VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

NEURONOVÝ STROJOVÝ PŘEKLAD PRO JAZYKOVÉ PÁRY S MALÝM MNOŽSTVÍM TRÉNOVACÍCH DAT

LOW-RESOURCE NEURAL MACHINE TRANSLATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **DENIS FILO**
AUTHOR

VEDOUCÍ PRÁCE **Ing. JOSEF JON**
SUPERVISOR

BRNO 2020

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2019/2020

Zadání bakalářské práce

Student: **Filo Denis**
Program: Informační technologie
Název: **Neuronový strojový překlad pro jazykové páry s malým množstvím trénovacích dat**
Low-Resource Neural Machine Translation
Kategorie: Zpracování řeči a přirozeného jazyka

Zadání:

1. Nastudujte literaturu týkající se strojového překladu pro takové jazykové páry, pro které existuje pouze málo nebo žádná paralelní data.
2. Připravte trénovací a evaluační datasety.
3. S pomocí zvoleného frameworku natrénujte překladové modely.
4. Proveďte experimenty s nastudovanými technikami .
5. Vyhodnoťte získané výsledky, navrhněte vylepšení zkoumaných technik.

Literatura:

- P. Koehn, Neural Machine Translation. 2017
- R. Sennrich, B. Zhang: Revisiting Low-Resource Neural Machine Translation: A Case Study. 2019.
- M. Przystupa, M. Abdul-Mageed: Neural Machine Translation of Low-Resource and Similar Languages with Backtranslation, ACL 2019, pp.224-235
- T.Kocmi: Exploring Benefits of Transfer Learning in Neural Machine Translation, MFF UK. Disertační práce. Praha 2019.
- dále dle doporučení vedoucího

Pro udělení zápočtu za první semestr je požadováno:

- Body 1, 2 a 3.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Jon Josef, Ing.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2019
Datum odevzdání: 31. července 2020
Datum schválení: 13. listopadu 2019

Abstrakt

Táto práca sa zaoberá neurónovým strojovým prekladom pre tzv. low-resource jazyky. Cieľom bolo pomocou experimentov vyhodnotiť súčasné techniky a navrhnúť ich vylepšenia. Prekladové systémy v tejto práci využívali architektúru neurónových sietí transformer a boli natrénované pomocou frameworku Marian. Vybranými jazykovými pármi pre experimenty boli slovenčina s chorvátčinou a slovenčina so srbčinou. V experimentoch boli predmetom skúmania techniky transfer learning a semi-supervised learning.

Abstract

This thesis deals with neural machine translation (NMT) for low-resource languages. The goal was to evaluate current techniques by using the experiments and suggest their improvements. The translation systems in this thesis used the neural network transformer architecture and were trained by the Marian framework. The selected language pairs were Slovak with Croatian and Slovak with Serbian. The subjects of the experiments were the transfer learning techniques and semi-supervised learning.

Kľúčové slová

neurónový strojový preklad, transformer, low-resource, transfer learning, semi-supervised learning, slovanské jazyky, slovenčina, chorvátčina

Keywords

neural machine translation, transformer, low-resource, transfer learning, semi-supervised learning, slavic languages, slovak, croatian

Citácia

FILO, Denis. *Neuronový strojový překlad pro jazykové páry s malým množstvím trénovacích dat.* Brno, 2020. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Josef Jon

Neuronový strojový překlad pro jazykové páry s malým množstvím trénovacích dat

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Josefa Jona. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....
Denis Filo
30. júla 2020

Poďakovanie

Moje poďakovanie patrí pánovi Ing. Josefovi Jonovi za cenné rady, ochotu a pomoc pri vypracovaní bakalárskej práce.

Obsah

Kapitola 1

Úvod

Ľudia v dnešnom svete sú v styku s technológiami na každodennom poriadku. Okrem iného ich využívajú aj na získavanie informácií a komunikáciu medzi sebou. Problém nastáva vtedy, ak sú texty v jazyku, ktorému dotyčný človek nerozumie. Z tohto dôvodu sa vynakladá veľké úsilie pri vyvíjaní strojového prekladu.

Vďaka intenzívnemu rozvoju strojového učenia sa za posledné roky spravil v preklade pomocou počítača veľký pokrok. Pri prekladaní medzi svetovými jazykmi je častokrát kvalita prekladu na vysokej úrovni. Ak sú prekladané vety jednoduché bez zložitej terminológie, tak výsledok sa dokonca môže približovať prekladu vytvoreným človekom. Je to vďaka veľkému množstvu dostupných viet, ktoré sú preložené do obidvoch jazykov, pomocou ktorých sa prekladový systém naučí prekladať.

Ak sa ale prekladá medzi jazykmi, ktoré nemajú k dispozícii na učenie taký objem dát, tak výsledné preklady strácajú na kvalite.

Cieľom bakalárskej práce je pokúsiť sa pomocou experimentov pri učení prekladových systémov dosiahnuť kvalitnejšie preklady pre takéto jazykové páry. Na základe výsledkov porovnať jednotlivé techniky trénovania a zistiť, ktoré faktory pozitívne alebo negatívne ovplyvňujú kvalitu vzniknutých prekladačov. Pomocou týchto pozorovaní následne navrhnúť vylepšenia pre otestované techniky učenia.

V nasledujúcej kapitole 2 je popísaný postupný vývin strojového prekladu, spoločne s opisom súčasných techník, ktoré riešia túto problematiku. V kapitole 3 sú popísané dáta, ktoré boli použité či už na trénovanie, validáciu alebo testovanie výslednej kvality prekladov. Kapitola 4 zahŕňa prípravu na samotné experimenty, t.j. voľba jazykov, nástrojov na trénovanie a spracovanie dát ako aj použitých techník. Kapitola 5 obsahuje popis priebehu jednotlivých experimentov a ich výsledky.

Kapitola 2

Low-resource neurónový strojový preklad

2.1 Strojový preklad

Jeden z prvotných cieľov v oblasti informačných technológií bol automatický preklad ľubovoľného textu z jedného jazyka do druhého. Strojový preklad je ale veľmi náročná úloha najmä kvôli variabilite a komplexnosti ľudského jazyka.[7]

Výsledkom vývoja v posledných desaťročiach, do ktorého sa zapájali ľudia z komerčnej aj akademickej sféry, boli tieto metódy[7]:

- Pravidlový strojový preklad,
- Štatistický strojový preklad,
- Neurónový strojový preklad.

2.1.1 Pravidlový strojový preklad

Najstaršia metóda strojového prekladu, označovaná ako RBMT (anglicky *Rule-Based Machine Translation*). Princípom tohto systému je vytvorenie jazykových pravidiel zdrojového a cieľového jazyka a slovníka medzi týmito jazykmi. Pravidlá sú často vyvíjané lingvistami a zahŕňajú sémantickú, syntaktickú a morfologickú stránku jazyka. Text sa preloží na základe informácií získaných zo slovníka a gramatických pravidiel.[7] V súčasnosti sa používa pre preklad *low-resource* jazykov napríklad *open-source* platforma *Apertium*[1][8].

2.1.2 Štatistický strojový preklad

Nástupom výkonnejších počítačov sa mohol začať naplno vyvíjať štatistický strojový preklad (SMT, anglicky *Statistical Machine Translation*). Tento systém potrebuje veľké množstvo paralelných dát - rôzne príklady textov v zdrojovom jazyku, ktoré sú preložené ľuďmi do cieľového jazyka. Na základe týchto dát si prekladový systém vytvorí slovník obsahujúci pravdepodobnosti jednotlivých prekladov. Existujú viaceré prístupy SMT, a to:

- Preklad založený na slovách,
- Preklad založený na syntaxi,

[1] https://apertium.org

- Frázový preklad (PBMT - anglicky *Phrase-Based Machine Translation*),
- Hierarchický frázový preklad.[6]

SMT bol desaťročia dominantnou prekladovou paradigmou. Najpoužívanejšie praktické implementácie SMT sú systémy PBMT, ktoré prekladajú na základe fráz – sekvencii slov.[39]

Pred príchodom priameho neurónového strojového prekladu boli neurónové siete ešte ako súčasť SMT systémov iba čiastočne úspešné. Išlo o zapojenie neurónového jazykového modelu namiesto n-gramového modelu. Avšak podstatou tohto prístupu je stále preklad založený na frázach, a preto zdedil aj nedostatky PBMT systémov. Boli vykonané aj ďalšie experimenty navrhovaných prístupov pre učenie frázových reprezentácií alebo *end-to-end learning* prekladov s neurónovými sieťami. V konečnom dôsledku priniesli však horšiu celkovú presnosť prekladov v porovnaní so štandardnými systémami PBMT.[39]

2.1.3 Neurónový strojový preklad

Pokusy v minulosti o koncept *end-to-end learning* strojového prekladu mali iba limitovaný úspech. Vďaka veľkému úsiliu zainteresovaných ľudí v tejto oblasti sa kvalita prekladu postupne blížila úrovni systémov PBMT. Pravdepodobne prvý úspešný pokus o prekonanie PBMT bolo dosiahnutie zlepšenia o 0,5 BLEU skóre (algoritmus na ohodnotenie kvality strojovo preloženého textu) v porovnaní s najmodernejším *phrase-based* systémom. Tento pokrok bol opísaný v práci *Neural Machine Translation of Rare Words with Subword Units* od Rica Sennricha a ostatných[31].[39] *Neural machine translation by jointly learning to align and translate*[3] a *Sequence to Sequence Learning with Neural Networks*[34] patria medzi ďalšie práce, v ktorých bolo rovnako ukázané dosiahnutie porovnateľných výsledkov neurónového strojového prekladu v porovnaní s PBMT.

Od tej doby bolo navrhnutých veľa nových techník pre ďalšie zlepšenie neurónového strojového prekladu. Presnosť prekladu týchto systémov bola sľubná, ale zatiaľ chýbali systematické porovnania s rozsiahlymi systémami PBMT.[39]

Neurónový strojový preklad (NMT, anglicky *Neural Machine Translation*) je treťou generáciou prístupov k strojovému prekladu. Po kombinovaných systémoch SMT a neurónových sietí nastalo úsilie zamerané výhradne na NMT. Prvé kroky zahŕňali použitie konvolúčnych a rekurentných sietí. Tieto modely dokázali produkovať kvalitné preklady pre krátke vety, ale s rastúcou dĺžkou vety boli menej presné. Pridanie mechanizmu pozornosti (anglicky *attention*) prinieslo konkurencieschopné výsledky. S niekoľkými ďalšími vylepšeniami sa NMT stal novou modernou v strojovom preklade.[20]

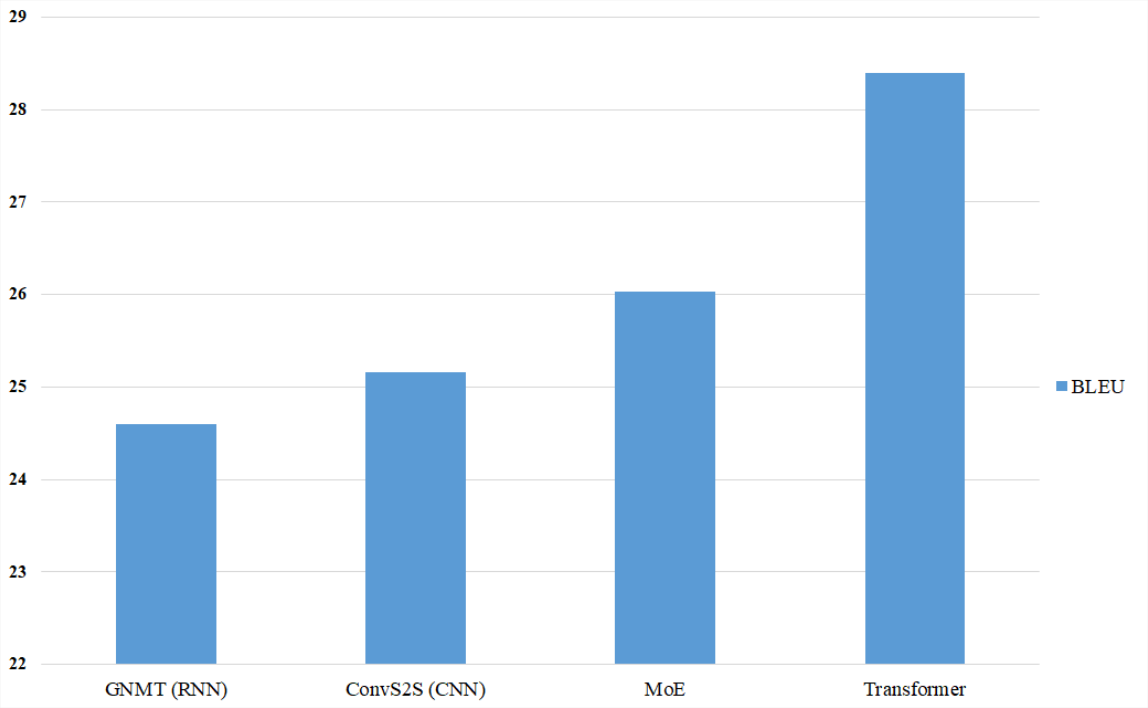
V priebehu jedného alebo dvoch rokov bola celá oblasť výskumu strojového prekladu zameraná na neurónové siete. Pre lepšiu predstavu o rýchlosti prechodu: Na zdieľanej úlohe, ktorú organizovala Konferencia o strojovom preklade (WMT – *Conference on Machine Translation*[1]), bol v roku 2015 predložený iba jeden NMT systém. Bol síce konkurencieschopný, ale prekonali ho tradičné SMT systémy. O rok neskôr systém NMT vyhral porovnávania takmer v každom jazykovom páre. V roku 2017 boli takmer všetci zúčastnené systémy založené na NMT.[20]

