Prevajanje naravnih jezikov upoštevajoč drevesa izpeljave

Jernej Vičič, Andrej Brodnik
Pedagoška fakulteta Koper
Univerza na Primorskem
Cankarjeva 5, 6000 Koper, Slovenija
jernej.vicic@pef.upr.si
andrej.brodnik@pef.upr.si

Prevajanje naravnih jezikov upoštevajoč drevesa izpeljave

Abstract

The article describes a method that enhances basic SMT (Statistical Machine Translation) idea with the introduction of source language formal grammar and target language formal grammar. New tools (grammars) enable usage of additional language knowledge. Along with basic word and word phrase translation (basic SMT translation), this method uses techniques and knowledge based on other language theory levels. Sentences are presented by parse tree of formal grammar. Method was applied to an example of word classes, formal grammar symbols were presented as word classes. Method was empirically tested using a bilingual, parallel, annotated corpus IJS-ELAN [1]. Corpus language pair is Slovene – English.

Povzetek

Članek predstavlja metodo, ki dopolnjuje osnovno idejo statističnega strojnega prevajanja z uvajanjem formalnih slovnic izvornega in ciljnega jezika in prevajanjem na ravni dreves izpeljave formalnih slovnic. Novi orodji nam omogočata upoštevanje dodatnega znanja o posameznem jeziku in jezikovnem paru. Poleg osnovnega statističnega prevajanja besed in besednih zvez, omogoča nov pristop tudi upoštevanje znanja ostalih jezikovnih ravni. Povedi so predstavljene z drevesi izpeljave formalne slovnice. Metoda je preizkušena na primeru uvajanja formalne slovnice besednih oznak. Empirično je preverjena s pomočjo dvojezičnega označenega korpusa IJS-ELAN. Jezikovni par korpusa je slovenščina in angleščina.

Uvod

Članek predstavlja nov pristop k statističnem strojnem prevajanju (Statistical Machine Translation - SMT) naravnih jezikov. Metoda, predstavljena v nadaljevanju, pri prevajanju, poleg samih besed, upošteva višje povedne značilke, kot na primer besedno vrsto, ki ji določena beseda pripada. Značilka definira simbol formalne slovnice. Delo je sestavljeno iz osnov strojnega prevajanja, nadaljevanje predstavi osnovne ideje, ki so vodile k snovanju metode. Struktura članka po poglavjih:

1. poglavje predstavlja osnovne pojme strojnega prevajanja ter motivacije za razvoj te panoge.
2. poglavje predstavlja osnovne pojme statističnega strojnega prevajanja.
3. poglavje predstavlja korpus IJS-ELAN, osnovo za empiričnem preverjanje predstavljene metode.
4. poglavje predstavlja samo metodo in motive, ki so privedli k snovanju predstavljene metode.
5. poglavje predstavlja metodo za samodejno učenje pravil prevajanja besednih vrst v povedih na osnovi verjetnosti.
6. poglavje naniza nekaj primerov tako samega korpusa, kot tudi prevodov.

1. Strojno prevajanje naravnih jezikov

Samodejno prevajanje iz enega jezika v drugi imenujemo strojno prevajanje (Machine Translation – MT). Strojno prevajanje na splošnih domenah trenutno še ne dosega ravni drugih računalniških področij. Za opravljanje predstavljenih opravil mora računalnik »poznati« izvorni ter ciljni jezik. Poleg osnovnih jezikovnih pravil kot so sintaksa, gramatika, sinonimi besed oziroma fraz v obeh jezikih, mora sistem poznati še semantiko oziroma pomen prevajanih sporočil.

