ASR-hibaterjedés vizsgálata a gépi beszédértés szemszögéből

Tündik Máté Ákos, Szaszák György

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
e-mail:{tundik,szaszak}@tmit.bme.hu

Kivonat Napjainkban a mesterséges intelligencia alapú megoldások egyre inkább a beszélt nyelv gépi megértésére törekednek. Ennek preferált megközelítése az, amikor automatikus beszédfelismerő (ASR) rendszerek használatával átiratokat hozunk létre, amelyek további, szövegalapú elemzésen mennek keresztül. A gépi átiratok szóhibákat is tartalmazhatnak; ezen hibák továbbterjednek a szöveges feldolgozási folyamatba, így a gépi központozásba, kivonatolásba is. Ugyanakkor szubjektív tesztjeink azt igazolták, hogy az emberek a gépi átiratokat a szóhibák és a központozási hibák ellenére is jól tudják értelmezni. Célunk az, hogy bemutassuk az ASR-hibaterjedésből adódó, szemantikai térben bekövetkező információveszteségeket, valamint az ASR-hibaterjedés automatikus összefoglalásra gyakorolt hatását is elemezzük. Bemutatjuk, hogy az egyes mondatreprezentációk a szóhibák hatására enyhén eltolódnak a szemantikai térben, de ez jócskán elmarad a dokumentum mondatainak átlagos szemantikai távolságától. Megmutatjuk azt is, hogy a központozás hibáinak nagyobb hatása van az összefoglalók kiértékelésére, mint a szóhibáknak, ami arra enged következtetni, hogy a feladathoz elengedhetetlen a megfelelő mondatszintű tokenizálás.
Kulcsszavak: szemantikai hasonlóság, hibaterjedés, gépi beszédértés, tartalmi összefoglalás

1. Bevezetés

A legmodernebb, beszédalapú összefoglaló rendszerek egy automatikus beszédfelismerő (ASR) eszköz segítségével szöveges átiratot készítenek, majd következő lépésként a szöveges dokumentum összefoglalása következik. Az utóbbi modul általában először mondatszintű tokenizálást hajt végre, majd ezt további, a szemantikai térben elvégzett műveletek követik. Az összefoglaló készítése kétféleképpen történhet; (1) ún. extraktív módon, amikor a gépi átirat mondatai kerülnek felhasználásra rangsorolást követően (Celikyilmaz és Hakkani-Tür, 2011), (2) absztraktív módon, amikor egy szemantikai kódolási algoritmus biztosítja a még tömörebb, 'újrafogalmazott' összefoglaló létrehozását (Genest és Lapalme, 2012; Paulus és mtsai, 2017). A szemantikai térbe történő projekció leggyakoribb módja a szóbeágyazások (szóvektorok) használata (Mikolov és mtsai, 2013b).

Az ASR kimeneten átadott átiratok feldolgozásakor központozási hibákkal és szóhibákkal is számolni kell; ezek a hibák továbbterjednek a feldolgozási folyamatban, így befolyásolják a beszéd tartalmi összefoglalását is.

Az első szövegfeldolgozási lépésként elvégzendő mondatokra bontás nehézsége, hogy az írásjelek és a nagybetűk hiányoznak a nyers ASR-átiratokból. A központozás megvalósítására vagy prozódiai alapú szegmentálást végzünk el, közvetlenül a beszédanyagon (Beke és Szaszák, 2016), vagy a gépi beszédátiratban állítjuk vissza az írásjeleket (Klejch és mtsai, 2017; Öktem és mtsai, 2017; Tündik és Szaszák, 2018). Az utóbbi megközelítés alkalmazásával nemcsak akusztikus, hanem nyelvi (szöveges) jellemzők is kiaknázhatók. A legkorszerűbb automatikus központozó rendszerek teljesítménye F1-mértéket tekintve 70-80%, tehát ezen megoldások esetében is még jócskán jelen vannak központozási (írásjelezési) hibák a központozott átiratban.

ASR vonatkozásában - feladattól és a környezeti feltételektől függően - az ipari hasznosítás szempontjából releváns alkalmazásokban a szóhibaarány (WER) 1-30% között van. Kevés tanítóanyaggal rendelkező, vagy nyelvi szempontból tekintve speciális nyelv - például morfológiailag vagy összetett szavakban gazdag, stb. - esetében a WER sokkal magasabb lehet, mint hasonló funkcionalitást nyújtó angol nyelvű alkalmazások esetén. A felhasználói élmény ugyanakkor általában kevésbé romlik le, mint azt a WER különbsége sugallná, sőt, ugyanazon mértékű szóhibaaránnyal működő angol ASR rendszert akár a végfelhasználók rosszabbra is értékelhetnek, mint egy finn (Kurimo és mtsai, 2006) vagy magyar (Tündik és mtsai, 2018) rendszert.

