A különböző modalitások hozzájárulásának vizsgálata a témairányítás eseteinek osztályozásához a HuComTech korpuszon

Kovács György[1], Váradi Tamás[1]

Magyar Tudományos Akadémia, Nyelvtudományi Intézet,
Budapest VI., Benczúr utca 33.
e-mail:gykovacs@inf.u-szeged.hu, varadi.tamas@nytud.mta.hu

Kivonat Az ember és gép közötti, valamint az emberek közötti interakció fontos kérdése a témairányítás. Gépi felismerésének vizsgálatakor nem csak az érdekes számunkra, hogy milyen pontosság- vagy fedésértékeket tudunk elérni, hanem az is, hogy mely jellemzők mennyiben járultak hozzá ehhez az eredményhez. Kísérleteink során egyéni neuronhálókat tanítottunk a különböző modalitásokból kinyert jellemzők felhasználásával, hogy lemérjük az így kapott neuronhálók teljesítményét a témairányítási címkék osztályozásában. Továbbá megvizsgáltuk, hogy a különböző neuronhálók kimeneteként kapott valószínűség-becslések mely súlyozásával érhetjük el a legjobb osztályozási eredményt. Két modalitás (multimodális, szintaktikai) emelkedett ki a többi közül, a helyes osztályozáshoz való hozzájárulásukkal Az ezen modalitásokból származó jellemzők megfelelő kombinációja ugyanolyan jó eredményt adott, mint az összes modalitás jellemzőinek kombinációja. Továbbá mindkét kombináció jobb eredményt adott mint az összes jellemzőt kombináció nélkül felhasználó neuronháló, sőt ez utóbbi teljesítményét a kizárólag multimodális jellemzőket felhasználó neuronháló is felülmúlta.[1]

Kulcsszavak: HuComTech, témairányítás, valószínűségi mintavételezés, jellemzőkiválasztás

1. Bevezetés

Az ember-számítógép interakció elősegítéséhez fontos, hogy a gép tudja, beszélgetőtársa mikor fejti ki az aktuális témát, mikor tér el attól (kis mértékben módosítva azt, az előzmények figyelembevételével, vagy teljesen eltérve attól), és mikor nem járul hozzá érdemben a témához. Ezért kutatásunk egyik célja, hogy beszélgetés-szegmentumokat témairányítás szempontjából különböző kategóriákba soroljunk. a HuComTech multimodális beszédadatbázisban ezek a kategóriák a következők:

[1] A szerzők köszönetüket fejezik ki az Országos Tudományos Kutatási Alapprogramok (OTKA) programnak, amely a K116938 számú projekt keretében az itt ismertetésre kerülő kutatást támogatta.

- Témakezdeményezés: a beszélő a korábban elhangzottaktól motiváltan új témába kezd, mely illeszkedik a társalgás addigi menetébe.
- Témaváltás: a beszélő oly módon kezd új témába, hogy az a korábbi beszélgetésbe kevésbé illeszkedik, az nem indokolja a téma választását.
- Téma kifejtése: a beszélő az aktuális témát taglalja.
- Hozzájárulás hiánya: szakaszok, melyek nem sorolhatók be egyik korábbi kategóriába sem. Meg kell jegyezzük, hogy ez inkább az egyéb címkék hiánya, mint önálló kategória.

Korábbi cikkünkben [1] kísérleteink többségében követtük ezt a felosztást, azonban jelentősen jobb eredményeket értünk el, amikor a témakezdeményezést (motivált témaváltást) és a motiválatlan témaváltást egyetlen kategóriaként, témaváltásként kezeltük. Ezért jelen cikkünkben ez utóbbi megközelítésre koncentrálunk, és kísérleteink többségében a témairányítási címkék osztályozását három osztály esetére vizsgáljuk.

A témairányítás kérdésköre nem csak az ember és gép közötti kommunikáció elősegítése miatt lehet hasznos, hanem az emberek közötti kommunikáció jobb megértéséhez is. Többek között ez okból nem csak annak lesz jelentősége számunkra, hogy gépi osztályozásában milyen pontosság- vagy fedésértékeket tudunk elérni, hanem hogy mely jellemzők/jellemzőcsoportok járulnak hozzá leginkább az osztályozási eredményekhez. Ezért jelen cikkünkben öt jellemzőcsoportot elemzünk, két különböző módszerrel. Először azt vizsgáljuk, hogy a különböző jellemzőcsoportokokat önmagukban használva milyen eredményeket kapunk, majd azt elemezzük, hogy a jellemzőcsoportok mely kombinációjával kapjuk a legjobb eredményt.

A témában született korábbi munkák főleg lexikális [2,3,4] és prozódiai [5,6] információra, vagy ezek egy kombinációjára támaszkodtak [7,8]. Egyebek mellett a prozódiai információ felhasználását is megvizsgáltuk, ám a lexikális információ közvetlen felhasználására nem volt lehetőségünk, az adatbázis annotációjának sajátosságai miatt. A következő fejezetben az adatbázis bemutatása során ezekről is említést teszünk. Majd az azt követő fejezetben ismertetjük a kísérletek során felhasznált módszereket, miután bemutatjuk az eredményeket, és végül ismertetjük konklúzióinkat, valamint terveinket a jövőbeni munkára.

