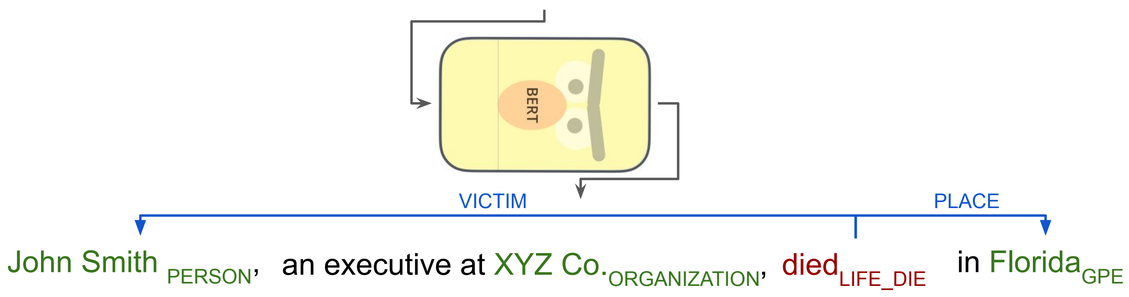
Hizkuntzen tipologiak hizkuntzen arteko transferentzian duen eraginaren analisia gertaera-erauzketa atazetan

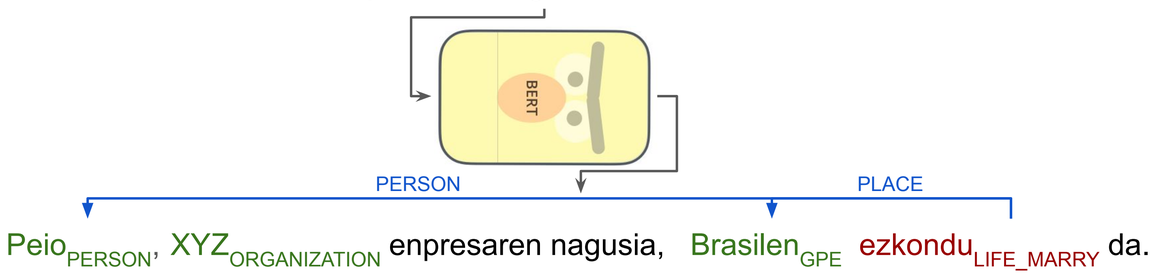
Mikel Zubillaga, Oscar Sainz, Ainara Estarrona, Oier Lopez de Lacalle, Eneko Agirre
HiTZ Hizkuntza Teknologiako Zentroa - Ixa Taldea
Euskal Herriko Unibertsitatea UPV/EHU

Testuingurua

John Smith, an executive at XYZ Co., died in Florida — **Entrenamendua**



Peio, XYZ enpresaren nagusia, Brasilen ezkondu da. — **Ebaluazioa**



- **Hizkuntzen arteko transferentzia** bidezko ikasketan **hizkuntza-eredu eleanitzak (HEE)** erabiltzen dira.
- HEE-ak hizkuntza jakin bateko datuekin entrenatzen dira, ondoren beste hizkuntza bateko datuetan erabiltzeko.
- Hizkuntzen arteko transferentzia oso erabilia da baliabide urriko hizkuntzetan, teknika honi esker baliabide ugariagoak dituzten hizkuntzatan dagoen ezagutza balibide urrietara transferitu daiteke eta. Adibidez, ingelesetik euskarara.
- Hala ere, **hizkuntza guztien arteko transferentzia ez da berdina**, zenbait hizkuntzen artean transferentzia hobea da beste batzuen artean baino.

Lan honetan:

- Artikulu honetan **hizkuntzen arteko transferentzian ezaugarri topologikoek duten eragina aztertu da**, euskara erabiliz proba hizkuntza nagusi bezala.
- Hau egin ahal izateko **EusIE izeneko datu-multzoa sortu da**, Euskarazko lehenengo Informazio Erauzketa datu-multzoa.