2. Statistično strojno prevajanje naravnih jezikov

Ročna izdelava pravil ali predlog je dolgotrajen proces, ki zahteva tudi velike denarne vložke. Izdelava takšnih sistemov ni finančno obvladljiva za manjše jezike oziroma za manj uporabljane jezikovne pare.
Alternativa ročni izdelavi pravile je, da prepustimo računalniku, da se sam nauči sintaktičnih in semantičnih pravil, tako da pregleda velike količine dvojezično-vzporednih besedil. Besedila morajo biti natančni prevodi iz izvornega jezika v ciljni jezik. Povedi morajo biti poravnane.
V zadnjih letih se je pojavilo že kar nekaj uspešnih projektov, ki uporabljajo statistično obdelavo podatkov za samodejno sestavljanje prevajalske baze iz vzporednih, poravnanih, dvojezičnih besedil. Tehnika statistične obdelave besedil je primerna za velike

količine besedil, ki edina ponujajo dovolj informacij o nekem jeziku oziroma dovolj informacij za prevod med dvema jezikoma. Naravni jezik je zelo kompleksna tvorba in le velika količina besedil lahko zajame dovolj pravil in izjem, ki opišejo pravila za prevajanje ter tudi vse izjeme, ki se pri samem prevajanju upoštevajo.
Statistično strojno prevajanje, Statistical Machine Translation (SMT), je bilo do sedaj le redko uporabljano. Statistični pristop je bil na področju MT le redko uporabljan, razloge pa gre iskati v precej zahtevni matematični podlagi, ki je potrebna za uporabo in razvoj statističnih MT metod. V zadnjih letih pa se ravno ta veja najhitreje razvija in dosega že kar zavidljive rezultate.

3. Korpus IJS ELAN in MSD oznake

V zadnjem času se je pojavilo kar nekaj kakovostnih vzporednih dvojezičnih označenih korpusov, ki vključujejo slovenski jezik [1], [2]. Zbrani in uniformno kodirani teksti (korpusi) že sami zase predstavljajo uporabno orodje. Z dodajanjem lingvističnih označb (primer MSD: označujejo vsako besedo z oznakami jezikovne teorije kot so besedna vrsta, sklon, število, spol, …) postanejo takšne zbirke neprecenljiva in trdna osnova za raziskovanje jezika ter pravil med jezikovnimi pari. Učljivi označevalci besednih oznak [3] so v zadnjem času napredovali do stopnje, ko jih lahko označimo za zrele in primerne za uporabo. Primer takšnega korpusa je IJS-ELAN [1], kodiran po standardu TEI P4, [2], zapisan z odprtimi standardi in samodejno označen s prosto dostopnimi orodji. Korpus obsega več različnih tekstov, ki naj bi čimbolj natančno povzeli pravila in posebnosti jezikovnega para slovenščina – angleščina.

4. Metoda prevajanja na osnovi dreves izpeljave

Do sedaj so se vse metode SMT osredotočale na besede ter besedne zveze ter s pomočjo velikih količin podatkov poskušale poiskati določena pravila prevajanja med besedami ter besednimi zvezami izvornega ter ciljnega jezika.
Označevanje elektronskih korpusov je doseglo že visoko raven, označbe niso koristne samo za lajšanje iskanja pravil v velikih količinah besedil. Označbe bi lahko s pridom izkoristili tudi pri učenju pravil za prevajanje. Iskanje višjih besednih in stavčnih značilk,, kot so iskanje besedne vrste, iskanje ostalih prilastkov posameznih besed, iskanje vloge besed v stavkih, je doseglo visoko raven, rezultate lahko s pridom uporabimo.
Pri opisu primera uporabe metode se bomo zaradi večje preglednosti omejili na uporabo informacije o besedni vrsti za posamezno besedo. Metodo lahko enostavno nadgradimo z upoštevanjem dodatnih prilastkov (npr. spol, sklon, število, …) posameznih besed, sama implementacija bo tako naravnana.