NMT podobne ako SMT potrebuje na trénovanie modelov korpusy – paralelné dáta zo zdrojového jazyka do cieľového jazyka. Výhodou tohto prístupu je, že stačí natrénovať iba jeden model a nie je potrebné mať hneď niekoľko prepojených špecializovaných modelov, ako to bolo pri SMT.[7]

Štandardné NMT systémy sa skladajú z kodéru, dekodéru a mechanizmu pozornosti. Existuje viacero rôznych prístupov založených na princípe kodér-*attention*-dekodér. V nasledujúcej sekcii bude popísaná architektúra neurónových sietí nazývaná transformer, ktorá je v NMT v súčasnosti najpoužívanejšia. Kodér je určený k vytvoreniu vektorov zo zdrojovej vety a ďalej dekodér používa túto reprezentáciu vety na predikciu symbolov v cieľovom jazyku. Mechanizmus pozornosti zlepšuje proces generovania prekladu určením, na ktorú časť zdrojovej vety je treba dať väčší dôraz.[24]

2.1.4 Transformer

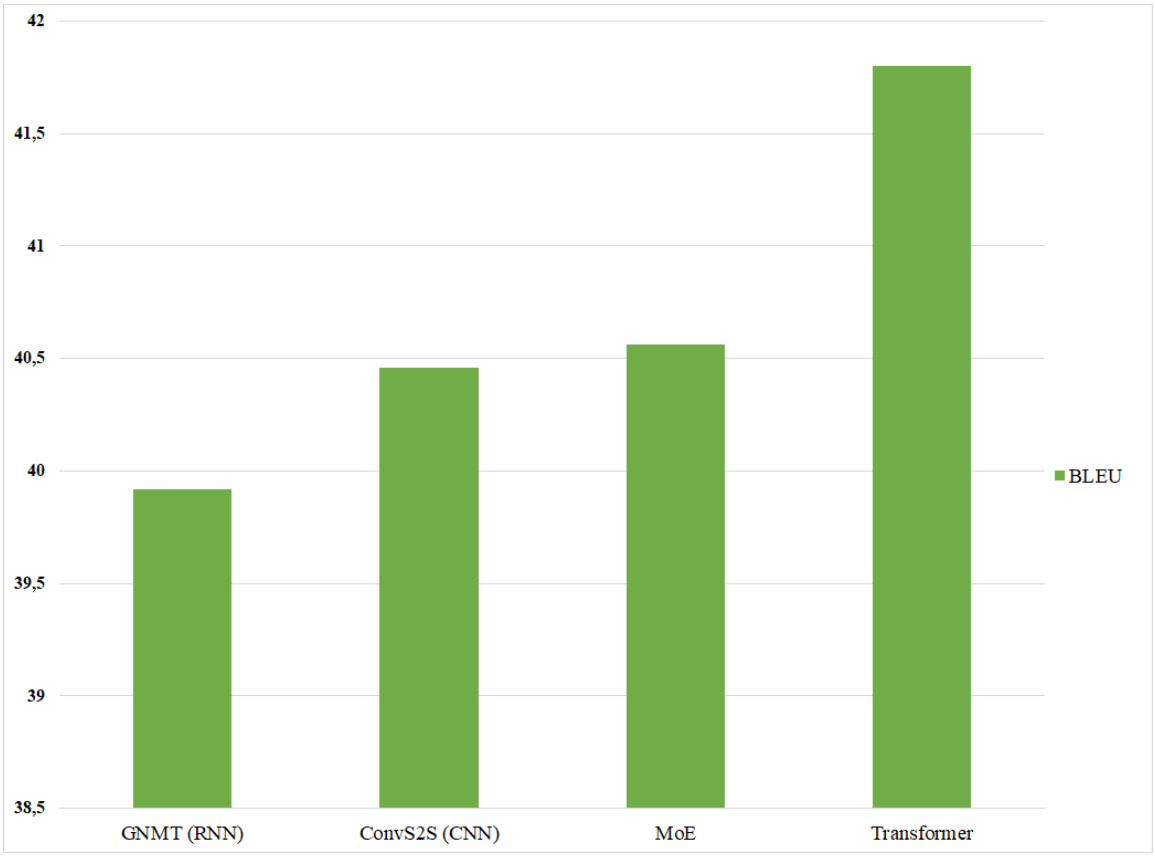
Google Brain a ich spolupracovníci publikovali článok[37], v ktorom predstavili novú architektúru, transformer, založenú iba na mechanizme pozornosti. Tento typ NMT prekonáva vo väčšine prípadov všetky ostatné NMT modely vrátane konvolúčnych a rekurentných sietí, ako aj PBMT systémy[11]. Dosahoval kvalitnejšie výsledky a trénovanie trvalo kratšiu dobu, vďaka možnosti lepšej paralelizácie[4].



Obr. 2.1: **Kvalita prekladu z angličtiny do nemčiny:** BLEU skóre (vyššie je lepšie) prekladových modelov na štandardných dátach *WMT newtest2014* z angličtiny do nemčiny[37].

Attention

Myšlienkou mechanizmu pozornosti je pozrieť sa na sekvenciu slov zo vstupu a rozhodnúť, ktoré slová sú dôležité pre generovanie konkrétneho výstupného slova. Pred transformerom sa používala *attention* iba medzi dekodérom a kodérom – pri spracovaní výstupu kodéra dekodérom nemali všetky pozície rovnakú váhu, ale určovala sa pomocou *attention* dynamicky pre každé generované slovo. V transformeri sa okrem tejto dekodér-kodér *attention* používa aj tzv. *self-attention* priamo pri vytváraní reprezentácií v kodéri a tiež v dekodéri



Obr. 2.2: **Kvalita prekladu z angličtiny do francúzštiny:** BLEU skóre (vyššie je lepšie) prekladových modelov na štandardných dátach *WMT newtest2014* z angličtiny do francúzštiny[37].

nad pozíciami už vygenerovaných slov. To pomáha modelu pochopiť slovo, ktoré aktuálne spracováva vďaka prepojeniu na ostatné relevantné slová vo vete, ktoré spoločne tvoria kontext. Napríklad pri spracovaní vety „Potom čo odišiel otec preč, bolo celej rodine za ním smutno.“, môže byť pri preklade potrebné zistiť, na ktoré slovo sa zámeno „ním“ odkazuje. *Self-attention* rieši tento problém zahrnutím tejto informácie do reprezentácie slova „ním“ v hlbších vrstvách kodéru.[37]

Funkcia *Attention* v transformeri používa tri vektory: dopyty (Q – *queries*), kľúče (K – *keys*), hodnoty (V – *values*). Výstupom je vážený súčet hodnôt, kde váhy sa počítajú z *queries* a kľúčov.

Funkcia *Attention* je definovaná nasledovne:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

kde d_k je odmocnina rozmerov vektorov kľúčov. Táto normalizácia pomáha stabilizovať gradienty pri trénovaní. Aplikovaním funkcie *softmax* získame distribúciu nad celou sekvenciou použitím skalárneho súčinu *queries* (ktoré reprezentujú skrytý stav všetkých pozícií v sekvencii) a kľúčov. Z tejto distribúcie sú získavané váhy jednotlivých hodnôt (ktoré tiež zaznamenávajú skrytý stav, podobne ako *queries*). To vedie k výslednému vektoru, kde sú zdôraznené relevantné slová alebo ich vlastnosti.[37]

Attention je použitá separátne v kodéri a dekodéri ako *self-attention*, kde všetky *queries*, kľúče a hodnoty pochádzajú z predchádzajúcej vrstvy. Taktiež sa používa kodér-dekodér *attention*, kde kľúče a hodnoty pochádzajú z kodéra a *queries* z dekodéra.[37]

Multi-Head Attention

Ak by NMT používal iba jednu *attention*, sústredil by sa výlučne iba na niektoré pozície z predchádzajúcej vrstvy, pričom by sa ignorovali ostatné relevantné slová. Pri modeli transformer je toto riešiené použitím niekoľkých hláv (anglicky *heads*) v každej vrstve, každá s vlastnou lineárnou transformáciou, čo vedie k súbežnému sledovaniu rôznych častí vstupu.[37]

Funkcia *Multi-Head Attention* je definovaná nasledovne:

$$MultiHead\left(Q,K,V\right) = Concatenate\left(head_1,...,head_h\right)W^O \tag{2.2}$$

$$\text{kde } head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

kde projekčné matice $W^{Q/K/V}$ sú trénovateľné matice odlišné pre každú *attention head* a h je počet hláv. Konkatenácia (zreťazenie) v *multi-head attention* sa potom lineárne premieta pomocou matice W^O.[37]

Positional Encoding

Keďže táto architektúra nepoužíva rekurentné alebo konvolúčne siete, je potrebné si nejakým spôsobom uchovať informáciu o pozícii každého slova. To je zabezpečené pridaním tzv. pozičného zakódovania (*positional encoding*) na všetky vstupné slová, čo umožňuje NMT identifikovať poradie slov.[37]

Absolútne pozičné zakódovanie slova *pos* je definované nasledovne:

$$PE_{(pos,2i)} = sin\left(pos/10000^{2i/d_{model}}\right) \tag{2.3}$$

$$PE_{(pos,2i+1)} = cos\left(pos/10000^{2i/d_{model}}\right)$$

kde i je rozmer (dimenzia) v pozičnom zakódovaní PE – každý rozmer PE zodpovedá sinusoide. Pozičné zakódovanie je pridané do *word embeddings* a je použité ako vstup pre prvú vrstvu transformera.[37]

Zhrnutie architektúry transfomer

Architektúra transformera sa spolieha na mechanizmus *self-attention*, ktorý odstraňuje všetky rekurentné operácie z predošlého prístupu. Kvôli absencii rekurentnosti bol pridaný tzv. *positional-encoding* na zachovanie si informácie o poradí slov. Vďaka paralelnému spracovaniu symbolov zo vstupu sa urýchlilo spracovanie vstupu a učenie. Rozšírením tradičného mechanizmu pozornosti sa namiesto jedného výpočtu pozornosti počíta hneď niekoľkokrát (*multihead attention*), čo rieši problém koreferencie[2] ako aj iné problémy[4]. Podrobnejší opis architektúry transformer je v originálnej práci *Attention Is All You Need*[37].

[2] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/ch08.html

2.2 NMT v low-resource

Rýchly vývoj systémov NMT viedol až k tvrdeniam, že preklady jazykových párov s veľkým množstvom trénovacích dát (anglicky *high-resource*), ako napríklad angličtina-čínština[12] alebo angličtina-čeština[5], sa blížia pri určitých podmienkach úrovni ľudského prekladu (anglicky *human parity*). Avšak na dosiahnutie takejto úrovne systémy NMT potrebujú dostatočné množstvo a kvalitu dát.

Presná definícia jazykových párov, ktoré patria do kategórie *low-resource*, neexistuje. Je potrebné zvážiť všetky aspekty dostupných jazykových zdrojov ako aj jazyk samotný. Jeden z aspektov je doména paralelných korpusov. Pod doménou sa rozumie zdroj dát - napríklad titulky, literatúra, spravodajstvo, lekárske správy, IT, dokumentácie a mnoho ďalších. Každý z nich používa iný slovník, štýl písania a má odlišný obsah. Ďalej možno rozlišovať formalitu a sentiment textu. Problém je ten, že tá istá informácia, môže byť napísaná pozitívnou, neutrálnou a negatívnou formou[26]. Všetky tieto faktory ovplyvňujú kvalitu korpusov. Ak špecifická doména obsahuje veľké množstvo paralelných viet, táto doména sa považuje za *high-resource*. Systém, ktorý je trénovaný na všeobecných dátach a prekladá vety zo špecifickej domény, nemusí dosahovať dostatočne kvalitné preklady. Napríklad spoločný zdroj paralelných viet pre *low-resource* jazyky je Biblia, ktorá je preložená do viac ako sto jazykov a obsahuje veľmi špecifický text[9].[17]

Vysoko ohybné jazyky (napríklad čeština, slovenčina) komplikujú definíciu *low-resource* jazykov, pretože prinášajú problém veľkej variabilnosti tvaru slov, a preto si vyžadujú viac paralelných viet v porovnaní s menej ohybnými jazykmi na dosiahnutie porovnateľných výsledkov[10].

Treba ešte poznamenať, že jazykový pár, ktorý je dnes považovaný za *low-resource*, nemusí byť v budúcnosti braný ako pár s malým množstvom trénovacích dát. Buď kvôli novým dostupným dátam alebo vďaka vylepšeniam trénovacích techník NMT.[17]

Existuje niekoľko techník NMT na zlepšenie výsledkov pre *low-resource* jazykové páry:

- Transfer learning,
- Unsupervised learning,
- Semi-supervised learning,
- Viacjazyčný model (anglicky *Multilingual model*) – jeden zdieľaný model pre viacero jazykových smerov: *one-to-many*, *many-to-many* a *many-to-one*[35],
- Preklad s pivotom (anglicky *Pivot-based translation*) – využitie pomocného jazyka (pivot) na medzipreklad, napríklad pre preklad z angličtiny (en) do slovenčiny (sk) sa použije ako pivot čeština (cs) a preklad bude prebiehať nasledovne: en→cs→sk[16],
- a iné.

2.2.1 Transfer learning

Hlavná myšlienka *transfer learning*-u je odovzdanie naučených znalostí z jedného modelu na druhý. Táto technika využíva znalosti z naučených úloh na zlepšenie výkonu na súvisiacej úlohe a obvykle slúži k zníženiu množstva požadovaných trénovacích dát. Najskôr sa predtrénuje rodičovský model (anglicky *parent model*) a pomocou neho sa natrénuje model pre *low-resource* jazykový pár, ktorý sa označuje ako potomkovský model (anglicky *child model*). Keď sa použije už predtrénovaný model na inicializáciu nového potomkovského

modelu, tak model nezačne trénovanie s náhodnými váhami, ale s váhami od rodičovského modelu.[17][18]

Torrey a Shavlik vo svojej práci[36] popísali, ako môže *transfer learning* zlepšiť celkový výkon. Konkrétne:

- zlepšenie počiatočného výkonu na začiatku trénovania v porovnaní s náhodne inicializovaným modelom, keď sú ich úlohy rovnaké,
- skrátenie času potrebného na dosiahnutie maximálneho výkonu,
- zlepšenie úrovne finálneho výkonu v porovnaní s modelom bez *transfer learning*-u.

