8.2.1.Műszaki leírás a pajzsmirigy-szcintigráfiás leletek diktálórendszeréhez

SZTE Informatikai Tanszékcsoport

Bevezetés

A 2. konzorciumi tag a pajzsmirigy-szcintigráfiás leletek diktálásához igazította felismerőjét. A két eltérő alkalmazási terület miatt az alkalmazott nyelvi modelleknek szükségszerűen különbözniük kellett, hiszen ennek a komponenst az adott témakörhöz kell igazítani. Végezetül, mivel a felismerőkhöz kialakított grafikus felületek is különbözőek, így lényegében a két csoport munkájának eredményeként két különböző felismerő-prototípus készült el. Az alábbi dokumentum a Szegeden készült, pajzsmirigy-szcintigráfiás leletek diktálását lehetővé tevő programot mutatja be, egész pontosan az a mögött húzódó technikai megoldásokat.

Az akusztikus modell

A felismerő akusztikus komponense beszédhang-egységekből építkezik, a beszédhangok modellezésére a jól ismert rejtett Markov-modell technológiát használja [1]. Az inputként beérkező beszédjelen kepsztrális analízist végez. Egész pontosan 25 ezredmásodperces adatkeretekre bontja a beszédjelet, és ezek mindegyikéből 13 MFCC-együtthatót nyer ki, valamint a hozzájuk tartozó Δ és $\Delta\Delta$ gyorsulási együtthatókat [1]. Az alkalmazott beszédhang-modellek a szokásos 3-állapotú, balról jobbra haladó topológiával rendelkeznek, állapotok átugrása nem engedélyezett. A rendszerben monofón modellezést alkalmaztunk, azaz minden beszédhanghoz 1-1 külön modell tartozik. Összesen 44 beszédhang-modell alkotja a rendszert, ami valamivel kevesebb, mint ahányra valójában szükség lenne, ugyanis a hosszú-rövid mássalhangzópárokat nem különböztettük meg. Tettük ezt egyrészt azért, mert a megcélzott orvosi diktálási feladatban nincsenek olyan szópárok, amelyek megkülönböztetése csak ez alapján lenne lehetséges, másrészt pedig azért, mert a rejtett-Markov modellek amúgy is gyengék ebben a feladatban.

Az akusztikus dekódolást a fent említett hangmodellek segítségével elvégző algoritmus az alábbi fő tulajdonságokkal bír. A felismerést jelentő keresés során az ún. Viterbi-keresés és a multi-stack dekódolási módszerek egyfajta keverékét alkalmazza. A keresés felgyorsítására többféle vágási kritériumot alkalmaz. Egyrészt használja az ún. nyalábolási csonkolást (beam pruning), ami nem engedi tovább azokat a hipotéziseket, amelyek rosszabbak, mint az aktuálisan legjobb hipotézis mínusz egy küszöbérték. Korlátozza továbbá az adott pillanatban kiterjeszthető hipotézisek számát is, ami lényegében a multi-stack keresés stack-korlátozás útján történő szűkítését jelenti. A kiértékelt maximális beszédhanghossz szintén korlátozható a rendszerben. Mindezekkel a hipotézistér-levágási trükkökkel a rendszer jelenleg valós időnél (azaz a beszéd kimondásánál) gyorsabban képes működni egy átlagos PC-n.

Az akusztikai beszédhang-modellek betanítására a pályázat első szakaszában összeállított MRBA-adatbázist használtuk. Mivel az adatbázisra vonatkozóan részletes leírást adtunk a említett pályázati szakasz lezárásakor, így erre itt most nem kívánunk kitérni.`

