ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 2 (6)

www.compling.tsuull.uz

MUNDARIJA

ISSN 2181-922X 6-19

MASHINA TARJIMASIDA MATNNI MOSLASHTIRISH USULLARI

Shahlo Hamroyeva [1],
Noila Matyakubova[2]

Annotatsiya

Matnni moslashtirish turli xil mashina tarjimasi tizimlarining muhim jarayonidir. Bu vazifa dastlabki matnning so‘z, jumla yoki paragraflari, ularning tarjimasi (parallel korpus) o‘rtasidagi moslikni aniqlashdan iborat. Parallel korpusni moslashtirishning ikki asosiy yondashuvi mavjud: statistik usul va lug‘atga asoslangan usul. Ushbu maqolada parallel korpusda keng qo‘llanilayotgan statistik va lug‘atga asoslangan usullar va sohada qilingan ishlar haqida to‘liq ma’lumot berib o‘tilgan.

Kalit so‘zlar: *matnni moslashtirish, mashina tarjimasi, statistik usullar, leksik usullar, tabiiy tilni qayta ishlash.*

Kirish

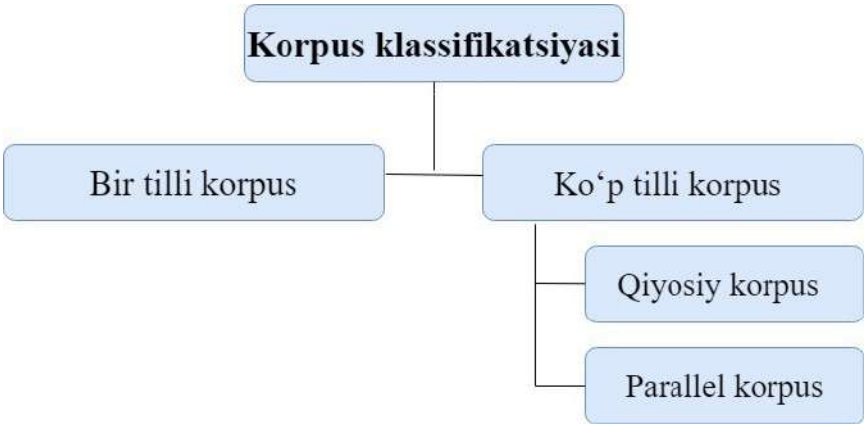
Tabiiy tilni qayta ishlash (NLP) – odamlar muloqot qilish uchun foydalanadigan tilni yaratish va tushunishga qaratilgan kompyuter fanining sohasi, uning juda ko‘p muhim vazifalari mavjud. Ulardan biri matnni avtomatik ravishda bir tildan ikkinchi tilga tarjima qilish maqsadida qo‘llaniladigan mashina tarjimasi(MT)dir. MTni amalga oshirish uchun lug‘at, statistika yoki misollarga asoslangan bir necha usullar mavjud. Ushbu usullarning afzallik va kamchiliklari bo‘lsa ham, ular matnni moslashtirish jarayoni kabi ba’zi usullardan keng foydalanadi. Matnni moslashtirish jarayoni paragraf, jumla, manba matnlarning so‘zlari va ularning tarjimalari o‘rtasida muvofiqlikni o‘rnatish uchun parallel korpusni tashkil qilishdan iborat. Parallel korpus turli tillardagi ikkita matn to‘plami sifatida belgilanishi mumkin, bu to‘plamlardan biri asliyatdagi matn, ikkinchisi esa uning tarjimasi hisoblanadi.

[1]*Hamroyeva Shahlo Mirdjanovna* – filologiya fanlari doktori, dotsent Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: shaxlo.xamrayeva@navoiy-uni.uz
ORCID: 0000-0002-5429-4708
[2]*Matyakubova Noila Shakirjanovna* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti tayanch doktoranti.
E-pochta: matyakubovanoila@navoiy-uni.uz
ORCID: 0009-0009-3154-723X

Asosiy qism.

Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: statistik usul; lug‘atga asoslangan usul. Leksikaga asoslangan yondashuvlar mavjud leksik bilimlarga tayanadi, masalan, antonim va sinonimlar, so‘zning tarjimalari va boshqalar. Statistik yondashuv esa lug‘atga bog‘liq bo‘lmagan ma’lumotlarga tayanadi. Masalan, jumla uzunligi, gapning o‘rni, takrorlanish chastotasi, ikki tilda gap uzunligi nisbati kabi holatlarga.

Matnni moslashtirish jarayonini amalga oshirish uchun biz yozma korpusdan foydalanishimiz kerak. Buni tadqiqot, ayniqsa, tarjima dasturlarini ishlab chiqish, tabiiy tilni qayta ishlash uchun foydalaniladigan matnlar to‘plami yoki katta hajmli matn sifatida aniqlash mumkin. Korpuslar teglangan yoki teglanmagan bo‘lishi mumkin. Teglangan korpuslar turli atributlarni yoki lingvistik ma’lumotlarni aniqlash uchun izohlanadi, masalan, korpus tarkibidagi hujjatlarning mavzulari yoki so‘zlarning turkumi va boshqalar. Masalan, "atirgullar" so‘zi uchun korpusda belgilangan atributlar ot, ko‘plik va boshqalar bo‘lishi mumkin. Teglanishi mumkin bo‘lgan lingvistik ma’lumotlar uning lemmasi, yani ma’lum bir lug‘at bo‘yicha so‘zning to‘g‘ri ma’nosi va boshqalar bo‘lishi mumkin. Rus va ingliz kabi tillarda, otning jinsini ifodalovchi belgilar qo‘shilishi mumkin. Teglangan korpus **/sentence/word/lemma/pos/id**ni aniqlaydigan maxsus belgi bilan belgilanadi [Moore, 2005, Iyun:1-8)]. Ammo teglanmagan korpus lingvistik ma’lumot va aniq tuzilishga ham ega emas. Bu ko‘pincha elektron pochta yoki tezkor xabar, hujjat yoki ijtimoiy media xabarlari kabi foydalanuvchilar tomonidan yaratilgan ma’lumotlardir.



1-rasm. Korpus klassifikatsiyasi

Korpusning har xil turlari mavjud bo‘lib, mashina tarjimasi sohasida bir tilli va ko‘p tilli korpusning tasnifi muhim ahamiyatga

ega. Ko‘p tilli korpuslar bir nechta tillardagi matnlar bo‘lib, ularni quyidagi kichik toifalarga bo'lish mumkin:

1. Parallel korpusni turli tillardagi ikkita matn to'plami sifatida qarash mumkin, bu to‘plamlardan biri manba matnlari, ikkinchisi esa ularning tarjimalari. Ushbu matnlarning har biri “bitext”lar deb nomlanadi [Harris, 1988: 8-10]. Parallel korpus bir yo‘nalishli, ikki tomonlama yoki ko‘p yo‘nalishli bo‘lishi mumkin. Masalan, Navoiyning “Xamsa” dostonlari va ularning turli tillardagi nusxalarini parallel korpus deb hisoblash mumkin.

2. Qiyosiy korpus – bu bir xil asosiy mavzuni yorituvchi, lekin uni ko‘rib chiqish uslubida farq qiluvchi turli tillardagi matnlar to‘plamidir. Bu shuni anglatadiki, qiyosiy korpus manba matni va ularning tarjimasi hisoblanmaydi. Misol tariqasida jurnal yoki OAV tizimlaridan olingan yangi maqolalar to‘plamlari olish mumkin, chunki ular turli tillarda bir xil voqeaga ishora qiladi va shuning uchun qiyosiy korpus deb hisoblanishi mumkin [Simões, 2004: 45-50].

Korpusni moslashtirish – bu manba matnlarining paragraf, gap yoki so‘zlari va ularning tarjimalari o‘rtasida muvofiqlikni o‘rnatish uchun parallel korpusni tashkil qilishdan iborat. Shunga qaramay, parallel korpuslarni avtomatik ravishda moslashtirish ba’zi til juftliklari uchun yechimini kutayotgan muammo hisoblanadi.