Valójában az emberek meglepően jól teljesítenek, ha hibákkal terhelt, automatikusan központozott gépi átiratokat kell olvasniuk és értelmezniük (Tündik és mtsai, 2018). Nyilvánvaló, hogy a gépi értelmezéssel szemben az emberek tágabb kontextusra és egyéb olyan aspektusokra is támaszkodhatnak, amelyek a gyakran nem is tudatosuló hibajavító mechanizmus működését segítik (Postma, 2000; Kröger és mtsai, 2016). A halláskárosodásban szenvedő személyek esetében korábban igazoltuk, hogy az ép hallású emberekhez viszonyítva jobban teljesítenek a szó-, és különösen az írásjelek hibáinak spontán javításában (Tündik és mtsai, 2018), valószínűsíthetően az ilyen hibák tudatos észlelésének küszöbértéke sokkal magasabb az esetükben.

A szemantikus térbe történő transzformációk, különösen a szóbeágyazások (Mikolov és mtsai, 2013a) nagyon népszerűvé váltak a természetes nyelvi feldolgozásban és a beszélt nyelv megértésében. Noha az ilyen szóvektor-ábrázolások a szemantikai vagy a szintaktikai konzisztencia és pontosság szempontjából messze nem tökéletesek, kiváló képességeket mutatnak az információ szemantikai feldolgozását magában foglaló (pl. következtetési, analógiai) feladatok esetében. A szóvektorok használata korszerűnek számít a tartalmi kivonatolásban is. Jelen cikkünkben az inspirált minket, hogy objektív mérések alapján felmérjük, mennyire torzul az információ a szemantikai térben a szó- és/vagy központozási hibák miatt az automatikus beszéd-szöveg átalakítást következtében. A szemantikai torzítást eddig elsősorban szubjektív szempontból vizsgálták (Kafle és Huenerfauth, 2016; Tündik és mtsai, 2018), ekkor az ASR-hibaterjedésének hatása

a szemantikai térben csekélynek mondható, ésszerű, ipari alkalmazást lehetővé tévő szóhibaarány mellett. Egyes kutatók megvizsgálták a szóhelyettesítési hibák hatását mondatbeágyazások szintjén (Voleti és mtsai, 2018), más munkák (pl. (Simonnet és mtsai, 2018)) az ASR hibák szimulációját javasolták az ilyen típusú elemzésekhez. Mivel a valós ASR átiratok előállítása nem bonyolult, amennyiben a hanganyag rendelkezésre áll, ezért nem szimuláltunk ASR hibákat, hanem valódi gépi átiratokat használtunk, ezzel is kiküszöbölve a szimulációval bevitt torzítást. Ezáltal lehetőségünk nyílt a helyettesítési hibák kizárólagos vizsgálata helyett az összes lehetséges szóhibát számításba venni (így a törléseket és a beszúrásokat is), csakúgy, mint a központozási hibákat, hogy a kísérleti beállítások a lehető legközelebb kerüljenek a valódi felhasználási helyzethez, körülményekhez.

Cikkünk a következőképpen épül fel: bevezetőnkben bemutattuk az ASR-hibaterjedés problémakörének jelentőségét, kifejtettük motivációnkat, és bemutattunk néhány, a témához kapcsolódó munkát. A következő fejezetek a felhasznált adatbázist, valamint a mondatszintű és a dokumentumszintű szemantikai hasonlóság méréséhez használt módszertant dokumentálják, az utóbbihoz egy népszerű, dokumentum-összefoglaló alapú megközelítést használva. Ezt követően bemutatjuk és megvitatjuk eredményeinket, mielőtt végső következtetéseinket levonnánk.

2. Adat, ASR és Központozás

2.1. Átiratok előkészítése

Kutatásunk során az ASR- és/vagy írásjelhibák által okozott szemantikai torzításokat vizsgáljuk. Ezáltal négy különböző, ámbár összehasonlítható átiratváltozatot készítettünk minden egyes beszédfájlra, az alábbiak szerint[1]:

MT-MP: Kézi Átirat - Kézi Központozás : emberek által készített referenciaátirat, amely az alábbi négy írásjelet tartalmazza: {. , ? !};

AT-MP: Gépi (ASR) Átirat - Kézi Központozás : gépi átirat felhasználása, melybe a referenciaátirat segítségével „visszacsempésztük" az írásjeleket[2];

MT-AP: Kézi Átirat - Automatikus Központozás: a referenciaátiratból eltávolítottuk az írásjeleket, majd azokat automatikus módszerrel prediktáltuk (Tündik és mtsai, 2018);

AT-AP: Gépi (ASR) Átirat - Automatikus Központozás: a gépi átiratok automatikus központozásához szintén a (Tündik és mtsai, 2018) cikkben ismertetett modellt használtuk.

[1] a rövidítésekben az angol megfelelőt használtuk, pl. Manual Transcript - Manual Punctuation

[2] Esetenként ez nagy kihívás, amennyiben a szóhibák miatt az eredeti írásjelezés értelmét veszti.