2. HuComTech multimodális korpusz

A HuComTech projekt keretében 111 beszélővel készült 222 interjú [9]. Minden beszélővel két beszélgetés, egy formális (szimulált állásinterjú), és egy informális beszélgetés került felvételre. A felvételeket aztán az adatforrást tekintve hat modalitás szerint (Multimodális, Szintaktikai, Prozódiai, Unimodális, Videó, Audió) összesen 39 szinten annotálták. Az annotáció elsősorban az interjúalanyra koncentrál, de több olyan eleme is van, amely az interjút készítő viselkedését is leírja. Bár a későbbiekben röviden minden modalitásról szót ejtünk, bővebb leírásra jelen cikk keretei között nincs lehetőség, az adatbázis részletesebb leírása azonban elérhető a projekthez kapcsolódó korábbi publikációkban [9,10,11].

2.1. Modalitások

Az adatbázis annotációja hat modalitásban történt, melyek összesen 221 jellemzőt adtak az osztályozáshoz. A jellemzőket oly módon használtuk, hogy az interjúkat 0,32 másodperces keretekre (frame) bontottuk, és az adott intervallumra jellemző címkét rendeltük az egész kerethez (a bináris jellemzők kivételével ezután az összes jellemzőt 0 átlagra és 1 varianciára standardizáltuk). A különböző modalitásokhoz az alábbi szintek és jellemzők tartoznak:

Multimodális annotáció. Az annotáció ebben a modalitásban a videó és audió adatok együttes felhasználásával készült a Qannot program segítségével. Itt minden információ kétszer jelenik meg: egyszer az interjúalanyra, egyszer az interjút készítőre vonatkozóan. A kategóriából származó információt 29 jellemzőben kódoltuk.

- Kommunikatív aktus: az interjúalany/interjút készítő kommunikatív aktusai, 14 (7-7) bináris (0 vagy 1 értékű) jellemzőben kódolva, a lehetséges címkéknek (none, other, acknowledging, commissive, constative, directive, indirect) megfelelően.
- Támogató aktus: az interjúalany/interjút készítő támogató aktusai, 8 (4-4) bináris jellemzőben kódolva, a lehetséges címkéknek (other, backchannel, politeness marker, repair) megfelelően.
- Témairányítás: az interjút készítő témairányítási aktusai, 3 bináris jellemzőben kódolva, a lehetséges címkéknek (témaváltás, témakezdeményezés, téma kifejtése) megfelelően.
- Információ: azt írja le, hogy az interjúalany/interjút készítő kapott-e olyan információt, amely új volt számára, vagy olyat, amelyet már ismert, esetleg nem kapott semmilyen információt. 4 (2-2) bináris jellemzőben kódoljuk.

Szintaktikai annotáció. A szintaktikai modalitásben egyetlen szint található, melynek 7 mezőjét 20 jellemzőben kódoltuk.

- Clause ID: az aktuális tagmondat helye a mondatban. 1 egész típusú jellemzőben kódolva.
- Alárendeltség: azon tagmondatok azonosítója, melyeknek a jelenlegi tagmondat alá van rendelve, 1 egész típusú jellemzőben (az azonosítók száma) kódolva.
- Egyeztetés: azon tagmondatok azonosítója, melyek egyeztetve vannak a jelenlegi tagmondattal, 1 egész típusú jellemzőben (az azonosítók száma) kódolva.
- Alárendelés: azon tagmondatok azonosítója, melyek a jelenlegi tagmondat alá vannak rendelve, 1 egész típusú jellemzőben (az azonosítók száma) kódolva.
- Beágyazás: azon tagmondatok azonosítója, melyek a jelenlegi tagmondatba ágyazódnak be, 1 bináris jellemzőben kódolva.
- Beágyazódás: azon tagmondatok azonosítója, melyekbe a jelenlegi tagmondat beágyazódik, 1 bináris jellemzőben kódolva.
- Hiányzó kategóriák: a tagmondatból hiányzó kategóriák. 14 bináris jellemzőben kódoljuk, a 14 lehetséges címkének megfelelően.

Prozódiai annotáció. A prozódiai annotáció a Prosotool [12] eszközzel történt. Az ezen modalitásból származó információt 37 jellemzőben kódoltuk.

