Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina

Gregor Donaj, Mirjam Sepesy Maučec

Fakulteta za elektrotehniko, računalništvo in informatiko
Univerza v Mariboru
Koroška c. 46, 2000 Maribor
gregor.donaj@um.si,
mirjam.sepesy@um.si

Povzetek

Strojno prevajanje z nevronskimi omrežji je najnovejši pristop k strojnemu prevajanju. V primerjavi s klasičnim statističnim prevajanjem, ki temelji na modelu šumnega kanala, sestavljenega iz množice neodvisnih komponent, pri nevronskem prevajanju učimo en sam model oziroma nevronsko omrežje, ki ga optimiramo v smeri čim kvalitetnejših prevodov. V članku predstavljamo naše prve izsledke nevronskega strojnega prevajanja za jezikovni par slovenščina-angleščina in rezultate primerjamo s klasičnim statističnim prevajanjem. Analiziramo prevajanje v obe smeri z različnimi hiperparametri učenja oz. modelov. Primerjava rezultatov kaže, da lahko z nevronskimi omrežji tvorimo prevode, ki so boljši za 6,2 točke BLEU (smer angleščina-slovenščina) oz. za 2,9 točke BLEU (smer slovenščina-angleščina) v primerjavi s statističnimi prevajanjem.

From statistical machine translation to translation with neural networks for the Slovene-English language pair

Neural machine translation is a newly proposed approach to machine translation. In comparison to the traditional statistical machine translation, which is based on the noisy channel model with many independent components, neural machine translation system is a single neural network trained to optimize the translation performance. In this paper, we present our f rst experiments with neural machine translation on Slovene-English language pair and compare the obtained results with classical statistical machine translation. Translation in both directions is analyzed with different model and learning hyperparameters. We found that neural machine translation outperforms statistical machine translation by 6.2 BLEU points in the translation from English to Slovene and by 2.9 BLEU points in the translation from Slovene to English.

1. Uvod

Statistično strojno prevajanje (SMT) je še do nedavnega veljalo za najuspešnejši pristop k strojnemu prevajanju. Vsaj 20 let smo lahko sledili kontinuiranemu izboljševanju kvalitete prevodov, ki so jih generirali frazni statistični prevajalniki različnih tipov: klasični frazni, faktorski, hierarhični ali temelječi na sintaktičnih strukturah (Koehn, 2010). V zadnjih nekaj letih pa lahko vidimo, da se je v raziskovalni srenji povečal interes za nevronsko strojno prevajanje (NMT), ki je dotlej veljajo za računsko preveč zahteven pristop. Že prvi rezultati so pokazali, da so NMT prevodi po kvaliteti primerljivi s SMT prevodi, za določene jezikovne pare pa občutno boljši (Junczys-Dowmunt et al., 2016). Zanimivo je, da je izboljšanje najbolj očitno pri najtežjih jezikovnih parih, kot je na primer prevajanje, ki vključuje kitajščino, arabščino in nemščino (Junczys-Dowmunt et al., 2016; Bentivogli et al., 2018). Hiter napredek nevronskega prevajanja izvira iz uporabe ponavljajočih nevronskih omrežij (tudi rekurentnih ali povratnih; ang. recurrent neural network – RNN) in arhitekture kodirnik-dekodirnik (ang. encoder-decoder), ki uporablja več nevronskih omrežij tipa RNN. Takšen pristop so med prvimi predlagali Cho et al. (2014).

Leto kasneje so Bahdanau et al. (Bahdanau et al., 2014) predlagali še rešitev za problem prevajanja dolgih povedi, ki so jo poimenovali mehanizem poudarka (ang. attention mechanism). Da je nevronsko strojno prevajanje "vroča" tema, kaže tudi osrednja konferenca na področju strojnega prevajanja WMT[1], kjer se je leta 2015 med tekmovalnimi MT sistemi prvič pojavil NMT, leto zatem pa je na NMT temeljila velika večina zmagovalnih sistemov. V letu 2017 so med tekmovalne naloge na novo uvrstili tudi učenje NMT.

Zakaj so nevronska omrežja tako učinkovita? Razlog je lahko v njihovi sposobnosti razločevanja in izločanja informacij iz zapletenih vzorcev, ki jih druge tehnike prevajanja ne zaznajo. Kot osnovno enoto uporabljajo poved, kar pomeni, da upoštevajo širši kontekst kot SMT, pri katerem je osnovna enota podatkovno def nirana fraza. Prednost je tudi v strukturi NMT sistema, ki je en sam velik model, v katerem se prevodi oblikujejo na kontekstno odvisen način, za razliko od SMT, ki je sestavljen iz povezane množice neodvisnih komponent (model prevajanja, jezikovni model, model preurejanja ipd.).