2.2. Adatbázisok

Kísérleteinket angol és magyar nyelven végeztük el. **Magyar nyelvre** 10 szöveges blokkot választottunk ki egy televíziós műsorok átiratait tartalmazó adatbázisból (Tarján és mtsai, 2016); sporthíreket, időjárás-jelentéseket és híradókat vizsgáltunk meg. Ez a részkorpusz összesen 500 mondatot, így megközelítőleg 8000 szót foglal magában. A felhasznált ASR rendszer (Varga és mtsai, 2015) szóhibaarány értékeit illetően rendre 6,8%-ot, 10,1%-ot és 21,4%-ot mértünk az időjárás-jelentések, a híradók és a sporthírek esetén. Automatikus központozáshoz a (Tündik és mtsai, 2018)-féle, magyar nyelvre adaptált modellt használtuk, melynek teljesítménye F1-mértéket tekintve 60-70% kézi átiratokon, gépi átiratokon pedig 45-50%.

Angol nyelvre az IWSLT2011 adathalmazban található TED előadások átiratai közül használtunk fel 9 szöveges blokkot (Federico és mtsai, 2012). Ez a részkorpusz összesen 800 mondatot, így megközelítőleg 12000 szót foglal magában. Az ASR átiratok a (Rousseau és mtsai, 2012) cikkben bemutatott módszerrel készültek, melyeken 18,7% -os szóhibaarányt mértünk. Automatikus központozáshoz a (Tündik és mtsai, 2018)-féle angol nyelvre adaptált modellt használtuk, melynek átlagos teljesítménye F1-mértéket tekintve 60-70% kézi átiratokon, gépi átiratokon pedig 50-55%.

A magyar és angol nyelvű referencia összefoglalók készítését 3 annotátor vállalta (minden szöveges blokkhoz 3 darab, 10-12 mondat terjedelmű összefoglaló készült), így a szóhibák és a központozási hibák által keletkezett szemantikai torzításokat egy dokumentum-összefoglaló feladat keretében is meg tudtuk vizsgálni.

3. Módszerek

Cikkünkben néhány olyan megközelítést ismertetünk és értékelünk ki, amelyek a szemantikai torzítások számszerűsítésére alkalmasak. Ezen mértékek esetén két alapvető szempont jön szóba: (i) kiszámítjuk az egyes mondatpárok (ugyanazon mondat kézi és gépi átiratának) szemantikai hasonlóságát, szóbeágyazások alapján, míg (ii) a gépi átiratból és írásjelezésből adódó hibák kölcsönhatásának elemzését tartalmi összefoglalási feladaton keresztül vizsgáljuk meg. A szemantikai torzításra vonatkozó összehasonlítást így mondat- illetve dokumentumszinten is elvégezzük.

3.1. Mondatszintű hasonlóság

Első lépésként meghatározzuk a mondatvektor-reprezentációkat egy adott mondat szóvektorainak segítségével. Angol nyelvre az előtanított GloVe (Pennington és mtsai, 2014) és word2vec (Mikolov és mtsai, 2013a) szóbeágyazásokat, magyar nyelvre pedig a „Makrai-féle" szóvektorokat (Makrai, 2016) használtuk fel vizsgálatainkhoz. Megfontoltuk a modernebb, kontextuális beágyazások és karakter N-gram sorozatokkal kiterjesztett szóvektorok használatát, de ezeket végül elvetettük, mivel nem álltak rendelkezésre magyar nyelvre a vizsgálat idején, illetve

a karakter N-gramok hozzáadását korábban kontraproduktívnak találtuk, valószínűleg a magyar nyelv extrém gazdag morfológiája és kötetlen szórendje miatt. (Azt tapasztaltuk, hogy a szóvektorok szépen megtanulják a morfoszintaxist, de összességében szinte teljesen elveszítik a szemantikus konzisztenciát).

Továbbá a mondatszintű kódolók (Cer és mtsai, 2018; Conneau és mtsai, 2017) alkalmazását is mellőztük, elsősorban azért, mert az általunk ismertetett, egyszerűbb megközelítések hasonló teljesítményt mutatnak ezekkel a nehéz és összetett megközelítésekkel (Ethayarajh, 2018). Ily módon nem kellett megküzdenünk olyan nehézségekkel sem, mint például a magyar nyelvre történő adaptálás; ehelyett inkább kihasználjuk a kevésbé bonyolult, felügyeletlen megközelítések összes előnyét. A következő vektorábrázolási formákat használjuk a szemantikai torzítás/hasonlóság mondatszintű vizsgálatára:

Szózsák (Bag-of-Words, BOW): a legegyszerűbb vektorizálási formában a mondat szavainak egyszerű átlagát vesszük. Esetlegesen stop-szó szűrést végzünk az NLTK könyvtárral.

Simított Inverz Gyakoriság (Smooth Inverse Frequency, SIF): A SIF mondatbeágyazások (Arora és mtsai, 2016), súlyozottan átlagolják a szóvektorokat. A súlyokat (W) az alábbi formulával számíthatjuk:

$$W(w_i) = \frac{a}{a + p(w)}, \tag{1}$$

ahol a a simítást befolyásoló paraméter (alapértelmezetten $a = 0,001$), $p(w_i)$ pedig a w_i szó referencia korpuszon számított relatív gyakorisága. Ily módon a gyakori szavak súlya kisebb, a szemantikailag relevánsabbaké pedig nagyobb lesz. Az ezt követő lépésben a SIF vektorokat konkatenáljuk egy mátrixba, amelyet szinguláris érték felbontással (SVD) felbontunk. A SIF mondatvektorok első szinguláris értékre vett projekcióját ezután kivonjuk a súlyozott átlagból, így csökkentve a szemantikailag nem odaillő szavak befolyását.

