Esaldi etenaldien ezarpena CART bidez Euskal Testu Ahots Bihurketarako

Eva Navas, Inmaculada Hernáez, Imanol Madariaga, Juan M. Sanchez, Nerea Ezeiza*

Elektronika eta Telekomunikaziak Saila
* Lengoaia Eta Sistema Informatikoak Saila
University of the Basque Country
{eva; inma; imanol; ion}@bips.bi.ehu.es

Abstract

This paper presents a prosodic phrasing method for the Basque language, to improve naturalness in text to speech synthesis. Binary classification trees are trained with morphological and syntactic information to predict locations of breaks. Overall score achieved by the prediction tree is 92.53%, which compares positively with the results published for other languages.

Laburpena

Artikulu honek testu ahots sintesiaren naturaltasuna hobetzeko prosodia ezartzeko metodoa aurkezten du. Sailkapen bitarreko zuhaitzak informazio morfologiko eta sintaktikoarekin entrenatzen dira etenaldiak non ezarri erabakitzeko. Aurreikuspen zuhaitzak lortutako puntuazio orokorra %92'53 da, beste hizkuntza batzuetarako publikatu diren emaitzen parekoa.

Berba gakoak: Ahotsaren sintesia, entonazioa, etenaldiak..

1. Sarrera

Esaldi etenaldi egokiak ezartzea funtsezkoa da ahots sintetikoaren naturaltasunerako, eta bere ulergarritasunerako ere bai, batez ere esaldi luzeetan. Zoritzarrez hau lan zaila da: ahots naturalean etenaldiak ezartzeko erabakia faktore askoren funtzio da, hala nola testuingurua, ahoskatze abiadura eta arnasa egiteko beharra. Behar denean etenaldirik ez egotea, edo leku desegokian egoteak TAB (testu ahots bihurketa) sistema naturaltasun gabekoa izatea eragiten du. Bestalde, etenaldiak TAB sistemen hainbat moduluk erabili ohi dituzte (iraupen, grafema-fonema eta intonazioa moduluak).

Etenaldiak arau bidez jartzen duten sistemak daude, funtzio/eduki berba sailkapena kontutan harturik (Karn 1996), eta beste sistema batzuk analisi estatistikoaren eskutik ezartzen dituzte etenaldiak (Busser et al. 2001) (Hirschberg 1991). Euskara hizkuntza atzizki-ezartzailea denez oso funtzio hitz gutxi daude, beraz metodo arruntak ez dira erabilgarriak. Lan honetan informazio morfologiko eta sintaktikoarekin entrenatutako sailkapen zuhaitza erabiltzen da etenaldiak ezartzeko.

2. Datuak

Lehen esperimentuetan aurretik eskuragai zegoen ahots datu basea erabili zen. Datu base hau gizonezko euskaldunzahar batek *Campusa* aldizkariko lau artikulu giro isilean irakurriz osatu zen. Zoritxarrez, datu basea txikia zen eta etenaldi kopurua ez zen nahikoa etenaldien aurreikuspen egokia egin ahal izateko. Datu base hau erabiliz egindako ikerketa oinarri baliotsua izan zen artikulu honetan azaltzen den lanaren abiapuntu bezala.

Ahots datu base berria grabatu eta etiketatzeko behar den lan itzela dela eta, testuzko datu basea erabili zen. Testu datu baseen abantailak (Hirschberg et al. 1996) erreferentzian azaltzen dira. Internet tokietan harturiko gai ezberdinetako 17 testurekin *Internet* izendatutako datu-basea osatu zen. I. Taulan agertzen dira datu base bien ezaugarri nagusiak.

