SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Fakulta informatiky a informačných technológii

Katedra informačního a znalostního inženýrství

Neurálne Modely na Generovanie Textu v Slovenskom Jazyku

Bakalárska práca

Dominik Vasko

Vedúcí práce: Ing. Samuel Pecár

máj 2019

Anotácia

V tejto práci sa budeme zaoberať generovaním textov pre ľudí, ktoré sú určený na čítanie a majú podobu prirodzeného jazyka. Práca obsahuje analýzu oblasti generovania textov prirodzeného jazyka a jednotlivých využití generovaných textov ako sumarizácia, simplifikácia, parafrázovanie a generovanie dialógov, ich vývoj a súčasný stav pre tieto úlohy. Súčasťou práce je aj návrh a implementácia modelu na generovanie textu na ktoré budú použité hlboké rekurentné neurónové siete, ktoré dosahujú v poslednej dobe úspech v tejto oblasti a zďaleka najlepšie výsledky. Samotné generovanie textu bude robené v jazyku slovenskom a anglickom, účelom implementácie modelu v dvoch jazykoch je hlavne porovnanie rôznych prístupov pre rôzne jazyky. Všetky experimenty budú robené na paralelných korpusoch podobnej veľkosti a z podobných voľne prístupných zdrojov ako napr. Wikipédia. Evaluácia výsledkov experimentov bude robená manuálne ľudskou silou metódou hodnotenia metrík generovaného text. Dôvodom pre manuálne hodnotenie je, že ľuďmi čitateľný text je komplexný a automatizované hodnotenie nie je vždy vhodné.

Annotation

The goal of this works is to generate text in a natural language which is readable to humans, it contains an analysis of the natural language generation field and its uses like text summarization, simplification, paraphrasing and dialog generation and their evolution. Part of this work is a design and implementation of a model for text generation. Deep recurent neural networks will be used for that task, the reason for that is their efficiency and the fact the most state-of-the art models use them. Texts will be generated in Slovak and English languages to demostrate differences between those languages from the context of NLG. All models used will be trained on a paralel corpuses of similar size for example from Wikipedia. The evaulation of the results will be done manualy with human power by evaluating metrics. The reason for that is mainly the complexity of natural languags where automated evaluation is not suficient.

Poďakovanie

Ďakujem Ing. Samuelovi Pecárovi za odborné rady a usmernenie pri vypracovávaní mojej bakalárskej práce.

Prehlásenie

Čestne prehlasujem, že záverečnú prácu som vypracoval samostatne s použitím uvedenej literatúry a na základe svojich vedomostí a znalostí.

V Bratislave dňa 27. marca 2019

Dominik Vasko

Obsah

Zoznam obrázkov

Zoznam tabuliek

Kapitola 1

Úvod

Zrak, hmat a sluch sú tromi našimi najdôležitejšími vnemami, pomocou ktorých prijímame veľké množstvo informácií a bez ktorých by sme sa ťažko v živote zaobišli. Práve preto text, či už vo viditeľnej, hmatateľnej alebo zvukovej podobe predstavuje dôležité médium na prenášanie, komunikáciu a organizovanie dát a informácií, ktoré má hlavne v dnešnej dobe internetu a sociálnych sietí veľmi dôležitú rolu [30] .

Čím viac sa spoločnosť blíži k viac digitálnej kde informácií je veľmi veľa bude treba tvoriť ale aj vstrebávať a organizovať viac a viac textov. Google, Wikipédia, protokoly ako html ktoré nám umožňujú prispievať obsah na internet, blogy, magazíny, knihy, noviny, väčšina obsahu na sociálnych sietiach a na internete atď. sú tvorené veľkým množstvom textov.[16]

Nie len digitálne ale aj mnoho klasických ľudských aktivít v sebe zahŕňa tvorbu alebo vstrebávanie nejakých druhov textov či je to obyčajná medziľudská komunikácia alebo písanie na papier. [17]

Práca s textom sa týka mnohých oborov od študentov, sekretárok až po softvérových architektov a manažérov. Aby veľké množstvo informácií nespôsobilo informačné preťaženie je treba prácu s textom automatizovať a vysporiadať sa s filtrovaním užitočných informácií od tých zbytočných, čo vedie k uľahčeniu práce mnohých ľudí. [26]

Väčšina týchto textov je v podobe prirodzeného jazyka(slovenčina, angličtina) a majú osobnú povahu, čo môže predstavovať prekážku pri automatizácii. Lúdia veľmi jednoducho vedia rozlíšiť zlý text od dobrého. [25]

Manuálne písanie textov môže byť náročné na realizáciu, jednak kvôli potrebe vedomostí v danej oblasti druhak, že to môže byť nudné a cenovo nevýhodné a v niektorých prípadoch, kedy je textu veľmi veľa, aj nemožné. Aj ľudia aj stroje by profitovali z automatizácie týchto činností. Zautomatizovanie takýchto činnosti sa zaoberá NLG alebo generovanie prirodzeného jazyka. [24]

V tejto práci sa zameriame na generovanie textov konkrétne v jazyku slovenskom a anglickom. Kvôli tomu, že väčšina automatizovaní tvorenia textov je robená v angličtine. Predstavuje to niekoľko prekážok. Slovenkých textov je pomerne málo a ešte ťažšie sa získavajú, ak chceme použiť nejaký novší prístup potrebujeme veľa vzorov na to ako text správne písať(nechceme to robiť manuálne). Navyše slovenčina je morfologicky bohatší jazyk, používa veľa prípon, predpôn a má viac časov ako angličtina. Taktiež exituje málo systémov, ktoré sa o generovanie slovenského textu pokúsili, preto budeme používať niečo, čo funguje na angličtinu ale vieme to použiť aj na sloven-