Nem felügyelt SIF (uSIF): az uSIF (Ethayarajh, 2018) módszer az előbb bemutatott SIF reprezentációhoz képest abban különbözik, hogy a értékét is közvetlenül becsüljük a gyakoriság szerint rendezett szótárból. Az első m szinguláris értéket őrizzük meg, rendre $\lambda_1 ... \lambda_m$ súlyokkal:

$$\lambda_i = \frac{\sigma_i^2}{\sum_{j=1}^{m} \sigma_i^2}, \tag{2}$$

ahol σ_i a mondatbeágyazó mátrix i-edik szinguláris értéke. Látható, hogy $m = 1$, esetén az uSIF a SIF-fel azonos, amennyiben a-t optimalizáltnak tekintjük. m leggyakrabban választott értéke 5.

A mondatok közötti hasonlóság mérésére páronként hasonlítjuk össze az egymáshoz illesztett mondatok szekvenciáit:

$$sim(a, b) = \frac{\sum_{i=0}^{S-1} a_i b_i}{\sum_{i=0}^{S-1} a_i^2 \sum_{i=0}^{S-1} b_i^2} \tag{3}$$

ahol a és b a két mondatbeágyazó vektor(melyek származtathatóak akár a BOW, a SIF vagy az uSIF eljárással) az S dimenziós mondatbeágyazó térben.

A mondatok közötti hasonlóságot egy negyedik módon, közvetlenül a szóvektorokból is származtathatjuk: a **Word Mover's Distance** (WMD) egy népszerű módszer dokumentumok / mondatok összehasonlítására (Kusner és mtsai, 2015). Alapja, hogy az összehasonlítandó dokumentumok (vagy esetünkben mondatok) között a szemantikus térben megadja azt a legkisebb költségű utat, amellyel a két dokumentum (mondat) egymásba átvihető. A WMD a népszerű Gensim python könyvtárban is implementált. A WMD alapján a hasonlóságot egyszerűen számíthatjuk két mondat közt:

$$WMS = \frac{1}{1 + WMD}. \tag{4}$$

3.2. Dokumentumszintű hasonlóság

A gépi beszédfelismerés egyik izgalmas felhasználási területe a beszélt nyelvi dokumentumok, rekordok tartalmi kivonatolása, összefoglalása. Ennek során beszédfelismerővel átírjuk a beszédet, majd az így nyert szövegen futtatjuk a tartalmi összefoglaló algoritmust.

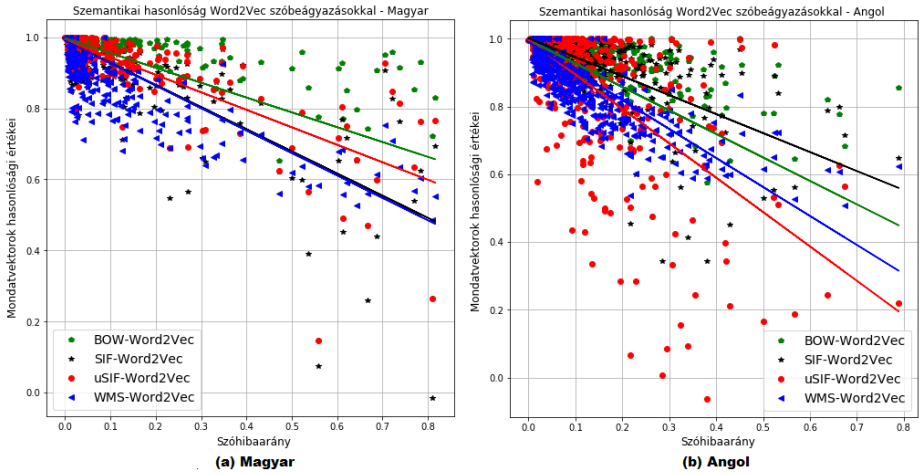
A kísérlethez az MT-MP, AT-MP, MT-AP és AT-AP eljárásokkal nyert szövegeket vesszük alapul, és valamennyire tartalmi összefoglalót generálunk. Az egyes összefoglalók közötti különbséget a Recall-Oriented Understudy for Gisting Evaluation eljárással, rövidebb nevén a ROUGE metrikákkal mérjük (Lin, 2004). A ROUGE többféle összehasonlítást is lehetővé tesz, ezek részletes ismertetése meghaladná jelen cikk kereteit, de kimerítő leírás található például a (Lin, 2004) irodalomban. Jelen munkában az alábbi ROUGE metrikákat használjuk:

- ROUGE-1: unigram (szavankénti) átfedést mér (felidézésben);
- ROUGE-2: bigram (szókettesek szerinti) fedést mér (a kérdéses összefoglaló milyen arányban idézi fel a referencia szóketteseit);
- ROUGE-L: leghosszabb közös szószekvencia;
- ROUGE-SU4: skip-bigram és N-gram alapján méri az együttes előfordulást (szinonimákat is kezeli a skip-gram révén).

