Ezeztapena identifikatzeko Murriztapen Gramatikako erregelak sentimenduen analisiaren testuinguruan

Jon Alkorta Koldo Gojenola Mikel Iruskieta
HiTZ Hizkuntza Teknologiako Euskal Zentroa - Ixa (UPV/EHU)
{jon.alkorta, koldo.gojenola, mikel.iruskieta}@ehu.eus

IKER GAZTE 2021
NAZIOARTEKO IKERKETA EUSKARAZ

University of the Basque Country
IXA GROUP

Sarrera eta motibazioa

· **Sentimenduen analisia:** testu bateko informazio subjektiboa erauztea.

· Bertako atazarik ohikoena testuak balorazio positiboa edo negatiboa egiten duen identifikatzea.

(1) Mugikor hori atsegin dut.

(2) Urak 100 gradutan irakiten du.

· Testuen balorazioa identifikatzeko ohiko metodoetako bat sentimenduen hiztegiak erabiltzea da.

Hitza lexikoian	Balentzia
bikain	+5
on	+2
eskas	−1
txar	−3

· Horrela, hitzei sentimendu balentziak esleitzen zaizkie.

(3) Antzezlan hori **ez**$_{-4}$ da [bikaina$_{+4}$] izan. = 0

(4) Ikusi dugun filma **ez**$_{+4}$ da [txarra$_{-3}$] izan. = +1

· Metodo honetan, ezeztapen-markak berak ere balentzia bat du (+/- 4) eta balentziadun hitzekin eragiketak gertatzen dira.

Aurrekariak eta helburuak

· Lanak ikuspegi ezberdinetakoak eta tratatutako hizkuntzak tipologia anitzetakoak.

· **Teknikak:**

1- Hitz-zaku ereduetan oinarritua.

· Pang et al. (2002): hitz bakoitzaren ondoren ezezkoa gehitu.

2- Balentzia-aldatzaileetan oinarritua.

· Polanyi eta Zaenen (2006): zenbait hitzek esaldiaren balentzia eraldatzen dute.

3- Erlazioetan eta dependentzietan oinarritua.

· Esuli eta al. (2008): termino ezberdinen (ezeztapena barne) eragin-eremua aztertu.

· **Gure helburua:**

· Ezeztapenaren irismena identifikatzeko Murriztapen Gramatikako erregelak sortu (2. teknika).

Metodologia

1- Marko teorikoa: Murriztapen Gramatika (Karlsson et al., 2011).

2- Erregelak sortu. Urratsak:

1- Euskarazko ezeztapen-marken zerrenda osatu.

2- Ezeztapen-marka hori duten esaldiak erauzi corpusetik.

3- Esaldietako hitzei sentimenduen balentziak esleitu.

(5) Pogostkinak ezin [hobeki]$_{+2}$ atera zituen. (MUS20)

(6) [Irabazi]$_{+2}$ ezinik jarraitzen du Eibarrek. (KIR17)

(7) Ikuspuntu [politikotik]$_{-1}$ ez ezik, [ekonomikotik]$_{+3}$ ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herria. (POL08)

4- Corpuseko esaldiak banan-banan eskuz aztertu.

(8) Pogostkinak [ezin hobeki$_{+2}$] atera zituen. (MUS20)

(9) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek. (KIR17)

(10) [Ikuspuntu politikotik$_{-1}$ ez ezik]. [ekonomikotik$_{+3}$] ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herria. (POL08)

5- Fenomenoak deskribatzen dituzten sasi-erregelak eskuz egin.

(11) (...) [ezin da baztertu$_{-1}$] ekaitz zaparradaren bat izatea. (EGU35)

(12) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek (...) (KIR22)

Adb.	Ezeztapen-marka	Erregelen egitura
(1)	ezin	PM *ezin* + [adjektiboa/adberbioa] (+ atzizki konp.) PM
(5)	ezin	PM [ad. nag. + *ezin*] PM PM [*ezin* (+ ad. lag.) + ad. nag.] PM
(10)	ez	PM [IS] *ez ezik* PM

6- Murriztapen Gramatikako erregelak sortu.

```
LIST PUNTUAZIOA = PUNT_PUNT PUNT_KOMA PUNT_BI_PUNT PUNT_GALD PUNT_ESKL PUNT_HIRU PUNT_PUNT_KOMA;

LIST EZ = ''ez'';
LIST BESTERIK = ''beste'';

# (2)
# besterik ez egiturak
MAP (!besterikezHAS) TARGET (DET) IF (0C BESTERIK)   (1C EZ);

(...)
```

3- Erregelak ebaluatu. Urratsak:

1- Euskarazko Iritzi Corpusa (Alkorta et al., 2018) prozesatu erregelak ebaluatu ahal izateko.

2- Murriztapen Gramatika formalismoan sorturiko 73 erregelak corpusaren test zatiko 144 testuetan aplikatu.

3- Gramatikak aplikaturiko erregelak eskuz ebaluatu.

```
''<,>''''<PUNT_KOMA>''
        PUNT_KOMA
''<batere>''
        ''batere'' ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG %SINT
''<harrotu>''   S:278/0
        ''harrotu'' ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG %ADIKAT S:278 !gabeAUR
''<gabe>''   S:141/0
        ''gabe'' ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> %SINT S:141 !gabe
''<$.>''''<PUNT_PUNT>''
        PUNT_PUNT
```

Emaitzak eta etorkizuneko lanak

1- Erregela-kopurua

Erregela		Kopurua		Kopurua
LIST motakoak				**24**
MAP motakoak				**73**
	Orokorrak	3	EDO/EDOTA/ALA EZ	4
	Egitura lexikalizatuak	10	EZ... EZ...	4
	EZIN	18	EZ	14
	SALBU	4	GABE	3
	IZAN_EZIK	5	EZTA	9
	EZ_EZIK	1		

2- Ebaluazio-emaitzak:

Eragina	Doitasuna	Estaldura	F1	Hitzak (instantziak): ✓/✗/+
Indartu	1,00	1,00	1,00	4 (2): 4/0/0
Ahuldu	0,93	0,80	0,86	1.050 (192): 784/63/203
Eraginik ez	0,97	1,00	0,98	36 (19): 35/1/0
Guztira	0,93	0,80	0,86	1.090: 823/64/203

	Doitasuna	Estaldura	F1	Hitzak (instantziak): ✓/✗/+
Ezeztapen-markak	1,00	0,96	0,98	195 (195): 188/0/7
Egitura lexikalizatuak	0,96	1,00	0,98	28 (16): 27/1/0
Irismena	0,91	0,75	0,82	867 (195): 601/63/196
Guztira	0,93	0,80	0,86	1.090: 823/64/203

3- Etorkizuneko lana

· Ezeztapen-marken corpusa handitu erregelak hobetzeko.

· Erregelak sentimenduak sailkatzeko euskarazko tresna batean inplementatu.

Erreferentzia bibliografikoak. Pang, Bo, Lillian Lee, eta Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machinelearning techniques.arXiv preprint cs/0205070. Polanyi, Livia, eta Annie Zaenen. 2006. Contextual valence shifters. In Computing attitude and affect in text:Theory and applications, 1–10. Springer. Esuli, A. 2008. Automatic generation of lexical resources for opinion mining: models, algorithms and applications. In*Acm sigir forum* (Vol. 42, No. 2, pp. 105-106). New York, NY, USA: ACM. Alkorta, J., Gojenola, K., & Iruskieta, M. 2018. SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis.*Computación y Sistemas*, *22*(4). Karlsson, F. 2011. Designing a parser for unrestricted text. In *Constraint grammar* (pp. 1-40). De Gruyter Mouton.