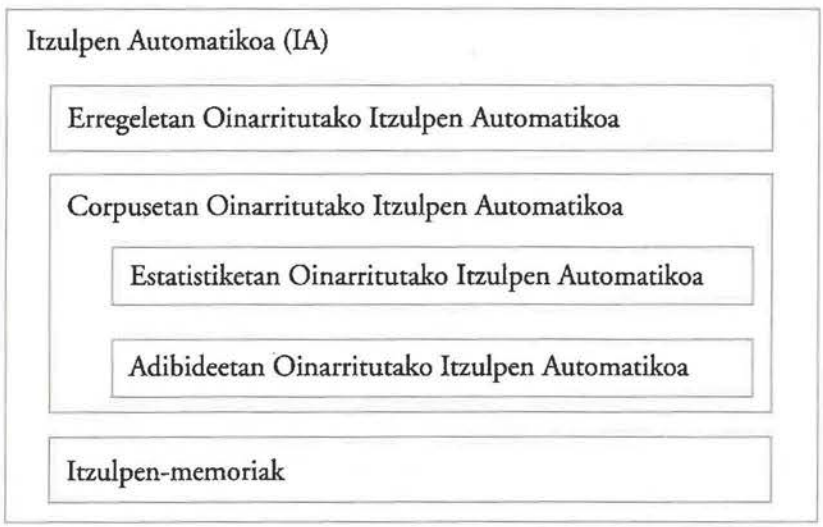
Adibideetan Oinarritutako Itzulpen Automatikoa (AOIA): azterketa bibliografikoa eta euskararekin lan egiteko proposamenak

Arantza Diaz de Ilarraza, Aingeru Mayor, Kepa Sarasola

1. Zer da Adibideetan Oinarritutako Itzulpen Automatikoa (AOIA)?

Itzulpen automatikoaren erronkari aurre egiteko oso estrategia ezberdinak daude. Horretan bi adar nagusi ikusten dira: erregela linguistikoetan oinarritzen direnak (EOIA) eta corpusetan oinarrituak. Azken hauetan bi hurbilpen oso ezberdin bereizten dira: estatistiketan oinarritutakoak eta adibideetan oinarritutako sistemak. Lan honetan azken hauei buruzko panorama orokor bat aurkeztuko dugu, inplementazio zehatz baten detaileak azalduko ditugu, eta euskararekin lantzeko zenbait ideia proposatuko ditugu.



Adibideetan Oinarritutako Itzulpen Automatikoa (AOIA, ingelesez *Example Based Machine Translation, EBMT*) burutzen duten sistemek ezaugarri hauek dituzte:

- Aldez aurretik itzulitako adibideen datu-basea edo corpusa erabiltzen dute.
- Sarrera berria adibideen datu-basearekin parekatzen da, egokiak diren adibideak erauzteko. Adibide hauek, ondoren, modu analogikoan bateratzen dira itzulpen zuzena erabakitzeko.

AOIA estrategia, askotan, Itzulpen Memoriekin (IM, ingelesez *Translation Memory TM*) lotu da. Eta bien ideia nagusia lehenago egindako itzulpenen adibideak berrerabiltzea bada ere, IM giza itzultzaileentzako tresna interaktibo bat da, AOIA itzulpen automatiko teknika bat den bitartean. Esaldi berri bat corpusarekin parekatzen denean, IM sistemek aurkitutako adibideak eskaintzen dizkio erabiltzaileari berak erabaki dezan haiekin zer egin, hauxe AOIAren hasiera besterik ez dela.

AOIAren ideia Makoto Nagaok aurkeztu zuen (Nagao, 1984), "adibideetan oinarritutako inferentziaren bidezko itzulpen automatikoa edo analogia printzipioaren bidezko IA" izendatu zuela:

> Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases..., then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference. (Nagao, 1984)

Bere esaldi horretan Nagaok AOIAren hiru osagai nagusiak identifikatzen ditu:

- esaldi zatiak benetako adibideetako datu-base batekin parekatzea,
- hauei dagozkien itzulpen zatiak identifikatu eta
- birkonbinatzea zati hauek itzulitako testua emateko

Ikus dezagun ondoko adibidea (Sato & Nagao, 1990), non (1)-en itzulpena lortzeko (2a) eta (2b)-tik zati egokiak hartzen dira, eta (3) izango da emaitza:

(1) He buys a book on international politics.
(Berak nazioarteko politikari buruzko liburu bat erosten du)

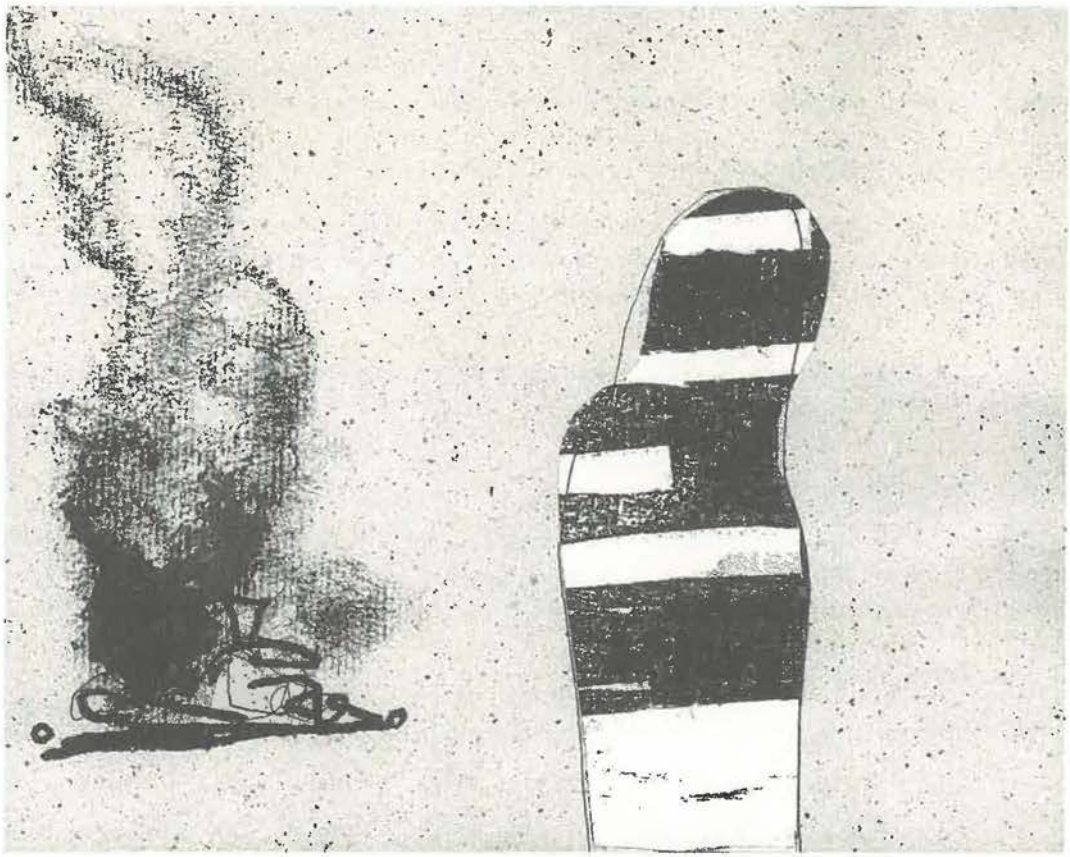
(2) 2a. **He buys** a notebook.
(Berak koaderno bat erosten du)
Kare wa noto **o kau.**