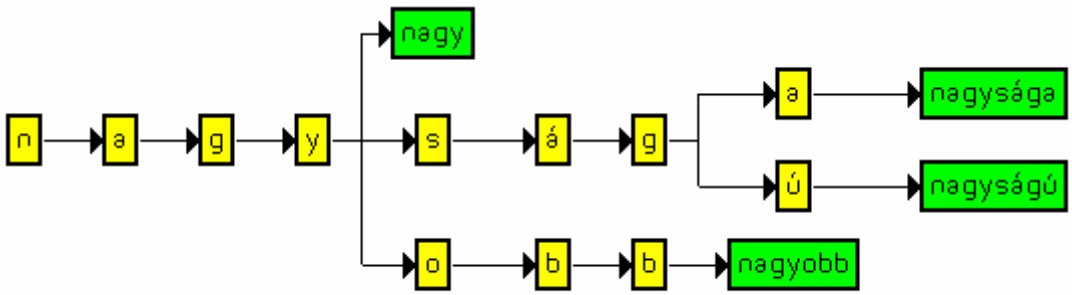
A nyelvi modell

A magyar nyelv statisztikai modellezésében különleges nehézséget okoz a nyelv erősen toldalékoló jellege. Az eredetileg angol nyelvre kitalált metodológiák ugyanis nem különböztetik meg a szó és a (toldalékolt) szóalak fogalmát, mivel az angolban az utóbbiból nincs nagyságrendileg több, mint az előbbiből. A németben, vagy pl. a szláv nyelvekben van

ugyan toldalékolás, de még mindig jóval kevesebb, mint a magyarban. Egy vizsgálat szerint egy nagyméretű szövegadatbázisban magyar nyelv esetében kb. ötször annyi szóalak fordul elő, mint a németben, és kb. hússzor annyi, mint az angolban [3]. Mivel a statisztikai nyelvi modellek az egyes szóalakok előfordulásainak összeszámolásán alapulnak, a magyar nyelv erősen ragozó jellege kétségessé teszi az ilyen módszerek (legalábbis egy az egyben történő) alkalmazását. A német avagy a szláv nyelvek esetében az a tapasztalat, hogy némi nehézségekkel ugyan, de ezek a technológiák átvihetők, főleg, ha nem általános alkalmazási területről, hanem valamely szókincs tekintetében leszűkített feladatról van szó. Azt gondoljuk ugyan, hogy a magyar nyelv esetében előbb-utóbb meg kell oldani a morfológiai modellezés beépítését a statisztikai alapú beszédfelismerőkbe, de a jelenlegi projektben egyelőre csak arra törekedtünk, hogy a hagyományos, angolra kitalált technológiákat átültessük – eddig ugyanis még ilyesmire sem volt példa. Azért választottuk ki alkalmazási területként az orvosi diktálást, mert ez egy olyan témakör, ahol egyrészt a szűkített szókincs lehetővé teszi az említett technikák bevetését, másrészt valódi, hasznos gyakorlati alkalmazás területet kínál.

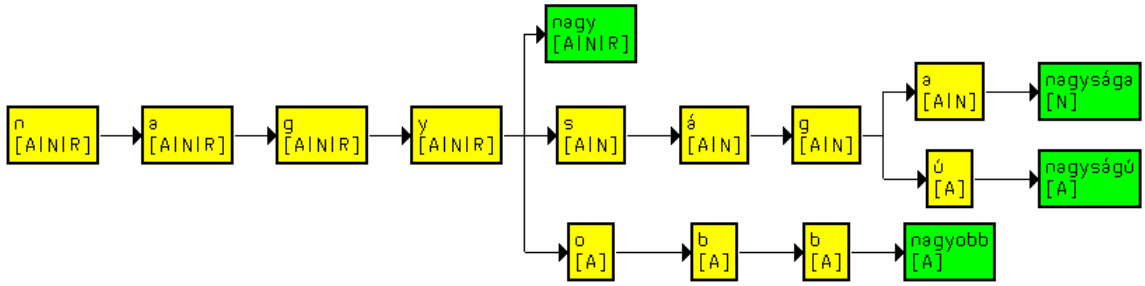
Az „orvosi íródeák” diktálóprogramunk pajzsmirigy-szcintigráfiás leletek diktálását lehetővé tevő változatának kidolgozásához – azon belül is a nyelvi modell kialakításához – 9231 szcintigráfiás leletet kaptunk meg a Szegedi Tudományegyetem Nukleáris Medicina Tanszékétől. Ez gyakorlatilag az összes leletet jelentette az 1998 és 2004 közötti időszakból. A leleteket ebben az időszakban különböző szoftverekkel rögzítették, sőt, ennyi év alatt az operációs rendszerek is változtak. Emiatt első lépésként a leleteket közös formátumra kellett hozni. Ezután többlépcsős hibakorrekció következett, például az elgépelések és a hiányos leletek kiszűrése, javítása. Ezután a szűrési fázis után 8546 lelet maradt meg, amelyek átlagosan 11 mondatból állnak, mondatonként átlagosan 6 szóval. A leletekből ezután a személyiségi jogok védelme érdekében eltávolítottuk azokat a mezőket, amelyek alapján a beteget vagy az orvost azonosítani lehetett volna (név, TAJ-szám). Speciális problémát jelentett a latin orvosi kifejezések egységes alakra hozása (pl. therapia – terápia). Föl kellett oldanunk a rövidítéseket is. Az ezen lépések után kapott szövegadatbázis körülbelül 2500 különböző szóalakot tartalmazott. Ez közepes szótárnak számít, ennyi szót a hagyományos technikákkal kényelmesen lehet morfológiai modellezés nélkül kezelni.