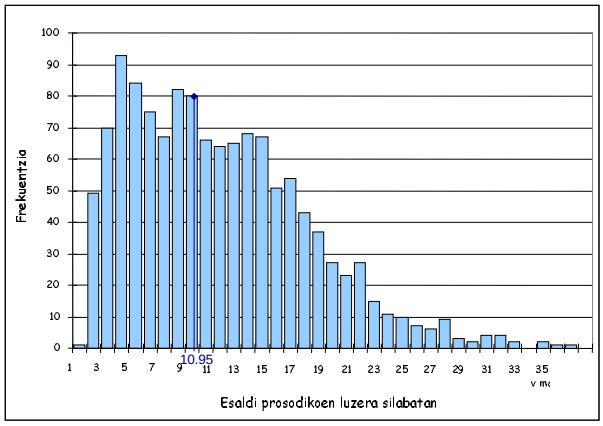
	Campusa	Internet
Mota	testu eta ahozkoa	testua
Grabaketaren luzera	10’	--
Testu kantitatea	7K	38.1K
Berba kopurua	899	4332
Esaldi kopurua	49	366
Etenaldi ortografikoen kopurua	93	605
Etenaldi ez-ortografikoen kopurua	162	665

1. Taula: Erabilitako taulen ezaugarri nagusiak

2.1. Datu basearen etiketatzea

Internet datu basea euskal hiztun euskaldunzaharrak markatu zuen, berba bikote bakoitzaren arteko espazioa etenaldi bezala etiketatuz muga ondo zetorrela pentsatzen zuenean. Ikur ortografiko guztiak etenaldi bezala etiketatu ziren. Ez zen etenaldien sailkapen zehatzagoa egin, mota gehiago erabiliz gero mota bakoitzeko etenaldi gutxiago egongo zirelako korpusean, eta emaitzak zentzudunak izateko mota bakoitzeko adibide kopuru handia behar litzakeelako.

Lortutako esaldi prosodikoen luzeren (silabatan neurtua) banaketa 1.Irudian agertzen da. Batezbesteko silaba kopurua ere adierazten da.



1.Irudia: Esaldi prosodikoen luzeraren banaketa Internet datu basean

Esaldi prosodiko laburrenak berba bakarra dauka (silaba 1), luzeenak 13 berba ditu (36 silaba), eta batezbesteko luzera 11 silabakoa da.

3. Etiketazio morfologikoa

Informazio sintaktiko eta morfologikoa etenaldiak ezartzeko hainbat hizkuntzatan oso baliagarri dela frogatua dago (Black et al. 1997), beraz ikerketa honetan erabilitako datu baseak morfologiko et sintaktikoki etiketatu ziren, IXA taldeak (http://ixa.si.ehu.es) garatutako tresnak erabiliz. Morfologikoki etiketatzeko MORFEUS (Ezeiza et al. 1998), euskararako aztertzaile morfologikoa, erabili da. Programa honek 15 etiketa nagusi ematen ditu, gehienak azpisailkapen maila bi gehiago dituztelarik, gramatikaren arabera. MORFEUSek ematen dituen etiketa kopurua handiegia da gure helbururako, korpusaren neurri txikia kontuan izanik batez ere. Horregatik etiketa multzo txikiagoa erabili dugu, 2. Taulan agertzen dena.

2. Taulan ikusten denez, eduki berba motak ez ziren azpisailkatu eta etiketa nagusia bakarrik erabili zen. Etiketa multzoan funtzio berba mota bi bakarrik zeuden: determinanteak (DET) eta konjuntzioak (LOT). Aurrekoak ez zuen azpisailkapenik behar , baina besteak bai, bigarren kategorizazio mailaren arabera, *Campusa* datu basearekin aurretik egindako ikerketen ondorioek gomendatzen zuten bezala. Etiketak ez ziren eskuz zuzendu.