2b. I read a book on international politics.
(Nazioarteko politikari buruzko liburu bat irakurtzen dut)
Watashi wa kokusai seiji nitsuite kakareta hon o yomu.

(3) **Kare wa** kokusai seiji nitsuite kakareta hon **o kau.**

Ixa taldean lehenengo saioak egiten ari gara itzulpen automatikoaren arloan. Orain arteko lanak izen-sintagma mailan bakarrik egin ditugu (Diaz de Ilarraza et al., 2001). Gure hurrengo urratsak diseinatzerakoan oso begi onez ikusten dugu Adibideetan Oinarritutako Itzulpen Automatikoaren bidea. Artikulu honetako hasieran arlo honetan guretzat interesgarri izan diren ideiak zabaldu nahi ditugu, eta gero, bukaeran, ideia horiek euskararen kasuan aplikatu ahal izateko guri bururatu zaigun hasierako proposamena aurkeztu nahi dugu.

AOIAren deskribapen zehatza aurki daiteke Harold Somers-ek 1999an idatzi zuen «*Review article: Example-based Machine Translation*» artikuluan. Artikulu hori izan da (oraindik argitaratu gabe dagoen bertsio eguneratu bat) gure aurkezpen honen iturri nagusia. Gai honetan sakontzeko interesik duenak bertara jo beharko luke.



Ildo horretatik, ondoko atalak bereizi ditugu AOIAren inguruko azalpen honetan. Hasieran funtsezko *arazoak* aipatuko dira. Gero bi galdera erantzuten jardungo dugu: AOIA eta erregeletan oinarritutako itzulpen automatikoa elkarren *kontra edo elkarri laguntzen?* Beraien arteko *hibridazioa* nola burutu? Ondoren AOIArako *testuinguru egokiak* identifikatzen saiatuko gara, AOIAren *abantailak* zeintzuk diren azaldu aurretik. AOIAn oinarritzen den *PanEBMT sistema*ren deskribapena egingo dugu orduan, eta bukatzeko, *euskararekin* lan egiteko gure hasierako *proposamenak* azalduko ditugu.

2. AOIAren funtsezko arazoak

Atal honetan itzulpen automatikoa egiteko adibideetan oinarritutako hurbilpenen oinarrizko arazo orokorrak azalduko ditugu

2.1. Corpus elebiduna

AOIAn behar den lehenengo gauza corpus elebidun lerrokatua da; hau da, testu bat bere itzulpenarekin, non bi testuak analizatuak izan diren bi hizkuntzetako segmentu berdinak erlazionatzeko.

Ikertzaileak sor dezake bere corpus lerrokatu propioa, edo bestela publikoa den bat aurkitu, orduan lerrokatzearen arazoari aurre egin beharko zaiola.

2.2. Adibideen "granularity"

Inplementazio gehienetan adibideen "ale-tamaina" esaldia da, batez ere abantaila praktikoak eskaintzen dituelako (esaldien mugak finkatzea erreza da, arlo askotan esaldiak sinpleak dira...). Baina badirudi esaldia handiegia dela erabilera praktikoetarako, eta parekatze eta birkonbinazio prozesuek zati txikiagoek behar dituztela:

> The potential of EBMT lies in the exploitation of fragments of text smaller than sentences. (Cranias et al., 1994:100).

2.3. Adibide kopurua

Zenbat eta adibide gehiago egon datu-basean, emaitza hobeagoak lortzen dira orokorrean (Sumita & Iida, 1991). Baina badirudi hobekuntza nahiko lineala bada ere, muga bat badagoela, eta zenbait kasutan ere prestazioak txarrera egiten dutela adibideak gehitzean.

2.4. Adibideen arteko interferentziak

Lerrokatutako corpus elebidun batek adibideen arteko bi interferentzia mota eduki ditzake:

- Adibide berdinek edo berdintsuek antzeko itzulpena dutenean, elkar indartzen dute.

Frekuentziaren informazioa erabili daiteke parekatze-fasean. Ez bada erabiltzen, adibide horiek zaborra izango dira, baina ez dute arazorik emango.

- Adibide berdinek itzulpen desberdinak dituztenean gatazka sortzen da. Gatazka saihesteko saia daiteke itzulpen okerrak ezabatzen. Beste hurbilpen bat litzateke adibide orokorrak eta "ezohikoak" bereiztea.

2.5. Adibideak gordetzeko modua

Kasurik sinpleenean adibideak *string* bikoteak bezala gordetzen dira, gainerako informaziorik gabe. Zenbait hurbilpenetan, berriz, adibideak etiketatutako zuhaitz egitura gisa gordetzen dira, eta beste hainbatetan adierazpide orokortu baten bidez:

- AOIAn egindako hasierako hurbilpenetan adibideak guztiz etiketatutako zuhaitz egitura gisa gordetzen ziren lotura esplizituekin (Sato & Nagao, 1990). Ondoren eraiki diren hainbat sistematan adibideak zuhaitz egitura bezala azaltzen dira, indizeen bidez egitura eta lexiko mailan adierazitako loturekin. Badira informazio tipografiko eta ortografikoa sartzen dutenak, eta baita *speech act* eta rol semantikoei buruzkoak ere. Sistema hauek kostu konputazional handia dute sorkuntza, gordetze, parekatze eta berreskuratze algoritmotan.
- Zenbait sistematan antzeko adibideak konbinatzen dira eta adibide "orokortu" bakar gisa gordetzen dira. Brown-ek (1999) adibideak *tokenizatzen* ditu baliokidetasun-klaseak (pertsona izenak, herrien izenak, datak...) erakusteko, adibideetako zatien ordez klase horiek jarriz.