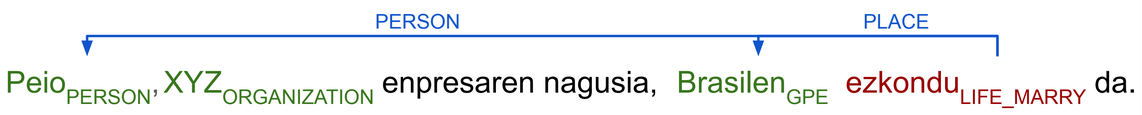
EusIE

Anotazioa

- EusIE anotatzeko aditu batek wikipediako esaldiak eskuz etiketatu ditu.
- 300 segmentu (1500 esaldi) anotatu dira: 150 garapenerako eta 150 ebaluaziorako.
- Kalitateari dagokionez, bigarren aditu batek 35 segmentu anotatu (%10a) ostean 0.92-ko adostasuna lortu da.

Entitateak: PER, ORG, GPE, LOC, FAC, VEH, WEA, CRIME, TIME, MON, POS, OBJ	
Gertaerak:	**Argumentuak:**
Life:Be-Born	Person, Time, Place
Life:Marry	Person, Time, Place
Life:Divorce	Person, Time, Place
Life:Injure	Agent, Victim, Instrument, Time, Place
Life:Die	Agent, Victim, Instrument, Time, Place
Movement:Transport	Agent, Artifact, Vehicle, Price, Origin, Destination, Time
Transaction:Transfer-Ownership	Buyer, Seller, Beneficiary, Price, Artifact, Time, Place
Transaction:Transfer-Money	Giver, Recipient, Beneficiary, Money, Time, Place
Business:Start-Organization	Agent, Organization, Time, Place
Conflict:Attack	Attacker, Target, Instrument, Time, Place
Conflict:Demonstrate	Entity, Time, Place
Contact:Meet	Entity, Time, Place
Contact:Phone-Write	Entity, Time
Personnel:Start-Position	Person, Entity, Position, Time, Place
Personnel:End-Position	Person, Entity, Position, Time, Place
Justice:Arrest-Jail	Person, Agent, Crime, Time, Place

Adibidea



Hizkuntza Tipologia

- Analisi tipologikoa egiteko **Euskara eta beste 8 hizkuntza** aztertu dira (72 hizkuntza pare guztira) **5 ezaugarri tipologiko**ren arabera sailkatuz.
- **Erregresio analisia** eginez, hizkuntzen arteko transferentzian ezaugarriek duten garrantzia neurtu da.
- **Hipotesiak**:
  - Antzeko hizkuntzek etekin handiagoa atera beharko lukete.
  - Ataza bakoitzak trebetasun desberdinak eskatzen dituztenez, ezaugarri desbedinetan oinarrituko dira transferentzia egiteko

Ezaugarri Taula

Hizkuntza	Morfologia	Morfosintaxia	Hitzen Ordena	Alfabetoa	Kokaleku Geografikoa
Euskara	Aglutinatzailea	Ergatibo-Absolutibo	SOV*	Latin	Europa Mendebaldea
Ingelesa	Fusionatzailea	Nominatibo-Akusatibo	SVO	Latin	Europa Mendebaldea
Gaztelera	Fusionatzailea	Nominatibo-Akusatibo	SVO	Latin	Europa Mendebaldea
Portugesa	Fusionatzailea	Nominatibo-Akusatibo	SVO	Latin	Europa Mendebaldea
Poloniera	Fusionatzailea	Nominatibo-Akusatibo	SVO	Latin	Europa Ekialdea
Turkiera	Aglutinatzailea	Nominatibo-Akusatibo	SOV	Latin	Europa Ekialdea/Asia Mendebaldea
Hindiera	Fusionatzailea	Ergatibo Banandua	SOV	Devanagari	India
Japoniera	Aglutinatzailea	Nominatibo-Akusatibo	SOV	Kanji eta Kana	Asia Ekialdea
Koreera	Aglutinatzailea	Nominatibo-Akusatibo	SOV	Hangul	Asia Ekialdea