Habár 2500 szóalak egyszerű listás formában („lineáris lexikon”) is tárolható lenne, a szavakat ún. lexikális fában tároltuk el (ld. 1. ábra). Ebben a reprezentációs módban a közös prefixek csak egyszer kerülnek eltárolásra, ami jelentős csökkenti a szükséges memóriát. Habár nem tudunk ilyen vizsgálatról, feltételezhető, hogy az olyan toldalékoló nyelvek esetében, mint a magyar, ez a fajta reprezentáció sokkal több előnnyel jár, mint az angol esetében. A reprezentáció további előnye, hogy az akusztikus modellek kiértékelését is gyorsítja a felismerés során, mert csökkenti a redundáns akusztikus kiértékelések számát [1].



1. ábra. Néhány magyar szó lexikális fába rendezve. Az előrehozott kiértékelési módszerrel a nyelvi modell már a fa elágazásainál képes kiadni az egyes szavak valószínűségét.

Tapasztalataink szerint a vizsgált orvosi leletek nemcsak szókincs tekintetében voltak jóval szűkebbek az általános magyar nyelvnél, hanem a bennük előforduló mondatszerkezetek is erősen korlátozottak voltak. E két okból feltételeztük, hogy a szövegeinkhez nagyon hatékony statisztikai nyelvi modelleket, azon belül is a beszédfelismerésben nagyon elterjedt N-gramm modelleket [1] leszünk képes építeni. Úgy tűnt továbbá, hogy nem kell figyelmet fordítanunk a modellépítés során nem látott ún. „out-of-vocabulary” szavak problematikájára sem, hiszen 6 évnyi időszak összes szóalakja a rendelkezésünkre állt. A rendszer a jelenlegi verziójában 3-grammokat alkalmaz, de képes ezekről visszalépni alacsonyabb fokú N-grammokra is, ha szükséges (ez az ún. „back-off” simítás). Az N-gramm modellek kiértékelése során a rendszer előretekintő kiértékelést (language model lookahead) alkalmaz, ami azt jelenti, hogy a nyelvi modell kiértékelése nem csak a szavak végére érve történik meg, hanem már hamarabb, amennyiben lehetséges [1]. Ezen technológia megvalósításához a lexikális fát faktorizált formában tároljuk, azaz a közös prefixekhez a hozzá tartozó szavak valószínűségének egy kombinációját rendeljük. Az előretekintő kiértékelés segítségével a rendszeren számottevően sikerült gyorsítani a korábbi felismerési pontosság megőrzése mellett [4].



2. ábra. Néhány magyar szó lexikális fába rendezve, MSD-kóddal felcímkézve. A fa elágazásainál a nyelvi modell azonnal képes kiadni a szó valószínűségét, akár a szó-N-gram, akár a csoport-N-gram értékek figyelembevételével.

