BERnaT: Euskal Hizkuntzaren Aniztasuna Modelatzen

Ekhi Azurmendi, Jaione Bengoetxea, Julen Etxaniz, Joseba Fernandez de Landa,
Maite Heredia, Aitor Soroa eta **Mikel Zubillaga**
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{maite.heredia, mikel.zubillaga}@ehu.eus

Motibazioa

- Hizkuntza aniztasunaren inpaktua aztertzea hizkuntza ereduak modelatzean
- Zehaztasun altuagoko euskararako kodetzaileen sorrera
  - Euskarazko erreferentziazko modelorik onena (RoBERTa-Euscrawl) oinarri-lerrotzat hartuta

Ekarpenak

- Hizkuntza aniztasuna barne biltzen duen corpus berria
  - Txioz osatutako **egile** (author) eta **data** (time) **corpus** erraldoiak
- Euskal hizkuntzarako modelo diskriminatibo berriak
  - Corpus tamaina handiagoarekin
  - Hizkuntza aniztasun zabalagoarekin

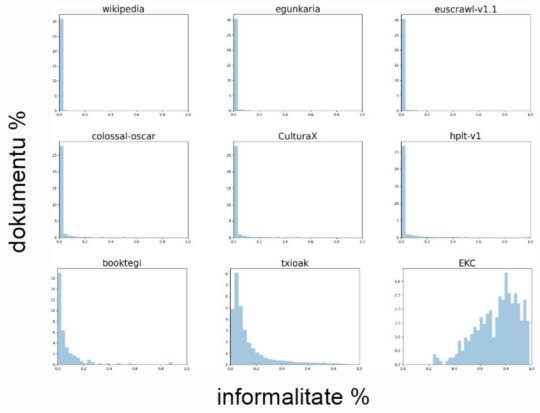
Entrenamendu Datuak

- **EusCrawl** **359M hitz**
- **Latxa Corpus** **1.220M hitz**
- **Euskal Klasikoen Corpusa** **21M hitz**
- **Txio-corpusa [berria]** **276M hitz**
  - txioak idatzi ziren **data**ren (time) arabera
  - txioak idatzi zituzten **egile**en (author) arabera

Informalitate Analisia

- Corpus atal bakoitzaren edukia informal-formal bezala sailkatu da textu sailkatzaile automatikoak erabilita

Source	Docs	Words	mean	mdn	std
wikipedia	409k	51M	0.001	0.00	0.016
egunkaria	176k	39M	0.003	0.00	0.023
euscrawl-v1.1	1.79M	359M	0.004	0.00	0.042
colossal-oscar	234k	105M	0.018	0.00	0.081
CulturaX	1.31M	541M	0.019	0.00	0.083
hplt-v1	375k	120M	0.030	0.00	0.109
booktegi	166	3M	0.073	0.03	0.125
txioak author	13k	188M	0.141	0.08	0.166
txioak time	11M	188M	0.100	-	-
EKC	338	21M	0.733	0.77	0.163



Esperimentuak

- **5 RoBERTa eredu:** corpus mota desberdinekin aurre-entrenatuak

Model/Dataset	EusCrawl 1.1	Latxa Corpus	Txioak	EKC
RoBERTa Euscrawl	X			
RoBERTa Latxa	X	X		
RoBERTa Txio			X	
RoBERTa Txio-Euscrawl	X		X	
BERnaT	X	X	X	X

Ebaluazioa

- **BasqueGLUE**
  - **Atazak:** Entitateen sailkapena (*NERC*), asmoen sailkapena (*intent* eta *slot*), gaien sailkapena (BHTC), QNLI, desanbiguazioa (*WiC*), Korreferentzien ebazpena (*coref*).
  - **Sare sozialetako** datuetan oinarritutako **atazak:**
    - Sentimenduen analisia: *BEC*
    - Jarreren detekzioa: *Vaxx*

Lehen Emaitzak

Ondorioak

- Corpus tamainaren eragina ez da determinantea emaitzetan, kontuan hartuta ereduaren parametro (base) mugatuak
- Hizkuntza aniztasunaren inpaktua ikusi daiteke, baina soilik ataza zehatzetan (BEC eta Vaxx)

Etorkizuneko lana

- Modelo diskriminatibo handiagoekin (large) ikerketak burutu
  - Corpusaren tamainaren eragina
  - Hizkuntza aniztasunaren inpaktua
- Ebaluaketa datu-multzo gehiagotan frogatu