Ikerketa hizkuntza-ingeniaritzan

Zer laguntza eman dezake makinak, euskaraz argitaratzen den guztia eskura edukita?

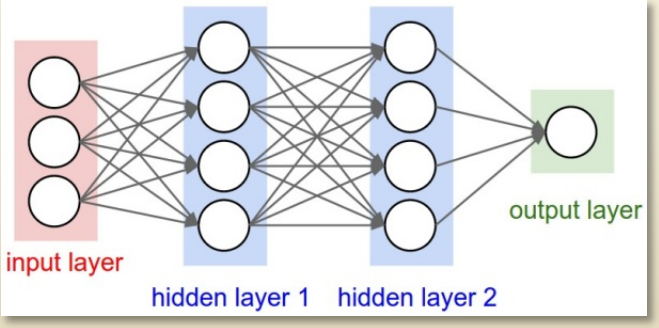
Arantza Díaz de Ilarraza Sánchez
Ixa Taldea
http://ixa.eus

Big Data eta hizkuntza-teknologiak

- "Datu Handien" garaian bizi gara.
- Egunero hainbat trilioi hitz ekoizten dira.
- Urtez urte gero eta gehiago.
- Euskaraz ere bai.

Big Data eta hizkuntza-teknologiak

- Teknika berriak daude informazio hori baliatzeko:
  - Hodeiko konputazioa
  - Ikasketa sakona (*deep learning*)
  - Neurona-sareak



Big Data, hizkuntza-teknologiak eta euskara

- Baina teknika berri horiek… baliagarriak al dira euskararako?
- … testu kopurua ingelesezkoa baino askoz txikiagoa izanik ere?

Edukiak

- Sarrera: Ixa Taldea
- Aplikazio-arloak
  - Testuen prozesaketa
  - Itzulpengintza automatikoa
  - Humanitate digitalak
  - Medikuntza
  - Hizkuntzen ikaskuntza

Ixa Taldea

2017ko ekaina

UPV/EHUko Ixa Taldea

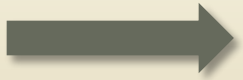
- Hizkuntzaren tratamendu automatikoan aritzen den ikerketa-taldea (50 pertsona baino gehiago).
- Duela 30 urte sortua.
- Informatikariak eta hizkuntzalariak elkarlanean.
- Hizkuntzak: euskara, ingelesa, gaztelania...
- Lankidetza: Elhuyar, Aholab, Tecnalia, Vicomtech, Langune, UEU, Iker (Baiona), UZEI... Microsoft, Google.
- Produktuak: zuzentzaileak, itzultzaileak, hiztegi elektronikoak, testu-corpusak...

Edukiak

- Sarrera: Ixa Taldea
- Aplikazio-arloak
  - Testuen prozesaketa
  - Itzulpengintza automatikoa
  - Humanitate digitalak
  - Medikuntza
  - Hizkuntzen ikaskuntza

