*Mirjam Sepesy Maučec**

Evalvacija avtomatskih prevodov v projektu SUMAT

POVZETEK
V članku bomo predstavili zaključno fazo projekta SUMAT, ki smo ga pregledno predstavili na konferenci PAZU leta 2011. V projektu smo razvili avtomatske prevajalnike podnapisov za 14 jezikovnih parov. Prevajanje podnapisov je kompleksna naloga, ki se v veliki meri razlikuje od drugih oblik prevajanja. Avtomatski prevajalniki, ki smo jih razvili v projektu, so danes v obliki spletne storitve na voljo profesionalnim prevajalcem kot pripomoček pri njihovem delu. V prispevku se bomo posvetili obsežni evalvaciji kvalitete prevodov, ki smo jo opravili ob sodelovanju profesionalnih prevajalcev. Izpostavili bomo najpogosteje odkrite napake v prevodih in primerjali kvaliteto prevodov za različne jezikovne pare.

Ključne besede: statistično strojno prevajanje, podnapisi, evalvacija, kvaliteta, produktivnost.

1. Uvod

Obseg multimedijskih vsebin, ki jih ponujajo različni viri, raste izredno hitro. Podnaslavljanje je priljubljen način za posredovanje tujejezičnih multimedijskih vsebin v veliko evropskih državah in za večino žanrov[1]. Vendar se podnaslavljanje hkrati srečuje z določenimi problemi, kot so kratki časovni roki, visoki stroški in z njimi povezana vprašljiva kvaliteta podnapisov. Na podlagi tega se je razvila ideja projekta SUMAT, vključiti tehnologijo statističnega strojnega prevajanja v prevajalski proces in s tem olajšati delo prevajalcu, predvsem pa skrajšati čas, potreben za izdelavo prevoda.

Statistično strojno prevajanje se je skozi številne raziskave pokazalo kot najučinkovitejši pristop k avtomatskemu prevajanju. Razloga za njegov uspeh sta dva. Prvi je velika količina jezikovnega gradiva, ki je na voljo v elektronski obliki in predstavlja osnovo statističnega prevajanja. Drugi razlog pa je, da za razvoj prevajalnika ni potrebno poglobljeno znanje o jezikih, med katerimi prevajamo. Zahtevnost strojnega prevajanja je odvisna od žanra in domene besedil, ki jih prevajamo. Sprva je kazalo, da je prevajanje podnapisov, s katerim smo se ukvarjali v projektu SUMAT, za statistično strojno prevajanje zelo hvaležno področje, saj so povedi praviloma kratke. Toda podnapisi prinašajo tudi številne probleme. Ker gre za podnaslavljanje video vsebin, so nekateri problemi blizu problemom govorjenega jezika. Še večji problem pa je, da so se mora dolžina besedila podrejati dolžinam podnapisa, kar privede do številnih postopkov krajšanja izvornega besedila.

Projekt SUMAT smo aprila 2014 zaključili. V nadaljevanju prispevka predstavljamo njegove rezultate. Razvili smo prevajalnike za 14 jezikovnih parov oz. smeri prevajanja. Vključeni so bili naslednji jeziki: angleščina, španščina, francoščina, nemščina, portugalščina, švedščina, srbščina in slovenščina. V zaključni fazi projekta nas je zanimala predvsem kvaliteta prevodov, ki jih generirajo prevajalniki, in produktivnost prevajalca, če le-ta pri svojem delu uporablja avtomatske prevode.