2.6. Parekatzea (ingelesez "matching")

AOIAko sistema baten lehenengo eginkizuna hauxe da: itzuli beharreko iturburu-hizkuntzako testua hartu eta hobekien parekatzen den adibidea (edo adibide multzoa) aurkitzea. Hauxe da Itzulpen Memoriako sistemen funtsezko eginbeharra. Parekatze prozesu guztietan distantzia edo antzekotasun neurri bat erabili beharko da.

Adibideak gordeta dauden moduaren arabera egin behar zaio aurre bilaketarako estrategiari:

1. **Karakteretan** oinarritutako parekatzea:
   Kasu sinpleena da eta ohiko patroi-ezkontzea (ingelesez *pattern matching*) erabiliko du.
2. **Hitzetan** oinarritutako parekatzea:
   Kasu honetan parekatze prozesuak posible egiten du sarrera testuetako hitzak sinonimo hurbilekin ordeztuak izatea adibide-basean bilatutako esaldietan. Honetarako distantzia neurri klasikoa funtsezkoa da: erabilera edo adieran oinarritutako thesaurus bat erabiltzen da hitzen antzekotasuna identifikatzeko (Nagao, 1984).

3. **Etiketatutako hitzetan** oinarritutako parekatzea:
   Hitzen funtzioa kontuan hartzen duen neurria erabiltzen da, kategoria sintaktikoaren etiketak erabiliz (Cranias et al., 1994).
   Pangloss sisteman parekatze-prozesuak bere baldintzak banan-banan (Nirenburg et al., 1994): hasieran parekatze zehatzak bilatzen ditu, ondoren hitzak kentzea edo jartzea onartzen du, gero aldaketak hitzen ordenan, aldaketa morfologikoak eta azkenik kategoria sintaktikoaren etiketetan diferentziak. Erlaxazio-fase bakoitzean zigor-neurri handiagoak ezarriz.
4. **Egituran** oinarritutako parekatzea:
   AOIA itzulpen automatikoko hurbilpen tradizional batean integratuta dagoen zenbait proposamenetan, adibideak egituradun objektu gisa gordetzen dira, eta, beraz, zuhaitz-parekatze konplexuagoa burutu behar da, kostu konputazionala dezente handituz.
5. **Parekatze partziala:**
   Parekatze funtzioak kasuak deskonposatzen ditu eta parekatutako zatien bilduma bat egiten du, estaldura handituz.

2.7. Moldagarritasun eta birkonbinatzea

Behin adibide multzo bat bere itzulpenekin parekatu eta eskuratu dela, AOIAren prozesuko urratsik zailenak datoz:

- Aurkitu diren antzeko testu itzulien zatiak iturburu-testuko zein zatiri dagozkien identifikatzea ("lerrokatzea" edo "moldatzea").
- Zati hauek modu egokian birkonbinatzea gramatikala den irteera bat emateko.

Adibideen berreskuratze prozesuak sistema gehienetan adibidearen barne egituraren moldagarritasuna soilik kontuan hartzen du, baina badago bere kanpoko testuinguruan ere kontuan hartzen duena. ReVerb (Collins & Cunningham, 1996) sistemako adibideak bere funtzio-etiketarekin gordetzen dira, beren baliokide lexikal zein funtzionalei lotuta; adibideak berreskuratzeko unean bi ezaugarri hartzen dira kontuan:

- sarrera-testua eta adibidearen arteko parekatzearen hurbiltasuna, eta
- adibidearen moldagarritasuna; adibidearen eta bere itzulpenaren errepresentazioak aztertzen dira neurri hau definitzeko.

3. AOIA vs. EOIA? Hibridazioa

AOIAren ideia zabaltzen joan zenean, beste metodoekin lehian arituko zen paradigma berri bat izan zitekeela pentsatzen hasi zen. Baina esperientziak kontrakoa esan du: AOIA hurbilpen tradizionalagoetan integratua izan da (eta *viceversa*) modu oso ezberdinetan, gehienetan erre-

geletan oinarritutako itzulpen automatikoa (EOIA, ingelesez *Rule Based Machine Translation RBMT*) burutzen duten sistemen lagungarri gisa erabili delarik (Hutchins et Somers, 1992).

Sumita et Iidak (1991) azaltzen dute zein kasutan AOIA den EOIA baino egokiagoa:

- Itzulpen erregelen osaketa zaila denean.
- Erregela orokorrak fenomeno bat zehazki deskribatu ezin duenean, kasu berezia adierazten duelako (adib. esaerak).
- Itzulpena ezin da burutu konposizio moduan xede hitzetatik.

Zalantzan jarri izan da AOIAk itzulpen prozesu osoa burutu zezakeenik, sistema hibridoek ikertzaile gehienen txaloa jaso dutela, ondoko paragrafoetan ikus dezakegun bezala:

> It is not yet clear whether EBMT can or should deal with the whole process of translation. We assume there are many kinds of phenomena. Some are suitable for EBMT, while others are suitable for RBMT. Integrating EBMT with RBMT is expected to be useful. It would be more acceptable for users if RBMT were first introduced as a base system, and then incrementally have us translation performance improved by attaching EBMT components. (Sumita et Iida, 1991)

> (...) This marks the advent of hybrid rule-based and example-based MT systems. The hybridization route is chosen in the hope that the resulting systems will have fewer practical shortcomings than the pure rule-based systems (a high complexity of processing plus a high price of knowledge acquisition) or the pure EBMT systems (a very ungraceful degradation curve when matches are bad). (Nirenburg et al., 1996)

> Current thinking in EBMT circles seems to be that a hybrid of EBMT and traditional rule-based MT is appropriate. (Somers, 1997)

> EBMT is certainly here to stay, not as a rival to rule–based methods but as an alternative, available to enhance and, sometimes, replace it. (Somers, 2001)

4. AOIArako testuinguruak

Oso ahalegin gutxi egin dira AOIA hutsa erabiliko luketen sistemen ikerkuntzan, hau da, corpusetik automatikoki jasotzen den informazio linguistikoez gain besterik ez erabiltzean, ikusi baita muga handia dela, eta askoz irtenbide hobea dela sistema hibridoena.