- F0-mozgás: a simított F0 mozgás az aktuális szegmensben. 5 bináris jellemzőként kódoljuk az öt mozgás-kategóriának (esés, csökkenés, stagnálás, növekedés, emelkedés) megfelelően.
- F0 szint: az alapfrekvencia szintje a jelenlegi szegmens elején és végén. 10 (5-5) bináris jellemzőben kódoljuk, a szegmens elején és végén álló címkék (L_2, L_1, M, H_1, H_2 ahol $L_2 < T_1 < L_1 < T_2 < M < T_3 < H_1 < T_4 < H_2$, és ahol a T_i értékeket küszöbként használjuk) alapján.
- F0 érték: az alapfrekvencia értéke a jelenlegi szegmens elején és végén, 2 valós típusú jellemzőben kódolva.
- Nyers F0 értékek átlaga: az alapfrekvencia értékek átlaga az adott keretre nézve, 1 valós típusú jellemzőben kódolva.
- Zöngés és zöngétlen invervallumok: a megadott intervallum zöngés, zöngétlen (vagy egyik sem), 2 bináris jellemzőben kódolva.
- I-mozgás: az intenzitás változás az adott szegmensben. A jellemzők kódolása ugyan olyan, mint az F0-mozgás esetén
- I-szint: az intenzitás szintje az aktuális szegmens elején és végén. A jellemzők kódolása ugyan olyan, mint az F0 szint esetén.
- I érték: az intezitás értéke az aktuális szegmens elején és végén. A jellemzők kódolása ugyan olyan, mint az F0 érték esetén.

Unimodális annotáció. Ebben a modalitásban az annotáció kizárólag a videó adatok felhasználásával készült, a HuComTech projekt keretében fejlesztett Qannot program segítségével. Az ezen modalitásból származó információt 15 jellemzőben kódoltuk.

- Fordulókezelés: a társalgási fordulók az interjúalany szemszögéből, 5 bináris jellemzőben kódolva.
- Figyelem: leírja, hogy az interjúalany az interjúkészítőre figyel-e, vagy figyelmet vár az interjúkészítőtől, 2 bináris jellemzőben kódolva.
- Egyetértés: az interjúalany által mutatott egyetértés szintje, 7 bináris jellemzőben kódolva.
- Újdonságérték: azt írja le, hogy az interjúalany kapott-e új információt, vagy nem, 1 bináris jellemzőben kódolva.

Videó annotáció. Ebben a modalitasban az annotáció két kategóriában – funkcionális és fizikai)– történt. Amikor az annotátorok a funkcionális szinten dolgoztak (érzelmek és emblémák, a videóhoz tartozó audió jelet is felhasználhatták. A kategóriából származó információt 111 jellemzőben kódoltuk.

- Arckifejezés: a beszélő arckifejezése által tükrözött érzelmek, 7 bináris jellemzőben kódolva.
- Tekintet: a beszélő tekintetének iránya, 6 bináris jellemzőben kódolva.
- Szemöldök: a beszélő szemöldökmozgása, 4 bináris jellemzőben kódolva.

– Fejmozgás: a beszélő fejének mozgása, 8 bináris jellemzőben kódolva.
– Kéz alakja: a beszélő kezei különböző alakzatokat formálhatnak a beszélgetés alatt. Itt ezen alakzatok kerülnek leírásra, 15 bináris jellemzőben kódolva.
– Érintés: annak a leírása, hogy a beszélő melyik kezével, milyen testrészén érintette/vakarta meg magát, 30 bináris jellemzőben kódolva.
– Testtartás: a beszélő testtartásának leírása, 10 bináris jellemzőben kódolva.
– Deixis: a beszélő deiktikus mozgása, 10 bináris jellemzőben kódolva.
– Érzelem: a beszélő látszólagos érzelmi állapota, 7 bináris jellemzőben kódolva. Fontos különbség az arckifejezéshez képest, hogy itt az annotátor az audió csatornát is használhatta a címke kiosztásakor.
– Embléma: a beszélőhöz kapcsolódó embléma címkék (agree, attention, block, disagree, doubt, doubt-shrug, finger-ring, hands-up, more-or-less, number, one-hand-other-hand, other, refusal, surprise-hands), 14 bináris jellemzőben kódolva.

Audió annotáció. Az audió annotáció a tagmondatok szintjén történt. Ez azzal járt, hogy az olyan információkat, mint az egyes szavak, hezitációk, ismétlések, a 25 századmásodpercet meghaladó szünetek, nem tudjuk időben elég pontosan elhelyezni, azaz nem tudjuk ezen jelenségeket a 0,32 másodperces keretekhez kötni. Így az audió annotációból egyedül az érzelmi címkéket használtuk fel, mivel ésszerűen feltételezhetjük, hogy ezek az adott tagmondatra nézve állandóak. Így az audió annotációból első kísérleteinkben egyetlen szintet tudtunk felhasználni, melyet 9 bináris jellemzőben kódoltunk, a megadott címkéknek (silence, overlapping speech, other, happy, neutral, surprised, recalling, sad, tense) megfelelően. Mivel a modalitás címkéinek egyelőre csak töredékét tudtuk jellemzőként hasznosítani, ezt a jellemzőcsoportot jelen cikkünk keretében nem vizsgáltuk.

2.2. Tanító/Validációs/Teszt felbontás

A modellek tanításához, paramétereiknek beállításához valamint a modellek kiértékeléséhez három különálló halmazra van szükségünk: egy tanító-, egy validációs- és egy teszthalmazra. Ezt a felosztást a HuComTech adatbázis esetére 75/10/15 arányban határoztuk meg. Ezt a korábban létrehozott felosztást [1] használtuk jelen munkákban is.