V sfére spracovania prirodzeného jazyka (NLP - anglicky *Natural Language Processing*) boli metódy *transfer learning*-u úspešne aplikované na rozpoznávanie reči, klasifikáciu dokumentov, analýzu sentimentu a mnoho ďalších[38]. Úspech použitia *transfer learning*-u však nie je vždy garantovaný. Napríklad ak prenáša učenie z málo súvisiacej úlohy, môže to obmedziť potenciálny finálny výkon na cieľovej úlohe[40].

Rodičovský aj potomkovský model môže použiť rozdielne jazykové páry. Pre rodiča, ktorý prekladá z jazyka XX do jazyka YY (XX→YY) existujú tieto 3 scenáre:

- Spoločný zdrojový jazyk - scenár, kde je zdrojový jazyk rovnaký pre rodiča aj potomka. Inak povedané, potomkovský model prekladá XX→AA.
- Spoločný cieľový jazyk - scenár, kde je cieľový jazyk rovnaký pre oba modely, a teda potomkovský model prekladá AA→YY.
- Žiadny spoločný jazyk - scenár so žiadnym zdieľaným jazykom, to znamená, že potomkovský model prekladá AA→BB.[17]

Ak chceme predtrénovať rodičovský model s iným jazykovým párom ako má potomok, je potrebné vyriešiť problém s nezhodou medzi rodičovským a potomkovským slovníkom. Existujú dva typy prístupov, ktoré riešia tento problém:

- *cold-start*[19],
- *warm-start*[18].

Cold-start

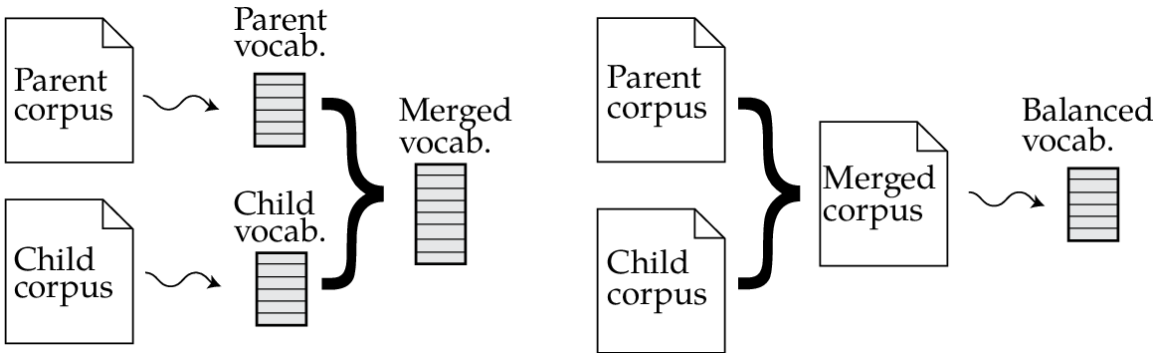
Cold-start prístupy používajú všeobecné rodičovské modely, ktoré nie sú počas trénovania nijak upravené podľa potomkovského jazykového páru. Ak rodičovský model používa slovník, ktorý sa z veľkej časti prekrýva s potomkovským slovníkom, tak môžeme ignorovať rozdiely a trénovať potomka pomocou rodičovského slovníka. Druhou možnosťou je transformovať rodičovský slovník pred trénovaním potomka rôznymi spôsobmi tak, aby vyhovovali potrebám zvoleného jazykového páru potomka.[17]

Všetky cold-start prístupy sa spoliehajú na schopnosť neurónových sietí rýchlo prispôsobiť rodičovské parametre novým podmienkam. Napríklad premapovanie rodičovských *subwords embeddings* na nesúvisiace potomkovské *subwords embeddings* (*subwords* v preklade podreťazce – bližšie popísané v podsekcii 4.3.1).[17]

Warm-start

Vo *warm-start* prístupoch sú trénovacie dáta potomka dostupné už v čase trénovania rodičovského modelu a vďaka tomu je možné vykonať kroky na prípravu rodiča na *transfer learning*. Nevýhody *cold-start* riešení sú také, že pri priamom použití rodičovského slovníka, môže byť v niektorých prípadoch nevyužitá veľká časť slovníka. Pri transformovaní slovníka je zas nutné znova natrénovať náhodne priradené potomkovské *embeddings*. *Warm-start* metódy dokážu tieto problémy vyriešiť pripravením rodičovského modelu v predstihu na nadchádzajúci *transfer learning* pre potomkovský jazykový pár.[17]

Základnou myšlienkou riešenia týchto problémov je priame použitie špecificky upraveného slovníka podľa potomka počas trénovania rodičovského modelu, ako je možné vidieť na priloženom obrázku 2.3. Použitie potomkovského slovníka pre rodičovský model oslabí jeho výkon, kvôli obmedzovaniu slovnej zásoby, avšak pri *transfer learning*-u neupriamujeme pozornosť na finálnu kvalitu rodičovského modelu, ale na výsledky potomkovského modelu.[17]



Obr. 2.3: **Proces vytvárania spoločných slovníkov:** Spojený slovník (naľavo) a Vyvážený slovník (napravo). Prevzaté z [17].

2.2.2 Unsupervised learning

Unsupervised learning v NMT využíva na trénovanie iba jednojazyčné (anglicky *monolingual*) korpusy, táto technika bola navrhnutá pre tie scenáre, kde zvolený jazykový pár má len veľmi málo alebo žiadne paralelné dáta. Aj keď bolo navrhnutých niekoľko prístupov, obvykle majú spoločné:

- trénovanie spoločného modelu pre oba smery, zdrojový jazyk→cieľový jazyk a cieľový jazyk→zdrojový jazyk,
- používanie iteratívnych spätných prekladov (anglicky *back-translations*),
- *denoising autoencoder*.[15]

Tieto metódy sú inicializované buď *cross-lingual word embeddings* alebo *cross-lingual* jazykovými modelmi[30]. Aplikovaním samostatného *denoising autoencoder*-u priamo na *word-by-word* (slovo po slove) preklady z *cross-lingual word embeddings* sa môžeme vyhnúť dlhému iteratívnemu trénovaniu[28]. Na zlepšenie výkonu, ale na úkor efektívnosti bola navrhnutá kombinácia *unsupervised* NMT s *unsupervised* PBMT[25].

Zdieľanie modelu medzi rozličnými prekladovými úlohami sa ukázalo pri preklade *low-resource* jazykových párov ako efektívne. Je to kvôli podobnosti prirodzených jazykov. Naučiť sa reprezentovať nejaký jazyk pomáha pri reprezentovaní iných jazykov, napríklad prenosom znalostí o všeobecných štruktúrach viet[2].

Na trénovanie obojsmerného modelu je potrebné mať bilingválne dáta, ktoré sa získavajú pomocou jednojazyčných textov. Metódy *unsupervised* NMT generujú syntetické dvojjazyčné dáta prostredníctvom spätných prekladov. Preložením jednojazyčných dát v zdrojovom a cieľovom jazyku vzniknú pseudo-paralelné dáta pre oba smery prekladu. Syntetické dáta by nemali byť vytvorené iba raz na začiatku, ale opakovane počas trénovania. V raných fázach trénovania je kvalita prekladov generovaných modelom nízka. Preto sa aktualizujú trénovacie dáta, vďaka čomu sa prekladový model zlepšuje. Zlepšený model pre smer zdrojový jazyk→cieľový jazyk spätne preloží jednojazyčné zdrojové dáta, čo zdokonalí model pre smer cieľový jazyk→zdrojový jazyk a naopak. Tento cyklus sa nazýva iteratívny spätný preklad, ktorý je znázornený na obrázku 2.4.[13]

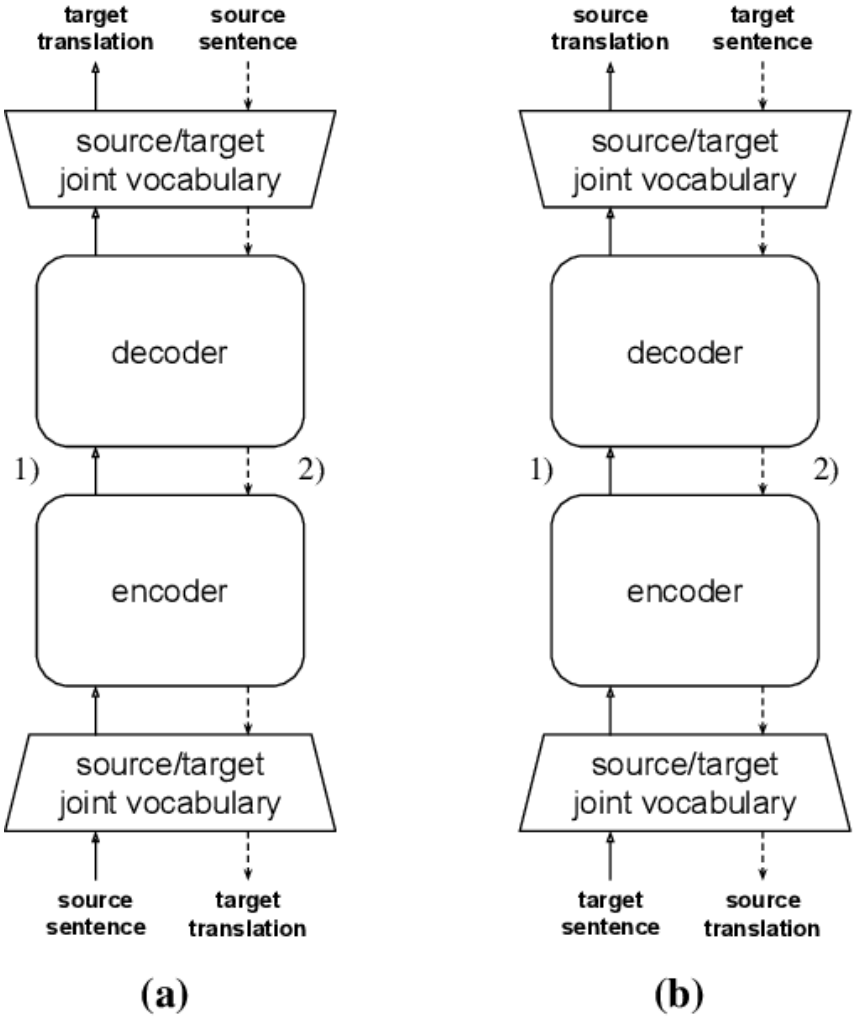
Počiatočný model má tendenciu generovať preklady s malým preskupením slov, ktoré nie sú veľmi presné a plynulé, ak je slovosled zdrojového a cieľového jazyka odlišný. Trénovanie na takýchto dátach bráni modelu meniť poradie slov, čo môže spôsobiť začarovaný kruh vytváraním ešte menej preskupených syntetických párov viet pri ďalších krokoch.[15]

Kvôli tomuto problému *unsupervised* NMT používa *denoising autoencoder*. Pôvodné vety sú umelo poškodené, napríklad vymazaním alebo permutáciou slov. *Denoising autoencoder* vezme poškodenú vstupnú vetu a trénuje model na preskupenie slov a vytvorenie originálnej vstupnej vety, čo pomáha pri generovaní plynulejších výstupov. Tento proces trénovania sa deje individuálne pre oba jazyky. Potom, ako je model naučený tieto poškodené vety zotavovať, je lepšie tento *denoising autoencoder* už ďalej nepoužívať alebo znížiť jeho váhu. V neskorších fázach trénovania sa model zlepšuje v preskupovaní slov a kvalite prekladu.[15]

Kim, Graça a Ney vo svojej práci ukázali, že aj jednoduchý *semi-supervised* alebo *supervised baseline* model s menej ako 50 tisíc paralelnými vetami dokáže prekonať ich najlepší *unsupervised* systém. Ak existuje aspoň malé množstvo bilingválnych dát, tak neodporúčajú použiť tento prístup.[15] Vzhľadom na to, že mojím zvoleným jazykovým párom je slovenčina→chorvátčina, kde existuje viac paralelných dát, ako je spomínaná hranica, som došiel k záveru, že táto metóda pre moju dvojicu nebude vhodná.

2.2.3 Semi-supervised learning

Metóda *semi-supervised learning* využíva paralelné korpusy spoločne aj s jednojazyčnými dátami. Idea tohto prístupu je taká, že najskôr sa natrénujú prekladové modely v oboch smeroch iba pomocou paralelných dát (*supervised learning*) a následne sa týmito modelmi vzájomne preložia jednojazyčné dáta, vďaka čomu vzniknú tzv. pseudo-paralelné dáta, ktoré sú pridané do trénovacej sady. Modely sú opätovne trénované na nových dátach spoločne aj s pôvodnými trénovacími dátami. Napríklad pre smer slovenčina→chorvátčina sa monolingválne dáta z chorvátčiny spätne preložia do slovenčiny, zo čoho vznikne dvojica viet, kde je preložená veta ako zdrojová veta a pôvodná prekladaná veta je cieľovou vetou. Ďalšou alternatívou je vytvoriť najskôr *unsupervised learning* model iba z jednojazyčných dát a následne vytvoriť trénovacie dáta zo pseudo-paralelných dát a dostupných paralelných dát. Väčšinou sa používa približne 50% paralelných a 50% pseudo-paralelných dát, ale úspech sa môže docieliť aj iným pomerom – či už väčším alebo aj menším množstvom monolingválnych dát. Napriek jednoduchosti princípu spätného prekladu pri *semi-supervised learning*-u sa táto technika ukázala vo veľa prípadoch ako efektívna.[32][15]



Obr. 2.4: **Iteratívny spätný preklad pre trénovanie obojsmerného *sequence-to-sequence* modelu:** Model najskôr preloží *monolingual* vety (plné šípky) a potom je trénovaný s preloženými vetami ako vstup a originálne vety sú použité ako referenčný preklad (prerušované šípky). Tento proces alternuje medzi **(a)** zdrojový jazyk→cieľový jazyk a **(b)** cieľový jazyk→zdrojový jazyk. Prevzaté z [15].