Naš cilj v tem članku je preizkusiti NMT prevajalnik na jezikovnem paru slovenščina-angleščina in ga primerjati z rezultati SMT prevajalnika. Članek je organiziran v naslednja poglavja: v poglavju 2 predstavimo splošne značilnosti NMT arhitekture in njene osnovne mehanizme. V poglavju 3 predstavimo zasnovo našega NMT sistema. Najprej podamo osnovne podatke učnega korpusa, sledi opis konf guracije NMT sistema in tudi SMT sistema, ki smo ga uporabili za primerjavo. Rezultati in analiza eksperimentov so v poglavju 4. Članek zaključimo v poglavju 5, kjer povzamemo ključne ugotovitve in podamo smernice za naprej.

[1] http://www.statmt.org/wmt17/

2. Prevajanje z nevronskimi omrežji

Nevronsko omrežje je model za obdelavo informacij, ki posnema delovanje živčnega sistema bioloških organizmov. Uči se iz primerov, zato je primeren model tudi za strojni prevajalnik, ki ga zasnujemo iz vzporednega korpusa poravnanih prevodov. Nevronsko omrežje sestavlja množica nevronov, ki jih v procesu učenja prilagodimo izbrani nalogi, v našem primeru nalogi prevajanja.

Obstajajo različne arhitekture nevronskih omrežij. Pri prevajanju se uporabljajo ponavljajoča nevronska omrežja RNN, pri katerih lahko signali potujejo v obe smeri: naprej in nazaj, kar dosežemo z vpeljavo povratnih zank. Ponavljajoča nevronska omrežja so zelo zmogljiva, a tudi zelo zapletena.

Nevronska omrežja so sestavljena iz treh plasti oz. skupin enot: vhodna plast, ena ali več skritih plasti in izhodna plast. Vhodna plast je povezana s skrito plastjo, le-ta pa z izhodno plastjo. Aktivnosti v vhodni plasti predstavljajo vhodno informacijo, ki jo vnesemo v omrežje. V skriti plasti določimo aktivnost vhodne plasti in uteži med vhodno in skrito plastjo. Kako bo odreagirala izhodna plast je odvisno od aktivnosti v skriti plasti in uteži med skrito in izhodno plastjo.

Delovanje nevronskega omrežja je, razen od uteži, odvisno tudi od aktivacijske (tudi pragovne ali prenosne) funkcije, ki povezuje vhod z izhodom. Navadno se uporablja hiperbolični tangens, ki iz neomejenega def nicijskega območja slika na interval [-1, 1]. V bolj izpopolnjenjih RNN, imenovanih RNN z vrati (ang. gated RNN), se uporabljajo GRU (ang. gated recurrent unit) enote, ki se prilagajajo različnim odvisnostim (Cho et al., 2014).

Nevronska omrežja so v splošnem omejena s f ksno dolžino vhodnega zaporedja, pri čemer je tudi izhodno zaporedje enake dolžine. Povedi, ki jih prevajamo, pa so različnih dolžin in tudi prevod se običajno v dolžini ne ujema z izvorno povedjo. Lahko ima več ali manj besed. Ta problem rešuje arhitektura kodirnik-dekodirnik (ang. encoder-decoder), kjer so dovoljena vhodna in izhodna zaporedja različnih dolžin. Kodirnik bere vhodno poved in jo kodira v vektorje f ksne dimenzije. Dekodirnik iz teh vektorjev tvori prevod. Kodirnik in dekodirnik za izbrani jezikovni par učimo sočasno, tako da maksimiziramo verjetnost pravilnega prevoda za izbrano vhodno poved.

Kodirnik je dvosmerno ponavljajoče nevronsko omrežje (ang. bidirectional RNN), sestavljeno iz naprej in nazaj usmerjene RNN. Naprej usmerjena RNN bere vhodno poved v pravilnem vrstnem redu, tj. od leve proti desni, in izračuna zaporedje naprej usmerjenih skritih stanj (ang. forward hidden states), medtem ko nazaj usmerjena RNN bere besede v obratnem vrstnem redu, tj. od desne proti levi, in generira zaporedje nazaj usmerjenih skritih stanj (ang. backward hidden states). Na ta način vsako besedo označimo s spetimi naprej in nazaj usmerjenimi skritimi stanji. To pomeni, da vsako besedo opremimo z levim in desnim kontekstom v povedi.