2.3. Az adatok kiegyensúlyozatlansága

A beszélgetések természete miatt sokkal többször fordul elő, hogy kifejtünk egy témát, vagy épp egyáltalán nem járulunk hozzá érdemben egy témához (beszélgetőtársunk viszi a szót) mint az, hogy témát váltunk, vagy új témát kezdeményezünk (a beszélgetések több mint harminc százalékában például egyáltalán nincs motiválatlan témaváltás az interjúalanyok részéről). És az előbbi esetek általában hosszabbak is, mint az utóbbi, ritkább esetek. Így az adatok olyan kiegyensúlyozatlansága lép fel, amely megnehezíti a tanítást és a kiértékelést is. A következő fejezetben leírt módszerekkel többek között erre keresünk megoldást.

3. Kísérleti módszerek

3.1. Súlyozatlan átlagolt fedés

Az osztályok kiegyensúlyozatlan eloszlása problémát jelenthet modelljeink kiértékelésénél. Teszthalmazunkban például az esetek mindössze 18 százaléka tartozik a (motivált vagy motiválatlan) témaváltás kategóriájába, ami azt jelenti, hogy akár 82 százalékos pontosságot is elérhetünk, anélkül, hogy a témaváltásnak akár csak egy esetét is helyesen felismernénk. Ez azt mutatja, hogy a nagyon kiegyensúlyozatlan eloszlású osztályozási feladatok esetén a pontosság nem feltétlenül megbízható mértéke a teljesítménynek. A modellek értékelésének egy népszerűbb mértéke (többek között annak köszönhetően, hogy gyakran használt az Interspeech kihívásokban [13]) a súlyozatlan átlagolt fedés (UAR).

Az UAR az osztályok fedésének súlyozatlan átlaga. Értéke kiszámítható az A tévesztési mátrixból, ahol A_{ij} jelzi j osztály azon elemeit, melyeket az i osztályba soroltunk. Ekkor az UAR értékét a következő képlettel kapjuk:

$$UAR = \frac{1}{N} \sum_{j=1}^{N} \frac{A_{jj}}{\sum_{i=1}^{N} A_{ij}}, \tag{1}$$

ahol N az osztályok száma.

3.2. Valószínűségi mintavételezés

Az osztályok kiegyensúlyozatlan eloszlása a kiértékelés mellett a tanítás során is problémát okozhat. Ha az algoritmusunk egyes osztályokból jelentősen többet lát a tanítás során, mint más osztályokból, az a ritkább osztályok rosszabb felismeréséhez vezethet [14]. Ez olyan extrém módokon nyilvánulhat meg, mint például bizonyos osztályok teljes figyelmen kívül hagyása. Ezt a problémát a különböző osztályokba tartozó elemek számának manipulálásával oldhatjuk meg. Ennek két útja képzelhető el: csökkenthetjük a gyakoribb osztályokba tartozó elemek számát, vagy megpróbálhatjuk növelni a ritkább osztályokba tartozó elemek számát. Az előbbi esetén értékes, nehezen megszerzett tanító adatokat dobunk el, az utóbbi pedig általában csak igen költségesen kivitelezhető. Azonban harmadik lehetőségként manipulálhatjuk úgy az egyes osztályokba tartozó elemek számát, hogy bizonyos elemeket többször felhasználunk a tanítás során. Erre a valószínűségi mintavételezés módszere két lépésben ad lehetőséget. Az első lépésben véletlenszerűen kiválasztjuk az osztályt, majd az adott osztályból véletlenszerűen választunk egy elemet [15]. Az osztály kiválasztását tekinthetjük úgy, mint mintavételt egy multinomiális eloszlásból, feltételezve, hogy minden c_i osztályhoz tartozik egy

$$P(c_i) = \lambda(1/N) + (1 - \lambda)Prior(c_i) \tag{2}$$

valószínűség, ahol N az osztályok száma, $Prior(c_i)$ a c_i osztály a priori valószínűsége, és $\lambda \in [0, 1]$ az eloszlás egyenletességét meghatározó paraméter. Ha $\lambda = 0$, az eredeti eloszlást kapjuk, míg $\lambda = 1$ esetén egyenletes eloszláshoz jutunk [16].

3.3. Mély neuronhálók

Kísérleteinkben egyenirányított mély neuronhálókat alkalmaztunk. Ezek olyan neuronhálók, melyeknek egynél több rejtett rétegük van, és rejtett rétegeikben a neuronok egyenirányítású (rectifier) aktivációs függvényt[2] alkalmaznak a standard szigmoid függvényt helyett. Az elmúlt években jelentősen nőtt ennek az architektúrának a népszerűsége, többek között a beszédfelismerés területén [17]. Az általunk használt neuronhálók három rejtett réteggel készültek, minden rejtett rétegben 250 illetve 1000 neuronnal (attól függően, hogy csak egy adott jellemzőcsoportot, vagy az összes jellemzőt használták bemenetükként). A neuronhálók tanítása a tanító halmazon történt, különböző λ paraméterek és kontextus-méretek mellett. Validációhoz, valamint a tanulási ráta meghatározásához a validációs halmazt használtuk, az UAR értéket használva kiértékelésre.