Kapitola 3

Dáta

Pre prekladový smer slovenčina→chorvátčina som použil paralelné korpusy JW300 v1, DGT v2019, QED v2.0a, bible-uedin v1, EUbookshop v2, TildeMODEL v2018 dostupné z OPUS[1], z ktorých som vytvoril trénovaciu sadu. Na vytvorenie validačného datasetu bolo použitých prvých 3000 viet z korpusu JW300 v1. Na testovanie pre všetky prekladové smery bola použitá testovacia sada TedTalks. Na detokenizáciu testovacej sady, ktorá bola už predom tokenizovaná, bol použitý detokenizér Moses[2]. Hodnoty v tabuľkách 3.1, 3.2, 3.3 a 3.4 sú zaokrúhlené. Trénovacie, validačné a testovacie sady sú navzájom disjunktné.

Korpus	**sk tokeny**	**hr tokeny**	**vety**
JW300 v1	15,2M	15,3M	0,9M
DGT v2019	13,8M	13,9M	0,7M
QED v2.0a	1,5M	1,5M	0,1M
bible-uedin v1	0,7M	0,7M	31,1k
EUbookshop v2	0,1M	0,1M	3,9k
TildeMODEL v2018	42,0k	0,1M	1,9k
Spolu	**31,3M**	**31,6M**	**1,7M**

Tabuľka 3.1: **Trénovacia sada slovenčina→chorvátčina:** počet jednotlivých tokenov a viet v použitých korpusoch, ktoré sa spojili do jednej trénovacej sady.

Korpus	**cs tokeny**	**hr tokeny**	**vety**
JW300 v1	16,4M	16,6M	1,0M
DGT v2019	13,9M	13,9M	0,7M
QED v2.0a	2,0M	2,0M	0,1M
bible-uedin v1	0,7M	0,7M	31,1k
EUbookshop v2	0,1M	0,1M	4,1k
TildeMODEL v2018	88,9k	42,0k	1,8k
Tatoeba v2020-05-31	99	110	27
Spolu	**33,2M**	**31,1M**	**1,8M**

Tabuľka 3.2: **Trénovacia sada čeština→chorvátčina:** počet jednotlivých tokenov a viet v použitých korpusoch, ktoré sa spojili do jednej trénovacej sady.

[1] http://opus.nlpl.eu/

[2] https://github.com/moses-smt/mosesdecoder

Pre prekladový smer čeština→chorvátčina, ktorý slúžil ako rodičovský model pri *transfer learning*-u som použil rovnaké typy korpusov, do trénovacej sady bol pridaný ešte paralelný korpus Tatoeba v2020-05-31. Validačná sada bola vytvorená rovnakým spôsobom – prvých 3000 viet z korpusu JW300 v1.

Trénovacie sady, kde bol zdrojový jazyk slovenčina, resp. čeština a cieľový jazyk srbčina bola vytvorená z korpusov QED v2.0a a bible-uedin v1. K dispozícií bola aj doména EUbookshop, ale rozhodol som sa ju nepoužiť, pretože obsahovala mnoho nekvalitných viet, v ktorých neboli slová oddelené, ale boli iba ako jeden celok. Ako validačná sada bolo tentokrát použitých prvých 2000 riadkov z korpusu QED v2.0a.

Korpus	**sk tokeny**	**sr tokeny**	**vety**
QED v2.0a	1,7M	1,7M	110,8k
bible-uedin v1	0,7M	0,7M	31,1k
Spolu	**2,4M**	**2,4M**	**141,9k**

Tabuľka 3.3: **Trénovacia sada slovenčina→srbčina:** počet jednotlivých tokenov a viet v použitých korpusoch, ktoré sa spojili do jednej trénovacej sady.

Korpus	**cs tokeny**	**sr tokeny**	**vety**
QED v2.0a	2,6M	2,7M	180,0k
bible-uedin v1	0,7M	0,7M	31,1k
Spolu	**3,3M**	**3,4M**	**211,1k**

Tabuľka 3.4: **Trénovacia sada čeština→srbčina:** počet jednotlivých tokenov a viet v použitých korpusoch, ktoré sa spojili do jednej trénovacej sady.

Tieto jazykové páry slúžili na demonštráciu najlepšie fungujúcich techník, ktoré boli najskôr otestované pre dvojicu slovenčina→chorvátčina.

Keďže niektoré srbské vety v spomenutých korpusoch ako aj v testovacej sade TedTalks boli napísané v cyrilike, bolo potrebné ich prekonvertovať na text v latinke. Na tento účel bol použitý modul *CyrTranslit*[3] v programovacom jazyku *python*.

[3] https://pypi.org/project/cyrtranslit/

Kapitola 4

Návrh

4.1 Slovenčina a chorvátčina

Na preskúmanie *low-resource* NMT som si vybral dvojicu slovenčina→chorvátčina, na ktorej môže byť odpozorované, ako kombinácia rôznych priaznivých a nepriaznivých vlastností jazykov ovplyvní celkovú kvalitu prekladu.

Oba jazyky patria do skupiny slovanských jazykov, presnejšie slovenčina, mnohými ľuďmi považovaná za slovanské „esperanto“[1], patriaca do západoslovanských jazykov a chorvátčina do južnoslovanských jazykov. Keďže sú zo spoločnej jazykovej skupiny, tento aspekt môže pozitívne ovplyvniť kvalitu prekladu, čo bolo opísané aj v práci *Neural Machine Translation of Low-Resource and Similar Languages with Backtranslation* od Michaela Przystupa a Muhammada Abdul-Mageeda[29]. Výhodou spoločnej jazykovej skupiny je to, že v daných jazykoch existujú slová, ktoré majú rovnaký koreňový základ slova. V prípade slovanských jazykov ide ale skôr o spoločný fonetický slovný základ.

Ak by som použil nejakú dvojicu z germánskych alebo románskych jazykov, ich výhodou by bolo to, že tieto skupiny jazykov majú tzv. viazaný slovosled. Viazaný slovosled má svoje členy pevne usporiadané na základe pravidiel daného jazyka a ich miesto sa až na výnimky (napr. citovo zafarbená reč) nemení. Slovanské jazyky majú oproti tomu pomerne voľný slovosled, čo je dané menej prísnymi pravidlami pre postavenie slov a slovných spojení vo vete, ako aj veľkým množstvom slovných tvarov vyskytujúcich sa aj na nezvyčajných miestach.[23]

Postavenie vetných členov v slovenskej a chorvátskej vete sa vo svojej podstate nelíši. Samozrejme v týchto jazykoch sú určité odlišnosti. Najviac rozdielov v slovoslede slovenčiny a chorvátčiny je v umiestňovaní kontextových slov, ktoré dotvárajú vetu naznačovaním kontextu, resp. okolností. Katarína Kösegiová, ktorá sa venovala porovnaniu slovosledu chorvátčiny a slovenčiny vyslovila túto myšlienku: „Pri prekladoch textov vznikajú často úskalia, ktoré je potrebné riešiť prihliadaním na úzus konkrétneho jazyka či dokonca intuíciu. Tá niekedy ako jediná napovie, ako text preložiť tak, aby v jazyku, do ktorého sa prekladá, vyznel prirodzene.“[23]. Z čoho vyplýva, že strojový preklad týchto jazykov môže byť celkom výzvou.[23]

Ďalším faktorom, ktorý treba zvážiť, je variabilita tvaroslovia slovanských jazykov. Ako príklad uvediem rozdiel medzi podstatnými menami v angličtine a chorvátčine. V chorvátskom jazyku pri podstatných menách rozlišujeme 7 pádov, 2 čísla a 3 rody, preto je nutné

[1] slovanské „esperanto“ - jazyk, ktorý sa vníma ako najzrozumiteľnejší aj pre používateľov iných slovanských jazykov. Podrobnejšie vysvetlené na https://slovake.eu/sk/intro/language/general.

brať do úvahy 42 potenciálnych tvarov paradigmy[2]. Na druhej strane pri angličtine môže mať podstatné meno iba 2 tvary (jednotné a množné číslo).

4.2 Framework

Na trénovanie modelov bol použitý voľne dostupný *framework* Marian[3] pre neurónový strojový preklad napísaný v programovacom jazyku *C++*, ktorý je z väčšiny vyvíjaný *Microsoft Translator*[4] tímom. Zvolenou architektúrou NMT bola architektúra transformer, opísaná v sekcii 2.1.4.[14]

4.3 Spracovanie dát

4.3.1 Subword

Z dôvodu obmedzenej pamäti je vhodné používať čo najmenší možný slovník. Značné množstvo slov je tvorené z niekoľkých častí, napríklad: predpony, prípony, spojené slová a podobne. Kvôli tomu sa namiesto slov ako najmenších jednotiek v slovníkoch používajú tzv. *subword units*, teda jednotiek menších ako slovo. Zároveň segmentácia na *subword*-y pomáha v prípade vzácnejších slov, ktoré nie sú zastúpené v korpuse v dostatočnom množstve, a tak neexistuje veľa príkladov k natrénovaniu kvalitného *embedding*-u celého slova. Rozdelením slov sa získa:

- menšia veľkosť slovníka – vďaka použitiu menších častí slov, z ktorých je možné skladať väčšie celky. Preto nie je potrebné, aby slovník obsahoval toľko slov,
- zredukovanie výskytu neznámych slov – slovník obsahuje menšie jednotky ako slovo, a tak je možné z nich poskladať všetky slová z trénovacích korpusov. Ak sa vyskytne neznáme slovo počas prekladu, je možné ho preložiť po menších častiach.[31]

4.3.2 SentencePiece

Na spracovanie dát a vytvorenie slovníkov bol použitý *SentencePiece*[5], ktorý je aj priamo zabudovaný vo *framework*-u Marian. *SentencePiece* je jazykovo nezávislý *subword* tokenizér a detokenizér určený aj pre NMT. Zatiaľ čo doposiaľ existujúce nástroje na *subword* segmentáciu predpokladajú, že vstup je predom tokenizovaný do sekvencie slov, tak *SentencePiece* dokáže natrénovať *subword* modely (slovníky) priamo z neupravených (anglicky *raw*) viet. Bolo ukázané, že je možné dosiahnuť porovnateľné presné výsledky aj s takýmto spôsobom trénovania.[22]

Implementuje 2 *subword* segmentačné algoritmy *byte-pair encoding* (BPE) a *unigram language model* s rozšírením o priame trénovanie z *raw* viet[22].

SentencePiece sa skladá zo štyroch hlavých častí: normalizér, tréner (*trainer*), kodér a dekodér. Normalizér je modul pre normalizáciu sémanticky ekvivalentných Unicode znakov. *Trainer* vytvára *subword* model (slovník) zo vstupných korpusov, pričom veľkosť slovníka je vopred špecifikovaná. Kodér tokenizuje vstupný text do sekvencie podreťazcov (*sub-*

[2] https://lingvo.info/sk/lingvopedia/croatian
[3] https://marian-nmt.github.io/
[4] https://translator.microsoft.com/
[5] https://github.com/google/sentencepiece

words) pomocou natrénovaného modelu. Dekodér konvertuje sekvenciu podreťazcov späť na text.[22]

SentencePiece používa pri vytváraní slovníkov tzv. id mapovanie, vďaka čomu je možné priamo konvertovať vstupný text na sekvenciu id a naopak[22].

Ukážka vytvorenia a použitia *SentencePiece* slovníka z korpusov v slovenskom a chorvátskom jazyku:

```
$ spm_train input=data/corpus.skhr model_prefix=skhr.spm vocab_size=32000

$ echo "Toto je ukážková veta." | spm_encode --model=skhr.spm.model
_Toto _je _ukáž ková _v eta .

$ echo "Toto je ukážková veta." | spm_encode --model=skhr.spm.model
--output_format=id
566 6 18380 10974 13 2092 3

$ echo "_Toto _je _ukáž ková _v eta ." | spm_decode --model=skhr.spm.model
Toto je ukážková veta.

$ echo "566 6 18380 10974 13 2092 3" | spm_decode --model=skhr.spm.model
--input_format=id
Toto je ukážková veta.
```

4.4 Použité techniky

V tejto práci bude vyskúšaných niekoľko techník, ktoré môžu potenciálne ovplyvniť finálny výkon prekladového modelu pre mnou vybraný *low-resource* jazykový pár, a to:

- vplyv veľkosti slovníka pri trénovaní základného *baseline* modelu,
- *transfer learning*:
  - *cold-start*,
  - *warm-start*,
- *semi-supervised learning*.

4.4.1 Baseline modely

V prvom experimente by som chcel odpozorovať, ako dokáže ovplyvniť veľkosť slovníka výslednú kvalitu prekladu *baseline* modelov pre moju špecifickú dvojicu jazykov. Jednotlivé veľkosti slovníkov, ktoré budú vyskúšané:

- 64000,
- 32000,
- 16000,
- 4000.

4.4.2 Transfer learning

Pri *transfer learning*-u budú odskúšané obidva typy prístupov, teda *cold-start* a *warm-start* prístupy, pričom oba typy budú následne porovnané.

Rodičovské modely pre *cold-start* prístupy:

- čeština→chorvátčina,
- angličtina→čeština,
- angličtina→slovanské jazyky (čeština, slovenčina, ruština, poľština, macedónčina, slovinčina, srbčina, chorvátčina, bulharčina).