činu. Z tohto vypláva aj ďalší cieľ práce a to porovnanie nejakých aktuálnych prístupov na generovanie textu v slovenčine a angličtine. [23]

Porovnanie chceme robiť kvôli tomu, že systémy majú svoj limit a tým pádom že slovenčina je pomerne expresívny a syntakticky bohatý jazyk ako angličtina. Preto prístupy vhodne pre angličtinu alebo iný jazyk nemusia nutne byt dobré pre slovenčinu, respektíve chceme sa dozvedieť ci je nejaký signifikantný rozdiel medzi rovnakým systémom, ktorý sa učí iný jazyk. [22]

Prínosom práce ma byť porovnanie toho či prístupy použite v iných jazykoch fungujú aj na slovenčine a či sa dá niečo zmysluplné vygenerovať. Na ohodnotenie budeme potrebovať ľudí, ktorý text ohodnotia a povedia či je dobrý na základe rôznych metrík. Dôvod pre použitie ludi je ten, že text je komplexný a automatické overenie správnosti je subjektívne, komplikované a nástroje, ktoré sú prístupné sú nepresné môžu považovať aj zlé texty za správne. [20]

Môžeme nájsť viacero definícií NLG jednou z klasických je, že NLG je jednou z oborov umelej inteligencie a NLP , ktorej úlohou je zvyčajne z nejazyčných vstupov vygenerovať text pre rôzne domény a oblasti ľudských činnosti . [18]

V poslednej dobre sa však na rôzne úlohy hlavne text-to-text používajú prístupy, ktorých vstupom nie sú len dáta ale aj text. Rozdiel medzi touto prvotnou definíciou NLG a reálnymi aplikáciami v súčasnosti značí, že NLG prešlo počas posledných dvoch desaťročí mnohými zmenami.

Vo všeobecnosti ide o generovanie textu, ktorý je priamo určený na čítanie pre ľudí. Táto definícia je všeobecná a preto má NLG široké využitie. Niektoré príklady využitia NLG zahŕňajú napr. generovanie článkov do novín ako napríklad futbalové reportáže, zdravotnícke správy, ale aj všeobecnejšie, nezávislé od domény využitia ako simplifikácia alebo sumarizácia textu, generovanie parafráz, prekladanie textov.

Kapitola 2

Generovanie Textu

Kým výstupom NLG je skoro vždy text. Vstupom môže byť spektrum údajov. Vstupom však musí byť niečo čo chceme koncovému čitateľovi výstupným textom sprostredkovať, informovať ho. Príkladmi vstupu môžu byť číselné údaje(data-to-text) alebo text samotný(text-to-text).

2.1. Podúlohy pri generovaní prirodzeného jazyka

Dosiahnutie transformácie vstupu na výstup môže byť vykonávaný rôznymi spôsobmi. Jeden z takých všeobecných prístupov opisujú Reiter a Dale vo svojej knihe. V tomto prípade je celý proces rozdelený na menšie podúlohy. Kde každá z týchto úloh vykoná istú časť transformácie vstupu. Tieto úlohy sú nasledovné Content determination, Text structuring, Sentence aggregation, reffering expression generation, lexicalization a surface realization . Toto rozdelenie sa stalo de-facto rozdelením v NLG , ktoré je dodnes používané v niektorých systémoch, Reiter k tomuto rozdeleniu dospel na základe pozorovaní a trendov v NLG systémoch.

2.1.1. Content determination

Alebo určovanie obsahu, slúži na určenie obsahu toho čo na konci generácie chceme čitateľovi sprostredkovať. Vstup, dataset z ktorého text generujeme môže obsahovať aj nadbytok informácií, ktoré sú irelevantné, pre naše špecifické použitie. Táto úloha je veľmi dôležitá lebo ovplyvňuje ostatné úlohy ktorú ju nasledujú, ak vyberieme málo informácií konečný text môže byť neúplný. Ďalšou s problémov je že určovanie obsahu je zvyčajne závislé od domény v závislosti čo generujem, chcem vybrať správne údaje.

Napr. pri generovaní článkov pre fanúšikov futbalu, sa na vstup vyberú len informácie, ktoré sa väčšinou vyskytujú v článkoch písaných ľuďmi.

Od starších prístupov prešlo na viac automatizované prístupy pomocou strojového učenia alebo rozpoznávania vzorov. Kde sa relevantný obsah vyberá na základe korelácie výskytu slov a variácie dát.

2.1.2. Text structuring

Štrukturovanie textu slúži na zoradenie informácií získaných výberom v predchádzajúcom kroku, do správneho poradia aby dávali zmysel. Nesprávne poradie môže viesť

k nezmyselnému textu, ktorý, čitateľ nebude vedieť sledovať. V Prípade futbalových článkov môžeme dáta znova organizovať do takého poradia v akom by sme to našli napr. v novinách. t.j. Nadpis, úvod, chronologický priebeh hry a beseda. Ak by sme začali rekapituláciu zápasu od zadu, dostali by sme článok ktorý je nezmyselný.