Etiketa	Azalpena
ADB	Adizlaguna
ADI	Aditz nagusia
ADJ	Izenlaguna
ADL	Aditz Laguntzailea
ADT	Aditz sintetikoa
BEREIZ	Puntuazio ikur berezia
DET	Determinantea
IOR	Izenordea
ITJ	Interjekzioa
IZE	Izena
LOT_JNT	Lokailu juntatzailea
LOT_LOK	Lokailua
LOT_MEN	Lokailu menderatua
PUNT	Puntuazio ikurra

2. Taula: Etiketa morfologikoak

Datuen anotazio sintaktikoa ere egin zen, oraindik garapenean dagoen anotazio automatikorako tresna batekin. Informazio sintaktikoa edukitzea, erroreekin bada ere, ezer ez edukitzea baino hobea da. Tresna honek berbak esalditan biltzen ditu eta beraien funtzio sintaktikoaren arabera sailkatzen ditu. Sailkapen honetan zeuden errore nagusiak eskuz zuzendu ziren: honela objektu zuzen eta subjektuen arteko anbiguotasunak ezabatu ziren, eta izendatu gabe gelditutako berba garrantzitsu batzuei (batez ere lokailuak) balio egokia eman zitzaien.

4. Analisi estatistikoa

Datuen analisi estatistikoa CART bidez egin zen (Breiman et al. 1984.). Entrenamendu datuetako berba bakoitzaren ondoren etenaldia zegoen edo ez aurreikusteko Sailkapen bitarreko zuhaitzak entrenatu ziren.

Korpusa entrenamendu datu (256 esaldi, 3482 espazio etenaldi barik eta 1140 etenaldiarekin) eta ebaluaketa datuetan (111 esaldi, 379 ez-etenaldi eta 130 etenaldi) banatu zen, eta multzo bien estatistikak kalkulatu ziren banaketa onargarria zela frogatzeko.

4.1. Aurreikuspen informazioa

Etenaldien kokapena erabakitzeko informazio hau eman zitzaion zuhaitzari:

- Aztertzen ari den berbaren inguruko bost berben etiketa morfologikoak.
- Aztergai dagoen berbaren eta inguruko bien funtzio sintaktikoa.
- Hurrengo berba oraingoaren sintagma berekoa den edo ez: sintagma bereko berbak ez omen dira prosodikoki etenda egongo.
- Aurreko etenalditik dauden berba eta silaba kopurua eta hurrengo puntuazio ikurrerainoko berba eta silaba kopurua. Etenaldi batetik

hurbil zailagoa da etenaldia egotea. Entrenamendurako datu hauek ezagunak dira; baina ezartzeko orduan zuhaitza ezkerretik eskumara aplikatzen da beraz azkeneko etenaldia zuhaitzak aurreikusitako azkenekoa da.

- Oraingo esaldiaren luzera, silabatan neurtua, esaldi luzeagoak etenaldi gehiago dutenaren hipotesia aztertzeko.

4.2. Aurreikuspen zuhaitza

Etete uneak erabakitzerakoan errore mota ezberdin bi bereiz daitezke:

- Behar ez den lekuan etenaldia sartzea. Hau errore txarrena dirudi erlazio sendoa daukaten sintagmak banatzen direlako.
- Etenaldia behar den lekuan etenaldirik ez sartzea. Hau ez da hain grabea, lortutako testuan ez delako naturaltasun handia galtzen etenaldiak kendu arren.

Etenaldiak erabakitzeko eraikitako lehen zuhaitza, predikzio errore minimoa kalkulatzerakoan errore mota biei garrantzi bera emanez entrenatu zen. Esperimentu honen ondoren beste zuhaitz bat entrenatu zen etenaldi faltsua sartzearen erroreari benetako etenaldia ezabatzeari baino %33ko kostu handiagoa emanez.

Zuhaitzaren lehen erabakiak etenaldia ezartzen du oraingo etiketa puntuazio ikurra bada. Hau espero genuen korpusean puntuazio ikur bakoitzaren ostean etenaldia markatua dagoelako. Ondoren hurrengo berbaren etiketa morfologikoa ebaluatzen da eta adjektiboa, adberbioa, aditz sintetikoa, lokailu menderatua edo puntuazio ikurra bada ez da etenaldirik ezartzen. Hurrengo berbaren morfologia zerrenda horretakoa ez bada orduan hurrengo berba oraingoaren sintagma berekoa den ebaluatzen da. Berba biak sintagma berekoak badira ez da etenaldirik sartzen, baina aurreko etenalditik 9 silaba edo gehiago badaude eta hurrengo puntuazio ikurrera 5 edo gehiago, etenaldia ezartzen da.