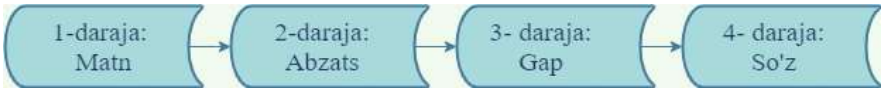
Kanadalik tilshunos E.Maklovich MTda qo’llaniladigan moslashtirish usullarini 4 darajaga bo‘lib, ularni quyidagicha tasniflagan [Macklovitch, Hannan, 1998: 41-57]:

1-darajali moslashtirish: matn yetarlicha uzun bo‘lmaganda butun matnni moslashtiradi.

2-darajali moslashtirish: bu daraja abzaslarni moslashtirishni takrorlaydi.

3-darajali moslashtirish: gaplarni moslashtirishni tavsiflaydi.

4-darajali moslashtirish: bitextlar orasidagi so‘zlarni moslashtiradi.



2-rasm. MTda qo‘llanadigan moslashtirish darajalari

Mashina tarjimasi kontekstida korpusning moslashuvi asliyat tilidagi jumlalarni tarjima tildagi tegishli tarjimalari bilan moslashtirish jarayonini anglatadi. Mashina tarjimasi modellarini o‘rgatish va takomillashtirish uchun foydalaniladigan asliyat va tarjima tillari o‘rtasida so‘z yoki ibora mosligiga erishish uchun bu moslashtirish zaruriy bosqich hisoblanadi. Korpusni moslashtirish jarayoni ket-

ma-ketligi manbalarda quyidagicha beriladi:

1. Oldindan ishlov berish. korpusdagi asliyat va tarjima tildagi matnlarni oldindan qayta ishlash. Bu odatda jumlalarni so‘z yoki so‘z birikmalar sifatida tokenizatsiya qilish, tinish belgilarini olib tashlash, matnni normallashtirish (masalan, kichik harflar, diakritik normalizatsiya) va har qanday tilga xos bo‘lgan dastlabki ishlov berish bosqichlarini qo‘llashni o‘z ichiga oladi.

2. Gaplarni moslashtirish. Boshlang‘ich jumla darajasidagi matnlarni yaratish uchun asliyat va tarjima til matnlaridagi gaplarni moslashtirish lozim. Bu bosqich gap uzunligi, o‘xshashlik o‘lchovlari asosidagi evristika yordamida yoki mavjud gaplarni moslashtirish vositalari yoki ochiq kutubxonalar yordamida amalga oshirilishi mumkin.

3. So‘zlarni moslashtirish modellari. moslashtirilgan gap juftlari o‘rtasida yanada aniqroq – so‘zlar o‘rtasida ekvivalentlikni aniqlash modellaridan foydalaniladi. IBM Modellari (1-5), Yashirin Markov Modellari (HMM) yoki neyron mashina tarjimasida ishlatiladigan turli xil moslashtirish modellari qo‘l keladi.

4. Moslash modellarini o‘rgatish. Tanlangan aligner modeli korpusda mavjud bo‘lgan moslashtirilgan gaplar yordamida o‘rgatiladi. Bu bosqich Expectation-Maximization (EM) algoritmi yordamida model parametrlarini baholash yoki xatolarni avtomatik to‘g‘rilovchi va gradient algoritmidan foydalangan holda neyron tarmoqqa asoslangan modellarni o‘rgatishni o‘z ichiga oladi.

5. Aligner vositasini yaratish. Butun korpus uchun so‘zlarni moslashtirish uchun o‘rgatilgan moslashuv modeli qo‘llaniladi. Bu jarayon asliyatdagi so‘zlar va ularning tarjima tildagi tarjimalari o‘rtasida moslashtirish aloqalarini belgilashni o‘z ichiga oladi.