Ondoko puntuetan AOIA erabili izan den testuinguru ezberdinak aurkeztuko ditugu.

4.8. AOIA kasu berezietarako

AOIAren lehenengo erabilpenetako bat izan zen, erregeletan oinarritutako hurbilpena zaila suertatzen zenean. Erabilera honetako kasu klasikoa japonierazko izen sintagma konplexuen itzulpena da. Japonieraz "N1 no N2" motako izen sintagmak normalean ingelesezko "N2 of N1" motakoei dagozkie, baina salbuespen asko dago hurrengo adibideetan ikus dezakegun bezala:

youka **no** gogo	the afternoon **of** the 8th
kaigi **no** mokuteki	the subject **of** the conference
kaigi **no** sankaryou	the application fee **for** the conference
kyouto–de **no** kaigi	a conference **in** Kyoto
kyouto–e **no** densha	the Kyoto train
isshukan **no** kyuka	one week's holiday
mittsu **no** hoteru	three hotels

Adibidez, erregeletan oinarritutako ATR (Sumita et Iida, 1991) sistema tradizionalean AOIA modulua deitzen da mota honetako eta antzeko kasu konplexu batzuetan.

4.9. Adibideetan oinarritutako transferentzia

Hainbat sistemetan iturburu-hizkuntzako sarrerak analizatzen dira modu tradizionalean, transferentzia modulua adibideetan oinarritzen da, erregeletan baino, eta ondoren xede-hizkuntzako irteera berriz modu arruntean burutzen da.

4.10. Adibideetatik transferentzia erregelak ateratzen

Eszenatoki honetan AOIA itzulpen teknika baino gehiago, erregela basea eraikitzeko ikerketa teknika bezala erabiltzen da. Corpusetik transferentzia erregelak izango diren patroiak "ikasten" dira, adibideak orokortuz kategoria sintaktiko eta informazio semantikoaren bidez. Patroi bakoitza aldagaiak izango dituzten pseudo-esaldien bikote elebiduna izango da.

Hurrengo adibideetan ikus dezakegu antzeko adibideak konbinatuz erregela orokorragoak inferitu daitezkela:

I took a ticket from Mary	↔	Mary'den bir bilet aldim
I took a pen from Mary	↔	Mary'den bir kalem aldim
I took a . . . from Mary	↔	**Mary'den bir . . . aldim**

The Commission gave the plan up	↔	La Comisión abandonó el plan
Our Government gave all laws up	↔	Nuestro Gobierno abandonó todas las leyes
. . . gave . . . up	↔	**. . . abandonó . . .**

Turkiera edo euskara bezalako hizkuntza aglutinatiboetan, azaleko forma soilik kontuan hartzen bada, horrelako orokortze asko galdu daitezke. Horren aurrean analisi morfologikoa beharrezkoa da, ondoko adibideetan dakusagun moduan:

He ido a casa	↔	Etxera joan naiz Etxe+[ALA] joan naiz
He ido a París	↔	Parisera joan naiz Paris+[ALA] joan naiz
He ido a . . .	↔	**. . .+[ALA] joan naiz**

4.11. AOIA beste estrategiekin paraleloan

Mota honetakoa dugu Pangloss sistema (Nirenburg et al., 1994; Brown, 1996), non AOIA paraleloan egiten du lan beste bi teknikekin: ezagutzan oinarritutako IA eta erregeletan oinarritutakoa. Sistema honen deskribapena 6. puntuan azalduko dugu sakonean.

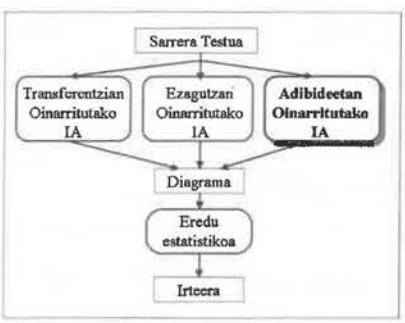
5. AOIAren abantailak

Metodo hau erabiltzen hasi zenean aldarrikatzen ziren abantaila guztiak benetakoak ez badira ere, badirudi ondokoak konprobatu direla (Somers, 2001):

- Adibideak benetako hizkuntzakoak dira, beraz hizkuntza errealean erabiltzen diren egiturak estaltzen ditu, erabiltzen ez direnak baztertu egiten direla.
- Ezagutza linguistikoa errazago aberats daiteke, adibide gehiago gehitu baino ez da behar.
- AOIA "datuek gidatua" da eta ez "teoriak gidatua", beraz, gramatika konplexurik ez dagoenez, gutxiago dira erregelen arteko konfliktoak eta interakzioak, eta ez da teoriaren ikuspegi orokorra eduki behar.
- Posible izan daiteke hizkuntza bikote berri baterako AOIA sistema bat azkar garatzea, lerrokatutako corpus elebidun bat oinarritzat hartuz. Hau oso interesgarria izan daiteke analizatzailea edo hiztegiak bezalako baliabiderik ez duen hizkuntza baterako.

6. AOIA inplementazio bat: PanEBMT

Atal honetan Pangloss sistemaren barruan AOIA burutzen duen PanEBMT motorra (Nirenburg et al., 1994; Brown, 1996, 1999, 2000) deskribatuko dugu. Pangloss barruan ere transferentzia-motorea (hiztegiak eta glosarioak) eta ezagutzan oinarritutako itzulpen automatikoaren motorea daude. Hauetako bakoitzak sarrerako zati ezberdinentzat itzulpen posibleak ematen ditu, diagrama (*chart*) batean konbinatzen direlarik. Ondoren eredu estatistiko batek erabakitzen du zein den diagrama horretako biderik hoberena eta hori izango da sarrera esaldiaren itzulpena adieraziko duen irteera. Sistemaren arkitektura ondoko irudian ikus dezakegu:



AOIA motorrak bi fasetan egiten du lan:

- Hizkuntza barruko parekatzea: corpuseko iturburu-hizkuntzako aldean puskak bilatzen ditu, indizeak erabiliz sarrera-testuko hitz kontsekutiboen agerpenak detektatzeko (intra-language matching).
- Hizkuntzen arteko parekatzea: corpuseko iturburu-hizkuntzako zatiari dagokion xede-hizkuntzako zatia aurkitzen du, hau da esaldi barruan lerrokatzen du (inter-language matching).