5. Emaitzak

Zuhaitzen arrakasta neurtzea ez da erreza. Zuhaitzaren puntuazioa ondo sailkatutako espazio kopurua eta espazio guztien kopurua zatituz kalkulatzen da.

Datu hau kontuz hartu behar da, jatorrizko korpusean dagoen etenaldi proportzioaren menpe dago eta. Ebaluaketarako erabilitako korpusean espazioen %74'46 etenaldibakoak etiketatu ziren, beraz etenaldirik ezartzen ez duen algoritmoaren arrakasta ia %75eko izango zen, ezer egin gabe. Arazo hau gainditzeko zuhaitzaren puntuazioa kalkulatzeko beste era proposatu da, kappa estatistika. Neurri hau sailkapen linguistiko lanetarako erabiltzea Carlettak proposatu zuen lehen aldiz (Carletta 1996) eta ordutik hona beste batzuk erabili dute (Sanders, E., 1995) puntuazioak testuko etenaldi bako espazio kopuruarekiko daukan menpekotasuna saihesteko.

Kappa estatistika 1. Ekuazioak adierazten duen bezala kalkulatzen da

$$\kappa = \frac{\Pr(A) - \Pr(E)}{1 - \Pr(E)} \qquad (1)$$

non Pr(A) zuhaitzak lortutako puntuazio orokorra da eta Pr(E) etenaldibako espazioen proportzioa datuen artean.

(1) espresioan zuhaitzak lortutako puntuazio orokorra datuen artean ez-etenaldi etiketa agertzearen probabilitatearekin konparatzen da, datuen egiturarekiko menpekotasuna ekidituz. Algoritmoak ez badu etenaldirik sartzen kappa estatistikaren balioa 0 izango da. Metodoak espazio guztiak zuzen aurreikusten baditu, κ=1. Balio negatiboak algoritmoak jarritako etenaldiak toki okerretan daudela adierazten du, beraz ez erabiltzea hobe dela.

3. Taulan sarrera eta ezabaketa erroreei kostu berdina ematen dien zuhaitza ebaluaketa datuetan aplikatzean lortutako emaitzak agertzen dira. Zatikako puntuazioak ondo ezarritako etenaldi (edo ez-etenaldi) proportzioa etenaldi (edo ez-etenaldi) kopuru osoarekiko adierazten du. Zuhaitz honek lortutako puntuazio orokorra %92'53 da. Kappa estatistika kasu honetan 0'71 da.

	Etenaldi gabe	Etenaldi	Zatikako puntuazioa
Erreferentzian etenaldi gabe	351	28	%92.61
Erreferentzian etenaldi	10	120	%92.31

3. Taula: 1.predikzio zuhaitzaren arrakasta

Sarrera erroreen kostua ezabatze erroreen kostua baino handiagoa daukan zuhaitzak lortutako emaitzak 4. Taulan agertzen dira. Zuhaitz honek lortutako puntuazio orokorra ere %92'53 da, eta kappa estatistikaren balioa 0'71, baina erroreen distribuzioa ezberdina da. Oraingoan aurrekoan baino ezabaketa errore gehiago eta sarrera errore gutxiago dago.

	Etenaldi gabe	Etenaldi	Zatikako puntuazioa
Erreferentzian etenaldi gabe	371	8	%97.89
Erreferentzian etenaldi	30	100	%76.92

4. Taula: 2.predikzio zuhaitzaren arrakasta

Taula bietako datuak zuhaitza entrenamendu eran erabilita lortu dira, hau da, aurreko etenaldirarteko