Pri týchto rôznych rodičovských modeloch bude odpozorované, či je vhodnejšie použiť ako rodiča model, kde zdrojový aj cieľový jazyk patria do rovnakej jazykovej skupiny ako potomkovská dvojica (čeština→chorvátčina) alebo model, ktorý má k dispozícií väčšie množstvo trénovacích dát (angličtina→čeština) alebo viacjazyčný model, ktorý prekladá do viacerých slovanských jazykov vrátane chorvátčiny.

Rodičovský model, vďaka ktorému sa dosiahne najlepší výsledok spomedzi týchto troch modelov, bude použitý aj ako rodič pri *warm-start* prístupoch (presnejšie bude použitý rovnaký smer, pretože pri *warm-start* prístupe, sa rodičovský model pripravuje vopred na potomkovský model).

Vo *warm-start* prístupoch bude rodičovský model natrénovaný so spojeným aj s vyváženým slovníkom. Na kvalitu prekladu rodičovských modelov sa nebude prihliadať a porovnané budú navzájom kvality prekladov potomkovských modelov.

4.4.3 Semi-supervised learning

V ďalšom experimente bude odskúšaná jednoduchá variácia *semi-supervised learning*-u, ktorá bude spočívať v tom, že pomocou natrénovaného *baseline* (*supervised*) modelu pre opačný smer (s cieľovým jazykom slovenčina) sa spätne preložia monolingválne dáta, z ktorých vzniknú pseudo-paralelné dáta. Tieto dáta sa následne pridajú do trénovacej sady modelu so zdrojovým jazykom slovenčina a bude pokračovať v trénovaní na týchto rozšírených trénovacích dátach.

4.4.4 Najlepšie fungujúce prístupy

Zo všetkých vyskúšaných prístupov sa vyberú tie, ktoré dosiahnú najlepšie výsledky. Tieto prístupy sa porovnajú na podobnom prekladovom smere slovenčina→srbčina. Srbčina a chorvátčina sú takmer rovnaké jazyky, iba s malými odlišnosťami[21]. Množstvo dostupných dát pre dvojicu slovenčina – srbčina je však ešte menšie, takže môže byť odpozorované, ako budú tieto vybrané prístupy fungovať na menšom objeme trénovacích dát.

Kapitola 5

Experimenty a výsledky

V tejto kapitole sú popísané jednotlivé experimenty, v ktorých boli otestované trénovacie techniky spomenuté v kapitole 4. Všetky experimenty boli vykonané na serveroch KNOT (Výskumná skupina znalostných technológií na VUT FIT)[1] pomocou *framework*-u Marian. Trénovanie bolo inicializované s implicitnými parametrami pri použití modelu transformer-base, ktoré sú uvedené aj na stránkach Mariana[2].

Parametre trénovania	
–enc-depth 6	počet vrstiev kodéra
–dec-depth 6	počet vrstiev dekodéra
–transformer-heads 8	počet hláv
–transformer-dropout 0.1	dropout medzi vrstvami transformera
–learn-rate 0.0003 –lr-decay-inv-sqrt 16000	meria učenia

Tabuľka 5.1: Výber niektorých parametrov trénovania, ktoré boli použité pri všetkých experimentoch.

Na evaluáciu výsledkov modelov bola zvolená metrika BLEU skóre, ktorá patrí k najpoužívanejším hodnotiacim metrikám a bola použitá aj vo všetkých referenčných prácach. Algoritmus BLEU (*bilingual evaluation understudy*) je založený na porovnaní strojového prekladu a referenčného ľudského prekladu na úrovní n-gramov. Skóre je počítané na úrovni celej testovacej sady, pričom sa vety porovnávajú s ich referenčným prekladom (porovnávajú sa zhody v n-gramoch). Výsledkom je skóre 0–100%. Čím je hodnota skóre vyššia, tým je preklad lepší a viac sa približuje kvalite ľudského prekladu. Na počítanie BLEU skóre bol použitý nástroj *SacreBLEU*[3].[27]

5.1 1. experiment – veľkosť slovníka

V tomto experimente boli natrénované *baseline* modely s rôznymi veľkosťami slovníkov za účelom zistenia, aká veľkosť slovníka by bola pre dvojicu ohybných jazykov slovenčina – chorvátčina najlepšia. Trénovacia sada obsahuje približne 1,7 miliónov paralelných viet.

[1] https://www.fit.vut.cz/research/group/knot/.cs
[2] https://marian-nmt.github.io/
[3] https://github.com/mjpost/sacrebleu

Každé trénovanie modelov prebiehalo s rovnakými hyperparametrami. Testovací dataset TedTalks obsahoval 2166 párov viet.

Veľkosť slovníka	**BLEU skóre**
64000	23,2
32000	24,4
16000	**24,9**
4000	23,6

Tabuľka 5.2: Výsledky *baseline* modelov s rozličnými veľkosťami slovníkov testovaných na testovacej sade TedTalks.

Výsledky tohto experimentu ukázali, že pre tento scenár je najlepšie zo štyroch vyskúšaných možností zvoliť veľkosť slovníka 16000, kde model s týmto slovníkom dosiahol BLEU skóre 24,9. Ako druhá najlepšia možnosť sa ukázala veľkosť 32000 (BLEU skóre 24,4), pričom táto veľkosť je aj štandardne prednastavená pri trénovaní modelu na *framework*-u Marian pri použití vytvorenia slovníkov pomocou *SentencePiece*.

Model so slovníkom s veľkosťou 64000 bol natrénovaný aj za účelom lepšieho porovnávania potomkovského modelu pri *transfer learning*-u, kde bol ako rodičovský model použitý model angličtina→slovanské jazyky, ktorého veľkosť slovníka je taktiež 64000.

5.2 2. experiment – cold-start transfer learning 1

Experiment č.2 spočíval v tom, že bol otestovaný tzv. *cold-start transfer learning* na vytvorenie potomkovského modelu pre prekladový smer slovenčina→chorvátčina. Tento prístup spočíva v tom, že váhy potomkovského modelu nie sú inicializované náhodne, ale na začiatku trénovania sú nastavené rovnako ako váhy rodiča. Tento rodičovský model ale nie je nijak vopred pripravený a upravený pre potreby potomkovského modelu.

Ako rodičovské modely boli použité 3 modely, každý zastupujúci určitú kategóriu modelov: čeština→chorvátčina (spoločná jazyková skupina a zároveň cieľový jazyk), angličtina→čeština (veľké množstvo paralelných dát), angličtina→slovanské jazyky (viacjazyčný model, prekladajúci do rovnakej jazykovej skupiny vrátane chorvátčiny). Pri tomto experimente boli pri trénovaní potomkovských modelov vytvorené nové potomkovské slovníky.

Všetky potomkovské modely boli trénované s rovnakými štandardnými parametrami transformera, ktoré boli použité aj pri trénovaní *baseline* modelov.

Model	**Rodičovský model**	**Veľkosť slovníka**	**BLEU skóre**
SK→HR	čeština→HR	32000	24,3
SK→HR	angličtina→čeština	32000	24,1
SK→HR	angličtina→slovanské jazyky	64000	22,5
SK→HR *baseline*	-	32000	24,4
SK→HR *baseline*	-	64000	23,2

Tabuľka 5.3: Porovnanie kvality potomkovských modelov, ktoré boli natrénované technikou *cold-start transfer leraning*. Tabuľka obsahuje informáciu o použitom rodičovskom modeli, veľkosti slovníka (rodičovský aj potomkovský model majú rovnaké veľkosti slovníkov) a výslednom BLEU skóre potomkovských modelov. V dolnej časti tabuľky sú pre porovnanie pridané *baseline* modely, ktoré používajú rovnakú veľkosť slovníkov: 32000 resp. 64000.

Všetky potomkovské modely použili zhodnú veľkosť slovníka ako majú rodičovské modely.

Z výsledkov tohto experimentu vyplýva, že použitie prístupu *cold-start transfer learning* pri trénovaní potomkovských modelov s rovnakými parametrami ako majú *baseline* modely, nie je dostačujúce. To môže byť zapríčinené použitím nového potomkovského slovníka a už pomerne vysokým BLEU skóre *baseline* modelov. Keďže vo veľa prípadoch v extrémnych *low-resource* jazykových párach majú *baseline* modely značne nižšie BLEU skóre a rodičovské modely majú v takýchto scenároch omnoho viac trénovacích dát v pomere s potomkom. Preto vtedy aj táto technika stačí na zlepšenie kvality prekladu.

Na druhej strane z porovnávania vysvitlo, že ako najvhodnejší rodičovský model sa javí rodičovský model čeština→chorvátčina, ktorý zaostával od *baseline* modelu iba o 0,1 BLEU skóre. Tento prekladový smer rodiča bol teda zvolený aj pre ďalšie experimenty.

5.3 3. experiment – warm-start transfer learning 1

V poradí tretí experiment bol zameraný na trénovanie pomocou prístupov patriacich do kategórie *warm-start transfer learning*. Pomocou experimentu 5.2 bol ako prekladový smer rodiča zvolený čeština→chorvátčina. Na trénovanie bol použitý spojený (*merged*) ako aj vyvážený (*balanced*) slovník.

Na vytvorenie spojeného slovníka bol použitý rodičovský slovník, ktorý bol vygenerovaný z trénovacích dát čeština→chorvátčina a potomkovský slovník vytvorený z trénovacích dát slovenčina→chorvátčina. Oba slovníky mali veľkosť 32000.

Spojený slovník vznikol preusporiadaním potomkovského slovníka tak, že *subwords*, ktoré sa vyskytovali aj v rodičovskom slovníku sa dali na rovnakú pozíciu ako u rodiča (to znamená, že spojený slovník obsahoval duplikované *subwords* iba raz a mali zhodné *embeddings* ako v rodičovskom modeli). Ostatné *subwords* potomka sa pridali do slovníka na zvyšné miesta, pričom sa zachovalo ich poradie (frekvencia).

Pri tvorbe vyváženého slovníka konkatenáciou rodičovských a potomkovských trénovacích dát vznikol jeden spojený korpus. Ten slúžil na vygenerovanie tohto slovníka pomocou *SentencePiece*. Veľkosť slovníka bola taktiež 32000. Spoločný korpus bol zostavený približne v takom pomere, aby každý jazyk bol v ňom obsiahnutý rovnomerne – 25% čeština, 25% slovenčina a 25% + 25% chorvátčina, čo predstavuje ≈ 1,7 milióna paralelných viet rodiča aj potomka. Pri tvorení spojeného korpusu sa neprihliadalo na to, koľko slov obsahujú jednotlivé vety, čiže tento pomer nemusí zodpovedať počtu slov jednotlivých jazykov obsiahnutých v korpuse.

Účelom tohto korpusu bolo iba vytvorenie vyváženého slovníka a nebol použitý na žiadne trénovanie prekladových modelov.

Spojený slovník S:	{_budeme, _cez, dych, ej, _cha, _na, niny, _od, ovať, _pr, _rodinn, te, zd, ...}
Vyvážený slovník V:	{_budeme, dy, ez, iny, _na, nej, _od, ovať, _rodin, prázdn, te, ...}
Slovenská veta:	Cez prázdniny budeme oddychovať na rodinnej chate.
Segmentácia slovníkom S:	_Cez _pr á zd niny _budeme _od dych ovať _na _rodinn ej _cha te .
Segmentácia slovníkom V:	_C ez _ prázdn iny _budeme _od dy ch ovať _na _rodin nej _ ch a te .

Tabuľka 5.4: Ukážka rozdielov v slovnej zásobe medzi spojeným a vyváženým slovníkom a následnej segmentácii príkladovej vety.

Rozdiel v jednotlivých slovníkoch je znázornený v tabuľke 5.4, kde je možné vidieť, že modely v tomto experimente budú používať odlišné slovníky, a teda každý model bude mať svoje rozličné *embeddings*. Vďaka tomu sa získajú dva iné modely, ktoré môžu byť vzájomne porovnané.

V oboch prístupoch bol spojený, resp. vyvážený slovník použitý pri trénovaní rodičovského modelu ako aj potomkovského modelu. Pri trénovaní všetkých typoch modelov (rodičovské modely, potomkovské modely, *baseline* modely) boli opäť použité rovnaké implicitné parametre. Pri transformácii skonvergovaného rodičovského modelu sa zmenili trénovacie datasety na datasety potomkovského jazykového páru a trénovanie pokračovalo ďalej s nezmeneným slovníkom ani parametrami.

Model	**Slovník**	**BLEU skóre**
SK→HR	spojený, veľkosť 32000	24,8
SK→HR	vyvážený, veľkosť 32000	**25,0**
SK→HR *baseline*	potomkovský, veľkosť 32000	24,4

Tabuľka 5.5: Porovnanie kvality potomkovských modelov slovenčina→chorvátčina, ktoré boli natrénované technikami *warm-start transfer learning*. V tabuľke je informácia o použitom slovníku a výslednom BLEU skóre potomkovských modelov. Oba potomkovské modely mali ako rodiča prekladový smer čeština→chorvátčina. V spodnej časti tabuľky je pre porovnanie *baseline* model s rovnakou veľkosťou slovníka.

Trénovanie potomkovských modelov pomocou techník *warm-start transfer learning* v tomto experimente prinieslo očakávané zlepšenie oproti *baseline* modelu. Pri použití spojeného slovníka dosiahol potomkovský model o 0,4 vyššie BLEU skóre a potomkovský model s vyváženým slovníkom o 0,6 vyššie BLEU skóre v porovnaní s *baseline* modelom.