Testuen prozesaketa

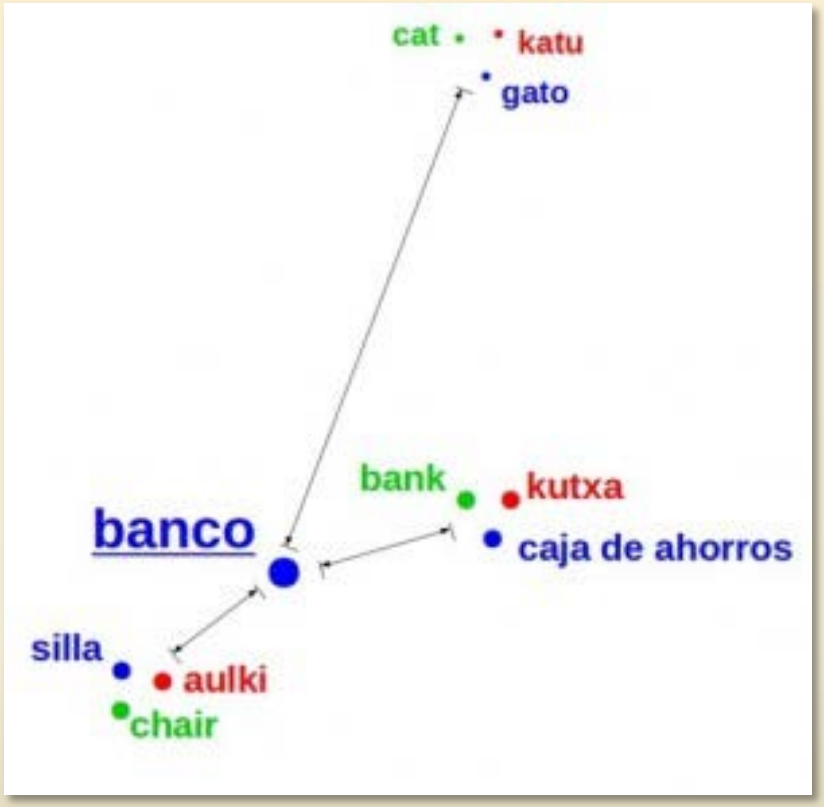
- Informazioa erauzi
  - Entitateak identifikatu: Carles Puigdemont, Mariano Rajoy, Madril, Katalunia… eta lotu baliabide eleaniztun eta zabalekin: Wikipedia eta bestelako baliabidekin.
  - Gertaerak (*events*), denbora-adierazpenak… Identifikatu.
- Semantikan oinarritutako bilaketa aurreratuak ahalbideratzeko



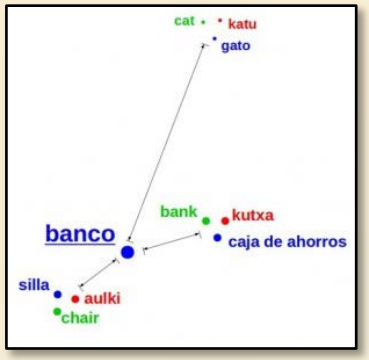
Informazioa errepresentatu behar

Testuen prozesaketa

- Hitzen esanahiak errepresentatu, beren arteko "distantziak" neurtu
- Zein dago gertu zeinetatik?



Testuen prozesaketa



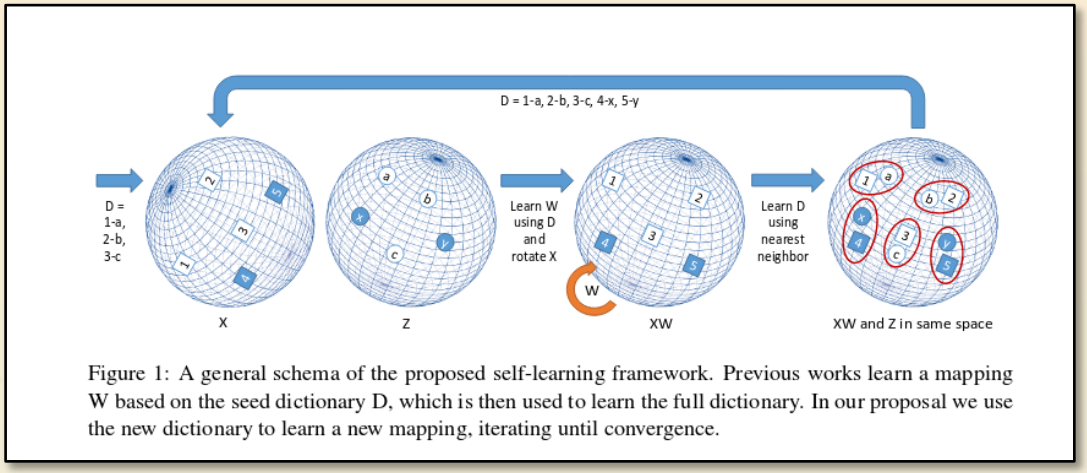
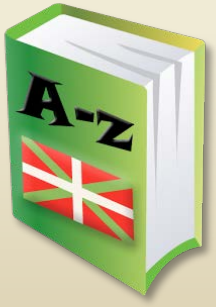


Figure 1: A general schema of the proposed self-learning framework. Previous works learn a mapping W based on the seed dictionary D, which is then used to learn the full dictionary. In our proposal we use the new dictionary to learn a new mapping, iterating until convergence.

