Gépi tanuló algoritmus automatikus címkézésre és alkalmazása beszédszintézis céljára

KISS GÉZA, NÉMETH GÉZA

BME Távközlési és Médiainformatikai Tanszék
{kgeza, nemeth}@tmit.bme.hu

Lektorált

Kulcsszavak: automatikus címkézési módszerek, gépi tanuló algoritmus, nyelvazonosítás, LID, TTS

A cikkben egy új, szöveg címkézési problémák megoldására használható tanuló algoritmust mutatunk be. Illusztráljuk a probléma fontosságát szövegfelolvasó rendszerek, ezen keresztül távközlési alkalmazások számára. Áttekintjük a jelenleg használt módszereket a szavankénti nyelvi címkézés problémájára. Bemutatjuk az általunk javasolt módszert, és demonstráljuk eredményességét a nyelvazonosításban, három különböző tanítóhalmazra, nagyméretű tesztkorpuszokon.

1. Bevezetés

Egyre szaporodik mind hazánkban, mind a nemzetközi színtéren a beszédalapú távközlési szolgáltatások száma, amelyekben a bejövő hívásokat IVR (Interactive Voice Response) rendszer fogadja. A legmodernebb rendszerekben a felhasználó már nem csupán billentyűzéssel, hanem beszéddel is közölheti mondandóját az ASR (Automatic Speech Recognition) rendszeren keresztül, és válaszul jó minőségű beszédet kap. Egyszerű esetekben a válasz előre felvett (prompt) illetve néhány elemből összeállított (kötött szótáras beszédszintézis) hangüzeneteket tartalmaz.

Amennyiben a kimondandó üzenet tartalma megjósolhatatlan, vagy nagyon nagy variabilitást mutat, akkor a beszédet szövegfelolvasó rendszerrel (TTS, Text-to-Speech) állítjuk elő. Néhány példa az utóbbira a hazánkban működő szolgáltatások közül: a T-Mobile telefonos e-mail felolvasó rendszere [1], számszerinti tudakozó rendszere [2], vagy a T-Com hangos SMS szolgáltatása.

A TTS rendszerektől elvárjuk, hogy pusztán az írott alakból jó minőségű, érthető és helyesen intonált beszédet hozzanak létre. Azonban elmondható, hogy a használt írásrendszerek (akár a magyar, akár más nyelvé) a beszéd információtartalmának csak egy tört részét tartalmazzák, a prozódiára csupán kis mértékben, néhány írásjeggyel utalnak. Az ember, ha olvas, a világról való ismerete valamint a szövegkörnyezet alapján egészíti ki elméjében a szöveget a hiányzó információkkal, ami segíti abban, hogy (szükség esetén) azt megfelelő kifejező erővel fel is olvassa. Pusztán a helyes kiejtés, hangsúlyozás megállapításához is szükség van olyan információk helyes felismerésére, mint a szöveg nyelve, az esetlegesen a szövegbe beékelt idegen szavak nyelve, az egyes szavak szerepe a mondatban (szófaj, nyelvtani szerkezet).

Az *1. ábrán* láthatunk néhány példát, ahol ez nem triviális. Egyező szóalakok esetén a helyes eredmény eldöntése nehéz, illetve a mondat nyelvétől eltérő nyelvű szó beékelődése esetén is.

A cikkben áttekintjük a szöveg alapú nyelvazonosításra használt módszereket, majd leírunk egy felcímkézett szövegből történő tanulásra szolgáló gépi tanuló algoritmust, amely használható különböző címkézési feladatok automatikus elvégzésére. A módszert nyelvazonosításra történő címkézés példáján mutatjuk be, utalva a szófaji címkézés lehetőségére is, amelyek például a távközlésben is egyre gyakrabban használt TTS rendszerek számára fontos információk; ezek mellett a módszer feltehetően számos más területen is használható.

2. Módszerek szöveg alapú nyelvazonosításra

A nyelvazonosítás megnevezéssel (Language Identification, LID) egyaránt illetik az ASR alkalmazása előtt a beszéd nyelvének megállapítására vonatkozó módszereket, és a szöveg nyelvének megállapítására vonatkozókat. A nyelvazonosítás tekinthető a címkézési, illetve osztályozási problémák egy speciális esetének, így a tanulságok más területen is alkalmazhatók - például szófaji címkézés esetére is.

Bár első ránézésre a legtöbb, a téma iránt kicsit is érdeklődő embernek vannak ötletei írott szöveg nyelvének automatikus megállapítására, valójában több olyan kérdés nehezíti a megoldást, ami miatt a feladat közel sem triviális. Míg viszonylag egyszerű módszerekkel nagy valószínűséggel a helyes nyelvet rendel-

1. ábra Példák nem triviális nyelvi illetve szófaji címkézési feladatokra

„a test” – magyar vagy angol kifejezés:	A lélek és a test.	This is a test.
idegen kiejtésű rész magyar szövegben:	A „Sok hűhó semmiért” Shakespeare műve.	
„egy” – számnév vagy határozatlan névelő:	Egy vagy két alma.	Egy alma esett le a fáról.

hetjük egy hosszabb szöveghez, a szöveg szavai nyelvének helyes megállapítása kevert nyelvű, több nyelvben is előforduló szóalakokat tartalmazó szövegen jóval nehezebb feladat, kiegészítve azzal, hogy a valóban hatékony megoldásnak nem csak pontosnak, hanem gyorsnak és viszonylag kis tárigényűnek kell lennie. Valamennyi követelménye egyidejű teljesítése már összetett, nehezen megoldható feladat.

