Magyar nyelvű beszédfelismerő rendszer diszkriminatív tanítása

Gyepesi György és Serény András

Alkalmazott Logikai Laboratórium
Budapest, Hankóczy J. utca 7. 1022
e-mail:{ggyepesi,sandris}@all.hu

A legtöbb modern beszédfelismerő rendszer az akusztikus jelből kivont, az adott hangrészletet jellemző feature-vektorok osztályozására rejtett Markov-modelleket (HMM) alkalmaz. A nagyszótáras rendszerekben általában az egyes fonémákat (esetleg környezettől függően) modellezik egy–egy HMM-mel, és alapvető feladat ezen HMM-ek paramétereinek beállítása, a tanítás. Ehhez tanító minta szükséges, ami egy (hosszú) beszéddarabból kivont feature-vektor sorozatból (ezt nevezzük megfigyelésnek) és helyesen átírt változatából áll.

Jelölje a megfigyelést o, a helyes átírást s_r, a tanítás során keresendő paraméterek összességét λ. A paraméterek beállítására a szokásos eljárás a maximum likelihood becslés (MLE), vagyis adott o megfigyelés és s_r helyes megoldás mellett keresendő λ azon értéke, melyre a modell által definiált $P_\lambda(o \mid s_r)$ valószínűség maximális; ennek megvalósítására egy hatékony eljárás a Baum–Welch algoritmus.

A diszkriminatív tanító eljárások az MLE módszer alternatíváit adják a paraméter-beállítás feladatának megoldására. Az az alapvető elképzelés, hogy a helyes megoldás mellett figyelembe vesszük a beszédfelismerő kimenetén megjelenő rossz megoldásokat is, és keressük λ egy olyan értékét, ami nem csupán azt biztosítja, hogy a helyes megoldás valószínűsége nagy legyen, de a helytelen megoldások valószínűségét is alacsonyan tartja. Innen adódik az eljárás neve: a helyes megoldást igyekszik megkülönböztetni a helytelenektől. Bár a diszkriminatív módszer kezdeti változatait HMM-ek tanítására már a nyolvanas évek végén kidolgozták, a magyar nyelvű beszédfelismerésben való alkalmazására – tudomásunk szerint – ez az első kísérlet.

A különféle diszkriminatív tanító eljárások közül igen ígéretes a Povey [1] által bevezetett minimum phone error (MPE) tanítás, ennek egy változatát valósítottuk meg és az alábbiakban ezt vázoljuk. Legyen S az összes lehetséges megoldás (mondat) halmaza, és valamely $s \in S$ esetén legyen $A(s, s_r)$ az s mondatnak a helyes megfejtéshez képest fonémákban mért pontossága, $A(s, s_r) = |s_r| - d(s, s_r)$, ahol $|\cdot|$ a fonémák száma és $d(\cdot, \cdot)$ az edit distance. Az optimalizálandó függvényünket úgy választjuk, hogy a mondatok valószínűségét pontosságukkal súlyozzuk, így a függvény maximalizálásakor a pontosabb mondatok jobban számítanak. Formálisan, keresendő a $\lambda \mapsto \sum_{s \in S} P_\lambda(s \mid o) A(s, s_r)$, az MPE célfüggvény, maximuma. Miután az összes lehetséges mondat felsorolása

és valószínűségeik kiszámítása nem megvalósítható, a következő eljáráshoz folyamodunk. Adott (o, s_r) tanító mintán először MLE tanítást hajtunk végre és az így nyert fonéma HMM-ket használva az o bemeneten felismerést futtatunk. A felismerés kimenete az első néhány legvalószínűbb mondat, ezek összességét kompakt formában, szóhálóként ábrázoljuk. A szóhálóban minden út egy mondatot reprezentál, az itt elő nem forduló mondatokat nem vesszük számításba.

Az MLE tanításhoz használt iteratív Baum–Welch eljárás adaptálható az MPE célfüggvény maximumának keresésére. Egy iterációs lépés elején a szóhálón futó forward–backward algoritmus hatékonyan számolja egyszerre a fonémákhoz tartozó statisztikákat és az $A(s, s_r)$ pontosság egy közelítését. A diszkriminatív tanítási elv szerint minden fonémához kétféle statisztikát gyűjtünk: az egyik azt tükrözi, hogy a fonéma milyen gyakran szerepel az átlagosnál pontosabb utakon („helyes megoldások"), a másik azt, hogy milyen gyakran szerepel az átlagosnál kevésbé pontos utakon („rossz megoldások"). Az újrabecslő egyenletekben mindkétféle statisztika által hordozott információ megjelenik, a fonémaparaméterek újrabecslésével egy iteráció véget ér. Az MLE esethez hasonlóan négy–öt iteráció elegendő a konvergenciához.

A fentiekben összefoglalt eljárást implementáltuk, annak futtatása, az eredmények értékelése és a baseline-nak tekintett MLE módszer eredményével való összevetése még folyik. Povey és Woodland [2] angol nyelvű korpuszokon végzett vizsgálatai a hibásan felismert szavak arányának három–öt százalékpontos csökkenését mutatják, magyar nyelvre is hasonló eredményt várunk.

Hivatkozások

1. D. Povey., Discriminative Training for Large Vocabulary Speech Recogniton. PhD thesis, University of Cambridge, 2003.
2. D. Povey and P. Woodland. Minimum phone error and I-smoothing for improved discriminative training. In Proceedings of the ICASSP, 2002.