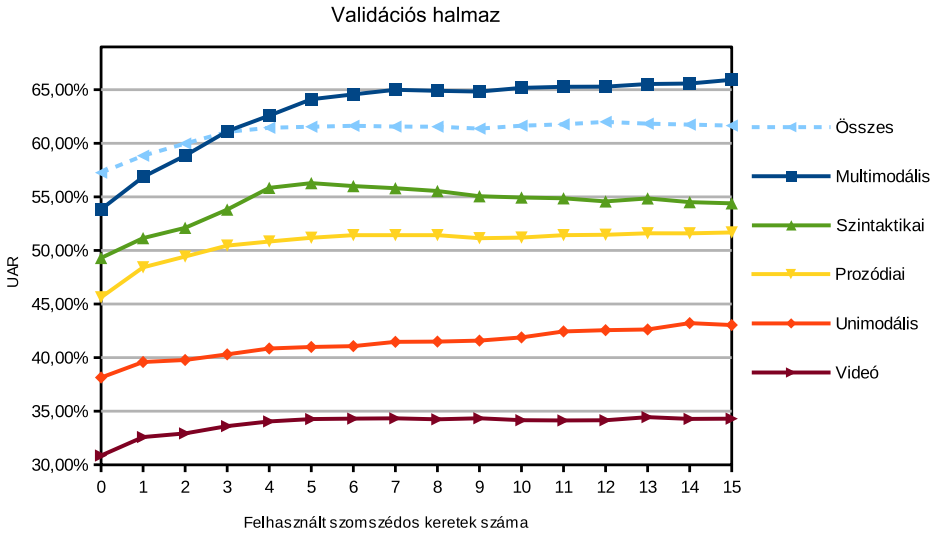
4. Kísérletek egyedülálló jellemzőcsoportokon

Először azt vizsgáltuk, milyen UAR értékeket érhetünk el az egyes jellemzőcsoportok felhasználásával tanított neuronhálók segítségével. Ehhez minden jellemzőcsoporthoz két paramétert kellett meghatároznunk, a bemenetként használt szomszédos keretek számát, valamint a valószínűségi mintavételezésnél használt λ paraméter értékét. Előbbit 0 és 15 (illetve mivel a neuronháló a szomszédokat szimmetrikusan használja, így valójában 0 és 30) között, utóbbit pedig 0 és 1 között (0,1-es lépésközzel) próbáltuk meghatározni. Minden paraméterpárra öt neuronhálót tanítottunk különböző súlyokkal inicializálva, majd megvizsgáltuk, hogy mely paraméterpárra kapjuk a legjobb átlagos UAR értéket a validációs halmazon. A kiértékelést a teszthalmazon ezzel a paraméterpárral végeztük el.

4.1. Eredmények négy osztály esetén

A validációs halmazon kapott eredmények jobb vizualizálása érdekében minden felhasznált szomszédos keretszám esetére kiválasztottuk azt a λ paramétert, amellyel a legjobb UAR eredményt értük el, és ezt az eredményt rendeltük az aktuálisan felhasznált keretszámhoz. Az eredményül kapott diagram a 1. ábrán látható. Az ábráról leolvashatjuk, hogy a különböző jellemzőcsoportok egymáshoz viszonyított teljesítménye meglehetősen stabil. Függetlenül a felhasznált keretek számától, a legjobb eredményt a multimodális jellemzőcsoporttal kapjuk, azt követi a szintaktikai és prozódiai jellemzőcsoport, majd az unimodális jellemzőcsoport, a legrosszabb UAR eredményeket pedig az egyébként legtöbb jellemzőt tartalmazó videó jellemzőcsoport adja. Az egyes jellemzőcsoportok és az összes jellemzőből álló csoport kapcsolata nem ilyen egyértelmű. Amikor a szomszédos kereteket nem használjuk fel a tanítás során, vagy csak keveset használunk közülük, az összes jellemzőt felhasználó neuronháló teljesít a legjobban, ahogy azt várnánk. Három felhasznált szomszédos keret után azonban a multimodális jellemzőcsoporttal jobb eredményeket kapunk.

