Dokumentu sailkapena euskaraz (DOKU)

Gorka Dabó Aizpurua

Abstract

Dokumentu sailkapena euskaraz (DOKU) proiektuaren helburua da euskarazko testuen sailkapena aztertzea eta hobetzea. Lan honek BasqueGLUE datu multzoa eta teknika desberdinak erabiliz, sailkapenaren errendimendua ebaluatu eta emaitza esanguratsuak lortzea proposatzen du.

1 Sarrera

Dokumentu sailkapena euskaraz (DOKU) lan honek testu sailkapeneko metodologia desberdinak aztertzen ditu euskarazko testuen kontextuan. Helburua da BasqueGLUE datu multzoa erabiliz sailkapen sistema eraginkorrak garatzea eta testuinguru elebidunean emaitza esanguratsuak eskaintzea. Lan honen proposatzailea, Oier Lopez de Lacalle, euskarazko testuen prozesamenduan esperientzia handiko ikerlaria da.

Testuen sailkapena hainbat esparrutan erabiltzen da, adibidez, hizkuntza naturalaren prozesamenduan, bilaketa motoreetan eta informazioaren berreskurapenean. Proiektu honek erronka bereziak aurkezten ditu, hizkuntzaren berezitasunak eta datu kopuruaren mugak kontuan hartuta. Lanaren helburu nagusia da euskarazko sailkapen metodoen etorkizunerako bideak argitzea eta testu sailkapena hizkuntzaren prozesamenduan lantzea.

2 Erlazionatutako Lanak

Lan hau hainbat oinarri teorikotan eta aurretik egindako lanetan oinarritzen da. Hauen artean daude bi PDF dokumentu hauek, eskuragarri egela.ehu.eus plataforman:

2.1 Transformers

- **3.3 Transformers**: 3.3.transformers.eu.pdf
- **3.4 Prompting**: 3.4.prompting.eu.pdf

Horrez gain, erabilitako laborategi praktiken artean **Prompting Laborategia (9. saioa)** nabarmendu behar da: Prompting Laborategia (9. saioa).

3 Sistema

Lan hau burutzeko, **Google Colab** plataformaren erabilera oinarrizkoa izan da. Hala ere, GPU baliabideen erabilera mugatua izan denez, hainbat kontu sortu behar izan ditut GPUren erabilera denboraren mugak gainditzeko. Plataforma honek eskaintzen dituen baliabide mugatuak aprobetxatzeko, proiektua modu egokian egokitu eta antolatu behar izan dut, beharrezko esperimentuak egiteko.

4 Datuak

Proiektu honetan bi datu multzo erabili dira, euskaraz eta ingelesez:

Euskarazko testu sailkapenaren datu-multzoa: BasqueGLUE proiektuaren barruan dagoen **BHTC** (Basque Headline Topic Classification) datu-multzoa erabili da, GitHub biltegitik deskargatu daitekeena. Datu-multzo honek, albisteen titularretan oinarrituta, gaikako sailkapenak egiten ditu. Zutabeak honakoak dira: **etiketa (label, gaia)** eta **testua**.

Etiketa posibleak 12 dira: [Ekonomia, Euskal Herria, Euskara, Gizartea, Historia, Ingurumena, Iritzia, Komunikazioa, Kultura, Nazioartea, Politika, Zientzia].

Datuak hiru azpidatu-multzotan banatuta daude: **train**, **val** eta **test**, eta banaketa honela geratzen da:

- Guztira: 12,296 sarrera.
- Etiketako banaketa:

```
Ekonomia Train: 801, Dev: 165, Test: 169
Gizartea Train: 2,411, Dev: 518, Test: 535
(gainerako kategoriak berdin antolatuta daude)
```

Ingelesezko datu-multzoa: BBC albiste sailkapenari dagokion BBC News Topic Dataset datu-multzoa erabili da, 2004 eta 2005 urteetako albisteekin. Kategoriak bost dira: **business, entertain-**

ment, politics, sport, tech. Datu multzo honek zutabe hauek dauzka: **text, label** eta **label text**.

5 Erabilitako Ereduak

Hainbat hizkuntza-eredu erabili dira testuen sailkapenerako:

Euskarazko BERT eredua: BERTeus base cased, euskarazko testu korpusetan aurreztrebatutako eredua. Testuinguru errepresentazioak sortzeko eraginkorra da, euskarazko datuetan oinarrituta.