6. Postprocessing. Yaratilgan aligner vositasini yaxshilash uchun keyingi ishlov berish usullari qo‘llaniladi. Bu qadam shovqinli moslashtirish holatlarini ko‘rib chiqish, moslashtirish chegaralarini aniqlashtirish yoki noaniq moslashtirish holatlarini hal qilishni o‘z ichiga oladi. Postprocessing, shuningdek, ma’lum mezonlar asosida xato moslashtirishlarni filtrlashni ham o‘z ichiga olish ehtimoli mavjud.

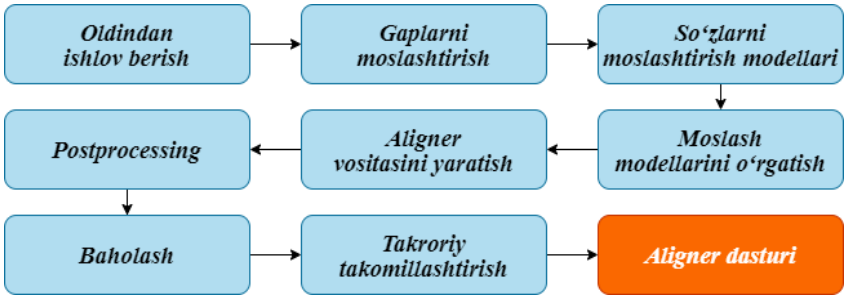
7. Baholash. Ushbu baholash bosqichida alignerlarni qo‘lda moslashtirilgan jumlalar bilan solishtirish yoki moslashtirishni baholash vositalaridan foydalanish orqali amalga oshirish mumkin.

8. Takroriy takomillashtirish. Baholash natijalari tahlil qilingach, ushbu bosqichda baholash jarayonidan olingan fikr-mulohazalarni o‘z ichiga olgan moslashtirish modeli takroriy takomillashti-

rish model parametrlarini sozlash, qo‘shimcha funksiyalarni kiritish uchun foydalaniladi.

9. Aligner dasturi. Yangi jumla juftlarini moslashtirish uchun aniqlangan aligner modelini qo‘llash yoki undan mashina tarjimasi modellari uchun o‘quv dasturining bir qismi sifatida foydalanish imkoniyati mavjud. Tegishli ma’lumotlardan tarjima jarayonini boshqarish, tarjima sifatini yaxshilash yoki ikki tilli lug‘atlar yoki iboralar jadvallari kabi til resurslarini olishda qo‘llanadi.

Shuni ta’kidlash kerakki, moslashtirish doimiy jarayon bo‘lib, alignerning sifati moslashtirilayotgan tillarning murakkabligi, o‘quv ma’lumotlarining mavjudligi va qo‘llaniladigan moslashtirish usullariga qarab farq qilishi mumkin.

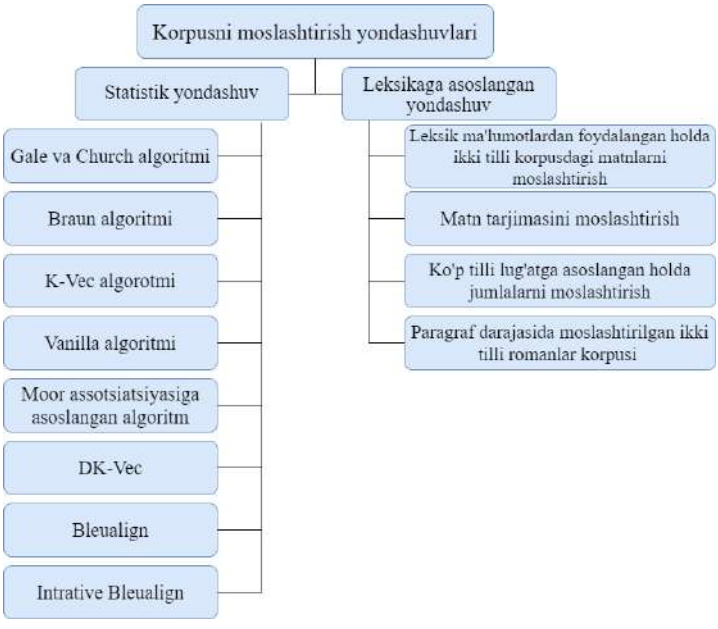


3-rasm. Korpusni moslashtirish bosqichlari.