2.1. Morfológiai elemzés

A módszerek egy csoportja a szószinten, sőt morféma-szinten való helyes azonosításra törekedve részletes morfológiai elemzést alkalmaz, például a DCG-k (Definite Clause Grammar) használatával [3], esetleg közvetve egy helyesírás-ellenőrző használatával [4]. Egy más, 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 [5]. Magyar nyelven elérhető morfológiai elemző a Humor [6], valamint a Hunmorph [7]. Azonban ha nem csak a magyar/nem magyar döntést kell meghoznunk, hanem több lehetséges nyelv közötti döntés is szükséges, akkor mindegyikhez szükséges morfológiai elemző, ami nehezen kivitelezhető feladatot jelent, valamint egyes alkalmazásokban problémát okozhat a szükséges nagy számítási kapacitás.

2.2. Szóalapú módszer

A szóalapú módszerek [8] azon a megfigyelésen alapulnak, hogy minden nyelvben van a szavaknak egy olyan, viszonylag kis halmaza, amelyet nagyon sokat használnak. Ezért egy nyelvhez tartozó ilyen szavak jelenléte nagy biztonsággal jelzi, hogy a szöveg az adott nyelven íródott. A leggyakoribb 1000 szó az összes előforduló szó 50-70%-át is kiteheti [9].

2. ábra
Példák öt európai nyelv leggyakoribb szavai között többen is előforduló szóalakokra

szóalak	angol	német	spanyol	lengyel	magyar
de			X		X
a	X		X		X
na				X	X
to	X			X	
in	X	X			
do	X			X	
el			X		X
is	X				X
es		X	X		
mit		X			X
was	X	X			
ha			X		X
so	X	X			
on	X			X	
mi			X	X	X
most	X				X
ja		X		X	
be	X				X
ma				X	X

A módszer hátránya, hogy minél rövidebb a szöveg, annál valószínűbb, hogy a halmaz egy szava sem jelenik meg benne. Emellett a szólista összeállításához is számottevő erőfeszítés kellhet, hiszen vannak olyan szavak, amelyeknek az írott formája több nyelven is ugyanaz, mint ahogy azt a 2. ábrán található példákból látjuk.

2.3. Vektortér módszerek

A vektortér módszerek alapgondolata, hogy a vizsgálandó dokumentumhoz és a lehetséges besorolási kategóriákhoz is egy-egy jellemzővektort rendel, amelyek bizonyos számszerűsíthető tulajdonságok leképzései, a kategóriák esetében a jellemző fontosságával súlyozva. A dokumentumnak az egy kategóriába való illeszkedését a jellemzővektorok skalár-szorzatával jellemzi, ahol a 0 érték ortogonalitást, az 1 érték megfelelést jelent. A [10]-ben leírt módszerben a jellemzők N-gramok N=2..5 értékekkel, valamint rövid, illetve korlátlan méretű szavak, melyekkel szövegek nyelvének megállapítását végzik.

2.4. Neurális hálók

A [11] megközelítésében egy többrétegű perceptront (Multi Layer Perceptron, MLP) tanítanak be, amelynek bemenetére a szöveg egy pozíciójára illesztett ablakba eső karaktereket helyezik, kimenetén pedig nyelv-valószínűség értéket ad. Ezeket a szó összes betűjére kombinálva szavankénti nyelvi döntést hoznak.

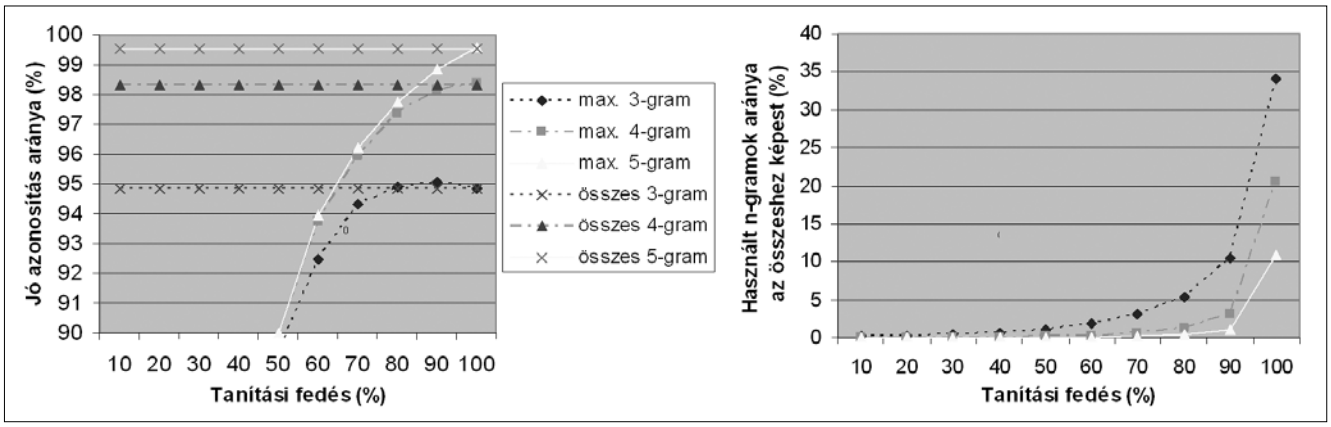
2.5. Döntési fa alapú módszerek

A 3. pontban javasolt módszerünktől eltérő döntési fa tanítást használ [12]: karakterenként külön döntési fát tanítanak, ahol az ágak a szomszédos karakterek azonosságára kérdeznek rá, a leveleken pedig a legvalószínűbb nyelv címkéje található. Szavanként hoznak döntést a szó karaktereire kapott nyelv-jelöltek közül a legtöbbször előfordulóra döntve.