Multilingual BERT: BERT multilingual base model (cased). 104 hizkuntzatan aurreztrebatutako eredua da, Wikipedia edukietan oinarritua.

Euskarazko GPT-2 eredua: GPT2 Basque small model Version 2. GPT arkitektura duen eredu txiki bat da, euskarazko testuak sortu eta ulertzeko gaitasunarekin. Erabilitako tokenizatzailea BPE da, 50,000 hitzezko hiztegi tamainarekin.

6 Emaitzak

Proiektu honetan, hiru zeregin nagusi ($Z1$, $Z2$, eta $Z3$) landu dira, eta $Z2$-n hiru estrategia desberdin aplikatu dira. Estrategia bakoitzean bi hurbilketa nagusi erabili dira emaitzak hobetzeko:

- **Mapeatu**: Datu-multzo baten etiketak (labels) beste datu-multzo bateko etiketen egiturara egokitzen dira.
- **Fine Tuning**: Eredu bat datu-multzo batean entrenatu eta azken geruza fintzen da ebaluazio-datu-multzoko etiketetara egokitzeko.

Lortutako emaitzak $F1$-score erabiliz neurtu dira, eta ondoko taulan laburbiltzen dira:

Zeregina	Estrategia	Mapeatu	FineTuning
Z1	-	74.76	
Z2	1	33.81	28.96
	2	36.24	32.06
	3	31.61	25.40
Z3	-	9.36	

Table 1: $F1$-score emaitzak.

6.1 Emaitzen azalpena

Z1: Euskararako eskuragarri dagoen *BERTeus* eredua erabiliz, dokumentu sailkapenean $F1 = 74.76$ lortu da. Emaitza hau aurreko lanekin alderatuta emaitza onen artean dago.

Z2: Euskararako entrenamendu-daturik ez dagoela suposatuz, hiru estrategia desberdin aplikatu dira:

- **Estrategia 1:** Ingelesezko datu-multzo antzeko batean eredu bat entrenatu da eta datuak euskarara itzuli dira ebaluatzeko.
- **Estrategia 2:** Euskarazko datu-multzo bat ingelesera itzuli da entrenatzeko, eta ebaluazioa euskaraz egin da.
- **Estrategia 3:** Eredu eleaniztun bat erabiliz, ingelesezko datuetan entrenamendua egin eta euskarazko test-datuetan ebaluazioa egin da.

Estrategien artean, emaitza onena $F1 = 36.24$ lortu da (**Mapeatu**, Estrategia 2). Hala ere, fine-tuning bidezko emaitzak oro har apalagoak izan dira.

Z3: Euskarazko *GPT-2* eredu bat erabiltzea esperimentatu da prompting bidez. Emaitzak apalak izan dira ($F1 = 9.36$), eta horrek adierazten du ereduak oraindik lan gehiago behar duela sailkapenlanetan emaitza esanguratsuak lortzeko.
Taulako emaitzek iradokitzen dute $Z1$ hurbilketak emaitza sendoenak eskaintzen dituela eta $Z2$-n itzulpenak emaitzak hobetzeko lagungarriak izan daitezkeela, baina $Z3$-n emaitza eskasak lortu dira prompting bidez.

7 Analisia

7.1 Tekniken Justifikazioa

Proiektu honetan erabiltzen diren datu-multzoek ez dute etiketa kopuru bera: ingelesezko BBC News dataset-ak bost etiketa ditu, eta euskarazko BasqueGLUE dataset-ak hamabi. Horrek erronka bat aurkezten du, izan ere, bost etiketa dituen eredu bat ezin da zuzenean hamabi etiketadun dataset batean ebaluatu. Hori dela eta, bi estrategia nagusi aplikatu dira:

- **Mapeatu etiketa-kopuruak:** Euskarazko datu-multzoko etiketa hamabiak ingelesezko bost etiketen egiturara mapatu dira, egokitasunaren arabera.
- **Fine-tuning:** Ingelesezko bost etiketa dituen eredu bat entrenatu ondoren, azken geruza aldatu da hamabi etiketa aitortzeko eta berriro fintzea (*fine-tuning*) egin da euskarazko datu-multzoarekin.

Fine-tuning teknikak emaitza hobeak lortuko zituela espero nuen, baina Mapeatuen bidez emaitza hobeak lortu dira $F1$-score metrikaren arabera.