Dekodirnik preiskuje izvorno poved in jo po principu veriženja pogojnih verjetnosti dekodira v prevod. Tudi dekodirnik je RNN, pri katerem si iterativno sledijo tri faze; look-update-generate. V fazi look je izbrano novo skrito stanje. Izračunano je iz treh podatkov: njegovega kontekstnega vektorja, predhodnega skritega stanja in predhodno generiranje besede v prevodu. Sledi faza update, v kateri se generira novi kontekstni vektor. Kontekstni vektor skritega stanja je odvisen od vseh oznacb, ki jih je generiral kodirnik za celotno izvorno poved. Izračunan je kot utežena vsota teh oznacb. Prehod v novo skrito stanje ima za posledico tudi generiranje nove besede v prevodu. To fazo imenujemo generate.

Problem omenjenega principa delovanja dekodirnika so dolge povedi in pridruženi vektorji f ksnih dolžin. Posebej problematično je prevajanje povedi, ki so daljše od povedi v učnem korpusu. Rešitev predstavlja mehanizem poudarka (ang. attention mechanism), ki nevronskemu omrežju omogoča učenje poudarka v izvorni povedi, tj. odsekov, ki vsebujejo pomembne informacije za generiranje posamezne besede v prevodu. Mehanizem poudarka je uporabljen med dvema GRU prehodoma dekodirnika (Miceli Barone et al., 2017; Sennrich et al., 2017).

Nevronska omrežja so se šele v zadnjih letih začela bolj pogosto uporabljati na področju strojnega prevajanja. Tako je tudi pred kratkim bilo izpostavljenih nekaj ključnih izzivov pri uporabi nevronskih omrežij (Koehn in Knowles, 2017). Med drugim je izpostavljena problematika dolžine stavkov, ki jih hočemo prevajat. Avtorja sta pokazala, da se kvaliteta prevodov z nevronskimi omrežji poslabša pri stavkih z več kot 60 pojavnicami. Drugi izpostavljen izziv je bilo prevajanje dokumentov izven domene učne množice.

V naši raziskavi smo tako dodali tudi primerjavo med sistemoma SMT in NMT glede na dolžino povedi in rezultate na množici izven domene učnega korpusa.

3. NMT sistem

3.1. Učni korpus

V eksperimentih smo uporabili Europarl korpus[2] (Koehn, 2005). Korpus je sestavljen iz besedil zbornika Evropskega parlamenta. Korpus pokriva 20 jezikovnih parov, pri katerih je en jezik v paru vedno angleščina, kot drugi jezik pa nastopajo jeziki držav članic Evropske unije. Za naše eksperimente smo uporabili vzporedni korpus za jezikovni par slovenščina-angleščina, ki obsega gradivo iz obdobja med letoma 2007 in 2011. Korpus vsebuje 623.490 stavkov, pri čemer je na slovenski strani 12,5 milijonov besed, na angleški pa 15 milijonov. Korpus smo razdelili na učni, razvojni in testni del. Razvojni in testni del obsegata vsak 2000 stavkov, ki smo jih izločili iz konca korpusa. Preostali stavki so v učnem korpus. Učni korpus vsebuje na slovenski strani 144.671 različnih besed, na angleški pa 66.604. Pred učenjem prevajalnikov smo korpus tokenizirali in normalizirali.

3.2. Konfiguracija NMT sistema

Tip modela je ponavljajoče nevronsko omrežje. Uporabljeni modeli temeljijo na arhitekturi omrežja kodirnik-dekodirnik (Bahdanau et al., 2014), kjer se vhodni podatki v nevronsko omrežje najprej preslikajo na enote v skritih plasteh omrežja (kodirajo) in nato preslikajo na izhodne podatke (dekodiranje).

[2] http://www.statmt.org/europarl/

	Angleško→slovensko	Angleško→slovensko (dtdn)	Slovensko→angleško	Slovensko-angleško (dtdn)
BLEU	35,5	29,8	44,1	39,1
METEOR	31,1	28,1	41,8	38,1
TER	46,0	52,1	38,9	43,3

Tabela 1: Rezultati prevajanja testne množice s sistemom SMT.

V postopku učenja smo določili hiperparametre modela oz. nevronskega omrežja. Prvi hiperparameter je dimenzija vgrajenih vektorjev v modelih. Ti vektorji predstavljajo besede, njihovo dimenzijo pa smo nastavili na 512. Drugi hiperparameter modela je dimenzija skritega stanja v RNN, ki smo ga nastavili na 1024.