2.6. N-gram alapú módszerek

Az N-gram alapú módszerek szórészeket használnak az azonosítás alapegységéül. Ezek két, három vagy több karakter sorozatából állhatnak; egy módszeren belül esetlegesen több különböző hosszt is használhatunk egyidejűleg. Az N-gram gyakoriságok statisztikáját a különböző nyelvekhez tartozó tanító szövegkorpuszból készíthetjük.

Az N-gramok jól kezelnek több problémát, amelyekben a szóalapú módszerek nem adnak megoldást. Ezek egyike az elektronikus szövegekben gyakori betűhibák (elgépelések, karakterfelismerési hibák), mivel ezeknek olyan nagy változatossága van, amit nem vagy nagyon nehezen lehet megoldani a szóváltozatok tárolásával, viszont az n-gram statisztikát



3. ábra
A rögzített és változó hosszúságú előzményen alapuló módszerek teljesítményének és méretének összehasonlítása a tanító halmazon ML becslés használatával

nem rontják el nagy mértékben. Egy másik szempont, hogy a „data sparseness" probléma (amely szerint gyakorlatilag soha nincs annyi adatunk, amely minden szóbajöhető elemről adna eloszlási információt; mindig találkozunk majd a tanítóhalmazban elő nem fordulókkal) jóval kisebb mértékben jelentkezik ebben az esetben, mintha szavakat vizsgálnánk, mivel egy szóban a szóhossz négyzetével arányos számú N-gram található. Egy jellegzetes példája ennek a megközletésnek Canvar és Trenkle módszere [13].

2.7. Markov-modell

Tudjuk, hogy egy *l* karakter hosszú szó előfordulási valószínűségét megkaphatjuk a láncszabály szerint:

$$\mathrm{P}(\text{szó} \mid \text{nyelv}) = \prod_{i=1}^{l+1} \mathrm{P}(c_i \mid c_0 \ldots c_{i-1}, \text{nyelv}) \qquad (1)$$

$c_1,\ldots,c_l$ ahol a szó karakterei, $c_i, i \leq 0$, illetve c_{l+1} speciális, szót kezdő, illetve bezáró jelek.

Ezt a valószínűséget közelíthetjük a Markov-modell segítségével, azaz feltételezve, hogy ez egy véletlen folyamat, amelyben a következő karakter valószínűségi eloszlása csak a jelenlegi állapoton múlik. Hagyományosan az állapoton az előző *n-1* karaktert értik a nyelvazonosítás esetén:

$$\mathrm{P}(\text{szó} \mid \text{nyelv}) \approx \prod_{i=1}^{l+1} \mathrm{P}(c_i \mid c_{i-n+1} \ldots c_{i-1}, \text{nyelv}) \qquad (2)$$

A karakterek egy adott környezethez tartozó feltételes valószínűségét az ML (Maximum Likelihood) becslés alapján közelíthetjük:

$$\mathrm{P}(c_i \mid c_{i-n+1} \ldots c_{i-1}) \approx \frac{\#\, c_{i-n+1} \ldots c_{i-1} c_i}{\#\, c_{i-n+1} \ldots c_{i-1}} \qquad (3)$$

Mivel gyakorlatilag bármekkora méretű tanítóhalmaznál kell számítanunk arra, hogy előfordulhatnak korábban nem látott N-gramok, ezért elkerülhetetlen valamilyen simító (smoothing) módszer használata, amelynek megválasztása jelentősen meghatározza a becslés minőségét.

A szó nyelven belüli valószínűségét ezután használhatjuk a nyelv valószínűségének becslésére:

$$\mathrm{P}(\text{nyelv} \mid \text{szó}) = \frac{\mathrm{P}(\text{szó} \mid \text{nyelv}) \cdot \mathrm{P}(\text{nyelv})}{\mathrm{P}(\text{szó})} \qquad (4)$$

A legvalószínűbb nyelv megállapításához a nevezőt (amely a vizsgált szövegre, és nem a nyelvre jellemző érték) figyelmen kívül hagyhatjuk. A nyelv valószínűségét vehetjük fix értéknek, vagy a kontextus alapján dinamikusan változónak.

$$\text{nyelv} = \arg\max_{\text{nyelv}} \mathrm{P}(\text{szó} \mid \text{nyelv}) \cdot \mathrm{P}(\text{nyelv}) \qquad (5)$$

Elméletileg a Markov-modellel való becslés megfelelő simító módszer választásával tetszőlegesen pontos becslést adhat, ha *n* elég nagy. (Ha *n* a maximális szóhossz, akkor a láncszabályt, ezen keresztül a pontos valószínűséget kapjuk.) A gyakorlatban azonban rendszerint $n = 2$ (bigrammok) vagy $n = 3$ (trigrammok) értéket használnak, két okból: a data sparseness probléma miatt, és mert nagy *n*-hez számottevő tárolási kapacitásra volna szükség.

Azonban ezek a hosszak nem teszik lehetővé a pontos osztályozást több nyelvre való döntés esetén, ahogy ezt a *3. ábrán* láthatjuk ML becslés használatával az *1. táblázat* első sorában leírt tanítóhalmazra; ez az elvi korlátot jelenti. Mint láthatjuk, az adott tanítóhalmazon való tanítás és ugyanazon való tesztelés esetén a helyes azonosítási arány még 5-grammok esetén sem éri el a 100%-ot, de ez már meglehetősen nagy adatbázist jelentene, valamint az ismeretlen szövegen feltétlenül használandó simítást még nem tartalmazza. A javasolt módszerrel létrehozott adatbázis mérete az előző 10-35%-a hasonló felismerési aránynál, valamint 100%-ra való tanításnál csekély mértékben jobb eredményt is ad.