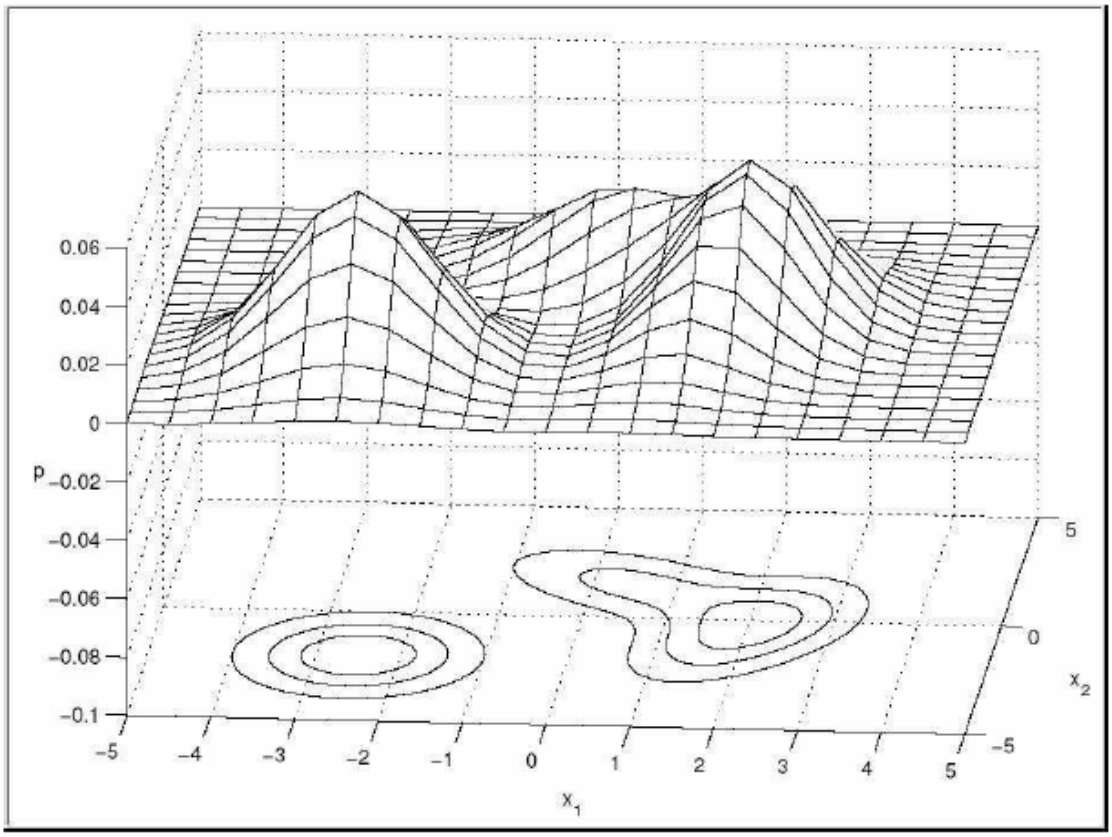
Az szó N-grammok mellett kísérleteztünk szócsoportokra épülő N-grammokkal is. Ehhez a szavakat csoportokra osztottuk az MSD-kódjuk alapján (2. ábra), amit az egyetemünkön készített PoS-tagger szoftverrel nyertünk ki [2]. A csoport-N-grammok segítségével a nyelvi modell robosztusabbá tehető olyan esetekben, amikor a szöveg egy adott pozíciójába egy korábban ilyen pozícióban nem látott, de nyelvtani szerepe alapján odaillő szó kerül. Kísérleteinkben azt tapasztaltuk, hogy a csoport-N-grammok 30%-os csökkentést hoztak a felismerési hibában [4].

A beszélőadaptációs algoritmus

A beszédfelismerők hatékonysága nagy mértékben képes javulni, ha lehetőséget kapnak arra, hogy adaptálódjanak az adott beszélő hangjához. Ennek természetesen csak akkor van értelme, ha az adott beszélő feltételezhetően sokáig fog beszélni, és az kell hozzá, hogy a beszélő felolvasson néhány legalább néhány tucat előre meghatározott mondatot. Az orvosi diktáló esetében feltehetjük, hogy a rendszert használó orvosok huzamosabb ideig fogják használni a rendszert, így a felismerési pontosság növelése érdekében mindenképpen megéri felolvasni azt a pár mondatot, ami szükséges a beszélőadaptációhoz.