Corpusean agertzen den esaldiko hitz sekuentzia guztietarako itzulpena sortzen saiatzen da eta horrela «optimal cover»ean ez egoteagatik puska posibleak baztertzea saihesten da. Testu osorako estaldura emateko zatiak konbinatzea eredu estatistikorako uzten da, Panglossen dauden beste itzulpen motoreentzat bezala.

Ondoko puntuetan AOIA motorearen ezagutza-iturria, bere jarduerako bi faseak eta adibideak orokortzeko egindako esperimentuak azalduko ditugu.

6.1. PanEBMT-ren ezagutza-iturriak

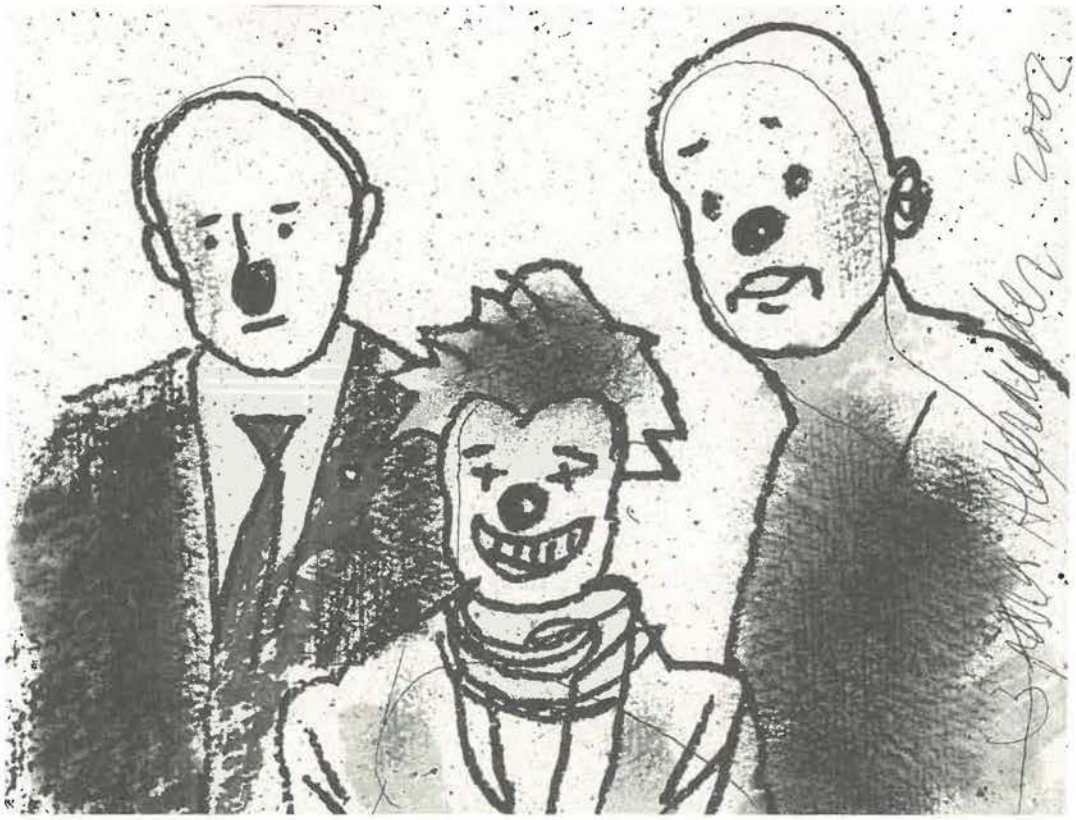
Lau ezagutza-iturri erabiltzen ditu:

- Esaldika lerrokatutako corpus elebidun paraleloa
- Hiztegi elebiduna
- Xede-hizkuntzako erro/sinonimoen zerrenda
- Hizkuntza zehatzaren informazioa (hautazkoa): hitzekin osatutako baliokidetasun-klaseak bere hitzekin (hilabeteak, asteko egunak, zenbakiak...), lerrokatzean eliditu daitezkeen hitzen zerrenda (artikuluak...), eta sartu behar diren hitzen zerrenda.

Erabilitako corpusa Nazio Batuen Erakundeko dokumentu elebidunetatik (gaztelania-ingelesa) hartua da. Corpusa lerrokatzeko algoritmo simetriko oso sinplea erabiltzen dute: testua paragrafoz paragrafo hartzen da, bi hizkuntzetan esaldi kopuru ezberdinak dituzten paragrafoak baztertzen dira (%5), eta besteetan esaldiz esaldi ordenan lerrokatzen da.

Corpusa iturburu/xede esaldi bikote multzoa da eta guztiz indexatuta dago iturburu-hizkuntzako esaldietan. Indizeak corpusean dauden iturburu-hizkuntzako esaldietako hitz eta puntuazio marka guztien agerpenak zerrendatzen ditu. Indexatu baino lehen *tokenizatzen* da testua, eta horrela baliokidetasun-klaseetan dauden hitzak klasearekin indexatzen dira. *Token* ezberdin bakoitzerako (hitz, puntuazio-marka edo baliokidetasun-klase), indizeak *token*aren agerpenen zerrenda bat gordetzen du, non agerpen bakoitzerako esaldiaren identifikadore eta esaldian okupatzen duen lekua adierazten baitira.

Hiztegi elebiduna eta xede-hizkuntzako erro/sinonimoen zerrendak hizkuntzen arteko esaldi barruko lerrokatzea burutzeko erabiliko dira.