[2] $recitifer(x) = max(0, x)$

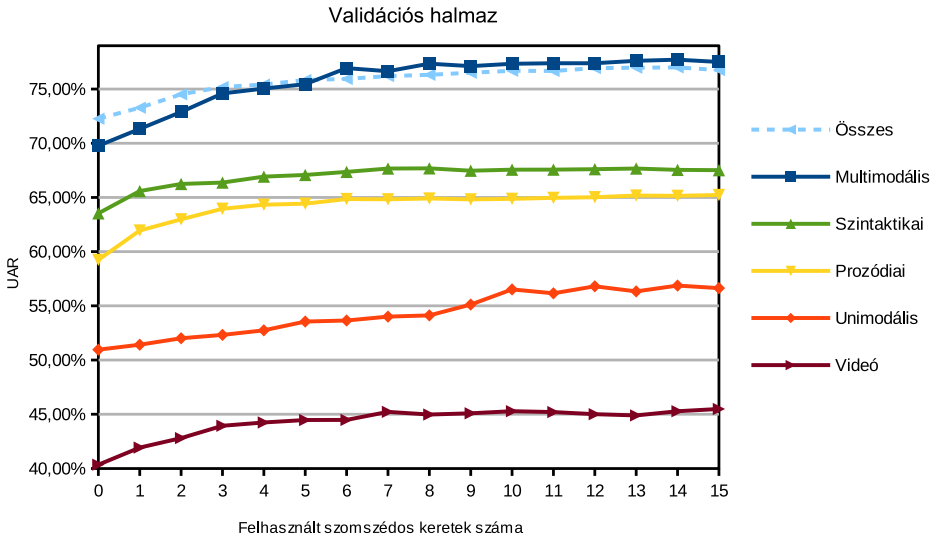


1. ábra. A legjobb elért UAR a különböző jellemzőcsoportokkal a felhasznált szomszédos keretek számának függvényében (öt neuronháló átlaga).

A validációs halmaz alapján minden jellemzőcsoporthoz megtaláltuk azokat a paramétereket, amelyekkel a teszthalmazon kiértékeljük őket. Az így kapott eredmények láthatók az 1. táblázatban. A validációs halmazhoz hasonlóan a teszthalmaz esetén is a multimodális jellemzőcsoport felhasználásával kapjuk a legjobb eredményt, valamint a jellemzőcsoportok sorrendje sem változik. Ám az unimodális és videó jellemzőcsoportok közötti különbség szinte teljesen eltűnik azáltal, hogy az unimodális jellemzőcsoporton tanított neuronhálók eredménye valamelyest romlik a validációs halmazhoz képest, míg a videó jellemzőcsoport eredménye nagy mértékben javul. Az így kapott eredmények továbbra is alacsonyak, ezért további kísérleteinkben a három osztályos esetre koncentrálunk.

1. táblázat. A különböző jellemzőcsoportokon, valamint az összes jellemzőn tanított neuronhálók teszthalmazon történő kiértékelésével kapott UAR eredmények (öt függetlenül tanított neuronháló átlaga).

Jellemző	Szomszédos keretek száma	λ	Validáció	Teszt
Összes	12	1,0	62,0%	62,6%
Multimodális	15	1,0	**65,9%**	**65,0%**
Szintaktikai	5	1,0	56,3%	55,0%
Prozódiai	15	1,0	51,7%	51,5%
Unimodális	14	1,0	43,2%	40,7%
Videó	13	1,0	34,4%	40,5%



2. ábra. A legjobb elért UAR a különböző jellemzőcsoportokkal a felhasznált szomszédos keretek számának függvényében (öt neuronháló átlaga).

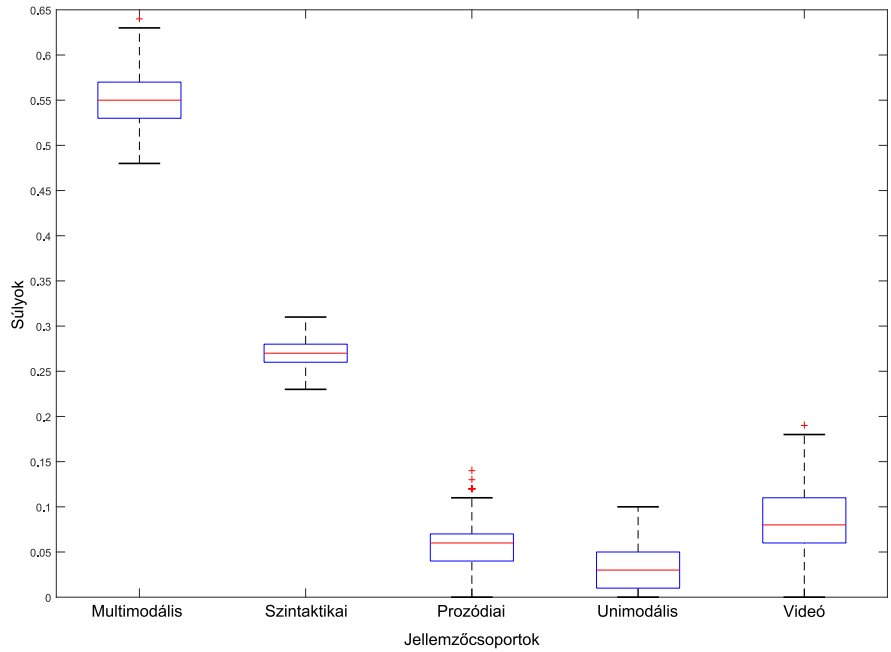
4.2. Eredmények három osztály esetén

A négy osztályra elvégzett kísérleteket megismételtük három osztály esetére. A validációs halmazon kapott eredmények leolvashatók a 2. ábráról. A négyosztályos esethez nagyon hasonló képet látunk: a különböző jellemzőcsoportok teljesítményének sorrendje változatlan, és ismételten azt látjuk, hogy amint a felhasznált szomszédos keretek száma átlép egy korlátot (ezúttal ez 5 keret), egyedül a multimodális jellemzőkkel konzisztensen jobb eredményeket kapunk, mint az összes jellemzővel. Mivel ez esetben a görbék lapultabbak voltak, mint négy osztálynál, a felhasznált szomszédos keretszámot az unimodális jellemzőcsoport alapján állapítottuk meg, 10 szomszédos keretben.