2.1.3. Sentence aggregation

Pri tejto úlohe sa stretávame s viacerými definíciami. ako odstránenie redundantných informácií alebo určenie blízkosti jednotlivých údajov(v jednej vete alebo vo viacerých). Výsledok je však rovnaký urobiť text kompaktnejším aby bol ľahšie a lepšie čitateľní. Môže to znamenať rozdiel medzi piatimi vetami kde sa mení iba predmet vety a jednou vetou kde všetky tieto predmety sú vyjadrené jedným slovom alebo spojené spojkami. V praxi to môže znamenať rozdiel medzi nasledujúcimi výstupmi.

- Tomáš mal v pondelok na raňajky jablko.
- Tomáš mal v nedeľu na raňajky jablko.
- Tomáš mal celý týždeň na raňajky jablká.

2.1.4. Lexicalization

Leksikalizácia znamená nájdenie správnych slov na vyjadrenie informácií v texte. Jazyky zvyčajne majú synonyma, pomocou ktorých môžeme text plynulejšie vyjadriť. Opakovanie rovnakých výrazov môže pôsobiť repetitívne a nudne.

2.1.5. Referring expression generation

Generovanie odkazujúcich výrazov(Reffering Expression Generation), slúži na tvorbu správnych odkazov pre jednotlivé veci v texte, hlavne kvôli rozlíšiteľnosti. Je to zvyčajne diskriminačná činnosti pri ktorej jednotlivé doménové entity špecifikujeme dovtedy, kým ich vieme od ostatných entít rozoznať.

2.1.6. Surface realization

Poslednou úlohou je samotné generovanie gramatický, syntaktický a morfologický správneho a konkrétneho textu. Slúži na pretavenie abstraktných reprezentácií ktoré sme doteraz zo vstupu získali do konkrétneho jazyka.

Najjednoduchšie sa dá spraviť šablónami alebo predurčenými statickými spravami(canned text) tieto prístupy sú však primitívne a nedostatočné. Prístupy, ktoré sa používajú sú zvyčajne založené na štatistických princípoch alebo na základe rôznych gramatík(CCG, SGS).

2.2. Krátka história NLG

Prvé systémy, ktoré sa objavujú v 60. rokoch slúžili hlavne na preklad textov z rôznych jazykov. Vo väčšine prípadov vykonávali iba surface realization. Až neskôr v 70. rokoch

sa objavujú systémy na generovanie textu z ne-lingvistických údajov. Tieto systémy poukazovali na to že NLG nie je len NLU odzadu a taktiež na vznik istých problémov pri NLG. 80. roky predstavujú obdobie rozvoja NLG, odstupovalo sa od stavania monolitických systémov a pristúpilo sa k skúmaniu jednotlivých podúloh pri generovaní. Koncom 90. rokov väčšina architektúr používala pipeline(rúrovú) architektúru s menšími zmenami. Tieto architektúry boli podobné aj naprie tomu, že mali rozdielne teoretické pozadia(východiská?). Taktiež rozdeľovali generovanie na podobné podúlohy. Táto podobnosť pravdepodobne súvisí s tým ako ľudský mozog vytvára hovorenú reč aj keď podobnosť k psycholingvistike nebol cieľom týchto systémov.

Rúrová architektúra sa teda stala de facto štandardom. Je mnoho dôvodov prečo rúrovú architektúru používať. Jedným z nich je jej jednoduchosť a jednoduchosť implementácie a následné odhaľovanie chýb. Vývoj takýchto systémov stojí menej námahy ako vytvárať komplexné systémy. Aj keď majú svoje nedostatky ako napríklad absencia spätnej väzby.

2.3. Úlohy v NLG

Okrem horeuvedených podúloh existujú aj iné prístupy, ktoré v sebe nezahŕňajú modulárne systéme. Sem patria metódy založené na strojovom učení a neuŕonových sietiach, ktoré sa učia vzťah medzi vstupnými a výstupnými údajmi. Nasleduje zopár populárnych využití NLG, sú to hlavne využitia pre úlohy typu text-to-text . Aj keď podobné prístupy môžeme použiť aj na klasické data-to-text generovanie.

2.3.1. Parafrázovanie

Parafrázovanie alebo prerozprávanie toho istého obsahu inými slovami. Je užitočná činnosť, ktorá má mnoho využití v kombináciou s ostatnými úlohami NLG ako sumarizácia, odpovedanie otázok(question answering), strojový preklad a.i . Prístupy k parafrázovaniu môžeme rozdeliť na monolingvistiské, pivotné metódy a neurónové prístupy. V porovnaní prístupov metóda založená na neurónovoých prístupoch mala lepšie výsledky ako klasické modely založené na frázach.

2.3.2. Simplifikácia

Simplifikácia textu znamená zmenu jazykovej štruktúry pričom informácie, ktoré sú v tomto texte obsiahnuté ostávajú rovnaké, zmysel textu sa nemení. Hlavnou motiváciou pre simplifikácia je sprístupnenie informácií menej vzdelaným ľuďom, deťom, cudzincom, alebo osobám s rôznymi poruchami ,ktoré im sťažujú pochopenie text ako napr. dyslexia alebo ľudia trpiaci hluchotou .

Slabší čitatelia môžu mať problémy s čítaním komplikovanejších textov keď sa mozog musí sústrediť na spracovávanie slov vyššie kognitívne činnosti trpia. Ľudia trpiaci takýmito poruchami musia použiť väčšiu časť pamäte na pochopenie textu. Rozdelenie viet na kratšie časti spôsobuje odbremenenie od pamätanie si, a zjednodušuje čítanie. Je dokázané, že metódy manuálnej simplifikácie textu pomáhajú slabším čitateľom. Práve tieto štúdia motivovali výskum automatizovanej simplifikácie textu.

