest)** and they clearly did n't have enough to compete with a potent Raptors squad . Robert Covington **had one of his best games of the season though** , tallying 20 points , five rebounds , two assists and two steals on 7 - of - 11 shooting . Jahlil Okafor **got the start for Embiid** and finished with 15 points and five rebounds . Sergio Rodriguez , Ersan Ilyasova , Nik Stauskas and Richaun Holmes all finished with 11 points a piece . **The Sixers will return to action on Wednesday , when they host the Sacramento Kings for their next game** . Toronto (11 - 6) left very little doubt in this game who the more superior team is . Kyle Lowry carried the load for the Raptors , accumulating 24 points , four rebounds and eight assists . **Terrence Ross was great off the bench , scoring 22 points on 8 - of - 11 shooting** . DeMar DeRozan finished with 14 points , five rebounds and five assists . Jonas Valanciunas recorded a double - double , totaling 12 points and 11 rebounds . **The Raptors next game will be on Wednesday , when they host the defensively - sound Memphis Grizzlies** .

Note: ₁ Points; ₂ Field Goal Percentage; ₃ Rebounds; ₄ Assists; ₅ Field Goals; ₆ Field Goals Attempted; ₇ Starting Position; N/A means undefined value

Figure 2.2: Príklad vstupných tabuliek a zlatého sumáru z datasetu. Žltou sú zvýraznené informácie nenachádzajúce sa v tabuľke, zatiaľ čo modrou sú zvýraznené informácie, ktoré z tabuľky a celého korpusu vyplývajú len implicitne.

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 back-propagation. 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 ro-towire 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 Bah-danau 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 Lebrete, 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.
- Lisa Tozzi. The great pretenders, 1999. URL http://weeklywire.com/ww/07-05-99/austin_xtra_feature2.html.
- Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. Challenges in data-to-document generation, 2017. URL <https://arxiv.org/abs/1707.08052>.

List of Tables

2.1	Príklad tímových štatistík z datasetu Rotowire	7
2.2	Príklad hráčskych štatistík z datasetu Rotowire	7