

Mirjam Sepesy Maučec

Jezikovne tehnologije, tihi spremljevalec IKT

POVZETEK

Jezikovne tehnologije imenujemo skupino tehnologij, ki so namenjene samodejnemu procesiranju naravnega jezika. V računalništvu obstaja veja, imenuje se računalniško jezikoslovje, ki se navezuje na jezikoslovje. Njen cilj je modelirati naravni jezik za širok spekter računalniških aplikacij. Jezik se rabi v dveh oblikah, govorjeni in pisni. Jezikovne tehnologije modelirajo, procesirajo ali generirajo jezik v obeh načinih. Modeliranje naravnega jezika je za računalnik težka naloga, saj ima ta lastnosti, ki jih človek zlahka razpozna, stroj pa mnogo težje (večpomenskost, parafraze, nedoločenost ipd.). V prispevku bomo pregledali temeljne jezikovne tehnologije, med katere se uvrščajo: razpoznavanje in sinteza govora; kategorizacija, povzemanje in indeksiranje besedil; govorni dialog z računalnikom; odgovarjanje na vprašanja v naravnem jeziku; tehnologije prevajanja ipd. Danes se srečujemo z zmeraj večjim naraščanjem količine informacij. Da so le-te za človeka še vedno obvladljive, je uporaba jezikovnih tehnologij nujno potrebna. V zaključnem delu bomo predstavili primer uporabe jezikovnih tehnologij v vsakdanjem življenju starejše osebe.

Ključne besede: jezikovne tehnologije, naravni jezik, govor, besedilo.

1. Uvod

Jezikovne tehnologije so informacijske tehnologije, specializirane za obravnavo najbolj kompleksnega informacijskega medija na svetu, tj. naravni jezik. Naravni jezik je zahteven informacijski medij, ki se je razvijal od nastanka človeka in se razvija oz. spreminja še danes.

Jezik v tehnologijah nastopa v dveh oblikah: v obliki govora (govorne tehnologije) in v pisni obliki. Čeprav je govor najstarejše in najbolj naravno komunikacijsko sredstvo, je večina znanja človeštva shranjena v pisni obliki. Jezikovne tehnologije procesirajo in generirajo naravni jezik v obeh oblikah (glej sliko 1). Govorjen in pisan jezik si delita določene vidike obravnave, kot so slovar besed, slovnica, semantika (pomen) ipd. Po drugi strani pa ima vsaka oblika določene specifične lastnosti. Določene besedne zveze so na primer zelo pogoste v govoru, v pisani obliki pa zelo redke [1].



Slika 1. Jezikovne tehnologije.

Človeka je že od nekdaj najbolj zanimala povezava med jezikom in znanjem. Potekajo intenzivne raziskave o tem, kako so jezik, znanje in misli predstavljeni v človekovih možganih. Bi lahko tem principom sledili tudi stroji? Zaenkrat je delovanje možganov še velika neznanka. V jezikovnih tehnologijah se poslužujemo formalnih, na statistiki temelječih metod, ki s principi v možganih nimajo dosti skupnega. Osnova so formalne predstavitve konceptov v jeziku, ki jih povezujemo z nalogami iz realnega sveta.

2. Povezovanje tehnologij

Človeku najbolj naravna je multimodalna komunikacija. To je komunikacija, ki vključuje več človekovih čutil. Pravimo, da komunikacija vključuje različne modalnosti. Razlikujemo senzorične in akcijske modalnosti. Senzorične modalnosti so vid, sluh, tip in voh. Med akcijske modalnosti pa uvrščamo govor, gestika obraza, premikanje oči, rok oz. prstov, glave in telesa. Multimodalne tehnologije proučujejo učinkovito uporabo različnih modalnosti na komplementaren način.

Govorne tehnologije se povezujejo z drugimi tehnologijami, npr. multimedijo. Multimedija ali večpredstavnost se ukvarja s sočasno uporabo različnih medijev: besedilo, zvok, video, animacija, ipd. Jezikovne tehnologije so lahko podporne tehnologije v tehnologijah znanja. Tehnologije znanja se ukvarjajo z zajemanjem, shranjevanjem, upravljanjem in odkrivanjem znanja (npr. podpora odločanju v zdravstvenih informacijskih sistemih).

Jezikovne tehnologije (predvsem govorne tehnologije) niso 100% zanesljive tehnologije, zato jih v kritičnih okoljih, ki zahtevajo 100% zanesljivost ne moremo uporabljati, oz. lahko nastopajo le kot podporne tehnologije.

