Skálázható szöveg-alapú nyelvazonosító módszer beszédszintézis céljára

Kiss Géza, Németh Géza

Budapest Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{kgeza, nemeth}@tmit.bme.hu

Kivonat: Szövegek nyelvének automatikus azonosítása nagyon fontos több alkalmazásterületen. E cikkben áttekintjük a szövegből történő nyelvazonosítása (language identification, LID) használt főbb módszereket és leírjuk legfontosabb tulajdonságaikat. Ezek egyes, nagyon rövid szövegekre helyes kezelését is igénylő alkalmazásterületeken – mint például a beszédszintézis – jelentkező hiányosságai kezelésére egy új módszert mutatunk be, amely változó hosszúságú N-gramok használatán alapuló, tisztán statisztikai módszer, emellett tetszőleges szöveg helyes azonosítására betanítható, jól skálázható, és viszonylag kis számítási kapacitást igényel az azonosítási fázisban. Bemutatjuk hatékonyságát a tanító- és attól független tesztanyagon, különböző méretű szövegtörzseken való tanítás esetén, kevés és nagyon nagy számú nyelven való működés esetén is. Az eredmények igazolják a megközelítés életképességét.

1 Bevezetés

A szövegből történő automatikus nyelvazonosítás (Language Identification, LID) még mindig fontos kutatási terület, bár sok fejlődés történt az elmúlt évtized folyamán.

A számítógépen tárolt szövegek nyelvének automatikus azonosítására számos alkalmazási területen szükség van. Ilyenek például a webes kereső motorok [1], valamint más webes alkalmazások, amelyek az internetet tudásbázisként használják, például ontológiák tanulásához [2], több nyelven elérhető szövegek gyűjtéséhez számítógéppel segített fordítás céljára [3]. Számos természetes-nyelv feldolgozási eljárás alkalmazásához is szükség van a szöveg nyelvének megbízható előzetes megállapítására, pl. kérdés-válasz (question answering) rendszerekben, automatikus fordításnál [4]. Az is igazolást nyert, hogy a nyelvazonosításra használható módszerek esetenként más szempont szerinti kategorizálásra is használhatónak bizonyulnak, mint például téma vagy szerző szerinti osztályozásra [5].

Egy más jellegű, de szintén fontos használati terület a többnyelvű vagy poliglott beszédszintézis, mivel kevert nyelvi környezetekben (pl. elektronikus levelek vagy webes szövegek felolvasásakor) szükséges, hogy pontosan ismerjük a szöveg nyelvét, különben a létrehozott beszéd érthetetlen vagy legalábbis rendkívül kellemetlen hangzású lesz. Ennél az alkalmazási területnél rövid szövegekre, mondatokra, sőt egészen

a szavak szintjéig megbízhatóan meg kell állapítanunk a nyelvet. Ennek oka egyrészt az, hogy esetenként nem áll rendelkezésre hosszabb szövegrész (pl. sms-felolvasás esetén), másrészt az, hogy mondandónk gyakran tartalmaz beékelt idegen nyelvű szavakat, kifejezéseket (pl. személyneveket, művek címét, idegen eredetű szavakat, szakkifejezéseket).

Jelen cikkben egy olyan újonnan kidolgozott nyelvazonosítási módszert mutatunk be, amely tisztán statisztikai alapon működik, de megfelelő méretű tanító szövegtörzs használatával tetszőleges megbízhatóságú nyelvazonosítás elérhető, a szavak szintjén is. Fontos tulajdonsága, hogy jól használható felismerési arány eléréséhez is csekély tárolási és számítási kapacitást igényel az azonosítási fázisban, és jól skálázható a két szempont bármelyike szerint.

