**EN>EU CORPUSAREN ITZULPENA**

Ikerketa honen helburua Neural Machine Translation (NMT) sistema aztertzea da, baliabide gutxiago dituzten hizkuntzentzat, esaterako, euskara. Azkeneko urtetan *deep learning*-aren gorakadarekin batera, NMT sistema desberdinak aurkeztu dira eta oso emaitza onak eman dituzte. Hala ere, itzulpen zehatzak lortzeko entrenamendu handia behar dute sistema hauek, eta ondorioz, korpus handiak behar dituzte entrenamendurako (4-6 milioi esaldi itzulita). Hizkuntza “handiek”, hau da, sarean presentzia handiena dutenek eta ikerketarako jende asko dutenek oso bizkor sortu ditzakete horrelako korpusak. Baliabide gutxiago dituzten hizkuntzek, ordea, zailagoa dute hau eta, ondorioz, itzulpen txarrak lortzen dira. Adibide moduan, ***Google Translate*** daukagu. Googleek iada erabiltzen du bere GNMT (Google Neural Machine Translation) sistema, baina ingelesetik euskararako itzulpenetan oso emaitza pobreak ematen ditu:

Ingelesezko esaldia: “an admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out diagnosis or a procedure, based on his status as a health care worker at a hospital”

Euskarazko Googleren itzulpena: “onartu pribilegioa bat medikua eskubidea ospitale edo osasun zentro batera gaixo bat aitortu diagnostikoa edo prozedura bat, oinarritutako bere osasun-laguntza langile gisa egoeraren burutzeko ospitale batean dago”

Gure ikerketa, itzultzaile profesionalek itzulitako korpus **txiki** batekin (ingelesa-euskara 2.000 esaldi) (***supervised learning***) funtzionatuko duen NMT bat sortzea da. Korpus txiki horretaz aparte, eskuragarri dagoen gainbegiratu gabeko informazioa (***unsupervised learning***) ere erabili dezakegu. Honen inguruan daukagun ideia bat da Googlek oker itzuli dituen euskarazko esaldiak erabiltzea nolabait sisteman. Argi dago euskarazko esaldi horiek ez dutela inongo zentzurik bere osotasunean, baina onartu beharra dago hitz ugari oso ondo asmatzen dituela. Uste dugu euskarazko esaldi oker hauek ematen diguten informazioaz balia gaitezkela NMT sistemaren entrenamendua errazteko.

NOIZ

Gure ideia irailaren 25erako artikulua prest izatea da LREC 2018 konferentziara aurkezteko. Bertan, korpus hau eta gure lehenengo emaitzak aurkeztuko genituzke. Horregatik**, korpusa abuztu amaierako prest egotea izango litzateke ideala**.

KORPUSAREN SORRERA

Gure datu basea hiru fitxategitan banatuko dugu: *train\_en\_euGo.csv*, *dev\_en\_euGo.csv* eta *test\_en\_euGo.csv*. Hiru taulek estruktura berdina izango dute (adibidea Taula 1). Guztira ingelesezko 2.000 esaldi dira .

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **English** | **Basque** | **Correct Basque** |
| Europarl-v7 | I thank you. | Eskerrak eman dut. | ........... |
| Europarl-v7 | In particular: | Zehazki: | ........... |

Taula 1

**Source** -> Ingelesezko esaldiaren iturria

**English** -> Ingelesezko esaldia

**Basque** -> Googlek itzulitako euskarazko esaldia

**Correct Basque** -> Euskarazko itzulpen zuzena (BETE BEHARREKOA!)

Datu baseari azkeneko zutabea betetzea geratzen zaio (**Correct Basque**). Itzultzailearen lana da zuzen itzultzea esaldia.

ITZULTZEARAKOAN KONTUAN EDUKI BEHARREKOA

1. Ingelesezko esaldiaren euskarazko itzulpen egokia idatzi behar da **Correct Basque** zutabean.
2. Itzulpena idazterakoan, kontuan eduki Googlek itzuli duen esaldia.
   1. Googlek itzulitako esaldia egokia baldin bada itzulpena kopiatu
   2. Googleren itzulpena okerra bada esaldi berria idatzi, baina Googleen itzulpenarengandik ahalik ete hurbilen ibiltzen saiatu.
   3. Ez du axola ezin baldin badira Googlek hautatutako hitz asko erabili. Garrantzitsuena itzulpena zuzena izatea da.
3. Saiatu itzulpen eredu berdina mantentzen 2000 esaldietan zehar. Hau da, beti modu berean itzuli datubasean zehar errepikatzen diren ingelesezko esaldiak.
4. Errespetatu komak, puntuak, puntu eta komak, galdera ikurrak, harridura ikurrak...
5. Batzuetan esaldia ez da ingelesez egongo. Kasu horretan ez itzuli esaldia, ingeleseko zutabean dagoen esaldia kopiatu besterik gabe
6. Train\_en\_euGo.csv fitxategian ikusiko duzu esaldi batzuk dagoeneko zuzen itzulita daudela (**Correct Basque**) zutabean. Hobe itzuli daitezkela iruditzen bazaizu zuzendu itzazu.

Iñigo Jauregi (University of Technology Sydney)

Lierni Garmendia (Itzultzaile autonomoa)

Massimo Piccardi (University of Technology Sydney)