Dodatno k temu smo še omejili dolžino stavkov v učnem korpusu na 50 (tukaj štejemo tako besede in vse ostale pojavnice, npr. ločila), kar pomeni, da se pri učenju daljši stavki odstranijo. V postopku učenja dodatno določimo še velikost mini serije (ang. mini-batch). To je število stavkov iz učne množice, ki se v vsaki iteraciji učenja uporabijo za učenje omrežja – njegovo posodobitev. Kot velikost mini-serije smo izbrali 64.

3.3. Trajanje učenja

Značilnost nevronskih omrežij je, da pri velikem številu iteracij učenja prihaja do prekomernega prilagajanja (ang. overf tting) modela na učno množico. Pri tem pojavu začne model vse bolje izražati primere v učni množici, s tem pa izgubi na splošnosti in slabše deluje na novih podatkih.

Z namenom iskanja optimalnega trajanja učenja za nevronsko omrežje uporabimo razvojni del korpusa. V prvem poskusu učenja smo kot trajanje učenja določilu 500 epoh (prehodov celotnega učnega korpusa). Po porabljenem času na primerljivi strojni opremi (približno 1 teden), je to učenje primerljivo s sistemi drugih raziskovalcev (Junczys-Dowmunt et al., 2016).

Med postopkom učenja smo modele shranjevali na vsakih 10.000 iteracij (posodobitev omrežja glede na eno mini serijo). Med testiranjem smo kasneje iskali optimalno število iteracij oz. epoh učenja.

3.4. Programska oprema

Za učenje modelov in prevajanje smo uporabljali orodje Marian (prej AmuNMT). Orodje AmuNMT (Junczys-Dowmunt et al., 2016) je bilo sprva razvito za hitro prevajanje z uporabo modelov, naučenih z orodjem Nematus (Sennrich et al., 2017). Kasneje je bila razvita ponovna implementacija orodja Nematus v jeziku C++, ki je bila nato združena z AmuNMT in imenovana Marian. Orodje je odprtokodno in prostodostopno[3].

Za tokenizacijo in detokenizacijo ter normalizacijo in denormalizacijo smo uporabljali skripte, ki so sestavni del orodja Moses (Koehn et al., 2003). Preveden tekst smo ocenjevali z orodjem Multeval (Clark et al., 2011) in pri tem vrednotili prevode z metrikami BLEU, METEOR in TER.

[3] `https://marian-nmt.github.io/`

3.5. Strojna oprema

Učenje modelov in prevajanje se izvajata le na graf čnem procesorju in graf čnem delovnem pomnilniku. Oba sta del graf čne kartice, v našem primeru Nvidia GeForce GTX 1080 Ti. Izkušnje kažejo, da je najpomembnejša lastnost graf čne kartice pri učenju nevronskih omrežij pasovna širina za prenos podatkov do graf čnega spomina. V primeru naše graf čne kartice je ta 484 GB/s.

Ostala strojno oprema ni bistvena za rezultate ali hitrost učenja oz. prevajanja.

3.6. SMT sistem za primerjavo

NMT sistem smo primerjali s fraznim statističnim prevajalnikom, ki smo ga zgradili na istem korpusu. Slovar prevajalnika je na slovenski strani vseboval 144.671 besed, na angleški pa 66.604. Pri gradnji prevajalnika smo uporabili orodje Moses (Koehn et al., 2003) in standardne nastavitve: besede smo poravnali v zaporedju iteracij IBM modela 1, HMM modela in modelov 3 in 4; uporabili smo "grow-diag-f nal-and" simetrizacijo; jezikovni model je bil besedni 3-gramski z modif ciranim Kneser-Ney glajenjem frekvenc. Za učenje jezikovnih modelov smo uporabili celoten učni korpus. Iz statistike smo izločili besede, ki se pojavijo le enkrat. Perpleksnost jezikovnega modela slovenskega jezika je bila 109, angleškega pa 62.