2.8. Tanulságok a nyelvazonosításról

Ebben a szakaszban a szövegből történő nyelvazonosítás néhány módszerét tekintettük át, melyeknek két nagy csoportja a részletes elemzés alapján való döntés és a statisztikai módszerek.

Összefoglalásként elmondható, hogy a jelenleg használt, tisztán statisztikai alapú megközelítések általában nem adnak eléggé pontos nyelvazonosítást rö-

vid szövegeken, így a szavak szintjén való azonosításhoz nem elég megbízhatóak. Emellett azok, amelyek a dokumentumot előzetesen tanító szövegtörzsből nyelvenként készített „nyelvi profilokhoz” hasonlítják a vizsgálandó részt, gyakran számottevő számítási kapacitást igényelnek az azonosítási fázisban is. A részletes morfológiai elemzés végzése viszont nehezen kivitelezhető, főként nagyszámú nyelvre, valamint problémát okozhat egyes alkalmazásokban a szükséges nagy számítási kapacitás.

3. A javasolt módszer

Az alábbi módszert elsősorban a szövegből történő nyelvazonosítás feladatára dolgoztuk ki, ezért ebből a szempontból tárgyaljuk. Azonban a „nyelv” kifejezés helyett mindenütt „osztályt”, a „szó” helyett „szövegegységet” behelyettesítve, egy általános címkéző rendszer leírásaként is olvashatjuk.

3.1. A módszer alapelve

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 egyben 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 – például elég pontosan ahhoz, hogy arra a nyelvre legyen a legnagyobb a becsült valószínűség, amelyre legnagyobb a tanítóhalmazból számolt valószínűség. Módszerünk másik összetevője, hogy a szavak kontextusa alapján számítjuk az adott szóra a nyelv-valószínűséget. Ezután az (5) egyenletnek megfelelő nyelvre döntünk minden szó esetén.

Ez a megközelítés elvileg lehetővé teszi, hogy szószinten helyes nyelvazonosítást kapjunk, még homomorf (esetünkben egyidejűleg több nyelvhez tartozó) szavak esetén is a szövegkörnyezetnek megfelelően, valamint hogy egynyelvű szövegbe beszúrt idegen nyelvű szó a valódi nyelvének megfelelő azonosítást kapja, szemben a környezet alapján determinisztikusan döntő naiv megközelítéssel. Ha szövegrészenként (pl. mondatonként) egy nyelvre való döntés szükséges, a szavakra meghatározott nyelvi címkék alapján dönthetünk, például a többségi szavazás elvének megfelelően.

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. A megközelítés sikerességéhez a kulcs a feltételes valószínűség és a nyelvi valószínűség megfelelő pontosságú becslése.

3.2. Feltételes valószínűségek becslése

A P(szó | nyelv) valószínűség becslésére általunk kidolgozott módszer változó méretű N-gramok használatán alapszik. Míg a szokványos Markov-modellt használó megoldásban rögzített hosszúságú előzményt használunk egy karakternek az előzőek után való következésének valószínűségi becslésére, a javasolt módszerben többféle hosszúságú előzményt használunk.

$$\mathrm{P}(\text{szó}\,|\,\text{nyelv}) \approx \prod_{i=1}^{l+1} \mathrm{P}(c_i \,|\, c_{i-n_i+1}\ldots c_{i-1}, \text{nyelv}), n_i \geq 0 \quad (6)$$

Az n_i hosszt minden környezetre egy tanítási folyamat során határozzuk meg. A tanítás 0 hosszúságú karakter-környezettel indul minden karakterre (ez a karakter előfordulásának valószínűsége), majd ezt a hosszt bizonyos környezetekben növeli a megcélzott valószínűség-becslési kritérium elérésére, amely lehet például a leggyakoribb szavak bizonyos százalékának helyes felismerése. A növelési folyamat korlát nélküli folytatása a láncszabályt adja, ezzel pedig nyelvenkénti szó-valószínűséget (megfelelő simítómódszer alkalmazása esetén), ezért a tanító folyamat tetszőleges tanító halmaz esetén jobb szó-valószínűség becsléshez, így a tanítóhalmazra a korrekt azonosításhoz konvergál. Hosszabb N-gramokat tartalmazó, nagyobb méretű felismerő adatbázis használatával pontosabb azonosítási eredmény érhető el. Ezáltal a módszer skálázható, mivel az adatbázis méret-kívánt felismerési arány-páros egyike szabadon megválasztható.

Az N-gram környezetekhez 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 és az attól független teszt-halmazon való eredmény, azaz az általánosító képesség mellett az alapján is vizsgáltunk, hogy mennyire tömör adatbázist eredményez. A tömörséget nem pusztán a mérettel jellemeztük – hiszen nem közömbös, hogy milyen felismerési arányt ad a tömörebb adatbázis – hanem a felismerési arány/méret hányadossal, amelyet a LID-adatbázis teljesítményének nevezünk. A legjobb hasznosság-függvényeket és az őket jellemző grafikonokat a *4. ábrán* láthatjuk. A legjobb általánosító képességű megoldáshoz és a legtömörebb adatbázishoz eltérő hasznosság-függvény szükséges.