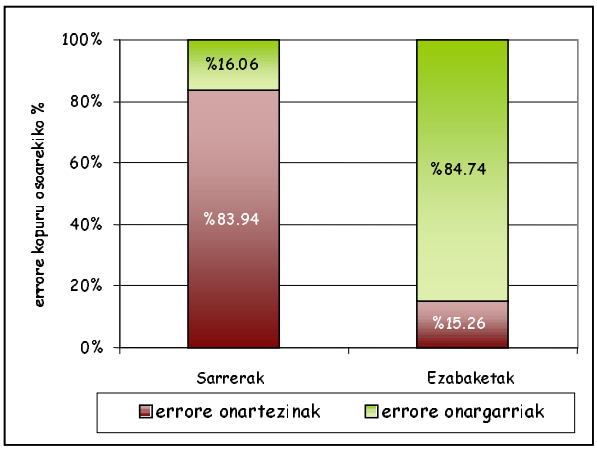
silaba eta berba kopurua etenaldien kokapen zuzena jakinik neurtuta.

6. Eztabaida

Etenaldiak ezartzeko entrenatutako zuhaitzek lortutako puntuazioak beste hizkuntza batzutan lortutako emaitzen parekoak edo hobeak dira:

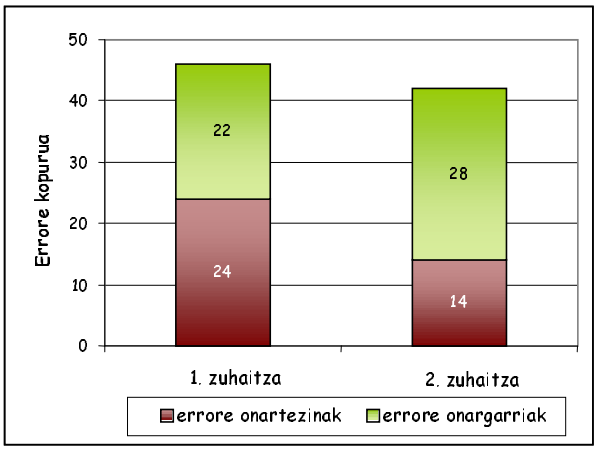
- Mexikar Española: %94.2ko puntuzio orokorra lortu zuen (Hirschberg et al. 1996) erreferentziako lanak, baina algoritmoa entrenamendu datu beraiekin ebaluatuz.
- Ingelesa: %90 (κ=0.5) lortu zuen (Sanders, E., 1995) lanekoak proposatutako metodo onenarekin, (Black et al. 1997), lanekoak %91.5 (κ=0.53) lortzen du sekuentzia morfologikoak eta Markov eredua erabiliz etenaldi sekuentzia egokiena emateko eta (Busser et al. 2001) lanak %90.8 (κ=0.59) lortzen du memorian oinarritutako ikasketarekin.
- Korearra: (Yeon-Jun et al. 1999) lanak %77'0 (κ=0'56) lortu zuen, (Sangho et al. 1999) lanak %84'9 (κ=0'62) CARTetan oinarritutako metodoarekin eta (Byeongchang et al. 2000) lanak %85.5 (κ=0.64).
- Japoniera: (Fujio et al. 1997) lanak %89'9 puntuazio orokorra lortu zuen testuinguru gabeko gramatika estokastikoa entrenatzen.

Zenbaki hauek algoritmo ezberdinak konparatzeko balio dute baina kontuz hartu behar dira. Errore guztiak eragin berdina daukate puntuazioa kalkulatzerakoan, baina ez daukate guztiak garrantzi berdina. Zuhaitzek egindako erroreak mota bitan banatu ditugu: errore onargarriak eta onartezinak. Errore mota bien banaketa sarrera eta ezabatze erroren artean 2. Irudian agertzen da. Aurrerago esan denez, sarrera erroreak ezabatze erroreak baino kaltegarriagoak izan ohi dira, %84 onartezinak dira. Ezabatze erroreak aldiz onargarriak dira proportzio berean.