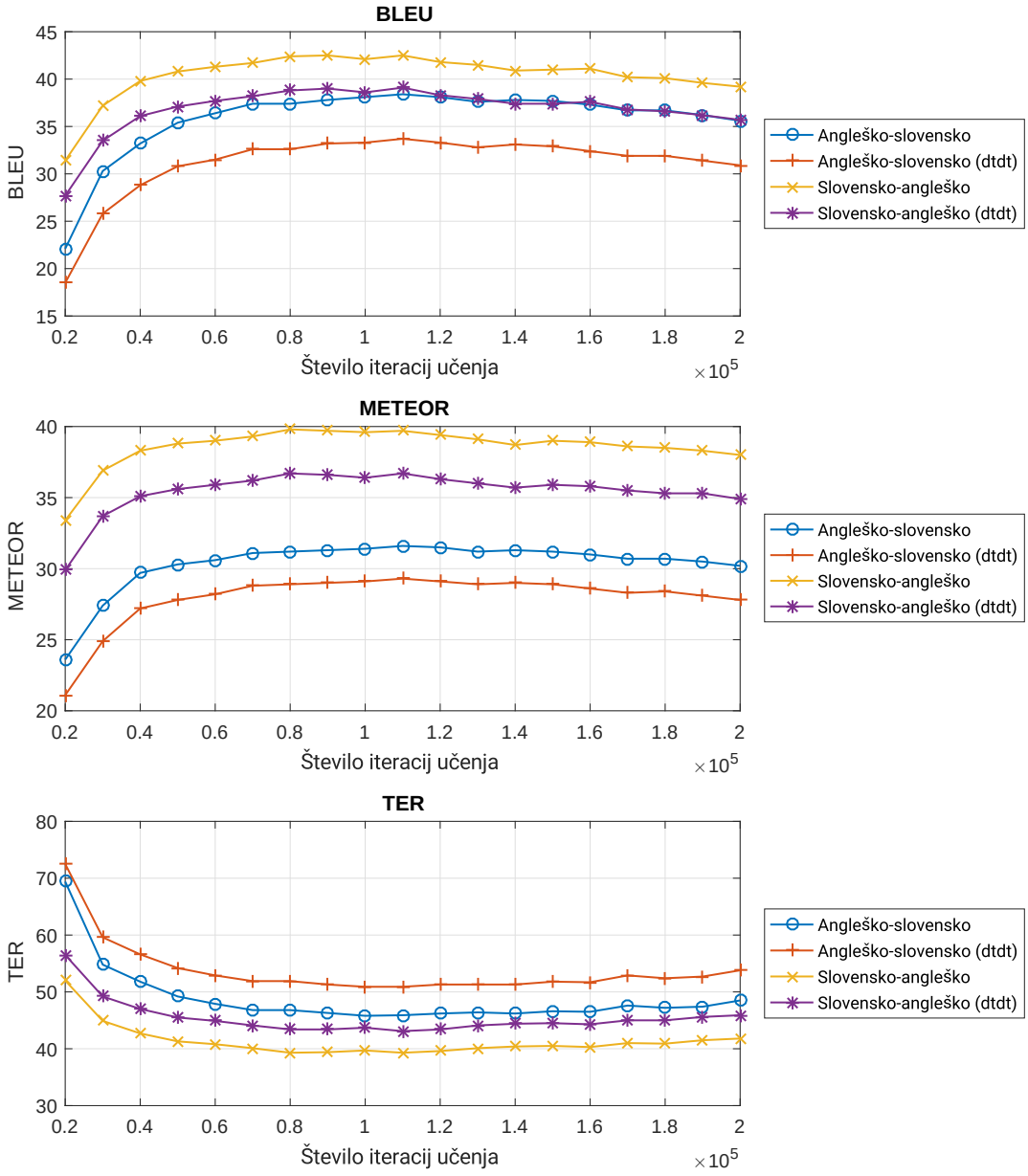
Ideja raziskave v tem članku je primerjava NMT in SMT prevajalnikov, ki temeljijo izključno na poravnanem korpusu, brez uporabe dodatnih jezikovnih ali kakršnihkoli drugih informacij. Več podatkov o SMT prevajalniku je v (Sepesy Maučec in Donaj, 2016), kjer smo uspešnost osnovnega SMT prevajalnika v nadaljevanju še izboljšali z uporabo jezikovno-specif čnih oznak, ki jih pa v tej raziskavi ne vključujemo.

4. Rezultati

4.1. Hitrost učenja

Hitrost učenja modelov je odvisna od hiperparametrov modela in učenja. Pri naših modelih je bila hitrost učenja modela v smeri slovensko-angleško 368 povedi na sekundo, za učenje modelov v smeri angleško-slovensko pa 390 povedi na sekundo. Za učni korpus Europarl je to pomenilo 2,3 epohe na uro (slovensko-angleško) oz. 2,5 epoh na uro (angleško-slovensko). Rezultate na razvojni množici bomo predstavili na prvih 22,5 epohah, za katere je potrebnih približno 9 ur učenja za vsakega izmed obeh modelov.