Korpusni moslashtirishga oid yondashuvlar

Parallel korpusni moslashtirishning ikkita asosiy yondashuvi mavjud: birinchi yondashuv statistik ma’lumotlarga asoslanadi, ikkinchisi esa qo‘shimcha lingvistik bilimlarni qo‘llaydi. Ushbu farqning asosi qayta ishlash usullaridan mustaqil ravishda qayta ishlanadigan ma’lumotlar turi bilan bog‘liq [Gelbukh, Sidorov, Vera-Félix, 2006: 16-23]. Ushbu yondashuvlar asosida bir nechta metodlar ishlab chiqilgan, ularning har biri o‘zining afzallik va kamchiliklariga ega.

Leksikaga asoslangan yondashuv mavjud leksik resurslar, masalan, katta hajmdagi ikki tilli lug‘atlar, tillar haqida ma’lumot olish uchun maxsus atamalar ro‘yxatiga, antonim va sinonimlar, so‘zning tarjimalari va boshqalarga tayanadi. Ushbu usul statistik ma’lumotga asoslangan usuldan ko‘ra sekinroq ishlaydi va tilga bog‘liq bo‘ladi. Mazkur usulning asosiy kamchiligi shundaki, ishlash jarayoni ko‘p jihatdan moslashtirishda qo‘llaniladigan leksik ma’lumotlarga bog‘liq. Biroq bu usul bilan statistik usuldan ko‘ra yaxshiroq natijalarga erishish mumkin. Statistik yondashuvlar leksik bo‘lmagan ma’lumotlarga tayanadi, masalan, gap uzunligi, gapning o‘rni, birgalikda sodir bo‘lish chastotasi, ikki tildagi jumla uzunligi nisbati. Bu usullar moslashtirish jarayonini tezlashtiradi.



4- rasm. Korpusni moslashtirish yondashuvlari

Biroq bu usulning asosiy kamchiligi shundaki, uning ishlashi maqsadli matn va bitextlarning manba matni o‘rtasidagi tarkibiy o‘xshashlikka bog‘liq [Gelbukh, Sidorov, Vera-Félix, 2006: 16-23]. 3-rasmda bugungi kunda korpus moslashuvida keng qo‘llanilayotgan usullar va algoritmlar keltirilgan.

Gale va Church algoritmi “bir tildagi uzunroq gap boshqa tildagi uzunroq gapga, qisqa gap qisqa gapga tarjima qilinadi” g‘oyasiga asoslanadi va moslashtirish tamoyili aynan shunga qaratiladi [Gale, Church, 1993: 75-102]. Ushbu algoritm faoliyati uchun paragraflarda allaqachon moslashtirilgan parallel korpus talab qilinadi. Ushbu algoritm matnlarni moslashtirish uchun gapning uzunligini belgilarda hisobga oladi. Bular har bir gap juftligi (manba matndan biri va tarjima matnidagi birlik) uchun masofa o‘lchovi deb ataladigan qiymatni hisoblash uchun ishlatiladi. Masofa o‘lchovi qanchalik past bo‘lsa, gaplarning o‘zaro mos kelishi ehtimoli shunchalik yuqori bo‘ladi.

Braun algoritmi “Braun klastering” deb ham nomlanadi, chunki u dastlab 1990-yillarda Braun universitetida ishlab chiqilgan. Gale va Church algoritmiga o‘xshab, bu algoritm matnlarni moslashtirish uchun gap uzunligini hisobga oladi, ammo farqi shundaki, o‘lchovda gap uzunligi so'zlarda hisobga olinadi. Unda Kanada Hansard korpusining TEX formatiga kiritilgan teglardan moslashish jarayonida foydalaniladi. Algoritm asosiy va kichik bog‘lanish nuqtalarini ko‘rib chiqadi va moslashtirishni ayni ushbu ikki bosqichda

amalga oshiradi [Brown, Lai, Mercer, 1991: 169-176].