2 A módszer bemutatása

2.1 A jelenleg használt nyelvazonosítási módok áttekintése

Az használt technikákban több csoportra oszthatók. Legnagyobb részük az egyes nyelvekre jellemző írásmód felszíni jelenségeit ragadja meg különböző statisztikus jellemzők használatával; ilyenek például a leggyakoribb szavak listájának használata[6], N-gram alapú módszerek [5], vektortér modellek [7], döntési fák [8], vagy neurális hálók [9]. Ezeknek gyakran hosszabb szövegrészre van szüksége a megbízható nyelvazonosításhoz, de a szavak szintjén való azonosításhoz nem elég megbízhatóak. Emellett azok, amelyek a dokumentumot előzetesen tanító szövegtörzsből nyelvenként készített „nyelvi profilokhoz" hasonlítják (pl. [5], [7]), gyakran számottevő számítási kapacitást igényelnek az azonosítási fázisban is. Ez lényeges szempont a felhasználhatóság szempontjából, míg a betanítási fázishoz szükséges számítási kapacitásnak, a tanító algoritmus futási idejének – ésszerű határok között – nincs jelentősége, főként ha az utóbbi rovására az előbbi csökkenthető.

Másik csoportjuk, főként a beszédszintézis említett jellemzői miatt a szószinten való helyes azonosításra törekedve részletes morfológiai elemzést alkalmaz, pl. DCG-k (Definite Clause Grammar) használatával [10], esetleg közvetve egy helyesírás-ellenőrző használatával [11]. Egy köztes megoldásban nem történik valódi morfológiai elemzés, hanem szótárak (szó és szóelem-listák) elemeire való illeszkedés alapján következtetnek a szavak nyelvére, kiegészítve ezt statisztikai módszerekkel [12].

Összefoglalásként elmondható, hogy a jelenleg használt, tisztán statisztikai alapú megközelítések általában nem adnak eléggé pontos nyelvazonosítást rövid szövegeken, és/vagy nagy számítási kapacitást igényelnek az azonosítási fázisban, míg a részletes morfológiai elemzés végzése nehezen kivitelezhető, főként nagyszámú nyelvre, valamint problémát okozhat egyes alkalmazásokban a szükséges számítási kapacitás.

2.2 A probléma választott megközelítése

A célunk olyan megoldás kidolgozása volt, amely lehetővé teszi nagyon rövid szövegek helyes azonosítását is, akár a szavak szintjéig, és amely kézben tartható abban az értelemben, hogy be lehet tanítani tetszőleges bemenet helyes azonosítására, de egy-

ben általánosító képességgel is rendelkezik, azaz nem látott szavak nyelvét is képes helyesen felismerni a tanítóhalmaz szavaihoz való hasonlóság alapján. Emellett célunk volt a működéshez szükséges adatbázis méretének korlátok között tartása is.

Ennek a célnak megfelel, ha a P (szó | nyelv) valószínűséget egy előzetesen rögzített kritériumnak megfelelő pontossággal becsüljük meg, majd arra a nyelvre döntünk, amelyhez tartozik, ill. homográfok (több nyelvben előforduló szóalak) esetén arra a nyelvre, amelyben a legnagyobb az előfordulásának valószínűsége. A szavakra meghatározott nyelvi címkék alapján dönthetünk a szövegrész nyelvére. A szavak kontextusa alapján számított nyelv-valószínűség figyelembe vételével akár szószinten helyes nyelvazonosítást is kaphatunk, még homomorf szavak esetén is. Ez azt is lehetővé teszi, hogy egy egynyelvű szövegbe beszúrt idegen nyelvű szó a valódi nyelvének megfelelő azonosítást kapja, szemben a környezet alapján determinisztikusan döntő naiv megközelítéssel.

Megfelelő valószínűség-becslési módszerrel az ismert szavak írásmódja alapján képesek lehetünk korábban nem látott szavakra is becsülni ezt a valószínűséget. Ez a megközelítés megőrzi a szó-alapú módszerek előnyét, a kézbentarthatóságot, kiterjesztve azt általánosító képességgel, és szóalapon is helyes működést tesz lehetővé.