Újítás még, hogy a nyelvek független szemlélése helyett a nyelvenkénti valószínűségek eltérésének elégségesen pontos becslésére törekszünk, amelytől kisebb adatbázis méretet várunk, hiszen így a tanítás során a két nyelvet megkülönböztető jellemzőkre való

„koncentrálásra” késztetjük az algoritmust. Ezáltal az algoritmus nem csak skálázható, hanem automatikusan skálázódik is a probléma nehézségének megfelelően. Például, ha két nyelvet a karakterkészlete is megkülönböztet, akkor a 100%-osan helyes azonosítás megcélzásakor is megáll a tanítás az unigrammok (1 karakteres N-gramok) használatánál.

3.3. Nyelvi valószínűségek használata

3.3.1 A nyelvi valószínűség fogalma

A „nyelvi valószínűség” kifejezésen az egyes nyelvek bizonyos környezetben való előfordulásának valószínűségét értjük; ezt használjuk (5)-ben a P(nyelv) helyén. E valószínűség becsült értékének kiszámításában megoldásunkban részt vehet a környező szavak nyelve és a közöttük lévő központozás. Azért választottuk ezeket a lehetséges vizsgálandó jellemzők közé, mert ezek szerepet játszhatnak az emberi értelmezés kialakításában is, de a lehetséges kérdéseknek más, szűkebb, illetve tágabb halmaza is elképzelhető.

A nyelvi valószínűség modellezését is tekinthetjük egy döntési-fa tanításnak, amelyben a 3.2 pontban adotthoz hasonlóan feltételes valószínűségek tárolására használjuk a fát, ám attól (és az általában használt módszerektől [15,16]) eltérően, nem csupán a szót megelőző, hanem az azt követő szavak azonosságát is felhasználhatjuk.

Ebben a megközelítésben nehézséget jelent egyrészt az, hogy a szó környezetében lévő szavak nyelvét is ismertnek tételezzük fel a szó nyelvének megállapításához, ami a valóságban nem teljesül, másrészt az, hogy a sok lehetséges kérdésfajta miatt, amelyek a fa ágain való továbbhaladást vezetik, nagyon nagy lehet a lehetséges döntési-alternatívák száma. Ezeknek a megoldását a következő két pontban tárgyaljuk.

3.3.2 A legvalószínűbb címke-sor megkeresése

Az első nehézség egy matematikailag megfogható problémát takar, bár a hatékony megvalósítás nem triviális. A feladat az, hogy megtaláljuk a címke-sort, amelyre az N szóból álló mondatra a valószínűség maximális lesz.

$$\{\text{nyelv}_{l_i} \mid i \in [1..N]\} = \arg\max \prod_{i=1}^{N} \text{P}(\text{nyelv}_{l_i} \mid \text{szó}_i) =$$

$$= \arg\max \prod_{i=1}^{N} \frac{\text{P}(\text{szó}_i \mid \text{nyelv}_{l_i}) \cdot \text{P}(\text{nyelv}_{l_i})}{\text{P}(\text{szó}_i)} \quad (7)$$

A $P(szó_i)$ tényezőt itt is figyelmen kívül hagyhatjuk, hiszen az nem függ a nyelvi címkétől, így az eredményt nem befolyásolja.

A legvalószínűbb címkesor kimerítő kereséssel történő megkeresése L számú lehetséges címke esetén L^N számítási lépést jelentene, ami a mondat hosszával exponenciálisan növekvő keresési teret határoz meg; ez kevés gyakorlati alkalmazásban megengedhető. Ezért az optimálist valamilyen módon közelítő módszer alkalmazása szükséges.

A jelenlegi rendszerben az alkalmazott közelítés a következő: először uniform nyelv-valószínűséggel kiszámítjuk a címkesor egy közelítését, majd az így kapott környezetet figyelembe véve újraszámítjuk a címkéket az egész sorra, balról jobbra haladva. Ha egy címke módosul a korábbihoz képest, akkor a tőle balra eső és általa (mint környezet által) esetleg befolyásolt címkéket újraszámítjuk, de csak akkor módosítjuk (részben az iteráció elkerülése végett) ha a szó nyelvére számolt valószínűség a korábbi nyelvnél nagyobb. A módszer tovább javítható, például szimulált lehűtés alkalmazásával (csökkenő mértékben engedve közvetlenül nem valószínűség növekedést eredményező címke-módosításokat is).

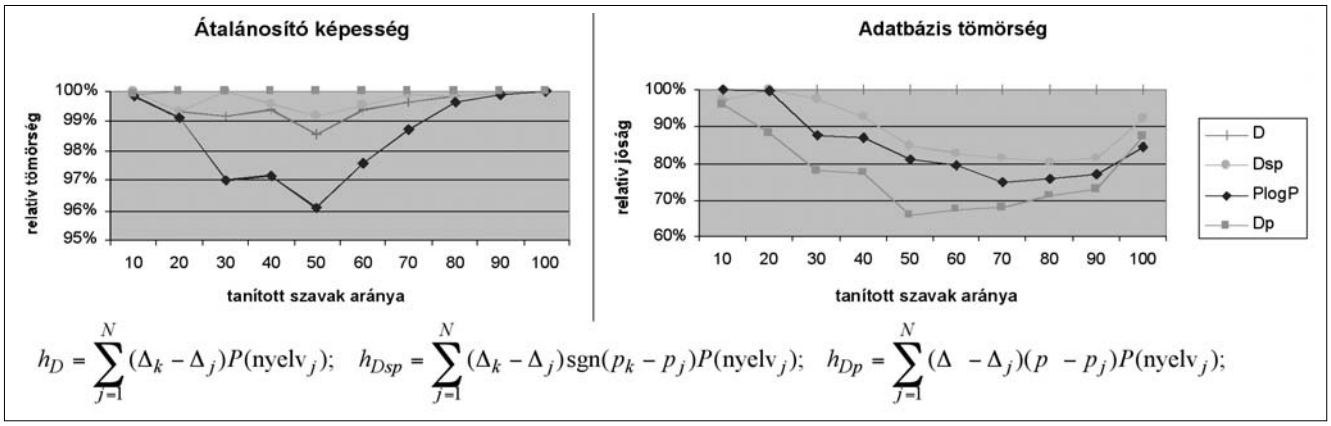
3.3.3 Szabály-sablonok használata

A második nehézség kiküszöbölésére mintarendszerünkben úgynevezett szabály-sablonokat alkalmazunk, amely teret ad a nyelvészeti tudás használatának, illetve heurisztikának, és ezzel számottevően csökkentheti a megvizsgálandó alternatívák számát.