5.4 4. experiment – cold-start transfer learning 2

Účelom tohto experimentu bolo natrénovanie potomkovského modelu slovenčina→chorvátčina pomocou *transfer learning*-u s využitím spoločného slovníka. Rodičovský model prekladajúci z češtiny do chorvátčiny bol natrénovaný s vygenerovaným slovníkom z trénovacích dát rodiča. Následne bol vytvorený spojený slovník z rodičovského a potomkovského slovníka (tvorba popísaná v experimente 5.3). Preusporiadaním rovnakých *subwords* v spojenom slovníku tak, aby sa nachádzali na tom istom mieste ako v rodičovskom slovníku, sa docieli to, že budú mať rovnaké *embeddings* ako v rodičovskom modeli.

Keďže sa rodičovský model nijako vopred nepripravuje pre potreby potomkovského modelu, tento *transfer learning* patrí do kategórie *cold-start*.

Veľkosti slovníkov pri všetkých modeloch bola 32000. Všetky modely boli natrénované s rovnakými parametrami.

Tento model dosiahol BLEU skóre 24,6, čo je možné vidieť v tabuľke 5.6. Oproti *baseline* modelu sa potomkovský model, natrénovaný v tomto experimente zlepšil o 0,2 BLEU skóre. V tabuľke 5.6 sú pre porovnanie zobrazené aj potomkovské modely, ktoré mali taktiež rodiča model čeština→chorvátčina. Rozdiel medzi nimi bol v použitej technike *transfer learning*-u.

Model	Slovník	BLEU skóre
SK→HR *warm-start*	spojený, veľkosť 32000	**24,8**
SK→HR *cold-start*	spojený, veľkosť 32000	24,6
SK→HR *cold-start*	potomkovský, veľkosť 32000	24,3
SK→HR *baseline*	potomkovský, veľkosť 32000	24,4

Tabuľka 5.6: Porovnanie výslednej kvality prekladov potomkovských modelov slovenčina→chorvátčina, ktoré boli natrénované pomocou *transfer learning*-u. Prekladový smer rodičovských modelov bol čeština→chorvátčina. V tabuľke je informácia, akým typom *transfer learning*-u bol daný model natrénovaný (*cold-start* alebo *warm-start*). V dolnej časti tabuľky je pre porovnanie *baseline* model s rovnakou veľkosťou slovníka ako majú potomkovské slovníky.

5.5 5. experiment – warm-start transfer learning 2

V ďalšom experimente boli vyskúšané 2 rovnaké prístupy ako v experimente 5.3, ale na inom jazykovom páre pre potomkovský model. Tentokrát pre dvojicu slovenčina→srbčina.

Hlavné dôvody výberu tohto jazykového páru:

- menšie množstvo paralelných dát – vďaka tomu môže byť odpozorované, ako efektívne budú *warm-start transfer learning* techniky pri rozdielnom počte trénovacích dát,
- veľká zhoda medzi chorvátčinou a srbčinou – chorvátčina a srbčina pochádzajú z toho istého jazyka (srbochorvátčina) a sú v mnohých ohľadoch takmer identické[21] (napríklad medzi týmito jazykmi existuje väčšia podobnosť ako medzi slovenčinou a češtinou[33]).

Pre prekladový smer slovenčina→srbčina bolo použitých približne 141000 paralelných viet na trénovanie, čo je asi len necelých 8% z celkového počtu trénovacích dát, ktoré boli k dispozícii pre dvojicu slovenčina→chorvátčina.

V tomto experimente rodičovské modely opäť prekladali z češtiny do chorvátčiny. Počet trénovacích dát pre rodiča ostal zachovaný. Kvôli nižšiemu objemu trénovacích dát pre potomkovské modely bola zvolená veľkosť všetkých slovníkov 8000.

Najskôr boli vygenerované slovníky s touto veľkosťou pre rodičovské aj potomkovské trénovacie dáta. Následne bol vytvorený spojený a vyvážený slovník, rovnakým spôsobom ako bolo popísané v experimente 5.3. Spojený korpus na generovanie vyváženého slovníka obsahoval približne 280000 paralelných viet (t. j. 25% viet v češtine, 25% v slovenčine, 25% v chorvátčine a 25% v srbčine).

S týmito slovníkmi boli najskôr natrénované rodičovské modely a po konvergencii týchto modelov, boli vymenené trénovacie sady rodiča za trénovacie sady potomka. Trénovanie potom pokračovalo ďalej s rovnakými slovníkmi aj parametrami trénovania.

Baseline model používal iba potomkovský slovník.

Podľa výsledkov v tabuľke 5.7 je možné vidieť, že trénovanie modelov týmito prístupmi pre jazykový pár s nižším počtom trénovacích dát prináša výrazné zlepšenie. Potomkovský model, ktorý použil ako rodiča model so spojeným slovníkom dosiahol oproti *baseline* modelu zlepšenie o 4,1 BLEU skóre. A ak bol použitý rodičovský model s vyváženým slovníkom, tak BLEU skóre bolo dokonca navýšené o 4,8.

Model	Slovník	BLEU skóre
SK→srbčina	spojený, veľkosť 8000	30,1
SK→srbčina	vyvážený, veľkosť 8000	**30,8**
SK→srbčina *baseline*	potomkovský, veľkosť 8000	26,0

Tabuľka 5.7: Porovnanie kvality potomkovských modelov slovenčina→srbčina, ktoré boli natrénované technikami *warm-start transfer learning*. V tabuľke je informácia o použitom slovníku a výslednom BLEU skóre potomkovských modelov. Oba potomkovské modely mali ako rodiča prekladový smer čeština→chorvátčina. V spodnej časti tabuľky je pre porovnanie *baseline* model s rovnakou veľkosťou slovníka.

Ďalším zistením v tomto experimente bolo to, že použitie techník *warm-start transfer learning* so spojeným, resp. vyváženým slovníkom je viac nápomocné pri jazykových pároch, ktoré majú len veľmi málo trénovacích paralelných dát (slovenčina→srbčina). Zvýšením objemu trénovacích dát, sa prírastok BLEU skóre zmenší. Porovnanie nárastov BLEU skóre je možné vidieť v tabuľke 5.8.

Model	Slovník	Trénovacia sada	nárast BLEU skóre
SK→HR	spojený, veľkosť 32000	1,7M	+0,4
SK→srbčina	spojený, veľkosť 8000	141,1k	+4,1
SK→HR	vyvážený, veľkosť 32000	1,7M	+0,6
SK→srbčina	vyvážený, veľkosť 8000	141,1k	+4,8

Tabuľka 5.8: Porovnanie výsledného nárastu BLEU skóre potomkovských modelov oproti *baseline* modelom v danom prekladovom smere. Prekladové smery sa od seba líšia množstvom trénovacích dát. Údaje o trénovacích sadách sú zaokrúhlené a predstavujú počet paralelných viet. Potomkovské modely boli natrénované technikami *warm-start transfer learning*. Vo všetkých prípadoch ako rodičovský model slúžil model čeština→chorvátčina.

5.6 6. experiment – warm-start transfer learning 3

Úlohou tohto experimentu bolo zistiť, aký rodičovský model je vhodnejší pri *transfer learning*-u s vyváženým slovníkom pre potomkovský model slovenčina→srbčina.

Jazykové páry pre rodičovský model:

- čeština→chorvátčina,
- čeština→srbčina.

Pri použití rodičovského modelu, ktorý prekladá do chorvátčiny, je výhodou väčšie množstvo trénovacích dát (≈ 1,8 miliónov paralelných viet). Tento model zároveň prekladá do veľmi podobného jazyka (chorvátčina) ako je cieľový jazyk potomkovského modelu (srbčina).

Na druhej strane, rodičovský model s prekladovým smerom čeština→srbčina má rovnaký cieľový jazyk ako potomkovský model. Nevýhodou je ale objem dostupných paralelných viet v trénovacej sade, a to približne 211000. Je to síce zhruba o 70000 viac, ako má trénovacia sada pre dvojicu slovenčina→srbčina, ale oproti prvému rodičovskému modelu je to výrazne zníženie.

Na vygenerovanie vyváženého slovníka, kde bol rodičovský model slovenčina→srbčina, bol vytvorený spojený korpus opäť s rovnakým pomerom viet ako v predchádzajúcich experimentoch. V tomto prípade: 25% čeština, 25% slovenčina a 25% + 25% srbčina.

Ako vyvážený slovník pre rodičovský prekladový smer čeština→chorvátčina bol použitý slovník opísaný v experimente 5.5.

Všetky rodičovské aj potomkovské modely boli trénované s rovnakými trénovacími parametrami a oba slovníky mali zhodnú veľkosť 8000.

Model	**Rodič**	**Slovník**	**BLEU skóre**
SK→SR	čeština→HR	vyvážený, veľkosť 8000	**30,8**
SK→SR	čeština→SR	vyvážený, veľkosť 8000	26,0
SK→SR *baseline*	-	potomkovský, veľkosť 8000	26,0

Tabuľka 5.9: Porovnanie kvality potomkovských modelov slovenčina→srbčina, ktoré boli natrénované technikami *warm-start transfer learning*. V tabuľke sú informácie o použitom slovníku a prekladovom smere rodiča. V spodnej časti tabuľky je pre porovnanie *baseline* model s rovnakou veľkosťou slovníka.

Potomkovský model, ktorého predtrénovaný rodičovský model bol čeština→srbčina dosiahol rovnaké BLEU skóre ako *baseline* model so zhodnou veľkosťou slovníka. Ale ak bol rodičovský model čeština→chorvátčina, tak potomkovský model dosiahol už spomenuté zlepšenie o 4,8 BLEU skóre.

Tento experiment ukázal, že je určite vhodnejšie, pokiaľ to je možné, použiť rodičovský model, ktorý ma väčšie množstvo trénovacích dát, ak aj prekladá do podobného cieľového jazyka. Ak sa ako rodič použil model, ktorý prekladal do rovnakého jazyka, pomohlo to iba na rýchlejšie skonvergovanie modelu, ale výsledné BLEU skóre to nevylepšilo (ale ani nezhoršilo).

5.7 7. experiment – semi-supervised learning

Ďalšou otestovanou technikou bola technika *semi-supervised learning*. Zdrojový jazyk prekladu bola slovenčina a cieľový jazyk srbčina.

Postup pri tejto technike bol nasledovný:

- natrénovanie *baseline* modelu pre smer slovenčina→srbčina,
- natrénovanie *baseline* modelu pre opačný smer,
- spätný preklad monolingválnych dát,
- vznik pseudo-paralelných dát – pridanie týchto dát do trénovacej sady,
- pokračovanie trénovania *baseline* modelu s rozšírenou trénovacou sadou.

Baseline modely použili na trénovanie približne 141000 paralelných viet. Parametre trénovania aj veľkosti slovníkov boli pri oboch modeloch rovnaké, a to 8000.

Ako monolingválne dáta bol použitý korpus SETIMES[4] v srbskom jazyku, ktorý obsahuje články z novín pre balkánske jazyky. Z tohto korpusu bolo použitých na spätný preklad

[4]SETIMES (alebo SETimes) – Southeast European Times. Korpus dostupný na: http://opus.nlpl.eu/SETIMES-v2.php.

prvých 200000 viet. Tento konkrétny počet nevychádza zo žiadnej štúdie a bol zvolený tak, aby počet pseudo-paralelných viet presahoval počet pôvodných trénovacích paralelných dát.

Spätný preklad bol v tomto experimente vykonaný dvoma spôsobmi – boli použité 2 rozličné modely na preloženie srbských monolingválnych dát:

- *baseline* model srbčina→slovenčina,
- *baseline* model chorvátčina→slovenčina.

Cieľom tohto experimentu teda nebolo iba zistenie efektívnosti samotnej techniky, ale aj toho, či je možné použiť na spätný preklad model, ktorý prekladá z veľmi podobného jazyka. Výhodou takéhoto modelu je to, že má niekoľkonásobne viac dostupných trénovacích dát v porovnaní s modelom, ktorý prekladá priamo zo srbčiny.

Baseline model chorvátčina→slovenčina bol natrénovaný s rovnakými parametrami ako ostatné modely a používal vygenerovaný slovník o veľkosti 32000.

Keďže tento model prekladá z takmer rovnakého jazyka, existuje predpoklad, že výsledné spätné preklady v slovenčine budú minimálne rovnako kvalitné ako tie, ktoré vyprodukuje model srbčina→slovenčina.

Prekladané vety v srbčine	
Veta č. 1	Ovogodišnji festival će prvi put uključivati prikazivanje filmova.
Veta č. 2	Ruski vojnici obezbeđuju kontrolni punkt u blizini gruzijskog sela Khurvaleti.
Veta č. 3	Nikaragva je kasnije postala jedina druga zemlja na svetu koja je to učinila.
Veta č. 4	„Grčki birači poslali su poruku da hoće dinamičnu i efektivnu vladu“, rekao je on.
Veta č. 5	U izveštaju je zaključeno da je situacija i dalje neadekvatna.
Referenčný preklad v slovenčine	
Veta č. 1	Tohtoročný festival bude prvýkrát zahŕňať premietanie filmov.
Veta č. 2	Ruskí vojaci zabezpečujú kontrolný bod v blízkosti gruzínskej dediny Khurvaleti.
Veta č. 3	Nikaragua sa neskôr stala druhou krajinou na svete, ktorá tak urobila.
Veta č. 4	„Grécki voliči poslali správu že chcú dynamickú a efektívnu vládu“, povedal.
Veta č. 5	V správe je dospeté k záveru, že situácia je naďalej neadekvátna.
Preložené vety modelom srbčina→slovenčina	
Veta č. 1	Tento ročný festival bude obsahovať premietanie filmov.
Veta č. 2	Ruskí vojaci strácajú kontrolu pri gruzínskej dedine Khurvalti.
Veta č. 3	Neskôr sa Nikaragua stala jedinou posvätnou krajinou.
Veta č. 4	„Grman by si poslali správu, že chce dynamickú a efektívnu vládu,“ povedal.
Veta č. 5	Je zistené, že situácia je stále neadekvátna.
Preložené vety modelom chorvátčina→slovenčina	
Veta č. 1	Toto je výročie festivalu prvýkrát pomocou vysielania filmov.
Veta č. 2	Ruský vojaci poskytujú ovládanie v blízkosti grófskej dediny Khurva.
Veta č. 3	Nikara sa neskôr stala jedinou krajinou na svete, ktorá to spôsobila.
Veta č. 4	„Grécki voliči poslali správu, že chce dynamickú a efektívnu vládu,“ povedal.
Veta č. 5	V správe sa dospelo k záveru, že situácia je stále nedostatočná.