7.2 Lortutako emaitzen analisia

Lortutako emaitzak aurrez definitutako zereginen ($Z1$, $Z2$, eta $Z3$) arabera konparatu dira, eta taula honen bidez laburbil daitezke:

Zeregina	Estrategia	Mapeatu	FineTuning
Z1	-	74.76	
3*Z2	1	33.81	28.96
	2	36.24	32.06
	3	31.61	25.40
Z3	-	9.36	

Table 2: $F1$-score emaitzak.

7.3 State of the Art-ekin konparazioa

Gaur egungo state-of-the-art emaitzak hauek dira, irudian erakusten den bezala:

- **RoBERTa-large (EusCrawl)**: $F1 = 77.6 \pm 0.5$
- **RoBERTa-base (CC100)**: $F1 = 76.2 \pm 0.4$
- **RoBERTa-base (EusCrawl)**: $F1 = 76.2 \pm 0.6$
- **RoBERTa-base (Wikipedia)**: $F1 = 70.0 \pm 0.8$
- **mC4**: $F1 = 75.3 \pm 0.7$

Emaitzak honela aztertzen dira:

- **Z1**: Proiektuko $Z1$-ko emaitza ($F1 = 74.76$) state-of-the-art emaitzei oso hurbil dago. Euskararako espresuki prestatutako *BERTeus* ereduak emaitza oso onak eman ditu, RoBERTa-base modelotik oso gertu, baina oraindik ez du *RoBERTa-large* gainditzen.
- **Z2**: Ingelesezko datu-multzoak erabiltzea eta etiketa-mapeaketa egitea fine-tuning baino emaitza hobeak lortzeko estrategia izan da. Hala ere, emaitzak state-of-the-art emaitzetatik urrun daude.
- **Z3**: GPT-2 oinarritutako prompting teknika oso emaitza eskasak ($F1 = 9.36$) eman ditu, eta horrek iradokitzen du eredu hau oraindik ez dela egokia euskarazko dokumentu sailkapenerako.

8 Ondorioak

Proiektu honetan lortutako emaitzak aztertuta, ondorio nagusi hauek atera daitezke:

8.1 Emaitzen Balorazio Orokorra

Oro har, proiektu honetako emaitzak nahiko baxuak izan dira, $Z1$ zereginean izan ezik. $Z1$-n $F1$-score emaitza on bat lortu dut ($F1 = 74.76$), eta horrek iradokitzen du Euskararako prestatutako *BERTeus* ereduak dokumentu sailkapenerako gaitasun sendoa duela. Hala ere, $Z2$ eta $Z3$ zereginen emaitzak oso urrun geratu dira espero zitekeenetik, eta horrek hausnarketa sakona eskatzen du.

8.2 $Z1$-ren Arrakasta eta Hipotesiak

$Z1$-n lortutako emaitza positiboa izan arren, zaila da ziurtatzea zein faktorek eragin duten emaitza hori, eta beste zereginen emaitza baxuak zergatik ez diren antzekoak izan. Honen inguruan hipotesi batzuk planteatu daitezke:

- **Ereduaren kalitate linguistikoa:** *BERTeus* ereduak euskararako egokitutako datu-multzo batean oinarritzen da (EusCrawl, adibidez). Litekeena da datu horiek euskarazko dokumentu sailkapenaren zereginetarako kalitate eta egokitasun handia izatea.
- **Datu-multzoen homogeneotasuna:** BasqueGLUE datu-multzoa sailkapeneko zereginerako kalitate handiko datu homogeneoak izan daitezke, eta horrek ereduaren errendimendua hobetzen lagundu dezake.
- **Fine-tuning metodoaren hobekuntza:** $Z1$-n erabiltzen den fine-tuning teknika, datu aski espezifikoekin entrenatuta, aukera eman dezake ereduak testuinguruan hobeto egokitzeko.

8.3 $Z2$ eta $Z3$-ren Emaitza Baxuak

Aitzitik, $Z2$ eta $Z3$-n emaitza oso baxuak lortu dira, eta horretarako hipotesi hauek proposa daitezke:

- **Itzulpenaren kalitatearen eragina:** $Z2$-n, itzulpen automatikoak testuingurua eta esanahia galtzeko arriskua dakar, bereziki topiko sailkapen zehatzetan. Horrek ereduaren errendimendua nabarmen kaltetu dezake.
- **Eredu eleaniztunen desegokitasuna:** Eredu eleaniztunak hizkuntza askotarako prestatuta daude, baina baliteke euskararako datu gutxi

izateak edo eredu horien arkitektura gehiegi orokortzeak emaitzak gutxitzea.