Okrem ľudí simplifikácia textu môže predstavovať rôzne výhodu aj pre systémy, ktoré s textom extenzívne pracujú ako napríklad systémy na strojový preklad.

Tak ako aj vo viacerých oblastiach NLG aj pri simplifikácii sa v poslednej dobe prechádza od klasických prístupov založených na gramatikách, ručne písaných pravidlách na používanie neurónové siete, konkrétne modely typu sequence-to-sequence, ktoré dosahujú lepšie výsledky ako doteraz známe systémy . Tieto systému sa učia end-to-end, a majú jednoduchšie architektúry ako klasické systémy založené na rôznych štatistických prístupoch, umožňuje to trénovanie modelov na základe znakov alebo aj vo viacerých jazykoch. Oproti prvotným systémov tieto pristupujú k celej úlohe simplifikácie ako k prekladu jazyka do jeho zjednodušenej podoby.

Populárnou trénovacou sadou pre simplifikáciu textov sú jednak manuálne simplifikované texty (gigaword duc etc.) alebo paralelné korpusy napr. Anglická wikipédia a jej simplifikovaná alternatíva.

2.3.3. Sumarizácia

Sumarizácia má za úlohu skrátiť text za účelom znížiť množstvo textu na čítanie pričom sa zachovajú dôležité informácie v texte. Hlavnou motiváciou pre takéto systémy je zníženie informačného preťaženia v dobe internetu a ľahko prístupných informácií.

Jedna z prvých pokusov o sumarizácia, ktorá sa zaoberala tvorbou abstraktov z vedeckých textov, fungovala na základe toho ako často sa jednotlivé slová vyskytujú a na základe tejto metriky sa vybrali vety, ktoré sa použijú v koncovom texte . Tento prístup je založený na tom, že autor ktorý dielo píše bude dôležité slová opakovať, a nebude mať dostatok synonym. Celá sumarizácia sa potom zakladá na extrakcii viet.

Okrem extrakcie poznáme aj ďalšie prístupy ako kompresia alebo abstrakcie. Avšak kvôli tomu, že väčšina ľudských sumárov je viac abstraktných, spôsoby kde sa vyberajú slová a vety z pôvodného textu sú nedostatočné. Je za potreby text pochopiť a následne abstraktnú myšlienku vyjadriť textom .

V prípade abstraktívnej sumarizácie dosahujú najlepšie výsledky neurónové siete typu sequence-to-sequence .

2.3.4. Generovanie dialógov

Dialógové systémy majú za úlohu generovať odpovede pre zadané otázky. Využitie takýchto systémov je možné napríklad pri chatbotoch, ktoré majú za úlohu komunikovať s užívateľom v prirodzenom jazyku.

Klasické prístupy v sebe zahŕňali výber odpovede z databázy odpovedí. Tieto prístupy majú malú úroveň generalizácie a možnosti odpovedí sú limitované tým pádom, že sa odpovedá na základe pravidiel.

Novšie prístupy zahŕňajú založené na štatistickom strojovom učení v sebe zahŕňajú učenie sa vzťahov medzi otázkami a odpoveďami na základe nejakého korpusu. Tieto prístupy sú automatizované a jediné čo potrebujeme sú páry - otázky, odpovede. Najlepšie na realizácie takýchto úloh sú znova sequence-to-sequence modely. Ktoré dosahujú vylepšené výsledky oproti starším spôsobom.

2.4. NLG pomocou deep learningu

Ako sme mohli vidieť mnoho state of the art systémov používajú neurónové siete. Tieto prístupy sa stávajú viac a viac populárnymi hlavne v dnešnej dobe keď je prístupným mnoho dát a korpusov.

Ako však použijeme neurónové siete? Tým pádom, že naším cieľom je použiť text ako vstup a z neho generovať text. budeme potrebovať architektúru neurónových sietí, ktorá je na túto úlohu vhodná. Klasické neurónové siete s kladnou spätnou väzbou nie sú dostatočné na spracovanie sekvenčných vstupov.

Pre spracovanie sekvenčných dát (hudba, text atď.) je vhodné používať rekurentné architektúry, hlavnou výhodnou je že tieto architektúry robia rozhodnutia aj na základe vstupov ktoré už dostali, nie len na základe aktuálneho vstupu. Klasické RNN siete trpia miznúcim alebo explodujúcim gradientom, ktorý je technický problém ktorý nastáva pri propagácia chyby a gradient jednoducho stráca, na adresovanie týchto nedostatkov boli vyvinuté architektúry ako **GRU** alebo **LSTM** s tzv. hradlami(gates) ktoré kontrolujú čo si neurón má zapamätať alebo zabudnúť. Všetky tieto neurónky majú za úlohy naučiť sa pravdepodobnostnú funkciu, ktorá zachytáva závislosť medzi vstupmi a výstupmi.

Viacero text-to-text NLG úloh majú na vstupe aj výstupe sekvenciu znakov, pre tieto vieme použiť špeciálne architektúry sequence-to-sequnce. Tieto architektúry sa zvyčajne skladajú z dvoch vrstiev jedného enkódera a dekodéra.

Kapitola 3

Návrh

Chceme pomocou neurónovej siete namodelovať funkciu ktorá nám povie, po istom počte písmen aké je najpravdepodobnejšie ďalšie písmeno. Takýmto spôsobom budeme vedieť generovať texty. Celý proces môžeme rozdeliť na dve podúlohy, prvá je trénovanie modelu, kedy sa model učí a druhá podúloha je generovanie, pri ktorom model na základe naučeného predpovedá.