*Fakulteta za elektrotehniko, računalništvo in informatiko UM, Smetanova 17, 2000 Maribor
E-naslov: mirjam.sepesy@um.si

2. Gradivo v projektu SUMAT

Pomemben korak pri izdelavi sistema za strojno prevajanje podnapisov je izdelava vzporednega korpusa podnapisov, potrebnega za učenje prevajalnika. Izvorno gradivo so iz svojih arhivov posredovala tri mednarodna podjetja, ki so specializirana za prevajanje podnapisov. Podjetja so zagotavljala, da gre za visoko kvalitetne podnapise, saj je vsak prevod pregledan na več nivojih, preden je posredovan naročniku. Poleg datotek s prevodi smo zbirali tudi samo enojezične datoteke, saj je pomembna komponenta prevajalnika tudi jezikovni model. Datoteke so pripadale različnim žanrom, kot so dnevno-informativne oddaje, serije, dokumentarni filmi ipd. Gradivo smo na koncu dopolnili še z materialom, ki smo ga zbrali iz prosto dostopnih spletnih virov[2]. Količina zbranega gradiva je zelo variirala glede na jezikovni par. Največ gradiva smo zbrali za par angleščina – nemščina, najmanj pa, skladno s pričakovanji, za par slovenščina – srbščina. Gradivo je potrebno ustrezno obdelati, preden ga lahko uporabimo za učenje prevajalnikov. Predpriprave izvornega gradiva vključujejo naslednje korake: pretvorbe v enoten format in enotno kodiranje znakov, identifikacijo jezika v datotekah, poravnavanje datotek, tokenizacijo, razcep po povedih in poravnavanje povedi ali podnapisov [3, 4].

3. SUMAT prevajalniki

Prevajalniki SUMAT so statistični prevajalniki s klasično strukturo. Kot osnovna enota prevajanja se običajno uporablja poved, v projektu pa je bilo opravljenih nekaj preliminarnih testov, ki so vodili v odločitev, da kot osnovno enoto uporabimo podnapis. Vsak prevajalnik sestavljajo 3 komponente: model prevajanja, model preurejanja in jezikovni model. Prvi dve komponenti smo zgradili s pomočjo Mosesovih skript, ki smo jih uporabili na poravnanem gradivu [5]. Jezikovni model pa

smo zgradili z orodjem SRI LM [6] in pri tem uporabili enojezične korpuse.

Učni vzporedni korpus izhaja iz različnih virov, zato smo modele prevajanja in preurejanja gradili za vsak vir posebej in jih potem sestavili po principu adaptacije na domeno. Kot vzorec ciljne domene smo uporabili razvojno množico, ki je obsegala 2000 podnapisov.

Uporabili smo 3-gramski jezikovni model z Good-Turingovim odštevanjem in sestopanjem po Katz. Uteži komponent prevajalnika smo optimirali po MERT [7] na razvojni množici 2000 podnapisov, ki smo jo uporabili tudi za sestavljanja komponent modelov prevajanja in preurejanja.

Prevajalnike smo v nadaljevanju na različne načine še izboljševali. Za določene jezikovne pare se je pokazalo, da je smiselno gradivo dopolniti z obliko-skladenjskimi lastnostmi besed. To je veljalo predvsem za visoko pregibne jezike, med katere sodita srbščina in slovenščina.

4. Evalvacija prevodov

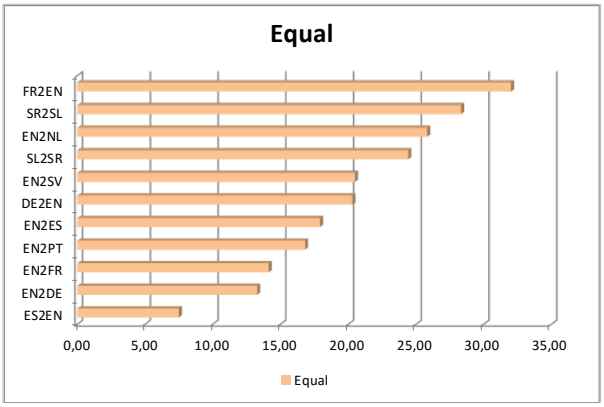
Evalvacijo avtomatskih prevodov smo izvedli v dveh fazah. Cilj evalvacije v prvi fazi je bil izboljšati sistem, v drugi fazi pa oceniti produktivnost prevajalskih procesov. V obe fazi evalvacije so bili vključeni profesionalni prevajalci. Predstavniki prevajalskih podjetij so najprej pripravili testne vzorce za evalvacijo. Vzorci so bili sestavljeni iz dokumentov realnega okolja. Vključevali so podnapise filmov, pogovornih oddaj, dokumentarcev ipd. Dokumente smo najprej prevedli z uporabo ustreznih strojnih prevajalnikov. Potem smo jih posredovali prevajalcem, ki so:

- popravili prevode do "običajnega" standarda kakovosti,
- ocenili kvaliteto prevoda: od 1 (neuporaben prevod) do 5 (brezhiben prevod),
- skladno s podano taksonomijo označili pogoste napake (v prevodih, ocenjenih s 3 ali več) in
- izpolnjevali vprašalnik, v katerem so podali tudi predloge za izboljšanje kakovosti prevodov.

Dokumente s popravljenimi prevodi smo uporabili kot referenčne dokumente, s katerimi smo primerjali izvorne avtomatske prevode in ocenjevali, kako podobni oz. različno so si.

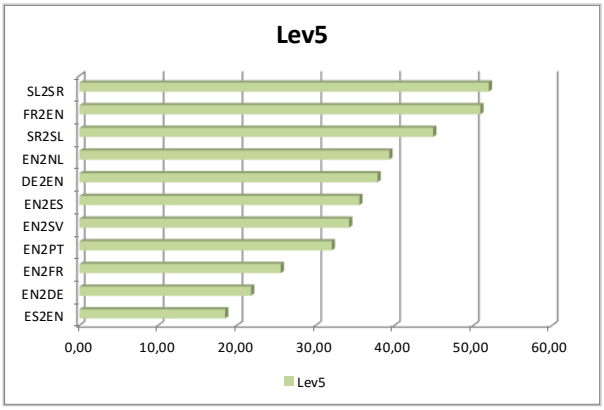
4.1 Avtomatska evalvacija

Najprej smo izvedli avtomatsko evalvacijo, v kateri smo prevajalnike vrednotili z metrikami avtomatske evalvacije. Zanimal nas je tudi delež podnapisov, ki se 100% ujemajo z referenco (Equal). Rezultati evalvacije so prikazani na sliki 1. Najboljši rezultat je bil dosežen za prevajanje iz francoščine v angleščino. Pri vrhu je tudi prevajanje med slovenščino in srbščino. Najslabši rezultat je bil dosežen za prevajanje iz angleščine v nemščino in iz španščine v angleščino. Znano je, da je strojno prevajanje v nemščino za avtomatske prevajalnike trd oreh, medtem ko je bil slab rezultat za španščino veliko presenečenje.



Slika 1. Rezultati metrike Equal za izbrane jezikovne pare.

Zanimal nas je tudi delež podnapisov, pri katerih je, da dosežemo ujemanje, potrebnih največ 5 korakov preurejanja (Lev5). Rezultati so na sliki 2. Vrstni red jezikovnih parov se je nekoliko spremenil, čeprav najboljši in najslabši pari ostajajo isti.



Slika 2. Rezultati metrike Lev5 za izbrane jezikovne pare.

4.2 Rangiranje prevodov

Prevajalci so vsak podnapis v strojnem prevodu rangirali glede na kvaliteto oz. zahtevnost popravljanja. Pri tem smo uporabili skalo, definirano v "WMT 2012 Shared Task on MT quality estimation", po kateri je vsak podnapis rangiran z vrednostjo od 1 do 5. Ocena 1 pomeni neuporaben in nerazumljiv prevod, ocena 5 pa brezhiben prevod, ki ne potrebuje nobenega popravka. 21% prevodov je dobilo oceno 3, 26% prevodov oceno 4 in 31% prevodov oceno 5. Oceno 1 ali 2 je dobilo le 22% prevodov.