K-Vec algoritmi Paskal Fung va Kennet Uord Cherxi [Harris, 1988: 8-10] tomonidan ishlab chiqilgan "sentence aligner" vositasi hisoblanadi. K-vec juda muhim xususiyatga ega; u gap chegaralariga bog'liq emas. Ushbu xususiyatdan foydalanib, algoritm o'xshash bo'lmagan tillarni moslashtirishni maqsad qiladi; masalan, ingliz va yapon, yevropa tillari kabi. Ushbu algoritm "agar ikkita so'z bir-birining tarjimasi bo'lsa, ular ikkita bo'lmagan so'zlarga qaraganda bir xil segmentlarda paydo bo'lish ehtimoli ko'proq" tamoyili asosida yaratilgan [Fung, McKeown, 1994: 12]. Bitekstdagi so'z boshqa so'zning tarjimasi ekanligini aniqlash uchun mazkur algoritm ularning tegishli matnda taqsimlanishining o'xshashligiga e'tibor qaratadi.

Vanilla Aligner 1997-yilda Pernilla Danielson va Daniel Ridings tomonidan taqdim etilgan, u Gale va Church algoritmining takomillashtirilganidir. O'zidan oldingisi kabi bu ham "sentence aligner" vositasi hisoblanib, gaplar chegarasiga bog'liq. Ushbu alignerning asosiy afzalligi SGML formatidagi bitextlar bilan ishlashga mosligidir [Danielsson, Ridings, 1997: 88-93]. Bitextlarni SGML formatida ishlatishning afzalliklaridan biri shundaki, standart shakl yoki tuzilma o'rnatilishi mumkin; gap chegaralarini osonroq aniqlashga yordam beradi.

Moor assotsiatsiyasiga asoslangan algoritm ikki asosiy tamoyilga asoslanib yaratilgan: birinchidan, ular "sentence alignerlar phrase aligner" – iboraga asoslangan alignerlarni yaratish uchun yaxshi debocha bo'lishi mumkinligiga qattiq ishonishgan; ikkinchidan, o'sha vaqtga qadar taqdim etilgan algoritmlar, masalan, Braun yuqori hisoblashda bir muncha murakkablik va kamchiliklarga ega, ammo past hisoblashda proportsional yaxshi aniqlik bilan ishlaydi. Moor so'zlarni moslashtirishning uchta turli strategiyasini taqdim etadi [Moore, 2005: 1-8]:

1) bir so'zni bitta ma'nosi bilan moslashtirish;

2) bir so'zni bir nechta ma'nosi bilan moslashtirish;

3) tokenlarni mosligini tanlash.

Har bir strategiyada bir necha muammoni yechish uchun ikki yoki undan ortiq usullar mavjud, ularning barchasi leksikon tarjimasini yaratishda qo'llanilgan Log-Likelihood-Ratio (LLR) assotsiatsiyasi o'lchoviga asoslangan.

DK-vec algoritmi – (Dinamik K-vec algoritmi) hozircha DK-vec o'zining ajdodi K-vec algoritmiga asoslangan, u ikki so'z bir-birining tarjimasi bo'lsa, bir segmentda paydo bo'lish ehtimoli ko'proq

degan faraz ostida ishlaydi. Biroq bu, odatda, o‘xshash bo‘lmagan tillarda sodir bo‘lmaydi. Bundan tashqari, k-vec algoritmi tilning eski ma’lumotlarini yoki uning ishlashini pasaytiradigan korpus xususiyatlarini hisobga olmaydi.

Bleualign sentence aligneri Riko Sennrich va Martin Volk tomonidan yaratilgan. Uning asosiy g'oyasi moslashtirish jarayonida yordam berish uchun MT tizimi va tarjima baholovchi BLEUdan foydalanishdir [Sennrich, Volk, 2010: 102-110]. Alignerni yaxshiroq tushunish uchun BLEU haqida bilish kerak. BLEU – bu MT tarjima natijasi sifatini baholovchi algoritm. Ushbu sifatni o‘lchash uchun BLEU MT tarjima natijasini bir yoki bir nechta inson tarjimalari bilan mosligini taqqoslash orqali baholanadi.