Na trénovanie budeme potrebovať dáta, z ktorých sa náš model vie naučiť pravdepodobnostné rozloženie písmen. Pre tento prípad sme si zvolili Wikipédiu [1] , kvôli svojej rozmanitosti a veľkosti. Model na vstup dostane sekvenciu znakov a bude musieť predpovedať sekvenciu znakov posunutú o jeden znak ďalej, takto sa naučí ako z aktuálnej sekvencie vygenerovať ďalšiu sekvenciu s pridaným znakom.

Na generovanie si môžeme vstupnú sekvenciu vymyslieť, môže to byť písmeno alebo celé slovo. A už trénovaná sieť bude vedieť povedať, pravdepodobnosti toho aké by malo byť ďalšie písmeno v poradí. Príkladom môže byť že mu na vstup zadáme sekvenciu "abc" a sieť vráti pravdepodobnosti pre celý slovník v tomto prípade iba a,b,c s prislúchajúcimi hodnotami ich pravdepodobnosti výskytu ako ďalšie písmeno(napr. 0.2, 0.2, 0.6) potom je už na nás, ktoré si vyberieme a ako.

3.1. Opis Modelu

Tým pádom že modelujeme sekvenciu znakov a závisí nám aj od predchádzajúcich vstupoch nie len na konkrétnom použijeme rekurentnú neurónovú sieť. Konkrétne **LSTM** siete, ktoré nemajú problém s dlhodobými závislosťami vstupov.

Na modelovanie a generovanie použijeme model znázornení na obrázku. Vstup do siete bude kódovaní ako one-hot vektor (každé písmeno má priradení index) tento vstup sa posunie do Embedding vrstvy, ktorá má za úlohu naučiť sa vektorové reprezentácie pomocou desatinnými číslami pre jednotlivé vstupy, táto reprezentácia sa potom posunie do **LSTM** vrstvy. Nakoniec výstup posunieme do lineárnej vrstvy ktorá nám zredukuje výstup z **LSTM** vrstiev na vhodnú veľkosť, ktorý potom môžeme interpretovať ako vektor pravdepodobnostných hodnoty pre jednotlivé možné vstupy.

Konfigurácia siete je nasledujúca: embedding vrstva nám vráti vektor s 200 dimenziami máme 2 vrstvy **LSTM** sietí, s veľkosťou 200 neuŕonov a na konci použijeme softmax

[1] Obsah celej Wikipédie v podobe textu je možné nájsť tu https://dumps.wikimedia.org/

3.2. Trénovanie Modelu

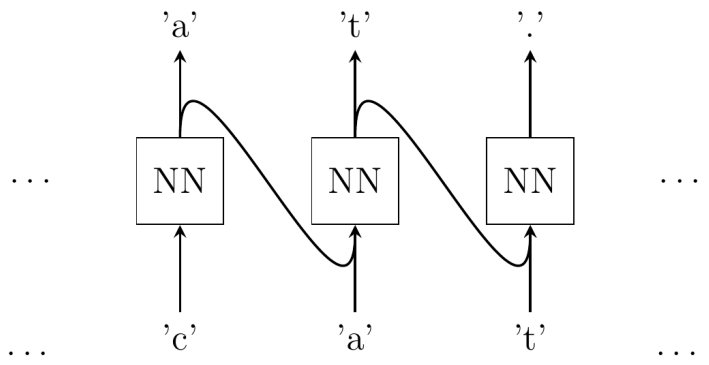
Náš model budeme trénovať na slovenskom a anglickom texte, korpus použijeme z wikipédie. Kvôli rozdielom vo veľkosti slovenskej a anglickej wikipédie použijeme voľne dostupné texty slovenskej a simplifikovanej anglickej wikipédie.

Trénovanie siete potom prebieha tak že každému písmenu pridelíme index(one-hot vektor) zo slovníka a následne ich pošleme do nášho modelu. Výslednú hodnotu porovnáme s ďalšou v poradí, váhy sa vhodne upravia

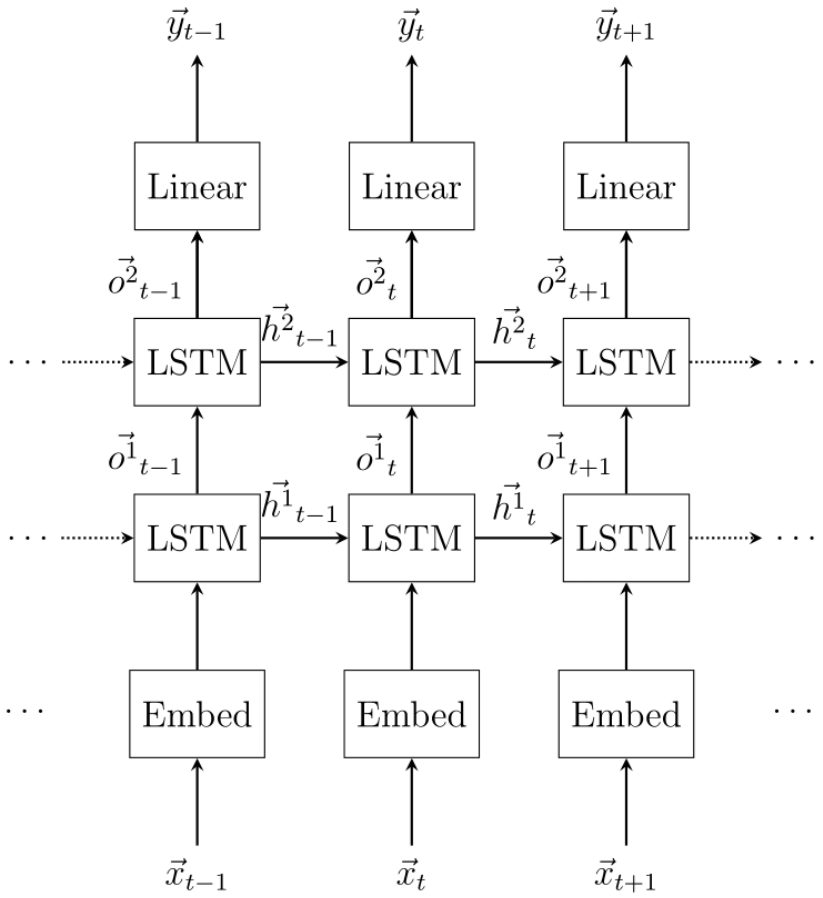
Celú sieť trénujeme na slovenskej a anglickej wikipédie. Kvôli rozdielom vo veľkosti a kvality slovenskej a anglickej wikipádii použijeme len simple english wikipédiu, ktorá má porovnateľnú veľkosť.