4.3 Klasifikacija napak

Prevajalci so napake v prevodih klasificirali v razrede:

- agr: slovnično neujemanje,
- miss: manjka polnopomenska beseda ali odsek,
- order: napačni vrstni red besed,
- phrase: večbesedna zveza napačno prevedena kot ločene, nepovezane besede,
- cap: napačen zapis velike/male črke,

- punc: napačno ločilo,
- spell: napačno črkovanje,
- length: predolg prevod glede na omejeno dolžino podnapisa,
- trans: napačen prevod.

Izkazalo se je, da je največ napak pripadalo razredu trans, torej napačni prevod. Veliko napak se je uvrstilo tudi v razreda agr in miss. Na osnovi klasifikacije napak smo sistemu dodali nekaj korakov postprocesiranja in tako izboljšali prevajalnike.

4.4 Subjektivne ocene prevajalcev

Prevajalci so na koncu izpolnili še vprašalnik, v katerem so izrazili svoje subjektivno mnenje o kvaliteti prevodov in podali ideje za popravke. Če se je izkazalo, da so popravki izvedljivi (to pomeni, da jih lahko implementiramo kot dodaten korak avtomatskega popravljanja prevodov), smo jih upoštevali. Reševanje določenih napak je bilo pogojeno z uporabo dodatnih jezikovnih virov, ki jih zaradi komercialne naravnanosti projekta nismo dodajali, saj je za vsak uporabljen vir potrebno dovoljenje za komercialno rabo.

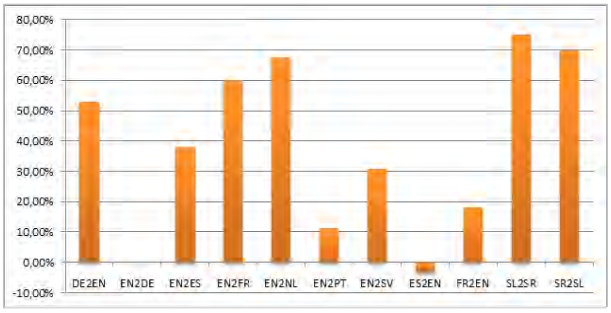
5. Merjenje produktivnosti

V drugi fazi evalvacije nas je zanimalo, ali avtomatski prevodi skrajšajo čas tvorjenja prevodov. Primerjali smo čas, ki ga potrebuje prevajalec, če neposredno prevaja dokument iz izvornega v ciljni jezik, s časom, ki ga potrebuje za popravljanje strojnih prevodov. Menimo, da je tovrstna primerjava zelo jasen in neposreden pokazatelj uporabnosti sistemov strojnega prevajanja.

Pred izvedbo drugega dela evalvacije smo v sistem prevajanja vpeljali še dodaten postopek filtriranja strojnih prevodov, v katerem smo izločili prevode slabe kvalitete. V razdelku 4.2 smo opisali rangiranje prevodov glede na kvaliteto. Na osnovi teh ocen smo učili binarni klasifikator, ki prevode klasificira v dva razreda, v razred dobrih in razred slabih prevodov. Za učenje klasifikatorja in klasifikacijo smo uporabili orodje QuEst, ki je podrobneje opisano v [9]. Strojne prevode, ki jih je klasifikator označil kot slabe, smo odstranili, kar je pomenilo, da jih mora prevajalec tvoriti iz podnapisa v izvornem jeziku. Ta korak smo dodali zato, ker je popravljanje slabih prevodov bolj zamudno kot neposredno prevajanje izvornega dokumenta.