Referenciaként a 3 független annotátor által az MT-MP szövegek alapján készített összefoglalókat használjuk (mivel többféle összefoglaló is készíthető, bevett gyakorlat nem egyetlen referenciával összevetni a kimenetet). A gépi tartalmi összefoglalást a Gensim modul (Mihalcea és Tarau, 2004) BM25 rangsoroló eljárásával (Barrios és mtsai, 2016) készítjük. Bár a BM25 több mint 10 éve ismert összefoglaló algoritmus, azért esett erre a választásunk, mert ipari alkalmazásokban is megtaláljuk, illetve mert nagyon egyszerűen használható, nem igényel adaptációt sem. Ugyanezen okokból mellőztük a beágyazásokon alapuló algoritmusokat is, illetve azért is, mert nem jellemző, hogy a felismerő szinonimára tévesszen, sokkal inkább hangzásában hasonló szóra. Mindazonáltal a jövőben mindenképp érdemes a kísérletet szemantikus reprezentációk alapján működő összefoglaló algoritmusokkal is elvégezni.

4. Eredmények és Diszkusszió

A mondatszintű kiértékelés esetében az MT-MP és az AT-MP átiratokat hasonlítottuk össze, mivel a kézi és az automatikus központozással készült dokumentumok mondatainak egymáshoz igazítása nem triviális feladat: az írásjelek megváltoztathatják a mondathatárokat, így a központozás típusai (MP és AP) szerinti összehasonlítás jobban illeszkedik a dokumentumszintű megközelítéshez. Az 1. ábra az MT-MP és az AT-MP átiratok mondatpárjain vett szemantikai hasonlósági értékeket (BOW, SIF, uSIF és WMS) ábrázolja, magyar (a) és angol (b) nyelvre. Így az x tengelyen lévő szóhibaarány is a mondat szintjén értendő.



1. ábra: Mondatszintű szemantikai hasonlósági értékek a szóhibaarány (WER) függvényében

Figyelembe véve a valós ASR-felhasználási eseteket, ahol a $WER < 30\%$ magyar nyelvre, angol nyelvre pedig $WER < 20\%$ értékű[3], a szemantikai térre gyakorolt hatás korlátozott, a hasonlósági értékek legtöbbje 0,8 és afölött van. Érdemes megvizsgálni a szórásokat is, melyek mértéke WER=20% felett látványos emelkedést mutat. Az MT-MP és AT-MP átiratok mondatain vett hasonlóságok átlagait az 1. táblázat mutatja, ahol a szóhibák ellenére nagyon magas szemantikai egyezést figyelhetünk meg. A magyar nyelvű kísérleteinkhez 300-dimenziós word2vec és 152-dimenziós GloVe szóbeágyazásokat használtunk. Mivel a SIF, az uSIF és a WMS kategóriák esetében a két megközelítés eredményei konzisztens trendeket mutattak, ezért csak a word2vec reprezentációkhoz tartozó eredményeket mutatjuk be.

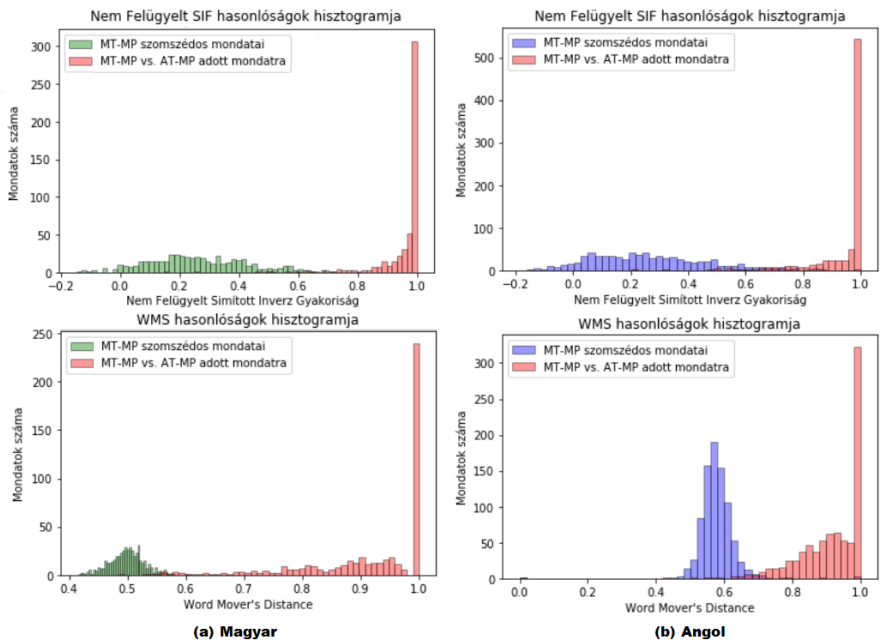
[3] A morfológiailag gazdag magyar nyelv esetén magasabb WER-érték mellett érzékeljük hasonlónak az ASR-teljesítményét (Kurimo és mtsai, 2006)

1. táblázat. Mondatszintű szemantikai hasonlósági értékek magyar és angol nyelvre

Mértékek	BOW	SIF	uSIF	WMS
Magyar	0,97	0,95	0,96	0,92
Angol	0,94	0,96	0,91	0,90

Ahogy az várható volt, nincs szignifikáns különbség a szóvektorok két típusa között. A BOW megközelítést illetően a szóvektorok két típusa kvázi-ekvivalenssé válik, amikor a mondatvektorok kiszámítása esetén egy előzetes stop-szó szűrést alkalmazunk az adott mondathoz tartozó GloVe szóvektorok átlagolásakor. Ez érthető, mivel a word2vec módszer esetén a stop-szavakat alulmintavételezik (Mikolov és mtsai, 2013b), míg a GloVe tanítása során megőrzik.