3.3. Generovanie Textu

V tejto fáze natrénovanej sieti pošleme písmeno a z výstupu siete zistíme písmeno s najvyššou hodnotou pravdepodobnosti, ktoré sa by sa mohlo vyskytnúť po vstupe tento výstup použijeme ako ďalší vstup do siete, takto pokračujeme, kým nemáme dostatočné množstvo textu viď obr. modelu generovanie textu.



Obrázok 3.1. Spôsob generovanie písmen na základe predchádzajúcich písmen



Obrázok 3.2. Model neurónovej siete na generovanie textu

3.4. implementácia

Na implementáciu použijeme programovací jazyk Python [2] a na uľahčenie práce s neurónovými sieťami knižnicu PyTorch , ktorá obsahuje väčšinu potrebných komponentov na poskladanie modelu.

[2] Väčšina kódu je dostupná na https://github.com/mmio/char-rnn.pytorch

Kapitola 4

Cieľ Práce

Cieľom tejto prace bude generovať text v prirodzenom jazyku, t.j. určený pre ludi na čítanie. Využitie takéhoto systému potom môže byt rôznorodé, slúži hlavne na automatizovanie prace pri písaní textov. A automatizovanie rôznych úloh na ktoré bolo treba robiť v minulosti manuálne sem spadajú už zmienené úlohy simplifikácie, sumarizácie, strojového prekladu, parafrázovania a. i.

Pozrieme sa na oblasť NLG ako sa vyvíjala a aké v nej majú využitia neuronové siete. Generovanie bude teda robene cez neuronové siete, hlavne kvôli dostupnosti veľkých množstiev dát a jednoduchosti ich použitia a trénovania. Ich výkon je tiež porovnateľný s inými metódami v NLG, navyše odstraňujú mnoho nevýhod starších prístupov.

Hodnotenie výsledkov generovania bude robené na základe metrík, ale kvôli komplexnosti prirodzeného jazyka a ťažkosti určenia objektívnej kvality textu text bude vyhodnotený aj manuálne.

Pre túto prácu sme si stanovili nasledovné ciele:

- Analyzovať oblasť NLG a možnosti na generovanie textu
- Zistiť čo v sebe generovanie textu zahŕňa
- Navrhnúť a implementovať model na generovanie
- Využitím NN natrénovať modely, ktoré sú schopné generovať text.
- vyhodnotenie výsledkov trénovanie na slovenskom aj anglickom texte z kvalitatívneho hľadiska
- Trénovanie na paralelných korpusoch

Kapitola 5

Vyhodnotenie

Výsledky experimentu budú vyhodnotené manuálne. Na vyhodnotenie sme každému testerovi dali 6 vygenerovaných textov, pričom sa vyhodnocovali tieto 4 kvalitatívne parameter textu:

- Správnosť jednotlivých slov
- Morfologický tvar slov
- Slovosled
- Zmysluplnosť celého textu

Tieto metriky sú v poradí od najkratšej závislosti písmen až po najdlhšie. Spôsob vyhodnotenia bude nasledovný, zoberieme všetky odpovede zvlášť pre slovenský a anglický text a vytvoríme z nich priemer.

5.1. Anglický text

Výsledky hodnotenie anglických textov sú nasledovné: Zhruba 79% vygenerovaných slov dávalo zmysel Slovosled bol správny v 40% textov Morfológia v prípade 66% A zmysel dávalo len 23% z celkových textov

V prípade angličtiny mala naša sieť problém hlavne zachovať nejakú zmysluplnú myšlienku, kým slová a ich tvary boli pomerne dobré, sieť nedokázala zachovať dostatočný súvis pre vety a celkové články.

5.2. Slovenský text

Výsledky hodnotenie slovenských textov sú nasledovné: Zhruba 69% vygenerovaných slov dávalo zmysel Slovosled bol správny v 58% textov Morfológia v prípade 54% A zmysel dávalo len 17% z celkových textov

Skoro v každej metrike bola slovenčina horšia, pravdepodobne kvôli rozdielu medzi komplexnosťami jazykov. Slovenčina má pestrejšiu morfológiu a slová sa dajú ohýbať do viacerých tvarov.