A szabálysablonok a következő formát vehetik fel (BNF leírásban):

<sablon> ::= { <címke-leírás> [<szeparátor-leírás>] }
<címke-leírás> ::= L{[<azonosító szám>][?|*] | "<címke-név>"}
<szeparátor-leírás> ::= S{[<azonosító szám>][*] | "<szeparátor>"}

4. ábra Hasznosság-függvények jellemzése
(a legjobb eredményhez viszonyított értékek a leggyakoribb szavak különböző százalékaira);
L a nyelvek száma, k a szó valódi nyelve, p a felveendő n-gram előfordulásának valószínűsége,
Δ az n-gram felvétele miatt a számolt valószínűség-értékben bekövetkező változás.



$$h_D = \sum_{j=1}^{N} (\Delta_k - \Delta_j) P(\text{nyelv}_j); \quad h_{Dsp} = \sum_{j=1}^{N} (\Delta_k - \Delta_j) \text{sgn}(p_k - p_j) P(\text{nyelv}_j); \quad h_{Dp} = \sum_{j=1}^{N} (\Delta - \Delta_j)(p - p_j) P(\text{nyelv}_j);$$

A címke ebben az esetben a nyelv Az egyforma azonosító számmal rendelkező címkéknek illetve szeparátoroknak meg kell egyezniük. A csillaggal („*”) jelölt leírások olyan elemeket jelölnek, amelyeknek a különböző értékei nem hoznak létre egymástól független szabályokat, csupán az egymás közötti (közös azonosítóval rendelkezők közötti) egyezésnek kell teljesülni, míg a csillaggal nem jelölt leírások a címke vagy szeparátor minden előforduló értékére külön szabályt hoznak létre.

A kérdőjellel („?”) jelölt címke-leírás az, amelyre az egyes címkék előfordulási valószínűségeit megfigyeljük a tanító halmazban. A használandó szabályok előállításához a szabálysablonokat ennél a megfigyelési pontnál fogva illesztjük a tanítóhalmaz minden szavára, és ahol a sablon illeszthető, a konkrét illeszkedő értékekkel kitöltött szabályt hozunk létre, amely a folyamat végén tartalmazza az egyes címkék előfordulási valószínűségét (az illeszkedések számából, és azon belül az egyes címke-fajták előfordulásának számából számítva).

Ha egy pozícióban több szabály is alkalmazható, akkor eltérő számú feltétel-résszel rendelkezők esetén a több feltételt tartalmazót alkalmazzuk (a döntési fa elvét követve). A szabálysablonokra és az azokból előállított szabályokra később láthatunk példákat az *5. ábrában*, a nyelvazonosítás témakörére.

4. Alkalmazás szöveg alapú nyelvazonosításra

4.1. Tanítóhalmaz gyűjtése

Elmondható, hogy jelenleg nem állnak rendelkezésre nagyobb méretű, szószinten helyes nyelvi címkékkel ellátott szöveganyagok, valamint valóban egynyelvű nagy méretű szövegtörzsek összeállítása is nehéz (már csak a minden nyelvben használatban lévő idegen eredetű nevek és kifejezések miatt is, melyek olyan, egy nyelvre specializált korpuszokba is bekerülhetnek, mint például a Project Gutenberg). Ezért a tanításhoz használható szöveghalmazok összeállítása is nehézségekbe ütközik, annak ellenére, hogy az Internetről nagyon nagy mennyiségű szöveg tölthető le gyakorlatilag tetszőleges nyelvre – melyek persze az említett kevert jelleget mutatják.

A probléma egyik áthidalása lehet, hogy a szövegeket egynyelvűnek tekintve betanítjuk a nyelvazonosítót több nyelvre, majd ezzel címkézzük a szövegeket. A tanító korpuszok névleges nyelvétől egyértelműen eltérő nyelvűnek megállapított szövegeket, illetve mondatokat kihagyva, a tanítás ismételhető, így már tisztább, az egynyelvűt jobban közelítő szöveggel taníthatjuk újra a rendszert. A folyamat ismételhető, amíg történik finomodás.

Viszonylag rövid méretű szövegeket tartalmazó korpuszon (ilyen lehet például egy újság cikkeinek archívuma) jelentősen torzíthatja a szó- illetve n-gram statisztikákat a szövegekben általában jelenlévő, nagyrészt ismétlődő fej- és láblécek miatt, hiszen ezek néhány (esetleg egyébként ritka) szó, kifejezés előfordulásainak számát megsokszorozzák. De az elvileg helyes működés érdekében hosszabb szövegek esetében is érdemes ezeket a szövegrészeket eltávolítani. Ez ahhoz is hozzájárul, hogy az idegen nyelvű összetevőktől való megtisztítás hatékony legyen, hiszen idegen szövegek esetében is általában a korpusz nyelvének megfelelő nyelvű fej- és lábléc található a fájlokban.