Intrative Bleualign sentence aligner algoritmi 2011-yilda Riko Sennrich va Martin Volk tomonidan moslashtirish jarayonida mashina tarjima qilish tizimidan foydalanishdagi kamchiliklarni chuqurroq tahlil qilish natijasida yaratilgan. Sennrich va Volk MT asosidagi alignerlar manba matnining to‘g‘ri tarjimasiga kuchli bog‘liqligini aniqladilar va MT tizimlari odatda moslangan matnlar bilan oziqlanganligini hisobga olsak, aylana bog‘liqligi mavjudligi aniq bo‘ladi. Ushbu qaramlikni bartaraf etish uchun ushbu algoritm moslashtirishni amalga oshirish uchun bootstrapping (yuklash) usulini taqdim etadi.

Ikki tilli korpusda jumlalarni moslashtirishning leksik yondashuvlari turli tillardagi tegishli jumlalarni aniqlash uchun leksik ma’lumotlardan foydalanishni o‘z ichiga oladi. Ushbu yondashuv turli tillardagi tegishli jumlalar o‘xshash leksik tarkibga ega bo‘ladi degan taxminga asoslanadi. Leksik asosda moslashtirishning keng tarqalgan usullaridan biri ikki tilli lug‘atdan foydalanishga asoslangan. Ushbu yondashuv boshqa tilda to‘g‘ridan-to‘g‘ri tarjimalari bo‘lgan bir tildagi so‘zlarni aniqlashni va keyin tegishli gaplarni moslashtirish uchun ushbu tarjimalardan foydalanishni o‘z ichiga oladi. Masalan, agar ingliz tilidagi "cat" so‘zining o‘zbek tilidagi "mushuk"-ga to‘g‘ridan-to‘g‘ri tarjimasi bo‘lsa, unda bu so‘zlarni o‘z ichiga olgan gaplarni moslashtirish mumkin.

Leksik yondashuvda moslashtirishning yana bir usuli statistik modellardan foydalanishni o‘z ichiga oladi. Bu modellar gaplar orasidagi yozishmalarni birgalikdagi so‘zlarning chastotasiga qarab aniqlash uchun ehtimollik algoritmlaridan foydalanadi. Misol uchun, agar turli tillardagi ikkita gapda umumiy so‘zlar ko‘p bo‘lsa, ularning mos kelish ehtimoli yuqori. Umuman olganda, leksikaga asoslangan yondashuv ikki tilli korpusdagi gaplarni moslashtirish uchun sama-

rali bo‘lishi mumkin, ayniqsa, birlashtirilayotgan tillar “yaqin qarindosh”, ya’ni o‘xshash shakl va ma’nolarga ega so‘zlar bo‘lsa, bu qoida amal qiladi. Biroq bu usul leksik tuzilmalari juda xilma-xil bo‘lgan tillar yoki idiomatik iboralar bilan ishlashda unchalik samarali bo‘lmasligi mumkin

Paragraf darajasida moslashtirilgan ikki tilli romanlar korpusi ikki tildagi matnlar to‘plami bo'lib, ular paragraflari asosida moslanlangan. Ushbu turdagi korpusda bir tildagi har bir paragraf boshqa tildagi tegishli paragraf bilan birlashtirilib, matnlarni oson taqqoslash va tahlil qilish imkonini beradi. Paragraf darajasida moslashtirilgan ikki tilli romanlar korpusini yaratish, odatda, bir necha bosqichlarni o‘z ichiga oladi. Birinchidan, har ikki tildagi romanlarni tanlab, raqamlashtirish kerak. Keyinchalik, har bir romandagi paragraflarni aniqlash va ajratish kerak. Keyin bir tildagi paragraflar boshqa tildagi tegishli paragraflar bilan mos kelishi kerak. Bu moslashtirish qo‘lda yoki aligner dasturi yordamida amalga oshirilishi mumkin.