- **Prompting estrategien heldutasuna:** $Z3$-ko emaitzak ($F1 = 9.36$) iradokitzen du GPT-2 ereduak oraindik ez dituela euskarazko testuinguru konplexuak behar bezala ulertzen. Baliteke euskararako pretraining datuetan hutsuneak egotea edo prompting teknika bera hobetzea beharrezkoa izatea.

- **Ereduen datu-prestakuntzaren eskasia:** Euskarazko datu-multzoak urritasun nabarmena dute beste hizkuntzekin alderatuta. Honek ereduaren prestakuntza eta egokitasun orokorra mugatzen du.

8.4 Proiektuaren Balorazio Pertsonala

Proiektu hau burutzeak esperientzia aberasgarria izan da niretzat, hainbat arrazoirengatik:

- **Ikaskuntza teknikoa:** Datuak tratatzeko, ereduak fine-tuning bidez egokitzeko eta itzulpen-estrategiak aztertzeko teknika ugari ikasi ditut. Honek hizkuntza-prozesamenduaren inguruko nire ezagutza eta trebetasun teknikoak nabarmen handitu ditu.

- **Ikerketa prozesuaren ulermena:** State-of-the-art emaitzak konparatu eta neurtzeak ikerketa prozesu baten logika eta metodologia sakonago ulertzen lagundu dit.

- **Motibazioa:** Proiektuak interesa piztu dit hizkuntza-teknologiaren munduan gehiago ikasteko eta espezializatzeko. Euskara bezalako hizkuntza gutxituetan aplikazio berritzaileak garatzea oso interesgarria iruditzen zait.

8.5 Etorkizuneko Lanak

Proiektu honetan ikasitakoan oinarrituta, etorkizunean zenbait hobekuntza proposatzen dira:

- Itzulpen automatikoaren kalitatearen hobekuntza eta estrategia berrien esplorazioa, adibidez, *back-translation* edo *data augmentation*.

- Eredu eleaniztunen euskararako fintzea eta egokitzapena, datu gehiago erabiliz edo pretraining prozesua optimizatuz.

- Prompting tekniken egokitzapena, testuinguru egokietan erabili ahal izateko.

- Datu-multzo berriak sortzea edo egokitzea, euskarazko hizkuntza-teknologiaren eremua indartzeko.

Oro har, proiektu hau burutzeak asko ikasteko aukera eman dit, eta emaitza zehatzak lortzetik haratago, hizkuntza-teknologiaren eremuan nire interesa eta motibazioa piztu ditu.

9 Erreferentziak

References

[1] Orai NLP Team. *BasqueGLUE: BHTC (Basque Headline Topic Classification) Dataset*. Disponible en: https://github.com/orai-nlp/BasqueGLUE/tree/main/bhtc

[2] SetFit. *BBC News Topic Dataset*. Disponible en: https://huggingface.co/datasets/SetFit/bbc-news

[3] IXA Group. *BERTeus Base Cased Model*. Disponible en: https://huggingface.co/ixa-ehu/berteus-base-cased

[4] Google. *BERT Base Multilingual Cased*. Disponible en: https://huggingface.co/google-bert/bert-base-multilingual-cased

[5] ClassCat. *GPT-2 Small Basque Model Version 2*. Disponible en: https://huggingface.co/ClassCat/gpt2-small-basque-v2

[6] Lopez de Lacalle, O. *Transformers (3.3)*. Acceso en plataforma eGela: https://egela.ehu.eus/pluginfile.php/9732390/mod_resource/content/2/3.3.transformers.eu.pdf

[7] Lopez de Lacalle, O. *Prompting (3.4)*. Acceso en plataforma eGela: https://egela.ehu.eus/pluginfile.php/9732405/mod_resource/content/5/3.4.prompting.eu.pdf

[8] Lopez de Lacalle, O. *Prompting Laborategia (9. saioa)*. Disponible en: https://colab.research.google.com/drive/1ZZ_Vwz5ZjcPwu0RQs6P2SFPzvUgxpMvT