Tabuľka 5.10: Ukážka spätných prekladov zo srbského monolingválneho korpusu SETIMES, ktoré boli využité na vytvorenie pseudo-paralelných trénovacích dát. Referenčný preklad k monolingválnym vetám bol vytvorený iba pre tieto ukážkové vety na demonštráciu a pochopenie zmyslu viet.

Ako je možné vidieť v tabuľke 5.10, spätným prekladom monolingválnych viet vznikli celkom nekonzistentné preklady. Napríklad niektoré vety preložené pomocou modelu chorvátčina→slovenčina sú až prekvapivo presné v porovnaní s referenčným prekladom (viď

veta č. 4 v tabuľke 5.10). Na druhú stranu v niektorých prípadoch sú ale presnejšie, resp. zrozumiteľnejšie preklady pomocou modelu srbčina→slovenčina ako tie preložené modelom chorvátčina→slovenčina. Inokedy sú zas oba preklady nepresné a majú odlišný význam ako pôvodné srbské vety – preložené vety sa medzi sebou líšia iba použitím vhodnejšieho alebo menej vhodnejšieho slova (viď veta č. 2 v tabuľke 5.10).

Z týchto prekladov vznikli pseudo-paralelné dáta, ktoré boli pridané do pôvodných trénovacích dát pre smer slovenčina→srbčina. *Baseline* modely pokračovali v trénovaní po výmene trénovacej sady.

Model	**Spätný preklad**	**BLEU skóre**
SK→srbčina *semi-supervised*	srbčina→SK	26,5
SK→srbčina *semi-supervised*	chorvátčina→SK	**28,7**
SK→srbčina *baseline*	-	26,0

Tabuľka 5.11: Porovnanie kvality modelov, ktoré boli natrénované pomocou spätného prekladu monolingválnych dát (*semi-supervised learning*). V stĺpci „Spätný preklad“ je informácia, akým modelom boli preložené srbské monolingválne vety. V spodnej časti tabuľky je pre porovnanie *baseline* model s rovnakou veľkosťou slovníka.

Evaluácie týchto pokračovaní trénovania je možné vidieť v tabuľke 5.11, kde je zaznamenané ich BLEU skóre.

Prvým zistením na základe výsledkov tohto experimentu je to, že už aj jednoduchá varianta *semi-supervised learning*-u stačí na to, aby vylepšila pôvodný *baseline* model. V oboch otestovaných prípadoch došlo k zlepšeniu, a to konkrétne o 0,5, resp. až 2,7 BLEU skóre.

Výsledky zároveň ukázali to, že použitím pseudo-paralelných trénovacích dát, ktoré vytvoril model chorvátčina→slovenčina sa dosiahne ešte kvalitnejší výsledok. Model, ktorý pokračoval v trénovaní s týmito dátami, dosiahol až o 2,2 vyššie BLEU skóre oproti modelu, ktorý použil monolingválne dáta preložené modelom srbčina→slovenčina. Vďaka tomu možno vyvodiť záver, že je možné na spätný preklad použiť aj model, ktorý prekladá z veľmi podobného jazyka. A ak tento prekladový smer má dostupných viac paralelných trénovacích dát, je pravdepodobné, že výsledky budú lepšie.

5.8 Vzájomné porovnanie výsledkov

V tejto časti sú porovnané navzájom všetky prekladové modely z jednotlivých experimentov. V tabuľke 5.12 sú hodnoty BLEU skóre modelov prekladajúcich do chorvátčiny. Zhrnutie výsledkov modelov so srbčinou ako cieľovým jazykom je v tabuľke 5.17.

Ukážky prekladov vybraných viet z testovacej sady pre chorvátčinu sú v tabuľkách 5.13, 5.14, 5.15, 5.16. Každá z týchto ukážok zastupuje určitú skupinu preložených viet.

V tabuľke 5.13 je možné vidieť, ako *baseline* preložil slovenské vety inak v porovnaní s referenčným modelom. Ak bol na preklad použitý najlepší model spomedzi všetkých experimentov, tak tieto vety už boli preložené správne. Vďaka tomu dosiahol tento model vyššie BLEU skóre ako *baseline* model.

Niekedy nastali aj také prípady, že najlepší model použil oproti *baseline* modelu správne slovo alebo slová, ale v porovnaní s referenčným prekladom nesúhlasil slovosled danej vety. Ukážku takéhoto prekladu je možné vidieť v tabuľke 5.14. V takýchto scenároch nastáva situácia opisovaná v sekcii 4.1. Keďže chorvátčina má pomerne voľný slovosled, môže byť

Model	Rodič	Slovník	BLEU skóre
SK→HR *cold-start*	EN→čeština	potomkovský, veľkosť 32000	24,1
SK→HR *cold-start*	EN→slovanské jazyky	potomkovský, veľkosť 64000	22,5
SK→HR *cold-start*	čeština→HR	potomkovský, veľkosť 32000	24,3
SK→HR *cold-start*	čeština→HR	spojený, veľkosť 32000	24,6
SK→HR *warm-start*	čeština→HR	spojený, veľkosť 32000	24,8
SK→HR *warm-start*	čeština→HR	vyvážený, veľkosť 32000	**25,0**
SK→HR *baseline*	-	potomkovský, veľkosť 4000	23,6
SK→HR *baseline*	-	potomkovský, veľkosť 16000	24,9
SK→HR *baseline*	-	potomkovský, veľkosť 32000	24,4
SK→HR *baseline*	-	potomkovský, veľkosť 64000	23,2

Tabuľka 5.12: Porovnanie výsledkov všetkých prekladových modelov pre jazykový pár slovenčina→chorvátčina, ktoré boli natrénované v jednotlivých experimentoch tejto práce. Všetky modely boli testované na testovacej sade TedTalks s počtom 2166 paralelných viet.

Referenčný preklad	Tamo su smješteni neuroni osjetljivi na lica.
Preklad *baseline* modelom 32k	**Tu se nalaze** neuroni osjetljivi na lica.
Preklad najlepším modelom	Tamo su smješteni neuroni osjetljivi na lica.
Referenčný preklad	To su vrlo snažni, živopisni osjećaji.
Preklad *baseline* modelom 32k	To su vrlo **upečatljivi, živi** osjećaji.
Preklad najlepším modelom	To su vrlo snažni, živopisni osjećaji.

Tabuľka 5.13: Ukážka č.1 prekladov viet z testovacej sady TedTalks v chorvátčine. Ako najlepší model bol použitý model natrénovaný *warm-start transfer learning*-om s vyváženým slovníkom.

niekedy náročné ohodnotiť preloženú vetu s iným slovosledom ako je v referenčnej vete. Napríklad v ukážke je slovosled vety preložený pomocou najlepšieho modelu tiež v poriadku.

Referenčný preklad	On samo zna kako voditi tvrtku.
Preklad *baseline* modelom 32k	On samo zna kako **prebaciti** tvrtku.
Preklad najlepším modelom	On *zna samo* kako voditi tvrtku.

Tabuľka 5.14: Ukážka č.2 prekladov viet z testovacej sady TedTalks v chorvátčine. Ako najlepší model bol použitý model natrénovaný *warm-start transfer learning*-om s vyváženým slovníkom.

Ukážka v tabuľke 5.15 predstavuje takú skupinu preložených viet, kde najlepší model síce upresnil preklad v porovnaní s *baseline* modelom, ale stále tieto vety obsahovali chyby a neboli zhodné s referenčným prekladom.

Samozrejme pri prekladoch testovacej sady oboma modelmi nastali aj také situácie, kedy obidva preklady boli chybné a najlepší model nepriniesol žiadne zlepšenie oproti preloženým vetám *baseline* modelom. Ukážka prekladov v tabuľke 5.16 obsahuje v oboch prípadoch zlé preklady. Preložené testovacie sady obsahovali aj také vety, kedy modely vyprodukovali identické preklady, ale oba boli nesprávne.

Referenčný preklad	Moji umjetnički horizonti se neprestano povećavaju.
Preklad *baseline* modelom 32k	Moji umjetnički horizonti **stalno se miču**.
Preklad najlepším modelom	Moji umjetnički horizonti se neprestano **pomiču**.

Tabuľka 5.15: Ukážka č.3 prekladov viet z testovacej sady TedTalks v chorvátčine. Ako najlepší model bol použitý model natrénovaný *warm-start transfer learning*-om s vyváženým slovníkom.

Referenčný preklad	Pa kako onda ti ljudi unovče sva ta zaražena računala?
Preklad *baseline* modelom 32k	**I** kako onda **oni svrstavaju te zaražene kompjutere**?
Preklad najlepším modelom	Kako onda **ti zaraženi kompjuteri spavaju**?

Tabuľka 5.16: Ukážka č.4 prekladov viet z testovacej sady TedTalks v chorvátčine. Ako najlepší model bol použitý model natrénovaný *warm-start transfer learning*-om s vyváženým slovníkom.

Model	**Rodič ○/ Spätný preklad •**	**Slovník**	**BLEU skóre**
SK→SR *warm-start*	čeština→HR ○	spojený	30,1
SK→SR *warm-start*	čeština→HR ○	vyvážený	**30,8**
SK→SR *warm-start*	čeština→SR ○	vyvážený	26,0
SK→SR *semi-supervised*	HR→SK •	potomkovský	28,7
SK→SR *semi-supervised*	SR→SK •	potomkovský	26,5
SK→SR *baseline*	-	potomkovský	26,0

Tabuľka 5.17: Porovnanie výsledkov všetkých prekladových modelov pre jazykový pár slovenčina→srbčina, ktoré boli natrénované v jednotlivých experimentoch tejto práce. Veľkosť všetkých slovníkov bola rovnaká: 8000. Všetky modely boli testované na testovacej sade TedTalks s počtom 2384 paralelných viet.

Kapitola 6

Záver

Cieľom tejto bakalárskej práce bolo preskúmanie neurónového strojového prekladu pre jazykové páry, ktoré majú k dispozícií iba málo paralelných dát. Pre tento účel bola zvolená dvojica slovanských jazykov slovenčina – chorvátčina. Všetky prekladové modely používali architektúru transformer a boli natrénované *framework*-om Marian. Trénovanie prebiehalo vždy s rovnakými implicitnými parametrami, pričom výsledné modely sa líšili v použitých metódach pre zlepšenie kvality prekladu *low-resource* jazykov.

V prvom experimente s *baseline* modelmi bolo zistené, že pre *low-resource* jazyky je určite vhodnejšie použiť slovníky menších veľkostí ako je štandardná prednastavená veľkosť 32000 jednotiek *subword*. A to aj v takom prípade, kedy oba jazyky patria medzi vysoko ohybné jazyky, kde slová môžu mať veľa rôznych tvarov. Model so slovníkom o veľkosti 4000 dosiahol lepší výsledok ako model, ktorý mal veľkosť až 64000 jednotiek (23,6 BLEU skóre oproti 23,2).

Pri testovaní techník *cold-start transfer learning* sa ukázala ako najlepšia možnosť použiť ako rodičovský model ten, ktorého jazykový pár bol z rovnakej jazykovej skupiny. V tomto prípade to bol model čeština→chorvátčina, čiže rodičovský aj potomkovský model mali rovnaký cieľový jazyk.

Použitie potomkovského slovníka pri *cold-start transfer learning*-u neprinieslo žiadne zlepšenie v porovnaní s *baseline* modelom, ktorý mal rovnakú veľkosť slovníka. Najlepší model natrénovaný týmto spôsobom zaostával o 0,1 BLEU skóre. Bol to práve model s rodičom čeština→chorvátčina. Preto sa dá považovať tento prístup za nedostačujúci.

Ak sa pre potomkovský model použil spojený slovník, pričom rodičovský model stále nebol natrénovaný pre potreby potomka (t. j. rodičovský model používal rodičovský slovník), tak takýto model už dosahoval lepšie BLEU skóre ako *baseline* model. BLEU skóre bolo v porovnaní s *baseline* modelom vyššie o 0,2.

Techniky *warm-start transfer learning* priniesli v experimentoch najväčšie zlepšenie. Tieto prístupy boli otestované aj na prekladovom smere slovenčina→srbčina, ktorý má k dispozícií ešte menšie množstvo paralelných dát. Modely so spojeným slovníkom dosiahli zlepšenie o 0,4 pri chorvátčine a až o 4,1 BLEU skóre pri srbčine. Pri použití vyváženého slovníka mali potomkovské modely o 0,6, resp. 4,8 vyššie BLEU skóre. Vo všetkých týchto prípadoch bol rodičovský model čeština→chorvátčina. Ak sa ako rodič použil model čeština→srbčina, ktorý má menej paralelných dát, tak výsledný potomkovský model mal rovnaké BLEU skóre ako *baseline* model.

Ďalším predmetom skúmania bola využiteľnosť veľmi podobného (skoro až identického) jazyka pri spätných prekladoch monolingválnych dát v tzv. *semi-supervised learning*-u. V tejto práci sa srbské vety preložili modelom prekladajúcim zo srbčiny do slovenčiny ako

aj kvalitnejším modelom z chorvátčiny do slovenčiny. V oboch scenároch nastalo zlepšenie v porovnaní s *baseline* modelom. Spätný preklad modelom srbčina→slovenčina priniesol zlepšenie o 0,5 BLEU skóre. Ak sa ale použil na preklad monolingválnych dát model chorvátčina→slovenčina, tak výsledný model dosiahol až o 2,7 vyššie BLEU skóre.