2.3 A kidolgozott módszer leírása

Az általunk kidolgozott módszer változó méretű N-gramok használatán alapszik. Míg a szokványos Markov-modellt használó megoldásban rögzített hosszúságú előzményt használunk egy karakternek az előzőek utána való következése valószínűségének becslésére, a javasolt módszerben többféle hosszúságú előzményt használunk, amely hosszt minden környezetre egy tanítási folyamat során határozunk meg. A tanítás 0 hosszúságú karakter környezettel indul minden karakterre (ez a karakter előfordulásának valószínűsége), majd ezt a hosszt egyes környezetekben növeli a megcélzott valószínűség-becslési kritérium elérésére, amely lehet pl. a leggyakoribb szavak helyes felismerése. A folyamat korlát nélküli folytatása a láncszabályt adja, ezzel pedig nyelvenkénti szó-valószínűséget, ezért a tanító folyamat tetszőleges tanító halmaz esetén jobb szó-valószínűség becsléshez, így korrekt azonosításhoz konvergál a tanítóhalmazra. Hosszabb N-gramokat tartalmazó, nagyobb méretű felismerő adatbázis használatával pontosabb azonosítási eredmény érhető el megfelelő tanítás esetén.

Az N-gram környzetekhez tartozó feltételes valószínűségeket fában tárolva úgy is felfoghatjuk a módszert, hogy egy fajta döntési fa tanítását jelenti a szóvalószínűségek becslése céljából. A fa bővítésének irányát a bővítésnek a becslési kritérium szempontjából meghatározott „hasznossága” szerint határozzuk meg. Több ilyen hasznosság-függvénnyel dolgoztunk, melyeket a tanító halmazon való helyes felismerési arány (recall) és az attól független teszt-halmazon való eredmény (precision), azaz az általánosító képesség mellett az alapján is vizsgáltunk, hogy mennyire tömör adatbázist. A tömörséget nem pusztán a mérettel jellemeztük – hiszen nem közömbös, hogy milyen felismerési arányt ad a tömörebb adatbázis – hanem a felismerési/méret hányadossal, LID adatbázis teljesítményének nevezünk. A legjobb adatbázis méretet adó függvény a feltételes valószínűségek logaritmusának nyelvek közötti eltérését, míg a legjobb általánosító képességet adó emellett az N-gram előfordulásának valószínűségét is figyelembe vevő, entrópia-jellegű mennyiség.

A módszerben újítás még, hogy a nyelvek független szemlélése helyett a nyelvenkénti valószínűségek eltérésének helyes becslésére törekszünk, amelytől kisebb adat-

bázis méretet várunk, hiszen így a tanítás során a két nyelvet megkülönböztető jellemzőkre való „koncentrálásra” késztetjük az algoritmust.

3 Eredmények

Több tesztet végezünk eltérő méretű tanító és felismerendő beszédkorpuszon. Először három nyelvre (angol, német, magyar) végeztünk betanítás nagyméretű korpuszon (British National Corpus, Project Guttenberg DE, Magyar Elektronikus Könyvtár), azoknak a hozzávegyült idegen nyelvű részektől való megtisztítása nélkül, a leggyakoribb szavak 90%-ának helyes felismerésére. A tesztet az előzőtől független szöveghalmazon végeztük (Project Guttenberg, online magyar újságok). Az [5]-ben bemutatott módszer egy web-en megtalálható implementációjához[81] használt, 77 nyelvhez tartozó kis méretű (5 kilobájt) szövegre is elvégeztük a betanítást.

A helyes azonosítás százalékos eredményeit az 1. táblázat tartalmazza. Az osztályozott szövegek áttekintése azt mutatta, hogy az első esetben a más nyelvűnek osztályozott szövegek gyakran valóban nem a csoportjuknak megfelelő nyelvhez tartoztak, vagy kevert nyelvűek voltak, valamint hogy valóban pontos szó-alapú működéshez szükség van egyes (formátumukat tekintve) nyelv-függetlennek tekinthető kifejezések azonosítására, melyekre példák a római számok, internet és e-mail címek, dátumok, nemzetközi szavak (pl. „tel.”, „fax.”), rövidítések, mértékegységeket tartalmazó kifejezések (pl. „2 cal”).

