OpenMT-2: Wikipedia eta Itzulpen Automatikoa biak elkarri laguntzen

Iñaki Alegria (Ixa taldea)
Unai Cabezón (Ixa taldea)
Unai Fernandez de Betoño (eu.wikipedia)
Galder Gonzalez (eu.wikipedia)
Mikel Iturbe (eu.wikipedia)
Gorka Labaka (Ixa taldea)
Kepa Sarasola (Ixa taldea)
Arkaitz Zubiaga (eu.wikipedia)

Jakina da Euskal Wikipediaren tamaina askoz txikiagoa dela gure inguruko erdarena baino. Itzulpen automatikoa ikertzen duen OpenMT-2 proiektuaren[1] barruan informatikari buruzko Wikipediako 50 artikulu luze gehitzeko iniziatiba bat martxan jarri dugu. Matxin itzulpen-sistemak sortuko ditu lehen zirriborroak espainierako Wikipediatik itzulita, eta ondoren hainbat boluntarioren artean, eta eu.wikipedia elkarteak koordinatuta, zirriborro horiek zuzendu eta argitaratuko dituzte.

Esperientzia aberasgarria izango da bi norabideetan. Wikipediarentzat esperientzia onuragarria izango da 50 artikulu berri sortuko direlako, eta itzulpen automatikoarentzat ere bai eskuz posteditatutako itzulpenekin 100.000 hitzeko corpusa batuko delako. Corpus hori, itzulpen-sistema automatikoaren kalitatea hobetzeko funtsezko baliabide izango da, teknika estatistikoak erabiliz.

Lankidetza 2010ko hasieran hasi zen, orduan finkatu genituen proiektuaren nondik-norakoak. Geroago 2010 urtean zehar lau lan burutu ditugu:

•Aztertu ditugu antzeko proiektuak; adibidez: Kanadako talde batek (Simard et al., 2007) emaitza ederrak lortu zituen 100.000 hitzeko postedizio-corpusa eta teknika estadistikoak erabiliz. Ixa taldean ere, Gorka Labakaren tesi-lanean (Labaka, 2010) esperimentu batean egiaztatu dugu teknika horrek hobekuntza dakarrela espainiera-euskara sistema batean (Labaka, 2010).

•Itzuli nahi ditugun wikipediako sarreren zerrenda zehaztu dugu. Gaztelaniazko Wikipedian dauden eta euskaraz ez dauden hainbat artikulu ez-motz hautatu ditugu lehen urrats baterako. Wikiproiektu bat[2] abiatu da dagoeneko horretarako, bertan ikus daiteke sarrera aukeratuen lista.

•Interfaze bat egokitu dugu boluntarioek postedizioan eroso lan egiteko. Aztertu ditugu hiru aukera: (1) World Wide Lexicon Translator (WWL[3]), Firefox-erako gehigarri bat webguneak itzulita ikusi ahal izateko, konbinatzen du giza-itzulpena eta itzulpen automatikoa, baina posteditatzeko intefazea ez dabil oso ondo; (2) Google Translation Toolkit tresnak Wikipediako sarrerak itzultzeko laguntza eskaintzen du baina software libre eta irekia ez denez ezin izan dugu moldatu gure beharretara; (3) OmegaT[4] software libreko itzulpen memoriak erabiltzeko tresna bat. OmegaT software librea izanda moldatzeko aukera ematen digu eta horregatik aukeratu dugu.

•Informatikako gaietan hobeto itzultzeko Matxin_Inf_2010 bertsio berri bat prestatu dugu.

Ikerketa honen azken esperimentua 2011ko bukaeran bukatu nahi dugu, bi urratsetan:
1.Wikipediako 5 sarrera (~5.000 hitz) itzuli euskarara hiru modutan:

1 http://ixa.si.ehu.es/openmt2
2 http://eu.wikipedia.org/wiki/Wikiproiektu:OpenMT2_eta_Euskal_Wikipedia
3 https://addons.mozilla.org/en-US/firefox/addon/13897
4 http://www.omegat.org/

a.Itzulpen automatikoa erabili gabe.
b.Matxin_Inf_2010 bertsio erabilita
c.Matxin_Inf_2010 eta postedizioko corpusa erabilita.
1.Konparatzea hiru sistemen emaitzak, eta identifikatzea gero ea zein baldintzetan b eta c aukerak diren hoberenak.
Gure hipotesia da baietz, hobekuntza lortuko dugula Matxin itzulpen-sisteman esperimentu honen emaitzei esker. Baina esperimentuarekin egiaztatu beharko dugu hipotesi hori. Bitartean lortuko dugu, batetik, 50 sarrera luze gehiago sortzea euskal wikipedian, eta bestetik, Matxin sistemaren erabilgarritasunaren azterketa praktikoa, hainbat boluntarioren laguntzarekin. Emaitzak onak balira wikipediako beste alorretan ere aplikatu ahal izango da teknika hau.

Erreferentziak

1. ↑ Alegria I., Arregi X., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. 2008. Strategies for sustainable MT for Basque: incremental design, reusability, standardization and open-source. Proceedings of the IJCNLP-08, pp: 235-243. Hyderabad, India.
2. Labaka, G. EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. Doktorego-tesia. *Lengoaia eta Sistema Informatikoak Saila (UPV-EHU). Donostia. 2010* http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1271852575/publikoak/GorkaLabaka.Thesis.pdf
3. ↑ Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. 2007. Rule-based translation with statistical phrase-based post-editing Proceedings of the Second Workshop on Statistical Machine Translation. pp:203-206. Prague, Czech Republic.
4. ↑ Díaz de Ilarraza A., Labaka G., Sarasola K. 2008. Statistical Post-Editing: A Valuable Method in Domain Adaptation of RBMT Systems MATMT2008 workshop: Mixing Approaches to Machine Translation. pp.35-40.