Z výsledkov hodnotenia vyplýva, že aj napriek jednoduchosti nášho modelu bol schopný sa naučiž krátkodobé závisloti ktoré nám dávali zmysluplné slová, čo sa týka komplikovanejších vecí ako súvis a závislosť jednotlivých častí textov, na ktoré sa siet potrebuje naučiť dlhšie závisloti, sieť nebola schopná zapamätať si vzťahy, pomohla by asi pozornosť.

|1|

Tabuľka 5.1. Ukážka vygenerovaného anglického textu.

John Cheerny John Leyn Show (August 24, 1940 0 February 23, 1987) was an American singer-songwriter and worldwide. He acted in the band for songs for the 1990s to the "Michael Albert Andrew & Ort". He was set during the mid-1990s. He won a former club which became a character for the business Adele Shelley dollar women's song book "Raw" as NXT War Railway from the 2001 census.

Tabuľka 5.2. Ukážka vygenerovaného slovenského textu.

Saint-Martin Saint-Maritre je francúzska obec, ktorá sa nachádza v departemente Orne, v regióne Dolná Normandia. Obec má rozlohu . Najvyšší bod je položený a najnižší bod Počet obyvateľov obce je (). Nasledujúci graf zobrazuje vývoj počtu obyvateľov v obci.

Dodatok A

Ukážka vygenerovaného textu v slovenčine

A čo mali prečítať protifičné miestna #nedelo :-) @user @user @user @user @user #hadlo #presov #volejbal #koleka #video @

A ty tentokrát pripomína na tejto srdce proti svojej politicke dom aj na strednej nápojenie v Babiáku!: Po rokoch rekord

Ako na nás partneria na predstavenie nebezpečného médií @user #kradla #presov #výkonnica #presov #vysoket @user @user @us

A je to tumbolle či nenechám, dnes večer brutálne z jeho hviezdych sociálnych hráčov s prechádzkou sa idu... @user @user

Aktuálne stretnutie prednáškou v rodine v českej školstave 2.12. Martin v S pozornom podvode v BA bol veľa svojmu... @use

Autor zacal sebavedidujúce miesto AngelsHorty Gáborín Chande, Martin, Josel Club (firme) a Prečo je to pokračujú... @user

A prosim to bolo sa aj na ebolu. A o vami ma aj o pomale... @user #demo #staratrznica #data #troca #17mster @user @user #

Ak si zastavoval ešte poslal stojí, výhercu je najlepšie a komunikácie ministerky už vie @user #got #foto #presov #volby

Ako precvičím nám dnes o 2000: Bolo to posledne sledovať ako po práve takto co by som si... na praci? @user @user #presov

Ako sa k výhramení v kole? Príbeh na kurze zoznam všetkým si prekvapíme ze to budeme možnosť" @user @user @user & @us

Apple s svadby - novinka @user (@ Železnicu) @user in Bratislava) @user @user a @user @user @user @user @user @user @user

A teraz sa pozreli o poriadku a počítač poradím od M2 sa pravda na nové veci. Tentoraz Kralovej a dobré a ctipy! @user #h

Ako môže tento piatok pre #SND? Všetci sme mali tvár v Bratislave komunita pre Vás dostal do tisíc vráť @user @user @user

A o tom, čo vám niekto mega zacinal na Slovensku? via @user @user @user @user @user ja :DViac ako uspevnení tomu, a

Aký tomu má takto som robí od Čelec (Estan postupne povenovanie už podobne súd rodič. @user #hokej @user #hokej #foto #mi

A je to bolo ani ne minimaly a teraz je bez. Top vy skutočne pod akodrajni realitailne 20.11 - dlhodobeje. #terming #base

Ako to zasiahol tretí farebe @user via @user @user @user , IFTTT, Nie Roja Idea Villain Deep Fico z formuly @user via @us

Aký je zasa zlogusáte? :) @user @user Vyčítal som, nie len na víkend na stránke na kontexte, baberné... @user @user #pres

Ak nechtáti by ste si získali veľký páro :) akujeme. @user @user @user @user @user @user až zabudli len 10 a... @user @us

Ako sa bude v Bratislave a pozrite sa aj vybavili. Majú v balkáqu, tak nič drsne viac. A potom nemôžete sa s fotosom na z

Ako sa nahotom s bielej noci v Klub Jozefa @user #hokej #presov #svk1992 @user @user @user @user @user @user @user @user

Bibliografia

[1] Dana Abu Ali a Nizar Habash. *Botta: An arabic dialect chatbot*. 208--212. http://aclweb.org/anthology/C16-2044. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. . 2016. The COLING 2016 Organizing Committee.

[2] Michele Banko a Lucy Vanderwende. *Using n-grams to understand the nature of summaries*. http://aclweb.org/anthology/N04-4001. Proceedings of HLT-NAACL 2004: Short Papers. . 2004.

[3] R. Chandrasekar, Christine Doran, a B. Srinivas. *Motivations and methods for text simplification*. http://aclweb.org/anthology/C96-2183. COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics. . 1996.

[4] Emilie Colin a Claire Gardent. *Generating syntactic paraphrases*. 937--943. http://aclweb.org/anthology/D18-1113. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. . 2018. Association for Computational Linguistics.

[5] William Coster a David Kauchak. *Simple english wikipedia: A new text simplification task*. 665--669. http://aclweb.org/anthology/P11-2117. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. . 2011. Association for Computational Linguistics.

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, a Yoshua Bengio. *Learning phrase representations using rnn encoder--decoder for statistical machine translation*. 1724--1734. http://www.aclweb.org/anthology/D14-1179. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. . October 2014. Association for Computational Linguistics.

[7] Hercules Dalianis a Eduard Hovy. *On lexical aggregation and ordering*. http://www.aclweb.org/anthology/W96-0508. Eighth International Natural Language Generation Workshop (Posters and Demonstrations). . 1996.