Za vsak jezikovni par oz. za vsako smer prevajanja sta sodelovala dva profesionalna prevajalca. Vsak prevajalec je tvoril tri datoteke. V prvi je prevajal iz izvornega jezika, v drugi je popravljal strojne prevode in v tretji je popravljal filtrirane strojne prevode. Pri tem je vsak prevajalec uporabil programsko okolje, ki ga tudi sicer uporablja pri svojem delu. Razlika je bila le v tem, da se je v ozadju meril čas efektivnega dela. Rezultati so zbrani na sliki 3. Vidimo, da je najbolj učinkovito popravljanje avtomatskih prevodov jezikovnega para slovenščina – srbščina. Razlog je najverjetneje velika podobnost jezikov. Produktivnost se je izrazito izboljšala tudi pri prevajanju iz angleščine v francoščino in iz angleščine v nizozemščino. Uporaba strojnih prevodov pri tvorjenju prevodov željene kakovosti se je izkazala kot neučinkovita pri prevajanju iz španščine v angleščino in pri prevajanju iz angleščine v nemščino. Slab rezultat za ta jezikovna para je bil, glede na rezultate evalvacije v prvi fazi, pričakovan.

Omenimo še en vidik uporabe strojnih prevodov. Za prevajalce popravljanje ni najbolj »všečen« proces in nekateri do tega čutijo določen odpor. V tem oziru so lahko prikazani rezultati do neke mere popačen prikaz, subjektivna percepcija strojnega prevajanja profesionalnih prevajalcev.



Slika 3. Rast produktivnosti pri uporabi strojnih prevodov v prevajalskem procesu.

6. Zaključek

V članku smo predstavili rezultate projekta SUMAT, katerega namen ni bil strojno tvoriti brezhibne prevode, ampak prevajalcu ponuditi prevode, ki mu skrajšajo čas, potreben za prevajanje. Glede na rezultate evalvacije smo zaključili, da so strojni prevodi lahko učinkovit pripomoček prevajalcev. Zaenkrat je popravljanje strojnih prevodov še relativno nepoznan postopek med prevajalci. Da bi bilo strojno prevajanje pozitivno sprejeto med njimi, bi bilo treba učenje tehnik popravljanja vključiti tudi v učne procese v prevajalstvu. V tej smeri potekajo aktivnosti v smislu izvajanja tečajev popravljanja na različnih univerzah Evrope.

Zahvala

Avtorica članka se za sodelovanje pri projektu zahvaljuje sodelavcem Laboratorija za digitalno procesiranje signalov, FERI, UM, ki so del slovenske skupine v projektu SUMAT: Marko Presker, Matej Rojc, Darinka Verdonik, Damjan Vlaj in Danilo Zimšek. Zahvala gre tudi koordinatorici projekta Arantzi del Pozo, ki nas je povabila k sodelovanju.

Literatura

1. European Commision (2010). Audiovisual Media Services Directive (AVMSD – 2010/13/EU). Official Journal of the European Union, 10 March 2010.
2. Tiedemann, J. (2009). News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: N. Nicolov, K. Bontcheva, G. Angelova,, R. Mitkov (eds.): Recent Advances in Natural Language Processing (vol. V) (pp. 237--248). Amsterdam, Philadelphia: John Benjamins.
3. Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages. In: Proceedings of the RANLP 2005 (pp. 590--596).

4. Maučec, M. S., Presker, M., Zimšek, D., Rojc, M., Vlaj, D., Verdonik, D., Kačič, Z. Izdelava slovensko-srbskega vzporednega korpusa podnapisov za razvoj strojnega prevajanja v projektu SUMAT. Zbornik Osme konference Jezikovne tehnologije, oktober 2012, str. 167-172.

5. Moses - statistical machine translation system, http://www.statmt.org/moses/, (dostop 24.10.2011).

6. Stolcke, A., 2002. SRILM: an extensible language modeling toolkit. Proceedings of the Int. Conf. on Spoken Language Processing, 901–904.

7. Och, F. J., 2003. Minimum error rate training in statistical machine translation, Zbornik 41st Annual meeting of the Association for Computational Linguistics, Sapporo, Japan.

8. Papineni, K., Roukos, S., Ward, T., Zhu. W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, 311–318.

9. Specia, L., Shah, K., de Souza, J. G., Cohn, T., Kessler, F. B., 2013. QuEst–a translation quality estimation framework. Zbornik 51st Annual meeting of the Association for Computational Linguistics : System Demonstrations, 79–84.