3. Uporaba jezikovnih tehnologij

Čeprav jezikovne tehnologije ne dosegajo zmogljivosti, ki jih dosega človek, so lahko učinkovita komponenta v številnih aplikacijah. Cilj je ustvariti programske izdelke, ki imajo določeno znanje o jeziku. Tovrstni izdelki lahko v veliki meri spremenijo naše življenje. V mislih imamo predvsem interakcijo človeka s strojem. Stroj ne razume jezika, ki ga govori človek. Obstajajo sicer stroju prilagojeni programski jeziki, a njihova uporaba ne ustreza konceptu človekovih misli. Običajno se v aplikacijah uporablja naravni jezik v okleščeni obliki. Že omejena uporaba naravnega jezika pri komunikaciji s strojem

lahko bistveno spremeni človekovo izkušnjo. Sodobne tehnologije človeka slišijo in mu odgovarjajo v naravnem jeziku. Pri tem imamo v mislih vmesnike to baze podatkov, pridobivanje informacij iz obsežnih besedil, različne ekspertne sisteme ter nadzor in upravljanje robotskih sistemov.

Jezikovne tehnologije lahko izboljšajo tudi komunikacijo med ljudmi. Ta težnja je veliko starejša od razvoja tehnologij komunikacije človeka s strojem. Ljudje govorimo različne jezike in avtomatsko prevajanje med različnimi jeziki je ena najstarejših idej umetne inteligence. Klub intenzivnim raziskavam je strojno prevajanje domensko neomejenega teksta še vedno zelo oddaljen cilj. Je pa strojno prevajanje lahko učinkovit pripomoček prevajalca, ki mu skrajša čas, potreben za generiranje prevodov in tako prispeva k njegovi boljši učinkovitosti. Nepopolni strojni prevodi so lahko v veliko pomoč tudi v drugih aplikacijskih okoljih, na primer pri iskanju informacij v zbirkah besedil v različnih jezikih.

4. Osnovne tehnologije

4.1. Govorne tehnologije

Jezikovne tehnologije delimo na govorne in tekstovne tehnologije. Osnova govornih tehnologij je govorni signal, tekstovnih pa pisano besedilo. Razpoznavnik govora govorni signal pretvarja v tekst [5]. Prvi korak razpoznavnika je digitalno vzorčenje signala. Sledi mu akustično procesiranje signala. Iz govornega signala izločimo značilke in jih primerjamo z akustičnimi značilkami vnaprej naučenih govornih enot. Akustično procesiranje je sestavljeno iz dveh korakov: predprocesiranja in izločanja značilk (parametrizacije govora). Pri akustičnem modeliranju se najpogosteje uporabljajo prikriti Modeli Markova z levo-desno topologijo in 16 Gaussovimi porazdelitvami na stanje. Značilnosti jezika so zapisane v ti. jezikovnem modelu. Najpogosteje uporabljeni jezikovni modeli so n-gramski modeli, ki statistično modelirajo zaporedja n besed. Zakaj je razpoznavanje govora tako težko? Govor je zelo spremenljiv. Isto besedilo dva govornika izgovorita zelo različno. Govor lahko obravnavamo kot biometrični podatek za avtorizacijo govornika. Tudi pri istem govorniku se npr. višina govora zelo spreminja, npr. v odvisnosti od emocij (ko smo jezni, presenečeni ipd). Po drugi strani pa imajo različne besede zelo podobno izgovorjavo (npr. spati, stati, stoti). Posebno trd oreh je spontan govor, v katerem se pojavljajo, mašila (npr. eerr, hmmm), ponovni ali napačni začetki ipd. Zaenkrat se v realnih okoljih uporabljajo razpoznavniki govora z omejenim slovarjem oz. z omejeno domeno (npr. razpoznavanje števk, vremenske napovedi, finančnih poročil).

Računalnik lahko tudi spregovori. Govorimo o sistemih sinteze govora. To so sistemi, ki poljuben tekst pretvarjajo v govorni signal. Proces je obraten kot pri sistemih za razpoznavanje govora. V želji, da bi umetno generiran govor zvenel čim bolj naravno, sintetizatorji sledijo naslednjim korakom. Prvi je pred-obdelava besedila, grafemsko-fonetična pretvorba, modeliranje trajanja in modeliranje intonacije in na koncu tvorjenje govornega signala. Tudi sinteza lahko temelji na statističnih pristopih, na primer na prikritih modelih Markova. Drug pristop je korpusna sinteza, pri kateri potrebujemo veliko število govornih posnetkov, med katerimi iščemo najbolj

primerne enote govora. Sintetizatorji govora so se izkazali kot primerni pripomočki slepih in slabovidnih, v obliki e-bralnikov.

Razpoznavanje in sinteza govora sta komponenti sistemov avtomatskega dialoga, pri katerih poteka komunikacija človeka s strojem izključno preko govora. Takšni sistemi so uporabni predvsem v telefonskih okoljih, kjer jih lahko kombiniramo le s tonsko izbiro, kot komunikacijsko alternativo, če govor ni razumljen. Običajno se to zgodi, če se govorec nahaja v preveč šumnem okolju.