Prevajalni sistemi privzemajo, da se vloga besednih vrst pri prevodu ohranja, besedne vrste naj bi bile tako v izvornem, kot v ciljnem jeziku enakovredne, spreminjala naj bi se samo vsebina. Ta trditev pogosto velja za sorodne jezike ter za besedila s togim stilom pisanja. Prevajanje takšnih jezikovnih parov je lažje in rezultati so veliko boljši kot pri prostih parih.
Pri tujih si jezikih je ta osnova zgrešena, pogosto se pri prevajanju zamenja vrstni red besed, porajajo se nove besede, določene izginjajo.
Očitna rešitev bi bila, če bi pravila prevajanja izbranih besednih vrst zbrali ter jih podali računalniku. Preprosto povedano, za vsako besedno vrsto bi morali postaviti pravilo, ki jo v odvisnosti od konteksta prevaja v ciljno besedno vrsto. Seveda bi lahko iskanje enostavno razširili tudi na ostale jezikovne ravnine ter dodatne besedne prilastke.
Takšno početje bi bilo zamudno, rezultati pa vprašljivi, pravila želimo določiti samodejno, s čim manjšo mero človeškega vmešavanja. **Sistem naj se sam nauči pravil**.
Pravila za prevajanje meta-besed (dreves izpeljave) v določenem kontekstu zgradimo iz dovolj velikih količin primerov s pomočjo zakonov matematične verjetnosti. Če se določen simbol formalne slovnice v izbranem kontekstu velikokrat prevaja v nek simbol, obstaja velika verjetnost, da bo to pravilo primerno za večino prevodov tega simbola v izbranem kontekstu.
Za prevode vseh simbolov, ki jih ponuja korpus, določimo verjetnost prevodov s številom pojavljanj v danem kontekstu.
Pri prevajanju izberemo pravilo, ki ustreza simbolu in kontekstu v izvorni povedi ter po pravilu prevedemo v simbol ciljnega jezika.
Oglejmo si metodo na primeru:

1) Izvorno poved prevedemo na morfosintaktične opise, dobili smo besedo "*FbesedaI*" izvornega jezika na nivoju formalne slovnice besednih vrst.
2) Za *FbesedaI* postavimo drevo izpeljave formalne slovnice v izvornem jeziku. Izberemo drevo izpeljave *FdrevoI,* ki z izbrano metodo obhoda enolično določa besedo *FbesedaI.*

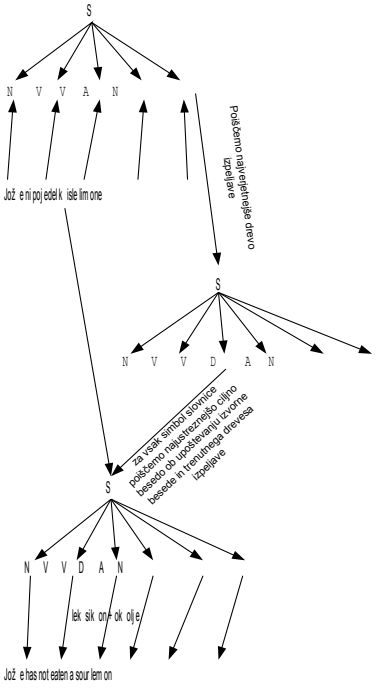
Slika1: Primer drevesa izpeljav v izvornem in ciljnem jeziku

3) Poiščemo najverjetnejše drevo izpeljave v ciljnem jeziku. Verjetnosti za posamezne pare dreves izpeljave smo določili v fazi učenja na podlagi števila pojavitev. Izbrali smo drevo izpeljave *FdrevoC* v ciljni slovnici.
4) *FdrevoC* je osnova za gradnjo ciljne povedi. Posamezne simbole *FdrevoC* prevedemo s pomočjo povezanih izvornih simbolov, vhodnega leksikona in konteksta *FdrevoC* v ciljne besede. Dobili smo ciljno poved.