2. Irudia: Erroreen garrantziaren distribuzioa errore motarekiko.

Ikerketa honetan eraikitako zuhaitz biek puntuazio orokor bera dute, baina sarrera eta ezabaketa errore proportzio ezberdina. 3. Irudian zuhaitz bakoitzak, etenaldiak ebaluaketa datuetan jartzerakoan, egindako mota bakoitzeko errore kopurua agertzen da. Errore kopuru osoa ez da 3 eta 4 Taulan agertzen dena, oraingoan aurreko etenalditik dauden silaba eta berba kopurua zenbatzeko zuhaitzak aurreikusitako etenaldiak erabili direlako. Honela emaitzak txarrago izateko joera dauka erroreak propagatzen direlako. Errore onartezinak konparatuz lehen zuhaitzaren emaitza bigarrena baino kaxkarragoa da. Beraz bigarrena erabili beharko genuke etenaldi prosodikoak ezartzeko.



3. Irudia Fujio S.; Sagisaka Y.; Higuchi N., 1997: Zuhaitz bakoitzaren erroreen sailkapena.

Metodoa hobetzeko asmoz errore txarrenen jatorria aztertu da, 5. Taulan agertzen diren emaitzekin.

Errore kopurua	Jatorria
7	Errorea anotazio sintaktikoan
3	Errorea zuhaitzak aurrerago jarritako etenaldiak erabiltzeagatik
1	Errorea silabak banatzeko algoritmoan
1	Errorea analisi morfologikoan
1	Txarto etiketatutako etenaldia jatorrizko datuetan
1	Silaba kopurua azkeneko etenalditik = 10

5. Taula: 2. zuhaitzaren errore onartezinen jatorriak

Errore onartezin erdiak etiketa sintaktiko okerretatik datoz, beraz analisi sintaktiko zuzena egitea funtsezko da.

Erabilitako korpusa txikia izan arren eta etiketetan erroreak egon arren, etenaldi prosodikoak ezartzeko aukeratutako ezaugarriak eta metodo estatistikoa oso baliagarriak dira, lortutako emaitzak frogatzen duten bezala.

7. Esker onean

Autoreek etenaldi prosodikoen ezarpenean ikerketaren hasieratik lan egin duen Patricia Perez-en laguntza eskertzen dute.

Baita Euskal Herriko Unibertsitateko eta Zientzia eta Teknologia Ministerioko dirulaguntzak (UPV147.345-TA066/98 eta TIC2000-1005-C03-03 proiektuak) eskertzen ditugu.

8. Aipamenak

Black, A.W.; Taylor, P. (1997). Assigning phrase breaks from part-of-speech sequences, In *Proceedings of Eurospeech'97,* Rhodes, 995-998 orr.

Breiman, L.; Friedman, J.H.; Olsen, R.A.; Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall.

Busser, B.; Daelemans, W.; van den Bosch, A. (2001). Predicting phrase breaks with memory-based learning. In *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Edimburgh.

Byeongchang K,; Geunbae L. (2000). Decision-Tree based Error Correction for Statistical Phrase Break Prediction in Korean. In *Porceedings of the 18th International Conference on Computational Linguistics*.

Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22 (2), 249-254 orr.

Ezeiza N.; Aduriz I.; Alegria I.; Arriola J.M.; Urizar R., 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In *Proceedings of COLING-ACL'98*, Montreal.

Fujio S.; Sagisaka Y.; Higuchi N. (1997). Prediction of Major Phrase Boundary Location and Pause Insertion Using a Stochastic Context-free Grammar. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*. New York: Ed. Springer.

Hirschberg, J. (1991). Using text analysis to predict intonational boundaries. In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, Genoa.

Hirschberg, J.; Prieto, P. (1996). Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, Vol. 18, 281-290 orr.

Karn, H.(1996). Design and evaluation of a phonological phrase parser for Spanish text-to-speech. In *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, vol. 3, 1696-1699 orr.

Sanders, E. (1995). *Using probabilistic methods to predict phrase boundaries for a text-to-speech system*. Master's thesis, University of Nijmegen.

Sangho L.; Yung-Hwan O. (1999). Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Communication* , vol. 28, 283-300 orr.

Yeon-Jun K.; Yung-Hwan O. (1999). Prosodic Phrasing In Korean; Determine Governor, and Then Split or Not. In *Proceedings of Eurospeech'99*, Budapest, 539-542 orr.