Hitzen errepresentazio grafiko horrekin hiztegi elebidunak sor daitezke automatikoki (orokorrak edo espezializatuak).



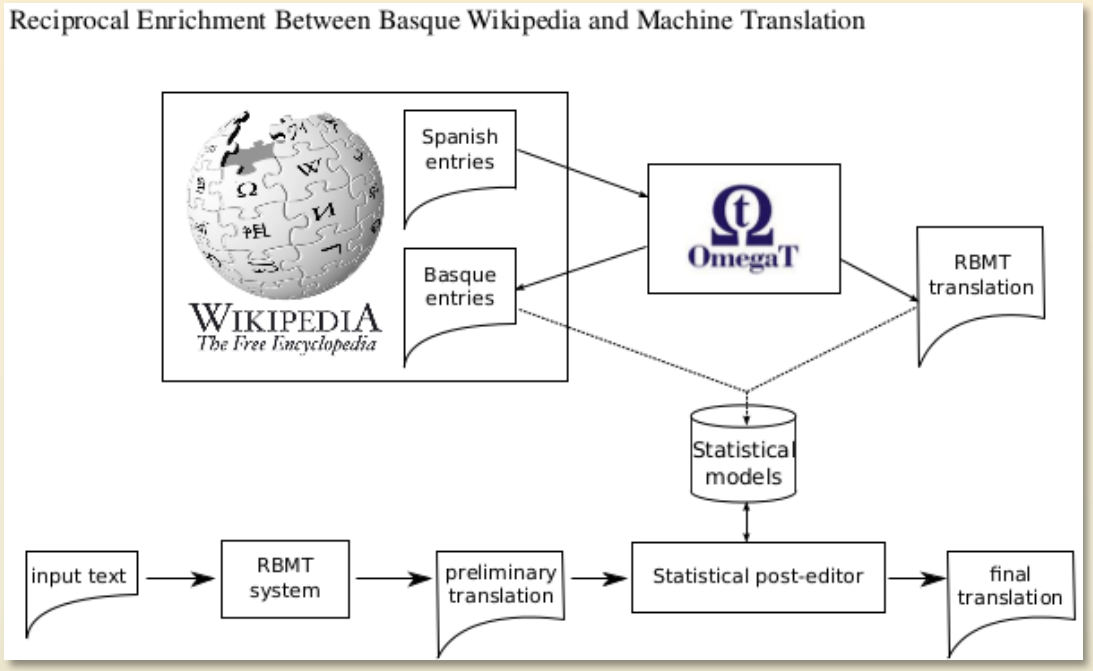
Edukiak

- Sarrera: Ixa Taldea
- Aplikazio-arloak
  - Testuen prozesaketa
  - Itzulpengintza automatikoa
  - Humanitate digitalak
  - Medikuntza
  - Hizkuntzen ikaskuntza

Itzulpengintza automatikoa

- Arlo interesgarria, ezagutzen ez ditugun hizkuntzetan idatzitako testuak ulertzen laguntzeko… Baina hainbat arazo daude:
  a) Lexiko-aukeraketa (desegokia batzuetan)
  b) Esaldiaren osagarrien ordena
  c) Kolokazioak, adierazpen idiomatikoak
  d) Gramatikaltasuna
  e) …
- Metodoak
  - Erregeletan oinarritutako hurbilpenetatik…
    - Datuetan oinarritutakoetara (corpus elebidun handiak)
    - Sare neuronaletara (ikasketa sakona)

Itzulpengintza automatikoa



- 2013: euskal Wikipediako 100 artikulu sortu ziren.
- % 10eko hobekuntza itzultzaile automatikoan.