Emaitzak

Domeinuko Emaitzak

Hizkuntza	Entitateak	Gertaerak	Argumentuak
Ingelesa	80.48 ± 0.19	78.47 ± 0.33	63.60 ± 0.09
Gaztelera	84.56 ± 0.38	63.86 ± 1.20	40.45 ± 1.91
Portugesa	80.42 ± 0.32	61.79 ± 0.63	68.18 ± 0.99
Poloniera	81.00 ± 0.40	69.09 ± 0.91	76.25 ± 1.26
Turkiera	70.83 ± 0.06	56.62 ± 0.43	24.08 ± 1.66
Hindiera	76.00 ± 0.41	48.21 ± 2.54	45.31 ± 1.55
Japoniera	47.43 ± 0.20	35.74 ± 1.92	52.93 ± 1.26
Koreera	71.31 ± 0.78	45.30 ± 0.49	34.71 ± 3.06
Denak	78.63 ± 0.17	68.01 ± 0.27	59.74 ± 0.99

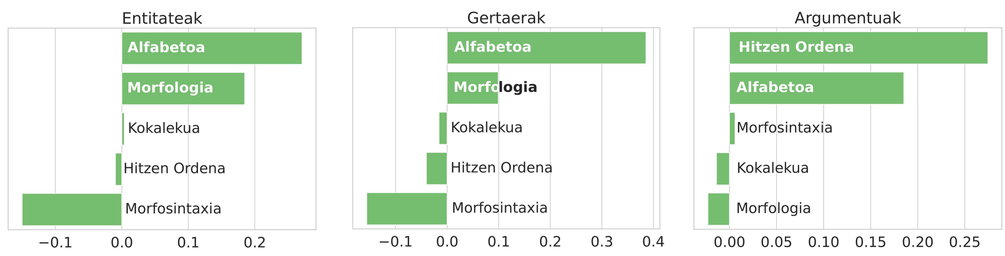
- Taula honek hizkuntza bakoitza bere ebaluazio datu-multzoarekin ebaluatzean lortutako emaitzak (F1) jasotzen ditu.
- "*Denak*" hizkuntza guztien datuekin entrenatutako eredu bat da.
- Emaitzak orokorrean altuak izan diren arren argumentuen kasuan nahiko baxuak dira, entrenatzeko datu gutxi baitzeuden ataza honetan.

EusIE-ko Emaitzak

Hizkuntza	Entitateak	Gertaerak	Argumentuak
Ingelesa	**59.65** ± 1.19	42.65 ± 6.57	13.93 ± 2.06
Gaztelera	56.56 ± 0.45	43.71 ± 2.63	2.88 ± 0.79
Portugesa	59.62 ± 1.67	24.30 ± 2.14	14.02 ± 0.96
Poloniera	59.48 ± 1.35	**46.37** ± 1.94	10.28 ± 0.29
Turkiera	55.72 ± 2.49	44.46 ± 2.73	14.84 ± 3.82
Hindiera	56.97 ± 1.44	34.70 ± 4.71	10.62 ± 0.70
Japoniera	47.17 ± 1.92	5.7 ± 4.03	10.96 ± 0.91
Koreera	46.67 ± 0.92	21.56 ± 7.62	**15.03** ± 0.81
Denak	56.92 ± 1.12	55.58 ± 0.80	28.05 ± 1.52

- Taula honetan hizkuntza guztiak EusIE-ko ebaluazioko datu-multzoarekin ebaluatzean lortutako emaitzak (F1) ageri dira.
- "*Denak*" hizkuntza guztiekin entrenatutako eredu bat da.
- Emaitzak erakusten dute **ez dagoela ataza guztietarako hoberena den hizkuntza bat.**

Analisi Tipologikoaren Emaitzak



- Bi ataza mota bereizi ditzakegu:
  - **Ataza lexikoak**: entitateak eta gertaerak.
  - **Ataza sintaktikoak**: argumentuak.
- Emaitzak ikusita ondorioztatu daiteke **ataza lexikoetan *alfabetoak* eta *morfologiak*** dutela eragin handiena.
- Bestalde, **ataza sintaktikoetan, *hitzen ordena*** bihurtzen da ezaugarri garrantzitsuena.