Vizsgáltuk a szavak környezete alapján számított nyelv-valószínűség figyelembevételének hatását is. Ehhez az azonosító által nyelvi címkézett szövegből kalibráltunk nyelv-valószínűség becslő szabályokat, amelyek a szomszédos szavak nyelvével való egyezés valószínűségét adták. Ezekkel a szószintű azonosítási eredménye a harmadik adatbázist használva a német korpuszon a korábbi 45%-ról 65%-ra növekedett, majd a folyamatot ismételve 70%-ra, igazolva az iteratív tanítás-finomítás létjogosultságát.

1. Táblázat: eredmények különböző tanító halmazok esetén,
egy azoktól független 3 nyelvű teszt szöveggel, szó és mondat szintű azonosításra

Tanító (szavak)	Nyelvek	Adatbázis	Tanító, szó	Teszt, szó	Teszt, mondat
20641974-89138922	3	54 kbájt	99,6%	94,2%	98,5%-99,5%
580-694 (5 kbájt)	3	7,4 kbájt	95,5%-97,8%	79,6%-87,4%	91,7%-97,2%
465-1681 (5 kbájt)	77	5,4 mbájt	70,0%-99,8%	30,1%-59,6%	71%-84,0%

4 Konklúziók

Bemutatottunk egy új szöveg-alapú nyelvazonosítási módszert, amely a nyelvenkénti szóvalószínűségek eltérésének döntési fa-alapú becslésén alapszik. A módszer statisztikai alapú, így viszonylag könnyen létrehozható nagy számú nyelvre működő változata, jól skálázható a felismerési arány és az adatbázis méret viszonylatában,

[81] http://odur.let.rug.nl/~vannoord/TextCat/Demo/

tetszőleges szó kívánt azonosítására tanítható, és az azonosítási fázisban relatíve kicsi számítási kapacitást igényel. További kutatást igényel a bemutatott, leggyakoribb szavakra való tanítás összehasonlítása a döntési fa adott pontosságú feltételes-valószínűség ill. ezek nyelvek közötti eltérésének becslésére való tanítás. A módszer várhatóan más kategorizálási feladatokban való felhasználása is lehetséges, például szófaj-címkézés (POS tagging).

Bibliográfia

1. Risvik, K. M., Michelsen, R.: Search engines and Web dynamics. Computer Networks, Vol. 39, Issue 3. (2002) 289-302
2. Kilgarri, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Computational Linguistics, 29(3) (2003) 333-348
3. Volk, M.: Using the Web as Corpus for Linguistic Research. In: Pajusalu, R., Hennoste T. (eds.): Tähendusepüüdja. Catcher of the Meaning. Festschrift for Professor Haldour Õim. University of Tartu, Estonia: Publications of the Department of General Linguistics 3 (2002)
4. Bond, F.: Toward a Science of Machine Translation. Proc. of the MT Roadmap Workshop at TMI-2002, Keihanna, Japan (2002)
5. Canvar, W. B., Trenkle, J. M.: N-gram based Text Categorization. Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas (1994) 161-176
6. Németh, G., Zainkó, Cs., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P.: The Design, Implementation and Operation of a Hungarian E-mail Reader. International Journal of Speech Technology, Kluwer Acedemic Publishers, Vol. 3, Numbers 3/4. (2000) 217-236
7. Prager, J. M.: Linguini: Language Identification for Multilingual Documents. Proc. of the Thirty-Second Annual Hawaii International Conference on System Sciences, Vol. 1. (1999) 2035
8. Häkkinen, J., Tian, J.: N-gram and Decision Tree-based Language Identification for Written Words. Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio Trento, Italy (2001)
9. Tian, J., Suontausta, J.: Scalable neural network based language identification from written text. Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 1 (2003) 48-51
10. Pfister, B., Romsdorfer, H.: Mixed-lingual text analysis for polyglot TTS synthesis. Proc. of Eurospeech 2003 (2003) 2037-2040
11. Halácsy, P., Kornai, A., Németh, L.,, Rung, L., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. Proc. of LREC 2004 (2004) 203-210
12. Marcadet, J. C., Fischer, V., Waast-Richard, C.: A Transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis. Proc. of Eurospeech 2005 (2005) 2249-2252