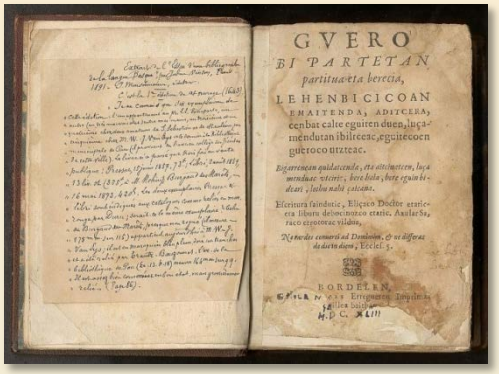
Edukiak

- Sarrera: Ixa Taldea
- Aplikazio-arloak
  - Testuen prozesaketa
  - Itzulpengintza automatikoa
  - Humanitate digitalak
  - Medikuntza
  - Hizkuntzen ikaskuntza

Humanitate digitalak

- Orain informatikariak eta hizkuntzalariak ari gara lankidetzan.
- Baina "ingurune digitalari" ekiteko, talde zabalagoak behar dira:
  - Historia
  - Kazetaritza
  - Soziologia
  - Psikologia
  - Zuzenbidea
  - ...
- Europan eta AEBn existitzen dira komunitate zabal horiek, lankidetzan ari direnak era naturalean.
- Estatu mailan, Ixak parte hartzen du Clarin azpiegitura-sarean (*Spanish Clarin Centre-K*).