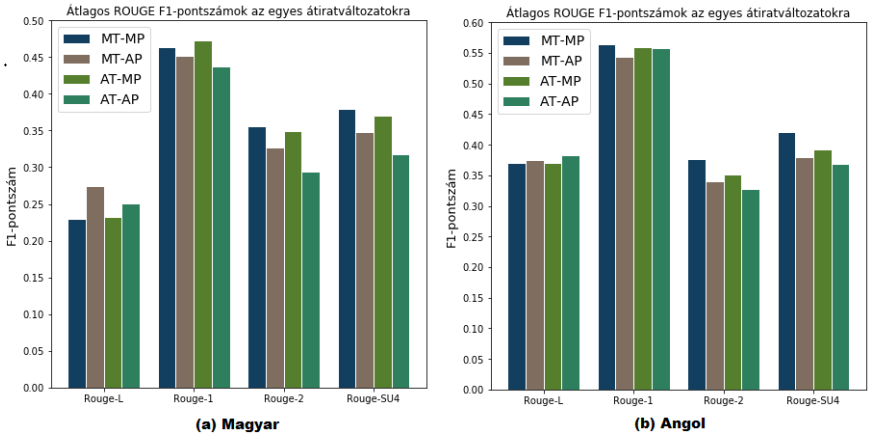
Egyfajta referenciaértékek felállítása érdekében – tekintettel az 1. ábrán látható MT-MP és AT-MP átiratváltozatok közötti hasonlósági értékre – az MT-MP típusú dokumentumban a szomszédos mondatok szemantikai hasonlóságainak eloszlását is meghatároztuk. Ennek a lépésnek az a célja, hogy össze tudjuk hasonlítani a szóhibákból származó mondatonkénti szemantikai változásokat a referenciadokumentum mondatai között megfigyelhető szemantikai hasonlósággal. A 2. ábra az uSIF és WMS mértékek eloszlását mutatja.



2. ábra: Szemantikai hasonlóságok (uSIF és WMS) eloszlásának ábrázolása hisztogrammal, szomszédos mondatok között a kézi átiratban, ill. ugyanazon mondat kézi és gépi átiratai között

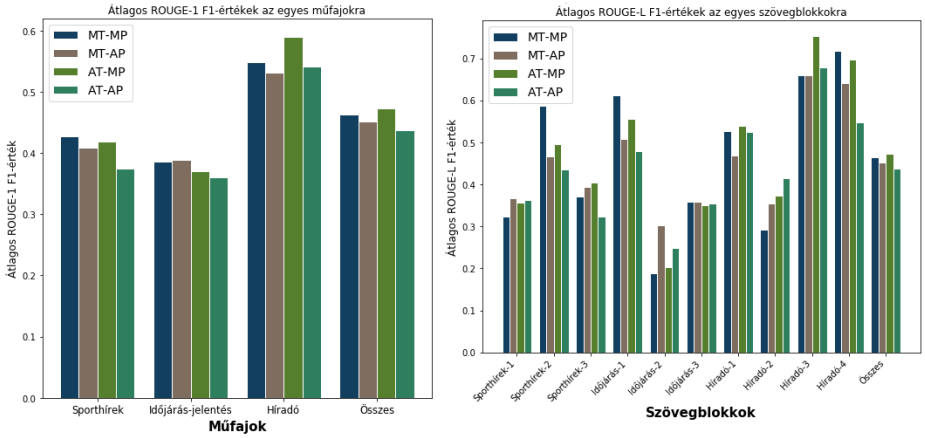
Mindegyik ábrán két hisztogram látható: a kézi és gépi átiratok közötti hasonlóságok eloszlását és a kézi átiraton belül, a szomszédos mondatok közötti hasonlóságok eloszlását. A két eloszlás között alig van átfedés, a magyar nyelv (lásd. 2. ábra 'a)' része) és az angol nyelv (lásd. 2. ábra 'b)' része) esetében. Ez azt jelenti, hogy a szóhibákból eredő szemantikai torzítás nem olyan magas, hogy egy tévesen felismert mondatot közelebb hozzon a szomszédos mondatok jelentéséhez, mint az eredeti jelentéshez. Figyelembe véve, hogy a szomszédos mondatok tipikusan közelebb állnak a szemantikai térben, mint ugyanazon dokumentumon belül a nem szomszédos mondatok, ez meglehetősen kielégítő eredmény, amely megmagyarázza azt is, hogy a tapasztalatokkal összhangban a jelentés kinyerése hogyan lehet kellőképpen robusztus azokból a mondatokból, amelyek szóhibákat tartalmaznak.