Ismét a validációs halmazon választott paraméterekkel értékeltük ki modelljeinket a teszthalmazon. A 2. táblázatból le tudjuk olvasni, hogy ebben az esetben is a multimodális jellemzőcsoport adta a legjobb eredményt. Láthatjuk továbbá,

2. táblázat. A különböző jellemzőcsoportokon, valamint az összes jellemzőn tanított neuronhálók teszthalmazon történő kiértékelésével kapott UAR eredmények (öt neuronháló átlaga).

Jellemző	Szomszédos keretek száma	λ	Validáció	Teszt
Összes	10	1,0	76,7%	75,7%
Multimodális	10	1,0	**77,3%**	**76,3%**
Szintaktikai	10	0,9	67,6%	67,4%
Prozódiai	10	0,9	64,9%	64,6%
Unimodális	10	0,9	56,5%	55,5%
Videó	10	1,0	45,3%	49,6%

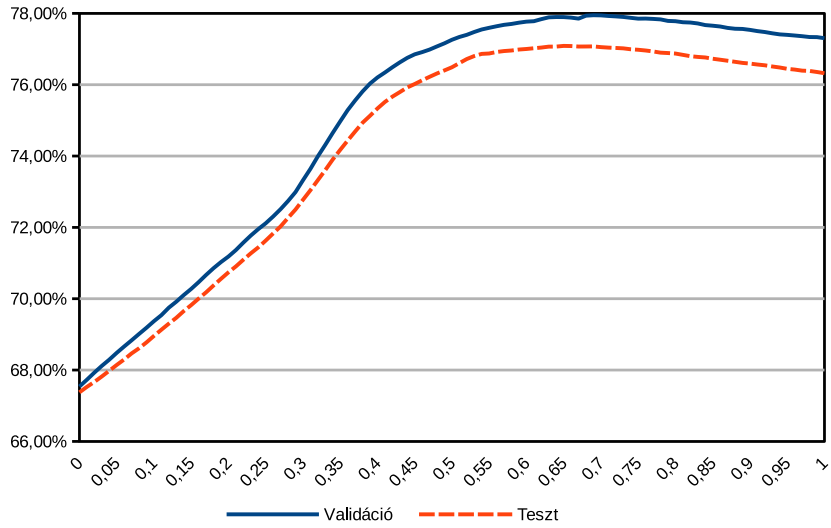


3. ábra. A különböző jellemzőcsoportokhoz tartozó súlyok doboz diagramja.

hogy a jellemzőcsoportok között a validációs halmazon kapott eredmények alapján felállított sorrend ezúttal sem változik a teszthalmaz eredményein.

5. Kísérletek jellemzőcsoportok kombinálására

Mivel a neuronhálók kimeneti rétegében a neuronok softmax függvényt valósítanak meg, így minden neuron kimenete a [0,1] intervallumba esik, és a kimenetek összege 1. Tehát az egyes neuronok kimenetét tekinthetjük az adott osztályba tartozás valószínűségének becsléseként. A különböző jellemzőcsoportokon tanított öt különböző neuronháló tehát öt különböző valószínűségi becslést ad az osztályainkra. A jellemzőcsoportokat úgy próbáljuk kombinálni, hogy ezeknek a valószínűségeknek a súlyozott összegét vesszük, és ez alapján hozunk döntést az osztályozásról. Ehhez 4,5 millió véletlen súlyvektort állítottunk elő 0,01-es lépésközzel, melyeket a validációs halmazon értékeltünk ki, és kiválasztottuk közülük a legjobb UAR eredményre vezető kétezret (itt a legjobb és legrosszabb súlyvektor átlagos teljesítménye között kevesebb, mint 0,05 százalékpontos különbség volt). Ezen kétezer súlyvektorban a különböző jellemzőcsoportokhoz rendelt súlyok terjedelmét, interkvartilis terjedelmét, valamint maximumát, minimumát és mediánját a 3. ábrán ábrázoltuk. Látható, hogy a validációs halmazon legjobban teljesítő súlyozások esetén a legnagyobb súlyokat a multimodális jellemzőcsoport kapta, medián értéke 0,55, míg a szintaktikai jellemzőcsoport medián értéke kevesebb, mint annak a fele (0,27). A súlyok mediánjának sorrendje ettől a ponttól kezdve azonban eltér a jellemzőcsoportok korábbi sor-



4. ábra. UAR eredmények a validációs és teszthalmazon, két jellemzőcsoport esetén a multimodális jellemzőcsoport súlyának függvényében.

rendjétől: a prozódiai jellemzőcsoportot megelőzve, a (korábban legrosszabbul teljesítő) videó jellemzőcsoport következik, és az unimodális jellemzőcsoport zárja a sort.