Formalna slovnica: S pomočjo dvojezičnega poravnanega korpusa sestavimo dve formalni slovnici; formalno slovnico izvornega in formalno slovnico

ciljnega jezika. Formalni opis slovnice izvornega jezika je enak kot pri ciljnem jeziku, različna je le vsebina množic.
Oglejmo si slovnico s pomočjo formalnih opisov, kot opisano v [6]:

- Γ končna množica simbolov (neterminali), v našem primeru oznake besed
- Σ končna množica črk abecede (terminali), v našem primeru besede izvornega ali ciljnega jezika
- Množici Σ in Γ sta si tuji
- Produkcijska pravila za slovnico ne zapišemo v eksplicitni obliki, kot je v navadi. Izražena so z množico dreves izpeljave za meta-besede.



Slika2: Primer prevoda povedi na osnovi prevoda drevesa izpeljav v izvornem in ciljnem jeziku

- Vsako poved izrazimo z drevesom izpeljave, izbran zapis drevesa izpeljave nam nedvoumno določa poved. Lahko izberemo prefix, infix ali postfix obhod.
- Meta-besede $mb \in \Gamma^*$ so s simboli predstavljene povedi iz korpusa, lahko so besedoslovne oznake (samostalnik, glagol, pridevnik, …), lahko so stavčni členi (osebek, povedek, …), lahko stavčne fraze (samostalniška fraza, glagolska fraza, …), ….

Tako predstavljena formalna slovnica posplošuje opis dreves opisov povedi (parse trees)[8]. Predstavitev je splošnejša, zajema tudi takšne oblike dreves izpeljave.
Zdaj imamo pripravljeno osnovo za vpeljavo prevajalne metode.
Prevajanje: Za poved v izvornem jeziku, namenjeno prevajanju, zgradimo drevo izpeljave, ki predstavlja ustrezno meta-besedo v izvorni formalni slovnici. Drevo izpeljave v izvorni formalni slovnici prevedemo v najverjetnejše drevo izpeljave v ciljni formalni slovnici. To drevo predstavlja kontekst, ki omogoča usmerjeno preiskovanje najustreznejših besed ciljnega jezika. Drevo izpeljave prevedemo s pomočjo verjetnostnih modelov, izbor in lastnosti modelov so predstavljeni v naslednjem poglavju. Posamezne simbole meta-besede ciljne formalne slovnice zamenjujemo z najustreznejšimi besedami ciljnega jezika. Za prevajanje potrebujemo še leksikon jezikovnega para izvornega in ciljnega jezika. Za večje jezikovne pare obstajajo odlični leksikoni v elektronski obliki, obstajajo pa tudi mnoge metode za gradnjo leksikona iz dvojezičnega korpusa [4], [5].
Podobno zgradimo še leksikon meta-besed. Oba leksikona morata biti urejena tako glede na verjetnost prevodov posameznih besednih parov kot glede na posamezen kontekst, kjer se je beseda nahajala v korpusu.

5. Predstavitev metode na primeru

Iz izbranega korpusa (v našem primeru IJS-ELAN [1]) izluščimo zapise za besedne vrste za vse besede združene po povedih. Korpus je poravnan po povedih, tako predelan korpus razpade na meta-besede v prostoru opisov besednih vrst, glej primeri, poglavje 6. Tako pripravljen korpus je osnova za učenje pravil prevajanja.
Sama metoda računanja najverjetnejših preslikav dreves izpeljave je povzeta po metodi IBM1 [4], ki postavi verjetnosti za prevode posameznih besed. V našem primeru. Algoritem prične z apriorno verjetnostjo vseh produkcijskih parov, ki je enaka $\frac{1}{|\Theta|}$, kjer je Θ množica vseh besed ciljnega jezika. Ta verjetnost se veča za vse ciljne produkcije, kjer se v korpusu pojavi izvorna beseda.
Tako naučen sistem je pripravljen za prevajanje na ravni zapisov besednih oznak (povedi prevajamo kot simbole naše formalne slovnice).
Pri prevajanju najprej predelamo vhodno poved, besede nadomestimo z ustreznimi oznakami za besedne vrste, dobimo vhodno meta-besedo formalne slovnice izvornega jezika, iz besede trivialno izpeljemo drevo izpeljave višine 1 za to besedo. Dobili smo množico produkcij, ki opisuje našo vhodno poved na ravnini naše formalne slovnice. Poiščemo najverjetnejše drevo izpeljave ciljnem jeziku, dobili smo prevod besedoslovnih oznak vhodne povedi.
Posamezne simbole ciljnih produkcij prevajane povedi pretvorimo v besede ciljnega jezika:
Poravnane besede vhodne povedi s pomočjo leksikona (glej poglavje 3) ter z upoštevanjem konteksta ciljnega drevesa izpeljave pretvorimo v besede ciljnega jezika.
Empirično preverjanje metode temelji na korpusu IJS-ELAN [1].