6.2. Hizkuntza barruko parekatzea

Urrats honetako sarrera itzuli nahi den gaztelaniazko esaldia eta corpus elebidunaren gaztelaniazko aldea dira. Irteerak honelako formatua du:

```
((input-substring-1
   ((corpus-string-1-1 score-1-1)
   (corpus-string-1-2 score-1-2)
   ...
   (corpus-string-1-5 score-1-5)))
(input-substring-2
...
(input-substring-10
   ((corpus-string-10-1 score-10-1)
   ...
   (corpus-string-10-5 score-10-5)))
```

Esaldiko hamar zati (input-substring-i) "hoberenak" ematen dira, non ez dagoen zati bat beste bat guztiz barne duenik. Hitz bakoitza (corpus-string-i) hiru indizerekin adierazten da: (fitxategi-indizea-i, esaldi-indizea-j, hitz-indizea-k)

Puskak bilatzeko prozedura ondokoa da:

- Esaldia puntuazio-marketan edo hitz ezezagunetan puskatzen da.
- Sarrerako hitz bakoitza sekuentzialki corpuseko indizean begiratzen da. Hitz bakoitzeko agerpenen zerrenda aurreko hitzaren agerpenen zerrendarekin eta aurreko hitza sartu den pusken zerrendarekin konparatzen da. Aurreko hitzaren agerpen baten albokoa den agerpen bakoitzerako puska berri bat sortzen da, edo dagoeneko sortua den puska bat hedatzen da.

Prozesuaren bukaeran corpusean dauden sarrerako bi hitz edo gehiagoko puska guztiak aurkitu dira. Maiztasun handiko hitz sekuentziak corpusean ehunka aldiz agertu daitekeenez, hitz sekuentzia bakoitzerako soilik azken bost agerpenak gordetzen dira. Horrela, azken agerpenak aukeratzean, corpusean gehitu diren esaldi berrienak ematen dira.

Sarrerako zati bakoitzarekin parekatu daitezkeen puskak bilatzean parekatzearen definizio "erlaxatu" bat erabiltzen da, non parekatze osoaz gain ondoko parekatzeak ere onartzen diren:

- hutsuneak daude (adib. *A X Y B Z C* parekatuko du *A B C* sarrera puska)
- hitzen ordena ezberdina da (adib. *B C A* parekatu dezake *A B C*)
- sarrera puskaren hitzen azpimultzo bat parekatzen da (adib. *A D C* parekatu dezake *A B C*)
- aldaketa morfologikoa daude (*gatos* parekatuko du *gato*)

Aurkitzen diren puska hautagaiak filtratzen dira atalase bat gainditzen duten puntuazioa jaso dutenak soilik gordetzeko. Hasieran parekatze puntuazioak banan-banan kalkulatzen dira goian azaldu dugun osatu gabeko parekatze bakoitzerako. Ondoren metatze-puntuazioa ematen da eta sistema honen atalasea 10 parekatze onenetan ezartzen da.

Puntuazio prozesua ondoko heuristikoek gidatzen dute:

- Parekatu gabeko hitzak: lehentasuna sarrera puskako hitz guztiak dituen esaldiei ematen zaie. Parekatu gabeko hitz bakoitzarengatik isuna 10ekoa da.
- Zarata: lehentasuna soberako hitz gutxien dituen esaldiei ematen zaie. Corpuseko esaldian dagoen soberako hitz bakoitzarengatik isuna 5ekoa da.
- Ordena: lehentasuna sarrera puskako hitzen ordenari hurbilen dauden esaldiei ematen zaie. Ordenan ez dagoen hitz bakoitzarengatik isuna 5ekoa da.
- Morfologia: lehentasuna hitzak berdin parekatzen dituen esaldiei ematen zaie, aldaketa morfologikoak dituztenei baino gehiago. Ezberdintasun morfologikoak dituen eduki-hitz bakoitzarengatik isuna 2koa da, eta maiz agertzen den funtzio-hitza bada, 1ekoa.

6.3. Hizkuntzen arteko parekatzea

Esaldi osoa parekatu bada itzulpen osoa jasoko da (Itzulpen Memoria modua) eta bestela esaldi barruko lerrokatzea burutuko da.

Fase honen sarrera hauxe da:

SP	Sarrera Puska
CIHE	Corpuseko Iturburu-Hizkuntzako Esaldia
CP	Corpus Puska (SPrekin parekatzen den CIHEko puska),
CHHE	Corpuseko Xede-Hizkuntzako Esaldia
PP	(Hizkuntza barruko) Parekatzearen Puntuazioa

adibidez:

SP	representa la nación
CIHE	las naciones unidas representan a todas las naciones
CP	representan a todas las naciones
CHHE	the united nations represent every country
PP	9'6

Hizkuntzen arteko parekatze fase honen helburua CHHEtik CPren itzulpena den zatia jasotzea da. Goiko adibiderako emaitzak *represent every country* izan beharko luke.

Hitzen arteko kidetasunak aurkitzeko erabiltzen den metodoak ondoko informazio-iturriak baino ez ditu erabiltzen:

- hiztegi elebiduna eta xede-hizkuntzako erro/sinonimoen zerrendaren bertsio elektronikoak (MRD)
- analizatzaile morfologikoa iturburu- eta xede-hizkuntzentzat

Hauek dira prozesuaren urrats nagusiak:

- CIHEko hitzen itzulpenak jaso hiztegitik
- CHHEko hitzak lematizatu
- CIHE eta CHHEen arteko hizkuntzen artean hitz mailako kidetasunak aurkitu hiztegiak erabiliaz
- CPren itzulpena izan daitekeen CHHEko puskarik luzeena aurkitu
- CHHEko puskarik luzeeneko puskarik hoberena aurkitzea, puntuazio metriko berezi bat erabiliaz non ondokoak neurtzen diren: xedean kiderik gabeko iturburu-hitzen kopurua, iturburuan kiderik gabeko xede-hitzak, eta xede-testuen arteko luzera diferentzia,...

6.4. Adibideak orokortzen

AOIAren arazorik handiena aurre-itzulitako adibideen beharra da (milioika hitz). Adibideak efektiboagoak izateko orokortzea izan daiteke bide bat, horretarako informazio linguistiko pixka batez hornitu behar dela. Demagun ondoko itzulpen adibidea (ingelesa-alemanera) daukagula:

John Hancock was in Philadelphia on July 4th.

John Hancock war am 4. Juli in Philadelphia.

Jakingo bagenu *John Hancock* pertsona bat dela, *Philadelphia* hiri bat, eta *July 4th* data, ondoko adibide bikotea gorde genezake gure datu-basean:

<PERSON> was in <CITY> on <DATE>.

<PERSON> war am <DATE> in <CITY>.