Külön odafigyelést igényel, hogy a gyűjtött szövegek karakterei többféle kódkészlettel lehetnek kódolva ugyanazon nyelv esetén is. A tanító szövegtörzs esetében szükséges a kódkészlet ismerete, hogy ennek megfelelően kezeljük. Ha várhatóan a felismerendő bemenet kódkészletét is ismerjük, akkor csak a közös kódkészletbe (célszerűen unicode) való konvertálásról kell gondoskodni. Ha nem ismert, akkor különböző kódolású szövegekkel taníthatjuk a rendszert, vagy a másmás kódolású tanítóhalmazokat eltérő nyelvűnek véve taníthatjuk a nyelvazonosítót, így a nyelv azonosításával egyidőben megtörténik a kódkészlet azonosítása is.

További probléma az, hogy egyes alkalmazási területeken (SMS-ek, e-mailek nyelvének meghatározása) az ékezetek hiányozhatnak a szövegekről, ami megzavarhatja a nyelv-azonosítást, ha nem szentelünk neki figyelmet. Lehetséges megközelítések a mindkét jellegű szöveget tartalmazó tanító halmaz használata, illetve a tanító szöveghalmaz és a felismerendő szövegek szűkebb (ékezet nélküli) karakterkészletre való konvertálása, a nyelv azonosításakor pedig a szó eredeti karakterkészlete alapján a számított nyelv-valószínűség módosítása [11].

4.2. Az alkalmazott teszthalmaz

Több tesztet végezünk eltérő méretű tanító és felismerendő szövegkorpuszon. Először három nyelvre (angol, német, magyar) végeztünk betanítást nagyméretű korpuszon (British National Corpus, Project Gutenberg DE, Magyar Elektronikus Könyvtár), azoknak a hozzávegyült idegen nyelvű részektől való megtisztítása nélkül, a leggyakoribb szavak 90%-ának helyes fe-

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
2-9 millió szó	3	54 kbájt	99,6%	94,2%	98,5%–99,5%
600-700 szó	3	7,4 kbájt	95,5%–97,8%	79,6%–87,4%	91,7%–97,2%
500-1700 szó	77	5,4 Mbájt	70,0%–99,8%	30,1%–59,6%	71%–84,0%

szabály-sablonok

L1* S* L? S* L1*

L1* S* L? S* L2*

a tanító korpuszból kapott szabályok

"": az összes alkalmazási lehetőség száma; "L": az egyéb címkék száma (sem L1, sem L2)

L1* S* L? S* L1*; { "": 13300305, "L1":13230575, "L":69730 }

L1* S* L? S* L2*; { "": 20188476, "L1":16625250, "L2":16625298, "L":168503 }

5. ábra A használt szabálysablonok és a kapott szabályok

lismerésére. A tesztet az előzőtől független szöveghalmazon végeztük (Project Gutenberg, online magyar újságok). Az [13]-ban bemutatott módszer egy web-en megtalálható implementációjához (http://odur.let.rug.nl/~vannoord/TextCat/Demo) használt, 77 nyelvhez tartozó kis méretű (5 kilobájt) szövegre is elvégeztük a betanítást.

4.3. Eredmények nyelvi valószínűség használata nélkül

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 a valóban pontos szó-alapú működéshez szükség van egyes (formátumukat tekintve) nyelv-függetlennek tekinthető kifejezések külön beazonosítására, melyekre példák a római számok, Internet és e-mail címek, dátumok, nemzetközi szavak (pl. „tel.", „fax."), rövidítések, mértékegységeket tartalmazó kifejezések (pl. „2 cal").

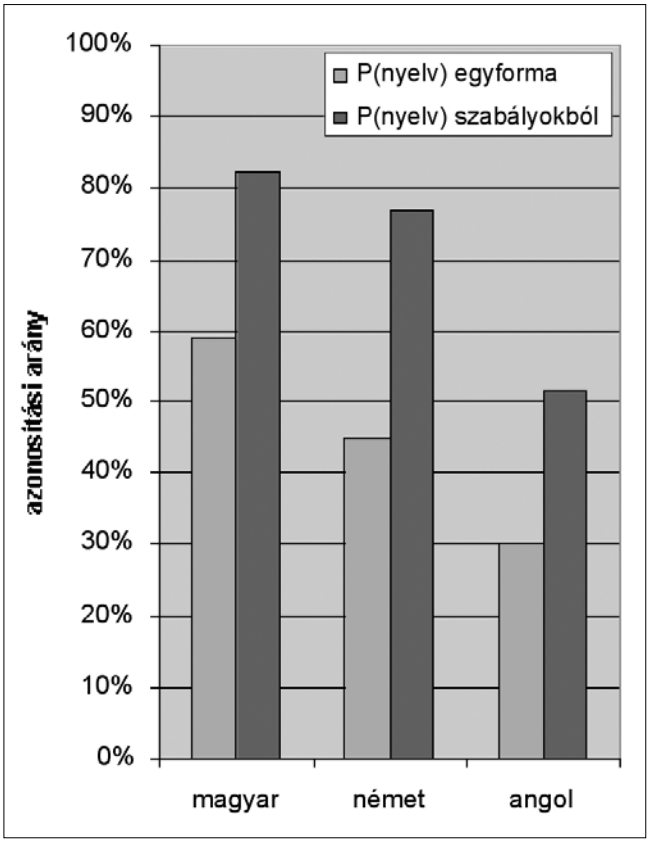
4.4. Eredmények nyelvi valószínűség használatával

Vizsgáltuk a szavak környezete alapján számított nyelv-valószínűség figyelembevételének hatását is. Ehhez az azonosító által nyelvi címkékkel ellátott szövegre alkalmaztuk a *5. ábrán* látható szabály-sablonokat, és az ugyanitt látható szabályokat. A P(szó | nyelv) valószínűségeket az 1. táblázat 3. sorára vonatkozó szöveges adatbázisból származtattuk, mivel a kis tanítóhalmaz és az ebből fakadó viszonylag gyengébb azonosítási arány nagyobb kihívást jelent a módszernek. A P(nyelv) értéket a kapott szabályokkal számoltuk, a 3.3.2 pontban leírt közelítő módszerrel.