6. Primeri

```
<s id="Oen.1.1.2.4">
<w lemma="Winston" ana="Np">Winston</w>
<w lemma="make" ana="Vmps">made</w>
<w lemma="for" ana="Sp">for</w>
<w lemma="the" ana="Dd">the</w>
<w lemma="stair" ana="Ncnp">stairs</w>
<c>.</c>
</s>
<s id="Osl.1.2.3.4">
<w lemma="Winston" ana="Npmsn">Winston</w>
<w lemma="se" ana="Px------y">se</w>
<w lemma="biti" ana="Vcip3s--n">je</w>
<w lemma="napotiti" ana="Vmps-sma">napotil</w>
<w lemma="proti" ana="Spsd">proti</w>
<w lemma="stopnica" ana="Ncfpd">stopnicam</w>
<c>.</c>
</s>
```

Slika3: Izvleček iz korpusa. Poravnana primera v obeh jezikih

Slika4: Drevesi izpeljave za primera s slike 3

7. Zaključek

Predstavljena metoda vnaša nov način razmišljanja v statistično strojno prevajanje. Uporabljena bo na več različnih implementacijah sistemov za prevajanje. Rezultati testiranj bodo med sabo primerljivi in primerjani. Uporabljena bo na sorodnih jezikih, togih jezikih ter končno na povsem splošnih besedilih dveh tujih si jezikov. Najboljše rezultate pričakujemo ravno tam, kjer so najbolj potrebni. Metoda ni primerna za jezikovne pare z veliko strukturno sorodnostjo, primerna je za čimbolj zapleten in tuje si jezike.

8. Literatura

[1] Tomaž Erjavec: The IJS-ELAN Slovene-English Parallel Corpus, International Journal of Corpus Linguistics, vol. 7, št. 1, str. 1-20, JBP, 2002

[2] Tomaž Erjavec: Compiling and Using the IJS-ELAN Parallel Corpus, Informatica, vol. 36, št 3, str. 299 – 307, ,2002

[3] Sašo Džeroski, Tomaž Erjavec, Jakub Zavrel: MS Tagging of Slovene: Evaluating PoS Taggers and Tagsets, LREC'00, str. 1099 – 1104, ,2000

[4] P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer: The Mathematics of Statisti-cal Machine Translation: Parameter …, C.L., 19 (2), 1993

[5] Jernej Vičič, Tomaž Erjavec: Vsak začetek je težak: avtomatsko učenje prevajanja slovenščine v angleščino, SDJT Ljubljana: str. 20-27, IJS, 2002

[6] J.C. Martin: Introduction to Languages and the Theory of Computation, 3rd ed., McGraw-Hill,

[7] Hans van Halteren: Syntactic Wordclass Tagging, Kluwer Academic Publishers, Dordrecht, 1999

[8] K. Simov, G. Popova, P. Osenova: HPSG-based syntactic treebank of Bulgarian (BulTreeBank); A Rainbow of Corpora: Corpus Linguistics and the Languages of the World; Munich ,, 2002