Rátérve a tartalmi összefoglalás feladatra, a 2.1. fejezetben ismertetett átiratváltozatokra vonatkozó ROUGE eredményeket a 3. ábra illusztrálja, magyar és angol nyelvre. Mivel a magyar nyelvű adatbázis különféle műfajú szövegeket tartalmazott, ezért eredményeinket a 4. ábrán műfaj szerinti bontásban, és az egyes blokkokat tekintve is bemutatjuk.



3. ábra: Tartalmi kivonatolás kiértékelése magyar és angol nyelvre

Az egyik legfontosabb szempont a „tökéletes" MT-MP és a valós felhasználást tükröző AT-AP átiratváltozatok eredményeinek összehasonlítása (utóbbinál mind az átirat, mind a központozás automatikusan történik). A különböző műfajokra vonatkozó magyar nyelvű tartalmi kivonatolási eredményeket szemlélve a 4. ábrán, az AT-MP átiratok eredménye szorosan korrelál az ASR pontossággal (sporthírek és híradók esetében), valószínűleg azért, mert az ASR rendszer nyelvi modelljének és a tartalmi kivonatoló szemantikai rangsoroló moduljának hasonló nyelvi komplexitású feladattal kell megbirkóznia. Az időjárás-jelentések kivételt képeznek; feltételezzük, hogy a gyakoriság alapú kivonatolási megközelítés kevésbé alkalmas ilyen típusú dokumentumokhoz.



4. ábra: Tartalmi kivonatolás magyar nyelvre, műfaji és blokkonkénti bontásban

A legjobb kivonatolási eredményeket a híradó kategóriájára kaptuk, annak ellenére, hogy a gépi átirat időjárás-jelentések esetében pontosabb volt. Az időjárás-jelentések esetében viszont a központozás pontatlanabb (Tündik és mtsai, 2018), a legkevésbé precíz automatikus központozás pedig a sporthírek kategóriájához társul (Tündik és mtsai, 2018).

Láthatjuk, hogy az írásjelekkel kapcsolatos hibák fontosabbak a kivonatolás szempontjából. Ez korrelál a mondatonkénti szemantikai vizsgálatainknál látottakkal: a szóhibák korlátozott torzítást eredményeznek a szemantikai térben a mondatok szintjén, feltéve, hogy a valódi mondathatárok ismertek (AT-MP). A 3. ábrán látható ROUGE-pontszámokra kitérve, a ROUGE-2 és a ROUGE-SU4 esetében megfigyelhető, hogy az MT-AP kategóriára vonatkozó értékek alacsonyabbak, mint az AT-MP esetében, valamint az, hogy az eredmények közötti különbség nagyobb, ha a központozás módját változtatjuk (kéziről automatikusra), mint amikor az átirat típusa változik (kéziről automatikusra).

Az AT-MP és az AT-AP kategóriák összefoglalóit összehasonlítva, a ROUGE-2 és a ROUGE-SU4 pontok szerinti különbség jelentős. Habár az AT-AP esetben a szóhibák már az automatikus központozásba is továbbterjednek, eredményeink azt igazolják, hogy a mondatszintű tokenizálási (központozási) hibák nagyobb mértékben befolyásolják a kivonatolást, mint a szóhibák. Az eredmények azt sugallják, hogy a mondatokra bontás esetében javallott a prozódiai jellemzőkre is támaszkodni, amelyek a szóhibákkal szemben jóval robusztusabbak, mint a szöveges jellemzők. A jövőben mind prozódiai alapú, közvetlen szegmentálási módszereket (pl. (Beke és Szaszák, 2016)), mind akusztikai-szöveges központozási megoldásokat (pl. (Szaszák és Tündik, 2019)) is érdemes megvizsgálni tartalmi kivonatoláskor.

5. Összegzés

Cikkünkben megvizsgáltuk a szóhibák és központozási hibák által kiváltott szemantikai torzítást. Az ASR rendszerekből származó szóhibák már az automatikus központozás feladatába továbbterjednek, amikor a nyers gépi átirat tokenizálása a cél; ezután pedig mindkét (szó- és központozási) hibatípussal számolni kell a tartalmi összefoglalók készítése esetében.

Egyszerű, mondatszintű hasonlósági metrikákkal bebizonyítottuk, hogy a szóhibák jelenléte kisebb mértékű torzítást eredményez a szemantikai hasonlóságban ugyanazon mondatot vizsgálva, mintha két, szomszédos mondat közötti szemantikai különbséget vizsgálnánk. Valójában a két eset hasonlósági eloszlása marginális átfedést mutatott, ami azt sugallja, hogy a szóhibák ritkán okoznak drámai eltolódást a szemantikai térben a mondatok szintjén (és ennélfogva magasabb szinteknél, pl. a dokumentumok szintjén).