Hitrost učenja velja za prvotne nastavitve hiperparametrov učenja in modela. Pri spremenjenih hiperparametrih (npr. povečana ali pomanjšana kompleksnost modelov) se čas učenja spremeni. Sprememba dimenzije vektorjev, ki predstavljajo besede ali pa dimenzije skritega stanja, skoraj



Slika 1: Rezultati metrik BLEU, METEOR in TER na razvojni množici za obe smeri prevajanja s sistemom NMT.

premosorazmerno vpliva na čas učenja. Povečanje števila povedi v mini-seriji pri učenju pa pohitri učenje, vendar je vpliv manj izrazit.

4.2. Rezultati SMT

Najprej smo izvedli učenje modelov in prevajanje s sistemom SMT. Rezultati BLEU, METEOR in TER so prikazani v tabeli 2.. Prikazani so rezultati vseh treh metrik za prevajanje v obe smeri, ki jih dobimo z ocenjevanjem pred detokenizacijo in denormalizacijo, kot tudi po detokenizaciji in denormalizaciji (dtdn). Rezultati nam služijo za primerjavo obeh sistemov.

4.3. Rezultati na razvojni množici

Na sliki 1 so prikazani rezultati metrik BLEU, METEOR in TER, ki jih dobimo na razvojni množici pri različnih trajanjih učenja. Optimalne vrednosti so maksimalni rezultati BLEU in METEOR oz. minimalni rezultati TER. Prikazani so rezultati za trajanja od 20.000 do 200.000 iteracij. Rezultati so prikazani za obe smeri prevajanja tako pred detokenizacijo in denormalizacijo kot tudi po detokenizaciji in denormalizaciji (dtdn). Iz rezultatov lahko razberemo, da imamo pri vseh 4 potekih maksimum metrike BLEU pri 110.000 iteracijah učenja, kar ustreza približno 12 epoham učenja.

	Angleško→slovensko	Angleško→slovensko (dtdn)	Slovensko→angleško	Slovensko→angleško (dtdn)
BLEU	40,8	36,0	46,4	42,7
METEOR	33,2	30,9	41,7	38,7
TER	43,0	48,1	37,1	40,9
Δ BLEU	5,3	6,2	2,3	2,9
Δ METEOR	2,1	2,8	-0,1	0,6
Δ TER	-3,0	-4,0	-1,8	-2,4

Tabela 2: Rezultati prevajanja testne množice s sistemom NMT in primerjava z rezultati, dobljenimi s sistemom SMT (Δ).

	Angleško→slovensko	Angleško→slovensko (dtdn)	Slovensko→angleško	Slovensko→angleško (dtdn)
Batch 16	40,4	35,6	45,2	41,7
Batch 32	40,0	35,0	45,6	41,9
Batch 64	40,8	36,0	46,4	42,7
Batch 128	39,4	34,5	45,6	41,8
Batch 256	39,7	34,8	44,4	40,7
EMB 256	40,2	35,3	44,8	41,0
EMB 512	40,8	36,0	46,4	42,7
EMB 1024	40,5	35,8	45,8	42,1
RNN 512	40,8	35,9	45,7	42,0
RNN 1024	40,8	36,0	46,4	42,7
RNN 2048	39,5	34,6	44,8	41,1

Tabela 3: Rezultati metrike BLEU za prevajanje testne množice s sistemom NMT in različnimi hiperparametri učenja oz. hiperparametri modela.

Čeprav lahko za optimalno trajanje učenja pri drugih metrikah opazimo manjša odstopanja, so idealni rezultati še vedno pri ali blizu 110.000 iteracijam učenja.

Na vseh graf h lahko vidimo slabšanje rezultatov pri večjem številu iteracij, kar je posledica prekomernega prilagajanja učni množici. Tako smo za idealni model določili model po 110.000 iteracijah, s katerim smo nato izvajali eksperimente na testni množici. Pripomniti velja, da bo to število iteracij veljalo le v primeru mini-serije z velikostjo 64. Ob večjih oz. manjših velikostih mini-serij se idealno število iteracij sorazmerno pomanjša oz. poveča. Število epoh učenja pa ostaja nespremenjeno.

4.4. Rezultati na testni množici

Rezultati metrik BLEU, METEOR in TER za testno množico so prikazani v tabeli 4.. Prav tako je prikazano izboljšanje rezultatov (pozitivna sprememba BLEU in METEOR oz. negativna sprememba TER) pri vseh metrikah v primerjavi s sistemom SMT.