Sistemak honetarako sarrera bereziak dituen ezagutza base bat dauka, non baliokidetasun-klaseak definitzen diren: pertsona izenak, hiriak, zenbakiak,.... Adibide-basea indexatzen denean eta sarrera berri bat corpusarekin parekatu baino lehen, sistemak sarrera *tokenizatzen* du, hau da, baliokidetasun-klaseetan dauden hitzak bilatzen ditu, eta hitz horien agerpen bakoitzean dagokion klasearengatik ordezkatzen du. Jatorrizko esaldia/hitza eta bere itzulpena ere gordetzen da beranduago erabili ahal izateko.

Baliokidetasun-klase bateko kideek ere tokenak izan ditzake, <DATE> definiziorako ikus dezakegun bezala:

<MONTH> <NUMBER>1 , <NUMBER>2 [ingelesa]
<NUMBER>1 . <MONTH> <NUMBER>2 [alemana]

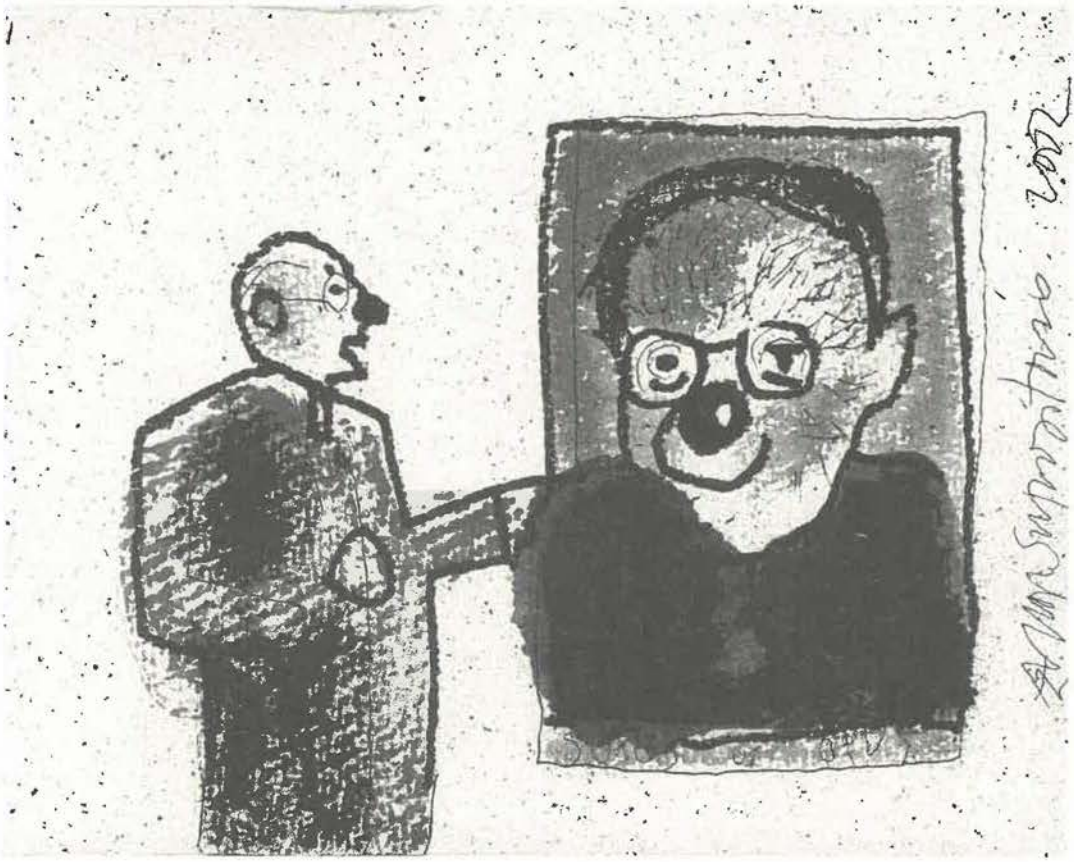
Honela, ingelesezko *July 4, 1776*, alemanera itzuliko dugu *4. Juli 1776* bezala.

Adibideak orokortzeko informazio linguistiko gehiago erabil daiteke: generoa eta numeroa, izen sintagmen identifikatzea,... erregela gramatikalak.

PanEBMTrekin egindako esperimentuetarako 94 baliokidetasun-klase sortu dira (asteko egunak, hilabeteak, animaliak, loreak, zuhaitzak, minerala, elementu kimikoak, enpresak, elkarteak, herriak, izenak, abizenak,...), guztira 5.397 itzulpen bikoterekin, eta hainbat erregela gramatikal gehitu dira.

Baliokidetasun-klaseak erabiltzean hobekuntza txiki bat ematen da, hobekuntza hori oso handi egiten dela orokortze errekurtsiboak onartzen direnean, bai estalduran zein parekatutako pusken luzeraren batez bestekoan. Behar diren adibideen kopurua dramatikoki jaisten da, itzulpenaren kalitatea modu marjinalean jaisten dela.

Gaztelania edo ingelesa bezalako hizkuntza handietarako hobekuntza hauek interesgarriak badira, corpus elebidunak ez dituzten hizkuntzetarako nahitaezkoak dira.



7. Euskararekin lan egiteko proposamenak

Itzulpen automatikoko transferentzia sistema bat eraikitzen ari gara (Díaz de Ilarraza et al., 2001), ingelesetik eta gaztelaniatik euskarara itzultzeko. Sistema hori aberasteko AOIA erabili daiteke modu ezberdinetan. Ondoko puntuetan ikertzeko lerro ezberdinak proposatzen ditugu.

7.1. Adibideetan Oinarritutako Transferentzia

Eraikitzen ari den transferentzian oinarritutako prototipoan, adibideak erabili daitezke transferentzia moduluan, justu transferentzia lexikoa egin baino lehen.

Momentuz ingeles-euskara prototiporako bi hitzetako adibideekin lan egiten duen modulu bat inplementatu dugu. Adibide-basea Morris hiztegiko adibideetatik jasoa da: guztira 40.440 adibide, haietatik 7.835 bi hitzez osatuta daudelarik.

Moduluak sintagmaren barruan adibide-basean aurkitzen diren bi hitz kontsekutiboak bilatzen ditu. Bi hitz horiek analisi fasetik datorren dependentzia-zuhaitzeko azpi-zuhaitz bat osatu

behar dute eta nodo umeak terminala behar du izan. Hala bada, nodo biak bakar batean biltzen ditu, adibide horren euskarazko ordainaren balio lexikoa adibide-basetik jasoz eta nodoan sartuz.