A rejtett Markov-modellezésen alapuló beszédfelismerőkben használatos adaptációs algoritmusok lényegében mind ugyanarra az alapötletre épülnek: a beszélőfüggetlen modellekből kiindulva azok paramétereit a használat során fokozatosan hozzáigazítják az

aktuális beszélőhöz, így tulajdonképpen fokozatosan előáll egy speciális, beszélőspecifikus modell.



3. ábra. Gauss-keverékeloszlás három komponenssel, két dimenzióban.

A rejtett Markov-modellek két fő komponensei a gaussos keverékeloszlással leírt kibocsátási valószínűségek és az állapotátmenet-valószínűségek (ld. 3. ábra). Korábbi kísérletek azt mutatták, hogy az átmenet-valószínűségek, illetve a gaussos komponensek súlyai kevésbé jelentős szerepűek a felismerés során, és az adaptáció alatt a legfontosabb a Gauss-eloszlások várható értékének áthangolása. Az általunk alkalmazott adaptációs módszer is a várható értékek beszélőspecifikus kiigazítására fókuszál.

Rendszerünk a Gauss-komponensek paramétereinek újrabecslését maximum a posteriori (MAP) technikával oldja meg. Az adaptációhoz szükséges adatot on-line, azaz a felismeréssel párhuzamosan nyeri ki. Ennek köszönhetően az adaptációs lépés nem különül el a felismerési lépéstől, hanem az abból kinyert releváns adatokon dolgozik. Egész pontosan az adaptáció a következőképpen történik [5]. A felismerés részét képező Viterbi-algoritmus eredményeképpen megkapjuk a legvalószínűbb beszédhang-modellek sorozatát, azokon belül is a legvalószínűbb állapotsorozatot, illetve az ezekhez tartozó Gauss-komponenseket. Mivel ez alapján ki tudjuk nyerni minden egyes input jellemzővektorhoz a megfelelő állapotokat és Gausskomponens-indexeket, az egyes Gauss-komponensek várható értékét frissíteni tudjuk az alábbi formulát használva:

$$\mu_{new} = \frac{N}{N+\alpha} m_{obs,ML} + \frac{\alpha}{N+\alpha} \mu_0 ,$$

ahol

$$m_{obs,ML} = \frac{1}{N} \sum_{i=1}^{N} x_i .$$

A fenti képletekben N az adott komponensre eső példák számát jelöli, α pedig a változtatás sebességét szabályozza.

A gyakorlatban a fenti képletnek az alábbi rekurzív változatát alkalmazzuk, ami folytonos, inkrementális adaptációt tesz lehetővé:

$$\mu_{d,N+1} = \frac{x_{N+1} + (N+\alpha)\cdot \mu_{d,N}}{N+1+\alpha}.$$

Az ismertetett adaptációs technikát mind felügyelt, mind felügyelet nélküli módban lehet alkalmazni. Felügyelet nélküli esetben a bemondott hanganyag fonetikai átirata a rendszer számára ismeretlen, így a feltételezett átiratot a saját felismerési eredményéből állítja elő.

Irodalomjegyzék

[1] X. Huang, A. Acero, H.-W. Hon: Spoken Language Processing, Prentice Hall, 2001.

[2] A. Kuba, A Hócza, J. Csirik: POS Tagging of Hungarian with Combined Statistical and Rule-based Methods, Proc. of TSD 2004, pp. 113–121, 2004.

[3] G. Németh, Cs. Zainkó: Word unit Based Multilingual Comparative Analysis of Text Corpora, Proc. of Eurospeech 2001, pp. 2035–2038, 2001.

[4] A. Bánhalmi, A. Kocsor, D. Paczolay: Magyar nyelvű diktálórendszer támogatása újszerű nyelvi modellek segítségével, Proc. of 3. Magyar Számítógépes Nyelvészeti Konferencia, pp. 337–347, 2005.

[5] E. Thelen: Long Term On-line Speaker Adaptation for Large Vocabulary Dictation, Proc. of ICSLP'96, Vol. 4, pp. 2139—2142, 1996.