Pri prevajanju iz slovenščine v angleščino vidimo minimalno poslabšanje rezultata metrike METEOR pred detokenizacijo in denormalizacijo. Vsi ostali rezultati kažejo izboljšanje rezultatov pri prehodu na sistem NMT.

Če kot najbolj uveljavljeno metriko smatramo BLEU, vidimo pomembna izboljšanja v obe smeri prevajanja, in sicer za 6,2 točki v smeri angleščina-slovenščina in 2,9 točke v smeri slovenščina-angleščina. Oba rezultata sta dobljena po detokenizaciji in denormalizaciji.

4.5. Rezultati pri različnih hiperparametrih

V tabeli 4. so prikazani še rezultati, ki jih dobimo z različnimi hiperparametri učenja in modelov. Naš osnovni model je bil učen z dolžinami mini serij 64 (Batch 64), dodali pa še smo modele, naučene z dolžinami serij 16, 32, 128 in 256. V osnovnem modelu smo uporabljali dimenzijo vektorjev za besede 512 (EMB 512), dodali pa še smo modele z dimenzijami 256 in 1024. Zadnji hiperparameter, ki smo ga spreminjali, je dimenzija skrite plasti, ki je bil v osnovnem modelu 1024 (RNN 1024), dodali pa še smo dimenzije 512 in 2048.

Primerjava vseh rezultatov kaže, da je naš osnovni model NMT, ki smo ga zgradili na priporočenih vrednostih hiperparametrov, v vseh primerih tudi v naših eksperimentih najuspešnejši. Pri ostalih modelih opazimo poslabšanja rezultatov do 2 točki BLEU.

4.6. Primerjava z rezultati na testni množici izven domene

Za primerjavo kvalitete prevodov teksta izven domene smo pripravili novo testno množico, ki smo jo dobili iz korpusa IJS-ELAN (Erjavec, 2002). Korpus je prosto dostopen[4] in vsebuje besedila iz različnih virov. Kot besedilo izven domene učnega korpusa smo izbrali leposlovje in sicer roman "1984" G. Orwella. Testno množico smo sestavili iz prvih 1000 segmentov poravnanega korpusa v slovenskem (elan-orwl-sl.xml) in angleškem (elan-orwl-

[4] http://nl.ijs.si/elan/c/

	Angleško→slovensko	Slovensko→angleško
SMT	9,0	10,5
NMT	8,5	10,0
Δ	0,5	0,5

Tabela 4: Rezultati metrike BLEU pri prevajanja v obe smeri za testno množico izven domene učnega korpusa.

en.xml) jeziku. Tako obe množici vsebujeta 16.000 oz. 18.000 besed.

Rezultati prevajanja so prikazani v tabeli 4.6.. Iz rezultatov vidimo, da oba sistema dajeta bistveno slabše rezultate na besedilu, ki ne spada v domeno učnega korpusa. Je pa razlika med obema sistemoma, saj daje prevajalnik SMT v obeh smereh prevajanja rezultate, ki so boljši za 0,5 BLEU točke.

Ta ugotovitev je skladna z ugotovitvami za jezikovni par angleščina-nemščina (Koehn in Knowles, 2017), vendar pa zaradi majhne razlike med našimi rezultati in dejstva, da smo preverili le eno testno množico iz druge domene, lahko zaključimo le, da oba prevajalnika dajeta primerljive rezultate na besedilih izven domene.

4.7. Rezultati glede na dolžine stavkov

Vrednotenje kakovosti prevodov smo ponovili tako, da smo testno množico razdelili na več podmnožic glede na dolžino stavka. Delitev se je izvedla za obe smeri prevajanja ločeno, pri tem pa smo vedno gledali število pojavnic v izvirnem jeziku. Delili smo na množice, ki vsebujejo:

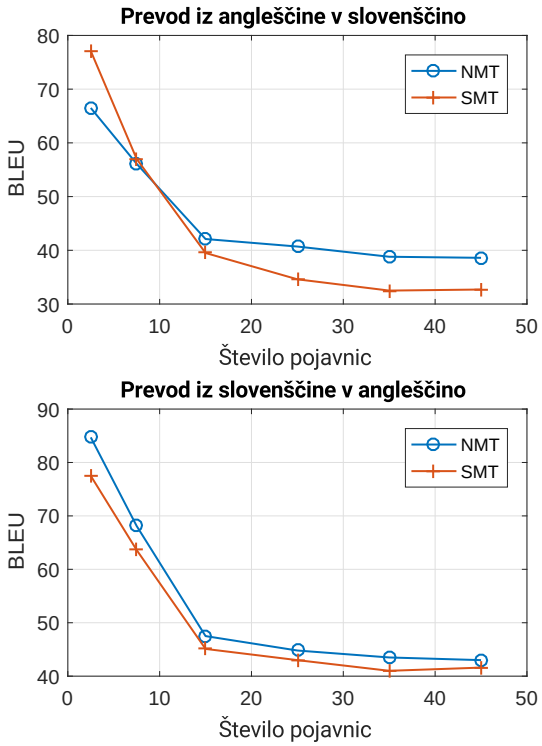
- od 1 do 5 pojavnic,
- od 6 do 10 pojavnic,
- od 11 do 20 pojavnic,
- od 21 do 30 pojavnic,
- od 31 do 40 pojavnic in
- 41 ali več pojavnic.

Rezultati kvalitete prevodov so prikazani na sliki 2. Iz rezultatov vidimo, da oba sistema dajeta boljše rezultate pri krajših stavkih in da celo prevajalnik SMT pri prevajanju iz angleščine v slovenščino daje boljše rezultate v prvih dveh množicah (stavkih do 10 pojavnic). V ostalih primerih pa vidimo, da daje prevajalnik NMT boljše rezultate.

5. Zaključek

Prve raziskave uporabe nevronskih omrežij za strojno prevajanje so pokazale, da lahko z NMT dosežemo boljše prevode, kot pa s klasičnimi statističnimi sistemi.

Prišli smo tudi do zaključka, da je izboljšanje kvalitete prevodov bolj izrazito pri prevajanju iz angleščine v slovenščino kot pri prevajanju v obratni smeri. Ta izsledek je posebej pomemben, saj prevajanje iz morfološko enostavnejših v morfološko kompleksnejše jezike velja kot zahtevnejša smer prevajanja. Zato so izboljšave v tej smeri bolj pomembne, še posebej za slovenski prostor.



Slika 2: Rezultati metrike BLEU glede na število pojavnic v izvirnem jeziku pri prevajanju testne množice v obe smeri s sistemoma SMT in NMT.

Če primerjamo naše ugotovitve glede časa učenja modelov z drugimi raziskavami, vidimo, da dobimo optimalne rezultate pri primerljivem številu epoh. To pomeni, da bo optimalno trajanje učenja odvisno od velikosti učnega korpusa.

V nadaljevanju bomo raziskave usmerili v vključevanje morfoloških informacij v modele prevajanja. Dodatno želimo preučiti uporabo klasičnih jezikovnih modelov za ponovno ocenjevanje hipotez prevajalnika, saj takšnega modela ni v osnovni zasnovi nevronskega omrežja. Preučevali bomo tudi adaptacijo modelov prevajanja pri prehodu v novo domeno besedil.

6. Zahvala

Raziskovalni program št. P2-0069, v okviru katerega je nastala ta raziskava, je sof nancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

7. Literatura

Dzmitry Bahdanau, Kyunghyun Cho in Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo in Marcello Federico. 2018. Neural versus phrase-based MT quality: An in-depth analysis on english-german and english-french. *Computer Speech & Language*, 49:52–70.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau in Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. V: *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*.

Jonathan H. Clark, Chris Dyer, Alon Lavie in Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. V: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, str. 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomaž Erjavec. 2002. Compiling and using the ijs-elan parallel corpus. *Informatica*, 26:299–307.

Marcin Junczys-Dowmunt, Tomasz Dwojak in Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. V: *International Workshop on Spoken Language Translation*, IWSLT '16.

Philipp Koehn, Franz Josef Och in Daniel Marcu. 2003. Statistical phrase-based translation. V: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, str. 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn in Rebecca Knowles. 2017. Six challenges for neural machine translation. V: *The First Workshop on Neural Machine Translation*, str. 28–39, Vancouver, Canada, August. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. V: *Conference Proceedings: the tenth Machine Translation Summit*, str. 79–86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, prva izd.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow in Alexandra Birch. 2017. Deep architectures for neural machine translation. V: *Proceedings of the Second Conference on Machine Translation*, str. 99–107. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry in Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. V: *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, str. 65–68, Valencia, Spain, April. Association for Computational Linguistics.

Mirjam Sepesy Maučec in Gregor Donaj. 2016. Morphosyntactic tags in statistical machine translation of highly inf ectional language. V: *Proceedings of the artificial intelligence and natural language conference (AINL FRUCT)*, str. 99–102.