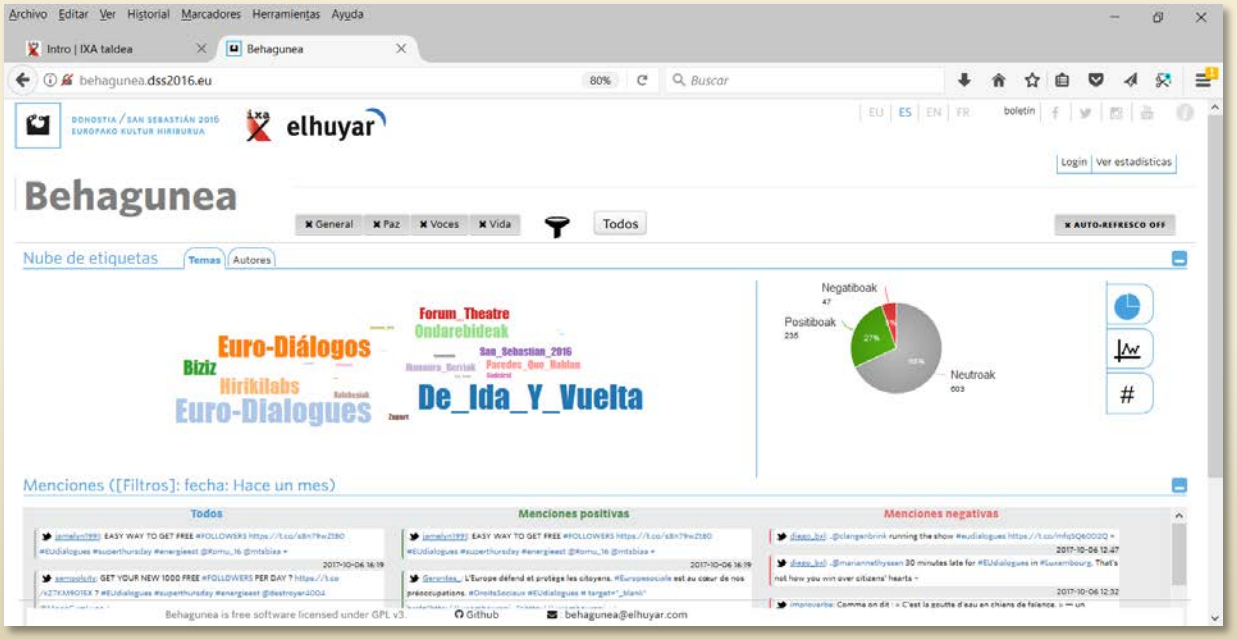
Humanitate digitalak: testu historikoak



ikhusiagatik

Testu historikoak, hizkuntza ez-normalizatua

- Sare sozialak
- Sentimentuen analisia: Behagunea proiektua

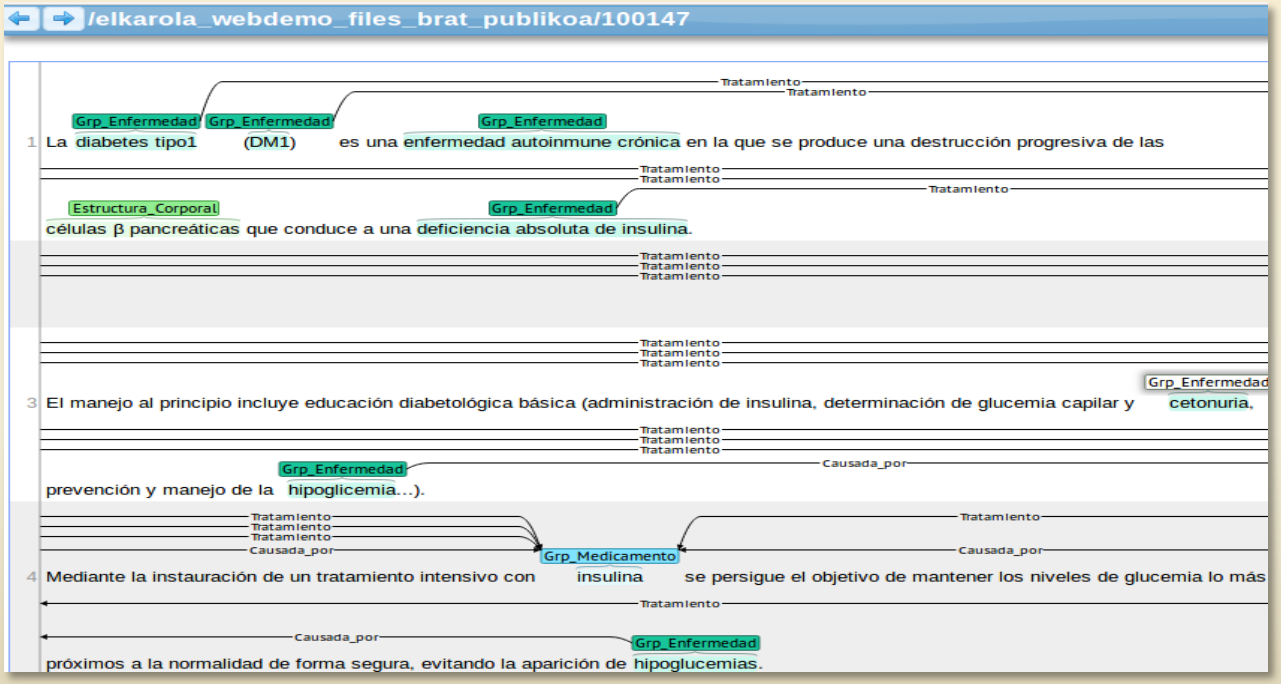


Edukiak

- Sarrera: Ixa Taldea
- Aplikazio-arloak
  - Testuen prozesaketa
  - Itzulpengintza automatikoa
  - Humanitate digitalak
  - Medikuntza
  - Hizkuntzen ikaskuntza

Medikuntza: testu medikoetatik informazioa erauzi

- Adibide bat: Botikek sortutako erreakzioak automatikoki identifikatzea osasun-txostenetan.



Medikuntza: testu medikoetatik informazioa erauzi

- Baina horretarako ….

- Osasun-txostenak euskaraz sortu.
- Terminologia finkatu behar da. Osasun-alorreko terminologia ez dago oraindik behar bezain landua.
- Ingelesezko 300.000 termino klinikotik gora ditu SNOMED CT datu-baseak; automatikoki euskaratua dago, eskuzko errebisioa falta da (Osakidetzan lantzen ari).

Edukiak

- Sarrera: Ixa Taldea
- Aplikazio-arloak
  - Testuen prozesaketa
  - Itzulpengintza automatikoa
  - Humanitate digitalak
  - Medikuntza
  - Hizkuntzen ikaskuntza

Hizkuntzen ikaskuntza

- Material didaktikoa eta ariketak sortzea, testu errealak erabilita.
- Idazlanen ebaluazioa: hiztegi-aberastasuna, espresioen erabilera, errore ortografiko eta sintaktikoak…
- Errore tipikoak identifikatzea. eta horiek zuzentzeko tresnak sortzea.
- Laburpenak egiten laguntzea.

Eskerrik asko!