[8] Roger Evans, Paul Piwek, a Lynne Cahill. *What is nlg?*. 144--151. http://www.aclweb.org/anthology/W02-2119. Proceedings of the International Natural Language Generation Conference. . 2002. Association for Computational Linguistics.

[9] Albert Gatt a Emiel Krahmer. *Survey of the state of the art in natural language generation: Core tasks, applications and evaluation*. 65--170. *Journal of Artificial Intelligence Research*. 61. 2018.

[10] Albert Gatt a Ehud Reiter. *Simplenlg: A realisation engine for practical applications*. 90--93. http://www.aclweb.org/anthology/W09-0613. Proceedings of the 12th

European Workshop on Natural Language Generation (ENLG 2009). . 2009. Association for Computational Linguistics.

[11] Yuqing Guo, Josef van Genabith, a Haifeng Wang. *Dependency-based n-gram models for general purpose sentence realisation*. 297--304. http://aclweb.org/anthology/C08-1038. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). . 2008. Coling 2008 Organizing Committee.

[12] Sepp Hochreiter a Jürgen Schmidhuber. *Long short-term memory*. 1735--1780. *Neural computation*. 9. 8. 1997.

[13] Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, a Tomoya Iwakura. *Text simplification for reading assistance: A project note*. http://aclweb.org/anthology/W03-1602. Proceedings of the Second International Workshop on Paraphrasing. . 2003.

[14] Irene Langkilde-Geary. *An empirical verification of coverage and correctness for a general-purpose sentence generator*. 17--24. http://aclweb.org/anthology/W02-2103. Proceedings of the International Natural Language Generation Conference. . 2002. Association for Computational Linguistics.

[15] Piji Li, Wai Lam, Lidong Bing, a Zihao Wang. *Deep recurrent generative decoder for abstractive text summarization*. 2091--2100. http://aclweb.org/anthology/D17-1222. 10.18653/v1/D17-1222. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. . 2017. Association for Computational Linguistics.

[16] Hans Peter Luhn. *The automatic creation of literature abstracts*. 159--165. *IBM Journal of research and development*. 2. 2. 1958.

[17] Gerasimos Lampouras a Andreas Vlachos. *Imitation learning for language generation from unaligned data*. 1101--1112. http://aclweb.org/anthology/C16-1105. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. . 2016. The COLING 2016 Organizing Committee.

[18] Jonathan Mallinson, Rico Sennrich, a Mirella Lapata. *Paraphrasing revisited with neural machine translation*. 881--893. http://aclweb.org/anthology/E17-1083. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. . 2017. Association for Computational Linguistics.

[19] Ani Nenkova, Kathleen McKeown, a others. *Automatic summarization*. 103--233. *Foundations and Trends in Information Retrieval*. 5. 2--3. 2011.

[20] Sergiu Nisioi, Sanja Stajner, Simone Paolo Ponzetto, a Liviu P. Dinu. *Exploring neural text simplification models*. 85--91. http://aclweb.org/anthology/P17-2014. 10.18653/v1/P17-2014. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). . 2017. Association for Computational Linguistics.

[21] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, a Bing Xiang. *Abstractive text summarization using sequence-to-sequence rnns and beyond*. 280--

290. http://aclweb.org/anthology/K16-1028. 10.18653/v1/K16-1028. Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. . 2016. Association for Computational Linguistics.

[22] Razvan Pascanu, Tomas Mikolov, a Yoshua Bengio. *On the difficulty of training recurrent neural networks*. 1310--1318. http://proceedings.mlr.press/v28/pascanu13.html. Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA. . Sanjoy Dasgupta a David McAllester. Proceedings of Machine Learning Research. 17--19 Jun 2013. PMLR.

[23] Rivindu Perera a Parma Nand. *Recent advances in natural language generation: A survey and classification of the empirical literature*. 1--32. *Computing and Informatics*. 36. 1. 2017.

[24] Ehud Reiter a Robert Dale. *Building applied natural language generation systems*. 57--87. *Natural Language Engineering*. 3. 1. 1997.

[25] Ehud Reiter a Robert Dale. *Building natural language generation systems*. Cambridge university press. 2000.

[26] Ehud Reiter. *Has a consensus nl generation architecture appeared, and is it psycholinguistically plausible?*. 163--170. Proceedings of the Seventh International Workshop on Natural Language Generation. Association for Computational Linguistics. . 1994.

[27] Advaith Siddharthan. *A survey of research on text simplification*. 259--298. *ITL-International Journal of Applied Linguistics*. 165. 2. 2014.

[28] Ilya Sutskever, Oriol Vinyals, a Quoc V Le. *Sequence to sequence learning with neural networks*. *Advances in Neural Information Processing Systems 27*. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, a K. Q. Weinberger. Curran Associates, Inc.. 2014. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

[29] Chris van der Lee, Emiel Krahmer, a Sander Wubben. *Pass: A dutch data-to-text system for soccer, targeted towards specific audiences*. 95--104. http://www.aclweb.org/anthology/W17-3513. Proceedings of the 10th International Conference on Natural Language Generation. Santiago de Compostela, Spain. . September 2017. Association for Computational Linguistics.

[30] Xianchao Wu, Ander Martinez, a Momo Klyen. *Dialog generation using multi-turn reasoning neural networks*. 2049--2059. http://aclweb.org/anthology/N18-1186. 10.18653/v1/N18-1186. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). . 2018. Association for Computational Linguistics.

Zoznam