Modu honetan lokuzioen itzulpen literalak saihesten dira. Adibidez, "*machine translation*" modu honetan itzuliko dugu: «itzulpen automatiko», eta ez *«makina itzulpena».

Ikerketa lerro hau hitz gehiagoko adibideetara orokortu behar da, eta metodoa zuzena dela ziurtatu.

7.2. Kasu berezietarako AOIA

Sumita et Iidak (1991) proposatu zuten ideia euskarara egokitu genezake. Haiek japonierazko "N1 no N2" motako izen sintagmak (normalean ingelesezko "N2 of N1") itzultzeko erabili zuten moduan, guk ingelesetik euskarara "N2 of N1" izen sintagmak itzul genitzake, euskaraz modu ezberdinetan gauzatzen baitira:

(1) the house of London	–>	Londres**eko** etxea	[lekuzko genitibo kasua]
(2) the house of the man	–>	gizon**aren** etxea	[genitibo kasua]
(3) the house of chocolat	–>	txokolate**zko** etxea	[instrumental kasua]

Mota honetako adibide-base batekin, itzulpen automatikoko sisteman, «*the house of Bera*» bezalako sarrera bat agertzean, modulu berezi bat deituko luke, non antzeko adibideak berreskuratuko liratekeen:

sintaxia	>	sarrera = (1), (2), (3)
semantika	>	Bera = London
		Bera /= the man
		Bera /= chocolat

egokia den itzulpena irteera bezala emanik: "Berako etxea".

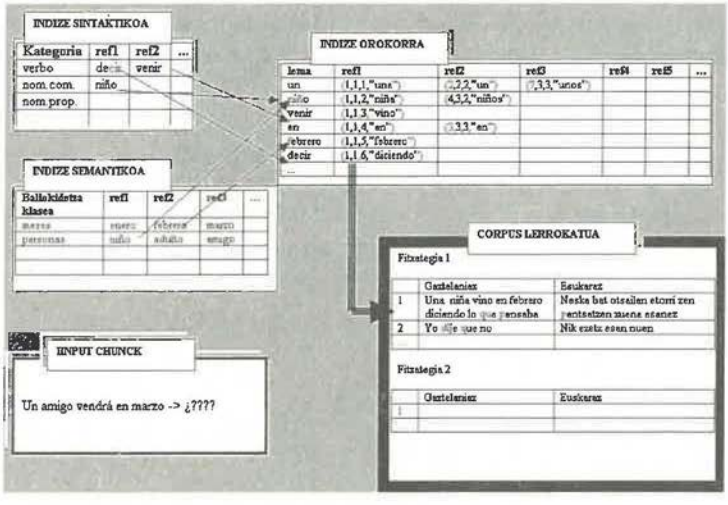
Antzeko adibideak berreskuratzeko neurri bat erabili beharko litzateke, eta horretarako WordNet-en oinarritutako distantzia semantikoa erabil genezake.

7.3. AOIA sistema oso bat eraiki, PanEBMTren moduan

Honelako sistema batek orain eraikitzen ari garen transferentzia-sistemarekin paraleloan lan egingo luke, baina noski, bien arteko integrazioa nola egin aztertu beharko litzateke.

EHAAko corpus elebiduna badaukagu. Corpus hau lerrokatzeko moduluak eraiki dira (Nieves 2000), baina egokitu behar dira corpus osoa lerrokatzeko.

Ondoren indexazio modua diseinatu beharko genuke, eta parekatze eta birkonbinatze moduluak sortu. Hurrengo irudian ikus dezakegu indexazioa egiteko lehenengo proposamen bat, non orokortze sintaktikoa eta semantikoa egiteko indizeak ere gordetzen diren.



BIBLIOGRAFIA

BROWN, R. D.: 1996. 'Example–Based Machine Translation in the Pangloss System', in Coling (1996), 169—174.

BROWN, R. D.: 1997. 'Automated Dictionary Extraction for «Knowledge–Free» Example–Based Translation', in TMI (1997), 111—118.

BROWN, R. D.: 1999. 'Adding Linguistic Knowledge to a Lexical Example–based Translation System', in TMI (1999), 22—32.

BROWN, R. D.: 2000. 'Automated Generalization of Translation Examples', in Coling.

BROWN, R. D., CARBONELL, J. G.: 'Generalized EBMT'. Web.

COLLINS, B. and P. CUNNINGHAM: 1996. 'Adaptation–Guided Retrieval in EBMT: A Case–Based Approach to Machine Translation', in I. Smith and B. Faltings (eds), Advances in Case–Based Reasoning: Third European Workshop, EWCBR–96, 91—104, Berlin: Springer.

CRANIAS, L., H. PAPAGEORGIOU and S. PIPERIDIS: 1994. 'A Matching Technique in Example–Based Machine Translation', in Coling (1994), 100—104.

Díaz de Ilarraza A., Mayor A., Sarasola K. 'Construcción de un prototipo de traducción automática multilingüe para el euskara'. 2001. SLPLT-2 . Jaén

Hutchins, W. J. and H. L. Somers: 1992. 'An Introduction to Machine Translation'. London: Academic.

Nagao, M.: 1984. 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in A. Elithorn and R. Banerji (eds) Artificial and Human Intelligence, 173-180, Amsterdam: North–Holland.

Nieves, N. «Generación automática de memorias de traducción». Proyecto fin de carrera. Facultad de Informática de la Universidad del País Vasco UPV-EHU. Junio 2000.

Nirenburg, S., S. Beale and C. Domashnev: 1994. 'A Full–Text Experiment in Example–Based Machine Translation'. International Conference on New Methods in Language Processing (NeMLaP), Manchester, England, 78— 87.

Sato, S. and M. Nagao: 1990. 'Toward Memory–Based Translation', in Coling (1990), Vol. 3, 247—252.

Somers, H. L.: 1997 'MT and Minority Laguages'.

Somers, H. L.: 1999 .'Review Article: Example-based Machine Translation'. Machine Translation 14 (1999), 113-158.

Sumita, E. and H. Iida: 1991. 'Experiments and Prospects of Example–Based Machine Translation', 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, 185—192.

Turcato, D., McFretridge P., Popowich F., Toole J.: 1999. 'A Unified Example-Based and Lexicalist Approach to Machine Translation'.