Mivel a gépi átiratban elveszik a valós mondatszint, automatikus központozást kell alkalmazni. A szemantikai torzítás tartalmi összefoglalások vizsgálatának szemszögéből történő értékelése lehetővé tette számunkra, hogy elemezzük az írásjelhibákat is a szóhibák mellett. Megállapítottuk, hogy az írásjelek miatt a ROUGE-2 és a ROUGE-SU4 pontszámok közötti relatív különbség nagyobb, mint a szóhibák esetén, bár a szóhibák az írásjelezési feladatra is hatást gyakorolnak. A teljesen automatikus (AT-AP) bemenetű összefoglalók elemzése azonban azt mutatta, hogy az ASR-feldolgozási lánc jelenlegi szűk keresztmetszete elsősorban a központozás okozta mondatszintű eltérésből fakad, nem pedig a szóhibákból, még a szóhibaarány 20%-hoz közeli szintjén is. Ezek a megállapítások extraktív tartalmi összefoglalásra érvényesek, absztraktív változat vizsgálatára a magyar nyelv korlátai miatt nem nyílt lehetőségünk.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak, amely az FK-124413 projekt keretében a cikkben ismertetésre került kutatást támogatta. Köszönjük továbbá az NVIDIA támogatását (GPU biztosítása a neurális hálózatok tanításához).

Hivatkozások

Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (2016)

Barrios, F., López, F., Argerich, L., Wachenchauzer, R.: Variations of the similarity function of textrank for automated summarization. arXiv preprint arXiv:1602.03606 (2016)

Beke, A., Szaszák, G.: Automatic summarization of highly spontaneous speech. In: International Conference on Speech and Computer. pp. 140–147. Springer (2016)

Celikyilmaz, A., Hakkani-Tür, D.: Discovery of topically coherent sentences for extractive summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 491–499. Association for Computational Linguistics (2011)

Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., és mtsai: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)

Ethayarajh, K.: Unsupervised random walk sentence embeddings: A strong but simple baseline. In: Proceedings of The Third Workshop on Representation Learning for NLP. pp. 91–100 (2018)

Federico, M., Stüker, S., Bentivogli, L., Paul, M., Cettolo, M., Herrmann, T., Niehues, J., Moretti, G.: The IWSLT 2011 evaluation campaign on automatic talk translation. In: International Conference on Language Resources and Evaluation (LREC). pp. 3543–3550 (2012)

Genest, P.E., Lapalme, G.: Fully abstractive approach to guided summarization. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 354–358 (2012)

Kafle, S., Huenerfauth, M.: Effect of speech recognition errors on text understandability for people who are deaf or hard of hearing. In: Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT). pp. 20–25 (2016)

Klejch, O., Bell, P., Renals, S.: Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5700–5704. IEEE (2017)

Kröger, B.J., Crawford, E., Bekolay, T., Eliasmith, C.: Modeling interactions between speech production and perception: speech error detection at semantic and phonological levels and the inner speech loop. Frontiers in Computational Neuroscience 10, 51 (2016)

Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., Saraclar, M.: Unlimited vocabulary speech recognition for agglutinative languages. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. pp. 487–494. Association for Computational Linguistics (2006)

Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: International Conference on Machine Learning. pp. 957–966 (2015)

Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)

Makrai, M.: Filtering Wiktionary triangles by linear mapping between distributed models. In: Proceedings of LREC. pp. 2776–2770 (2016)

Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 404–411 (2004)

Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013a)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (szerk.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013b)

Öktem, A., Farrús, M., Wanner, L.: Attentional parallel RNNs for generating punctuation in transcribed speech. In: International Conference on Statistical Language and Speech Processing. pp. 131–142. Springer (2017)

Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304 (2017)

Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of EMNLP. pp. 1532–1543 (2014)

Postma, A.: Detection of errors during speech production: A review of speech monitoring models. Cognition 77(2), 97–132 (2000)

Rousseau, A., Deléglise, P., Esteve, Y.: TED-LIUM: An automatic speech recognition dedicated corpus. In: LREC. pp. 125–129 (2012)

Simonnet, E., Ghannay, S., Camelin, N., Estève, Y.: Simulating ASR errors for training SLU systems. In: LREC 2018 (2018)

Szaszák, G., Tündik, M.Á.: Leveraging a character, word and prosody triplet for an ASR error robust and agglutination friendly punctuation approach. Proc. Interspeech 2019 pp. 2988–2992 (2019)

Tarján, B., Varga, Á., Tobler, Z., Szaszák, Gy., Fegyó, T., Bordás, Cs., Mihajlik, P.: Magyar nyelvű, élő közéleti- és hírműsorok gépi feliratozása. In: XII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2016. pp. 89–99. Szeged (2016)

Tündik, M.Á., Szaszák, G.: Joint word- and character-level embedding CNN-RNN models for punctuation restoration. In: 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 000135–000140. IEEE (2018)

Tündik, M.A., Szaszák, G., Gosztolya, G., Beke, A.: User-centric evaluation of automatic punctuation in ASR closed captioning. In: Proc. Interspeech 2018. pp. 2628–2632 (2018)

Varga, Á., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T., Bordás, C., Mihajlik, P.: Automatic close captioning for live Hungarian television broadcast speech: A fast and resource-efficient approach. In: International Conference on Speech and Computer. pp. 105–112. Springer (2015)

Voleti, R., Liss, J.M., Berisha, V.: Investigating the effects of word substitution errors on sentence embeddings. arXiv preprint arXiv:1811.07021 (2018)