4.2. Tekstovne tehnologije

Drugo skupino jezikovnih tehnologij predstavljajo tekstovne tehnologije, katerih osnova je tekst. Mednje se uvršča kategorizacija teksta, pri kateri avtomatsko razvrščamo besedila v vnaprej definirane kategorije. Poseben primer je razvrščanje v dve kategoriji. Na primer besedila izražajo mnenja, ki so lahko pozitivna ali negativna.

Danes se soočamo z zelo obsežnimi besedili, ki jih nimamo časa v celoti prebrati. Sistemi avtomatskega povzemanja nam ponudijo povzetke ključnih trditev oz. misli.

Tekstovne dokumente za učinkovito hrambo shranjujemo v indeksirane zbirke. Uporabnik lahko glede na svojo zahtevo, ki jo formulira v naravnem jeziku, iz zbirke pridobi le relevantne dokumente. Včasih želimo le relevantne informacije in ne celotne dokumente. Govorimo o luščenju informacij. V sistemih avtomatskega odgovarjanja na vprašanja nastopata tako vprašanje kot odgovor v obliki, kot ju poznamo v medčloveški komunikaciji.

4.3. Povezovanje govornih in tekstovnih tehnologij

Sistemi simultane prevajanja govora v govor kombinirajo govorne in tekstovne tehnologije [2, 6]. Predstavljajo možnost presejanja jezikovnih pregrad, saj omogočajo komunikacijo med govornicema, ki ne govorita istega jezika. Si lahko predstavljamo telefon, ki prevaja govor? Omogočal bi, da bi brez zadržkov poklicali kogarkoli in se ne ubadali z mislijo, da sogovornika ne razumemo oz. on ne razume nas. Sistemi prevajanja govora so ugledali luč sveta, a le za omejene domene in le za določene jezikovne pare.

5. Tehnologije prihodnosti

Jezikovne tehnologije so tehnologije prihodnosti. Postajajo nepogrešljiva komunikacijska komponenta v različnih aplikacijskih okoljih. Navedimo le primer razvoja podpornih tehnologij za starejše. Statistične raziskave kažejo, da se prebivalstvo stara. Povprečna pričakovana življenjska doba je trenutno preko 80 let, do leta 2020 pa naj bi bilo 25% evropskega prebivalstva starejšega od 65 let. Predvsem za starejše ljudi je televizija osrednji informacijski medij. Zakaj ne bi postala tudi osrednji komunikacijski medij, ki bi omogočal socialno interakcijo v bivalnem okolju? Starostniku bi lahko ponudila govorno prožen telefonski klic, opozarjala na redno jemanje zdravil in v kombinaciji z alarmnim sistemom na prisotnost dima, na izliv vode ipd [3]. Vključuje lahko tudi telemedicino, predvsem telediagnostiko (npr. spremljanje podatkov o krvnem tlaku in srčnem utripu na daljavo). Vse to in še več na uporabniku prijazen način, to je z uporabo komunikacije v

naravnem jeziku. Na ta način bi bila starejšim osebam omogočena večja samostojnost in varnost pri bivanju v domačem okolju.

6. Zaključek

Jezikovne tehnologije omogočajo razvoj IKT storitev z visoko stopnjo dodane vrednosti. Brez jezikovnih tehnologij ne bomo mogli doživeti resnično učinkovite interaktivne, večpredstavne in večjezične uporabniške izkušnje v bližnji prihodnosti. Za razvoj tovrstne tehnologije so nujno potrebni jezikovni viri govorjenega in pisanega jezika. Če želimo, da je tudi slovenski jezik vpet v razvoj jezikovnih tehnologij, moramo za to poskrbeti mi, saj je, žal, v mednarodnem prostoru jezikov slovenski jezik (predvsem komercialno) manj zanimiv [4].

Literatura

1. Uszkoreit, Hans: Language Technology. A First Overview. Dostopno na spletnem naslovu: <http://www.dfki.de/~hansu/LT.pdf>, [28.3.2016].
2. A. Waibel et al.: SpeeChalator: two-way speech-to-speech translation in your hand. in HLT-NAACL 2003, Edmonton, Canada, May-June 2003.
3. López-de-Ipiña, Diego et al.: ElderCare: An Interactive TV-based Ambient Assisted Living Platform. Activity Recognition in Pervasive Intelligent Environments, Atlantis Press, str.11-125.
4. Krek, Simon: Slovenski jezik v digitalni dobi, Zbirka bela knjiga, Springer-Verlag Berlin Heidelberg, 2012.
5. Žgank, Andrej in Kačič, Zdravko. Primerjava treh tipov akustičnih osnovnih enot razpoznavnika slovenskega govora, Elektrotehniški vestnik, 2005, vol. 72, no. 5, str. 273-278.
6. Sepesy Maučec, Mirjam, Kačič, Zdravko. Statistical machine translation from Slovenian to English. CIT, ISSN 1330-1136, Mar. 2007, vol. 15, no. 1, str. 47-59.