Ezzel például a német korpuszon a korábbi 45%-ról 65%-ra növekedett a helyes azonosítási arány (a szöveget teljesen német nyelvűnek feltételezve), majd a címkézés – szabály generálás folyamatát ismételve 70%-ra, majd 72%-ra emelkedett a helyes azonosítás.

A javulás mértékét a *6. ábrán* láthatjuk a három nyelvre. Az angol nyelvre a hibák egy jelentős része (10% illetve 14%) a nagyon hasonló skót nyelvre való tévesztésből fakadt. Figyelemre méltó, hogy annak ellenére növekedett ilyen mértékben a helyes azonosítási arány, hogy nem a szövegek tényleges nyelvére vonatkozó valószínűségeket használtunk, hanem az egymás mellett lévő szavak nyelvének egyezésére vonatkozó valószínűségeket. Az 1. táblázat alapján ez a szószintű azonosítási arány 80-90%-os helyes mondatszintű azonosítást tesz lehetővé a többségi döntés használatával.

Az Interneten elérhető különböző nyelvű szövegek nagy mennyisége miatt természetesen nem vagyunk rászorítva hogy ilyen kisméretű tanító halmazt használjunk, ezért gyakorlati alkalmazásokban a táblázat első sorában láthatónál is jobb, 100%-ot erősen közelítő helyes azonosítással számolhatunk.



6. ábra
A nyelvi valószínűségek figyelembevételével elért javulás

5. Összefoglalás

A cikkben rámutattunk az automatikus címkézési módszerek, mint például a nyelvazonosítás és szófaji címkézés, használatának jelentőségére. Illusztráltuk a probléma fontosságát a szövegfelolvasó rendszerek, ezen keresztül a távközlési alkalmazások számára.

Áttekintettünk néhány, a nyelvazonosításra használt módszert, majd ezek egyes gyengeségeinek kezelésére bemutattunk egy új, kétféle feltételes valószínűséget használó, ezek értékét döntési fa tanításával becslő eljárást. A módszer hatékonyságát demonstráltuk a nyelvazonosítás címkézés feladatán. Az eredmények igazolják a megközelítés életképességét. A módszer várhatóan jól használható más problémák kezelésére, például szófaji címkézés morfológiai elemző nélküli közelítő megoldására.

Irodalom

[1] 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, Vol. 3, Numbers 3/4, December 2000, pp.217–236.

[2] G. Németh, Cs. Zainkó, G. Kiss, M. Fék, G. Olaszy, G. Gordos: "Language Processing for Name and Address Reading in Hungarian", Proc. of IEEE Natural Language Processing and Knowledge Engineering Workshop, Oct. 26-29, Beijing 2003, China, pp.238–243.

[3] Pfister, B., Romsdorfer, H., "Mixed-lingual text analysis for polyglot TTS synthesis", Proc. of Eurospeech 2003, pp.2037–2040.

[4] Halácsy, P., Kornai, A., Németh, L., Rung, L., Szakadát, I., Trón, V., "Creating open language resources for Hungarian", Proc. of LREC 2004, pp.203–210.

[5] 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, pp.2249–2252.

[6] Prószéky, G., "Humor: a Morphological System for Corpus Analysis. Language Resources for Language Technology.", Proc. of the First European TELRI Seminar, Tihany 1995, Hungary, pp.149–158.

[7] Németh, L., Halácsy, P., Kornai, A., Trón, V., "Nyílt forráskódú morfológiai elemző" In: Csendes D, Alexin Z. (eds.): II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged 2004, pp.163–171.

[8] Ted Dunning, "Statistical Identification of Languages", Computing Research Laboratory, New Mexico State University, 1994.

[9] G. Németh, Cs. Zainko: "Multilingual statistical text analysis, Zipf's law and Hungarian Speech Generation", Acta Linguistica Hungarica 2002, Vol. 49 (3-4), pp.385–405.

[10] Prager, J. M.: Linguini, "Language Identification for Multilingual Documents", Proc. of the 32nd Annual Hawaii International Conf. on System Sciences, 1999, Vol. 1, p.2035.

[11] Tian, J., Suontausta, J., "Scalable neural network based language identification from written text", Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, 2003, Vol. 1, pp.48–51.

[12] 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.

[13] W. B. Canvar, J. M. Trenkle, "N-gram based Text Categorization", Symposium on Document Analysis and Information Retrieval, University of Nevada, Las Vegas, 1994. pp.161–176.

[14] Sproat, R., Riley, M., "Compilation of weighted finitestate transducers from decision trees" In: Association for Computational Linguistics, 34th Annual Meeting, Santa Cruz, Canada, 1996. pp.215-222.

[15] Suendermann, D., Ney, H., "Synther - a New M-Gram POS Tagger", Proc. of the NLP-KE 2003, Int. Conf. on Natural Language Processing and Knowledge Engineering, Beijing, China, 2003.

[16] Halácsy P, Kornai A., Varga D., "Morfológiai egyértelműsítés maximum entrópia módszerrel", Magyar Számítógépes Nyelvészeti Konferencia, 2005.