Vo viacerých experimentoch v tejto práci sa ukázalo využitie podobných jazykov alebo jazykov z rovnakej jazykovej skupiny ako nápomocné. Preto ďalším cieľom pri pokračovaní skúmania tejto problematiky môže byť rozšírenejšie prebádanie benefitov takýchto jazykov. Napríklad pri slovanských jazykoch aj využitie ich fonetickej podobnosti.

Literatúra

[1] Workshop on Statistical Machine Translation (WMT). *ACL Anthology* [online]. 2020 [cit. 2020-07-02]. Dostupné z: https://www.aclweb.org/anthology/venues/wmt/.

[2] Aji, A. F., Bogoychev, N., Heafield, K. a Sennrich, R. In Neural Machine Translation, What Does Transfer Learning Transfer? In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* [online]. Association for Computational Linguistics, Júl 2020, s. 7701–7710 [cit. 2020-07-14]. Dostupné z: https://www.aclweb.org/anthology/2020.acl-main.688.

[3] Bahdanau, D., Cho, K. a Bengio, Y. Neural machine translation by jointly learning to align and translate. [online]. 2014, [cit. 2020-07-14]. Dostupné z: https://arxiv.org/pdf/1409.0473.pdf.

[4] Benaffane, Y. Transformer vs RNN and CNN for Translation Task. *Medium* [online], 13. augusta 2019 [cit. 2020-07-02]. Dostupné z: https://medium.com/analytics-vidhya/transformer-vs-rnn-and-cnn-18eeefa3602b.

[5] Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B. et al. Findings of the 2018 Conference on Machine Translation (WMT18). In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* [online]. Brusel, Belgicko: Association for Computational Linguistics, Október 2018, s. 272–303 [cit. 2020-07-02]. Dostupné z: https://www.aclweb.org/anthology/W18-6401.

[6] Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F. et al. A Statistical Approach to Machine Translation. *Computational Linguistics* [online]. 1990, zv. 16, č. 2, s. 79–85, [cit. 2020-07-02]. Dostupné z: https://www.aclweb.org/anthology/J90-2002.

[7] Brownlee, J. A Gentle Introduction to Neural Machine Translation. *Machine Learning Mastery* [online], 29. decembra 2017. Revidované 7.8.2019 [cit. 2020-07-02]. Dostupné z: https://machinelearningmastery.com/introduction-neural-machine-translation.

[8] Cherivirala, S., Chiplunkar, S., Washington, J. a Unhammer, K. Apertium's Web Toolchain for Low-Resource Language Technology. In: *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)* [online]. Boston, MA, USA: Association for Machine Translation in the Americas, Marec 2018, s. 53–62 [cit. 2020-07-14]. Dostupné z: https://www.aclweb.org/anthology/W18-2207.

[9] Christodoulopoulos, C. a Steedman, M. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation* [online]. Jún 2014, zv. 49, s. 1–21, [cit. 2020-07-02]. DOI: 10.1007/s10579-014-9287-y. Dostupné z: https://www.researchgate.net/publication/269334313_A_massively_parallel_corpus_the_Bible_in_100_languages.

[10] Denkowski, M. a Neubig, G. Stronger Baselines for Trustable Results in Neural Machine Translation. In: *Proceedings of the First Workshop on Neural Machine Translation* [online]. Vancouver, Kanada: Association for Computational Linguistics, August 2017, s. 18–27 [cit. 2020-07-02]. DOI: 10.18653/v1/W17-3203. Dostupné z: https://www.aclweb.org/anthology/W17-3203.

[11] Garg, S., Peitz, S., Nallasamy, U. a Paulik, M. Jointly Learning to Align and Translate with Transformer Models. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* [online]. Hong Kong, Čína: Association for Computational Linguistics, November 2019, s. 4453–4462 [cit. 2020-07-14]. DOI: 10.18653/v1/D19-1453. Dostupné z: https://www.aclweb.org/anthology/D19-1453.

[12] Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J. et al. Achieving Human Parity on Automatic Chinese to English News Translation. [online]. Microsoft AI Research. 2018, [cit. 2020-07-02]. Dostupné z: https://arxiv.org/pdf/1803.05567v2.pdf.

[13] Hoang, V. C. D., Koehn, P., Haffari, G. a Cohn, T. Iterative Back-Translation for Neural Machine Translation. In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* [online]. Melbourne, Austrália: Association for Computational Linguistics, Júl 2018, s. 18–24 [cit. 2020-07-14]. DOI: 10.18653/v1/W18-2703. Dostupné z: https://www.aclweb.org/anthology/W18-2703.

[14] Junczys Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K. et al. Marian: Fast Neural Machine Translation in C++. In: *Proceedings of ACL 2018, System Demonstrations* [online]. Melbourne, Austrália: Association for Computational Linguistics, Júl 2018, s. 116–121 [cit. 2020-07-14]. Dostupné z: http://www.aclweb.org/anthology/P18-4020.

[15] Kim, Y., Graça, M. a Ney, H. When and Why is Unsupervised Neural Machine Translation Useless? [online]. Aachen, Nemecko: RWTH Aachen University. 2020, [cit. 2020-07-07]. Dostupné z: https://arxiv.org/pdf/2004.10581.pdf.

[16] Kim, Y., Petrov, P., Petrushkov, P., Khadivi, S. a Ney, H. Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* [online]. Hong Kong, Čína: Association for Computational Linguistics, November 2019, s. 866–876 [cit. 2020-07-22]. DOI: 10.18653/v1/D19-1080. Dostupné z: https://www.aclweb.org/anthology/D19-1080.

[17] Kocmi, T. *Exploring Benefits of Transfer Learning in Neural Machine Translation* [online]. Praha. [cit. 2020-07-02]. Dizertačná práca. Univerzita Karlova,

Matematicko-fyzikální fakulta, Ústav formální a aplikované lingustiky. Vedúci práce RNDr. Ondřej Bojar, P. doc. Dostupné z: https://arxiv.org/pdf/2001.01622.pdf.

[18] Kocmi, T. a Bojar, O. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In: *Proceedings of the Third Conference on Machine Translation: Research Papers* [online]. Brusel, Belgicko: Association for Computational Linguistics, Október 2018, s. 244–252 [cit. 2020-07-03]. DOI: 10.18653/v1/W18-6325. Dostupné z: https://www.aclweb.org/anthology/W18-6325.

[19] Kocmi, T. a Bojar, O. Efficiently Reusing Old Models Across Languages via Transfer Learning. [online]. 2020, [cit. 2020-07-03]. Dostupné z: https://arxiv.org/pdf/1909.10955.pdf.

[20] Koehn, P. Neural Machine Translation. In: *Statistical Machine Translation* [online]. Cambridge University Press, August 2015, revidované 25.9.2017 [cit. 2020-07-02]. Dostupné z: https://arxiv.org/pdf/1709.07809v1.pdf.

[21] Kovačić, M. Serbian and Croatian : One language or languages? *Jezikoslovlje* [online]. 2005, zv. 6, č. 2, s. 195–204, [cit. 2020-07-15]. ISSN 1848-9001. Dostupné z: https://hrcak.srce.hr/30869.

[22] Kudo, T. a Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* [online]. Brusel, Belgicko: Association for Computational Linguistics, November 2018, s. 66–71 [cit. 2020-07-13]. DOI: 10.18653/v1/D18-2012. Dostupné z: https://www.aclweb.org/anthology/D18-2012.

[23] Kösegiová, K. *Slovosled v chorvátčine v porovnaní so slovenčinou* [online]. Brno, 2014. [cit. 2020-07-12]. Bakalárska práca. Masarykova univerzita, Filozofická fakulta. Vedúci práce Mgr. Pavel Krejčí, P. Dostupné z: https://is.muni.cz/th/pa5vy/Bakalarska_praca_KATARINA_KOSEGIOVA_2014.pdf.

[24] Lakew, S. M., Cettolo, M. a Federico, M. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. [online]. Fondazione Bruno Kessler. 2018, [cit. 2020-07-02]. Dostupné z: https://arxiv.org/pdf/1806.06957v2.pdf.

[25] Marie, B., Sun, H., Wang, R., Chen, K., Fujita, A. et al. NICT's Unsupervised Neural and Statistical Machine Translation Systems for the WMT19 News Translation Task. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* [online]. Florencia, Taliansko: Association for Computational Linguistics, August 2019, s. 294–301 [cit. 2020-07-14]. DOI: 10.18653/v1/W19-5330. Dostupné z: https://www.aclweb.org/anthology/W19-5330.

[26] Mohammad, S. M., Salameh, M. a Kiritchenko, S. How Translation Alters Sentiment. *J. Artif. Int. Res.* [online]. El Segundo, CA, USA: AI Access Foundation. Január 2016, zv. 55, č. 1, s. 95–130, [cit. 2020-07-02]. ISSN 1076-9757. Dostupné z: https://www.svkir.com/papers/Mohammad-et-al-ArabicSA-JAIR-2016.pdf.

[27] Post, M. A Call for Clarity in Reporting BLEU Scores. In: *Proceedings of the Third Conference on Machine Translation: Research Papers* [online]. Brusel, Belgicko:

Association for Computational Linguistics, Október 2018, s. 186–191 [cit. 2020-07-17]. DOI: 10.18653/v1/W18-6319. Dostupné z: https://www.aclweb.org/anthology/W18-6319.

[28] Pourdamghani, N., Aldarrab, N., Ghazvininejad, M., Knight, K. a May, J. Translating Translationese: A Two-Step Approach to Unsupervised Machine Translation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* [online]. Florencia, Taliansko: Association for Computational Linguistics, Júl 2019, s. 3057–3062 [cit. 2020-07-14]. Dostupné z: https://www.aclweb.org/anthology/P19-1293.

[29] Przystupa, M. a Abdul Mageed, M. Neural Machine Translation of Low-Resource and Similar Languages with Backtranslation. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)* [online]. Florencia, Taliansko: Association for Computational Linguistics, August 2019, s. 224–235 [cit. 2020-07-12]. DOI: 10.18653/v1/W19-5431. Dostupné z: https://www.aclweb.org/anthology/W19-5431.

[30] Ren, S., Wu, Y., Liu, S., Zhou, M. a Ma, S. Explicit Cross-lingual Pre-training for Unsupervised Machine Translation. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* [online]. Hong Kong, Čína: Association for Computational Linguistics, November 2019, s. 770–779 [cit. 2020-07-14]. DOI: 10.18653/v1/D19-1071. Dostupné z: https://www.aclweb.org/anthology/D19-1071.

[31] Sennrich, R., Haddow, B. a Birch, A. Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [online]. Berlín, Nemecko: Association for Computational Linguistics, August 2016, s. 1715–1725 [cit. 2020-07-15]. DOI: 10.18653/v1/P16-1162. Dostupné z: https://www.aclweb.org/anthology/P16-1162.

[32] Skorokhodov, I., Rykachevskiy, A., Emelyanenko, D., Slotin, S. a Ponkratov, A. Semi-Supervised Neural Machine Translation with Language Models. In: *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)* [online]. Boston, MA, USA: Association for Machine Translation in the Americas, Marec 2018, s. 37–44 [cit. 2020-07-14]. Dostupné z: https://www.aclweb.org/anthology/W18-2205.

[33] Sokolová, M., Musilová, K. a Slančová, D. *Slovenčina a čeština : synchrónne porovnanie s cvičeniami* [online]. 1. vyd. Bratislava: Vydavateľstvo UK, 2005 [cit. 2020-07-22]. ISBN 8022321508. Dostupné z: https://is.muni.cz/el/1421/podzim2011/SKB503/um/SLOV-CES.pdf.

[34] Sutskever, I., Vinyals, O. a Le, Q. V. Sequence to Sequence Learning with Neural Networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. a Weinberger, K. Q., ed. *Advances in Neural Information Processing Systems 27* [online]. Curran Associates, Inc., 2014, s. 3104–3112 [cit. 2020-07-22]. Dostupné z: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

[35] Tan, X., Chen, J., He, D., Xia, Y., Qin, T. et al. Multilingual Neural Machine Translation with Language Clustering. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* [online]. Hong Kong, Čína: Association for Computational Linguistics, November 2019, s. 963–973 [cit. 2020-07-22]. DOI: 10.18653/v1/D19-1089. Dostupné z: https://www.aclweb.org/anthology/D19-1089.

[36] Torrey, L. a Shavlik, J. Transfer Learning. In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* [online]. IGI Gloabal, 2010, s. 242–264 [cit. 2020-07-02]. Dostupné z: https://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf.

[37] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. Attention Is All You Need. [online]. Google Brain. 2017, [cit. 2020-07-02]. Dostupné z: https://arxiv.org/pdf/1706.03762v5.pdf.

[38] Wang, D. a Zheng, T. F. Transfer Learning for Speech and Language Processing. [online]. Peking, Čína: [b.n.]. 2015, [cit. 2020-07-03]. Dostupné z: https://arxiv.org/pdf/1511.06066v1.pdf.

[39] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M. et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. [online]. September 2016, [cit. 2020-07-02]. Dostupné z: https://arxiv.org/pdf/1609.08144v2.pdf.

[40] Zoph, B., Yuret, D., May, J. a Knight, K. Transfer Learning for Low-Resource Neural Machine Translation. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* [online]. Austin, Texas, USA: Association for Computational Linguistics, November 2016, s. 1568–1575 [cit. 2020-07-03]. DOI: 10.18653/v1/D16-1163. Dostupné z: https://www.aclweb.org/anthology/D16-1163.

Príloha A

Obsah priloženého pamäťového média

- text práce vo formáte PDF
- zdrojové súbory LaTeX bakalárskej práce
- jednotlivé zložky prekladových modelov obsahujúce trénovací skript, trénovací log a preklad testovacej sady – prípadne aj najlepší model *model.npz.best-bleu.npz*
- zložka so skriptami na vytvorenie spojeného slovníka a úpravu textu na latinku