A 3. ábrán látható, hogy a prozódiai, unimodális és a videó jellemzőcsoport minimális súlya a legjobban teljesítő súlyvektorok között 0. Ezen megfigyelés alapján megvizsgáltuk, milyen UAR eredményeket kaphatunk a validációs halmazon kizárólag a multimodális és a szintaktikai jellemzőcsoportok használatával. Az így kapott eredményeket vizualizálja a 4. ábra. A validációs halmazon akkor kaptuk a legjobb eredményt, ha a multimodális jellemzőcsoport súlya 0,69, a szintaktikai jellemzőcsoport súlya pedig 0,31 volt. Az ezekkel a súlyokkal (2 csoport) kapott eredményt összehasonlítása az összes jellemzőcsoportot használó súlyvektorok közül a validációs halmazon legjobban teljesítő súlyvektorral kapott eredménnyel (5 csoport) és az összes jellemzőt felhasználó (Összes) neuronháló eredményével látható a 3. táblázatban. A két kombináció eredménye között sem a validációs, sem a teszt halmazon nincs szignifikáns különbség, és mindkettő szignifikánsan jobb eredményt ad, mint az összes jellemzőt felhasználó neuronháló magában.

3. táblázat. Az összes jellemzőt felhasználó neuronháló eredményének összehasonlítása a jellemzőcsoportok kombinációjával elért eredményekkel.

Típus	Validáció	Teszt
Összes	76,7%	75,7%
5 csoport	78,1%	77,1%
2 csoport	77,9%	77,1%

6. Konklúzió és jövőbeni munka

Kísérleteink alapján úgy tűnik, hogy a neuronhálós osztályozás eredményességéhez leginkább a multimodális és a szintaktikai jellemzőcsoportok járulnak hozzá. Csak ezen két csoport felhasználásával el tudunk érni az összes csoport kombinációjával kapott eredménnyel egyező eredményt, amely szignifikánsan jobb az összes jellemzőt kombináció nélkül felhasználó eredménynél. A jövőben tervezzük az audió jellemzőcsoport vizsgálatát is, miután a szószintű annotáció rendelkezésünkre áll. Valamint tervezzük, hogy az osztályozási feladatról felismerési feladatra lépünk tovább, HMM/ANN hibrid modell használatával.

Hivatkozások

1. Kovács, Gy., Grósz, T., Váradi, T.: Topical unit classification using deep neural nets and probabilistic sampling. In: Proc. CogInfoCom. (2016) 199–204
2. Sapru, A., Bourlard, H.: Detecting speaker roles and topic changes in multiparty conversations using latent topic models. In: Proc. Interspeech. (2014) 2882–2886
3. Holz, F., Teresniak, S.: Towards automatic detection and tracking of topic change. In: Proc. CICLing. (2010) 327–339
4. Schmidt, A.P., Stone, T.K.M.: Detection of topic change in irc chat logs. http://www.trevorstone.org/school/ircsegmentation.pdf (2013)
5. Baiat, G.E., Szekrényes, I.: Topic change detection based on prosodic cues in unimodal setting. In: Proc. CogInfoCom. (2012) 527–530
6. Zellers, M., Post, B.: Fundamental frequency and other prosodic cues to topic structure. In: Workshop on the Discourse-Prosody Interface. (2009) 377–386
7. Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G.: Prosody-based automatic segmentation of speech into sentences and topics. Speech Commun. **32**(1-2) (2000) 127–154
8. Tür, G., Hakkani-Tür, D.Z., Stolcke, A., Shriberg, E.: Integrating prosodic and lexical cues for automatic topic segmentation. CoRR (2001) 31–57
9. Abuczki, A., Baiat, G.E.: An overview of multimodal corpora, annotation tools and schemes. Argumentum **9** (2013) 86–98
10. Pápay, K., Szeghalmy, S., Szekrényes, I.: Hucomtech multimodal corpus annotation. Argumentum **7** (2011) 330–347
11. Hunyadi, L., Szekrényes, I., Borbély, A., Kiss, H.: Annotation of spoken syntax in relation to prosody and multimodal pragmatics. In: Proc CogInfoCom. (2012) 537–541
12. Szekrényes, I.: Prosotool, a method for automatic annotation of fundamental frequency. In: Proc. CogInfoCom. (2015) 291–296
13. Rosenberg, A.: Classifying skewed data: Importance weighting to optimize average recall. In: Proc. Interspeech. (2012) 2242–2245
14. Lawrence, S., Burns, I., Back, A., Tsoi, A.C., Giles, C.L. In: Neural Network Classification and Prior Class Probabilities. Springer Berlin Heidelberg, Berlin, Heidelberg (1998) 299–313
15. Tóth, L., Kocsor, A.: Training HMM/ANN hybrid speech recognizers by probabilistic sampling. In: Proc. ICANN. (2005) 597–603
16. Grósz, T., Nagy, I.: Document classification with deep rectifier neural networks and probabilistic sampling. In: Proc. TSD. (2014) 108–115
17. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: Proc. ICASSP. (May 2013) 6985–6989