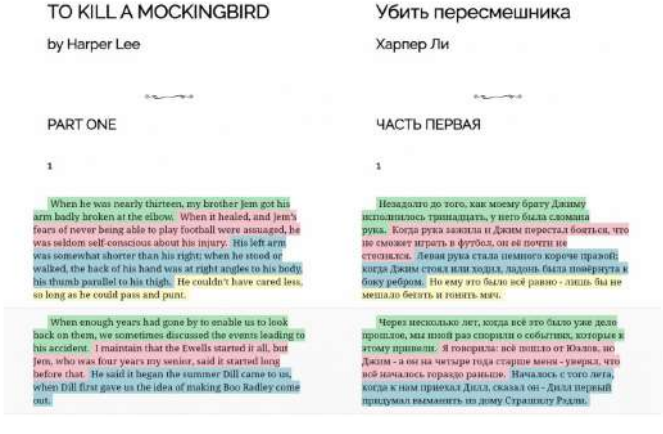
Moslash tugallangach, ikki tilli korpus turli maqsadlarda ishlatilishi mumkin. Masalan:

1. Qiyosiy adabiy tahlil. Tadqiqotchi turli tillardagi romanlarning uslub, mavzu va hikoya strukturalarini solishtirishi mumkin.

2. Tilni o‘rgatish va o‘rganish. O‘qituvchi korpusdan til o‘rganuvchilarning o‘qishni tushunish va tarjima qilish ko‘nikmalarini yaxshilash uchun materiallar ishlab chiqishda foydalanishi mumkin.

3. Mashina tarjimasi. Korpusdan mashina tarjimasi modellarini ularning aniqligi va ravonligini oshirish uchun o‘rgatish uchun foydalanish mumkin.

4. Parallel matnli kitoblar yaratish uchun qo‘llanilishi mumkin. Ikki tilli matnlardan tashkil topgan adabiyotlarni shakllantirishga yordam beradi.



TO KILL A MOCKINGBIRD

by Harper Lee

PART ONE

1

When he was nearly thirteen, my brother Jem got his arm badly broken at the elbow. When it healed, and Jem's fears of never being able to play football were assuaged, he was seldom self-conscious about his injury. His left arm was somewhat shorter than his right; when he stood or walked, the back of his hand was at right angles to his body, his thumb parallel to his thigh. He couldn't have cared less, so long as he could pass and punt.

When enough years had gone by to enable us to look back on them, we sometimes discussed the events leading to his accident. I maintain that the Ewells started it all, but Jem, who was four years my senior, said it started long before that. He said it began the summer Dill came to us, when Dill first gave us the idea of making Boo Radley come out.

Убить пересмешника

Харпер Ли

ЧАСТЬ ПЕРВАЯ

1

Незадолго до того, как моему брату Джиму исполнилось тринадцать, у него была сломана рука. Когда рука зажила и Джим перестал бояться, что не сможет играть в футбол, он её почти не стеснялся. Левая рука стала немного короче правой; когда Джим стоял или ходил, ладонь была повёрнута к боку ребром. Но ему это было всё равно - лишь бы не мешало бегать и гонять мяч.

Через несколько лет, когда всё это было уже дело прошлое, мы иной раз спорили о событиях, которые к этому привели. Я говорила: всё пошло от Юэллов, но Джим - а он на четыре года старше меня - уверял, что всё началось гораздо раньше. Началось с того лета, когда к нам приехал Дилл, сказал он - Дилл первый придумал выманить из дому Страшилу Рэдли.

5-rasm. LingTrain aligner dasturiy vositasi yordamida parallel matnli paragraph darajasida moslashtirish jarayoni.

Umuman olganda, paragraf darajasida moslashtirilgan ikki tilli romanlar korpusi tadqiqotchi, til o‘rganuvchi va mashina tarjimasi tizimlarini ishlab chiquvchi mutaxassislar uchun qimmatli manba bo‘lishi mumkin.

Xulosa

Korpus tilshunosligida statistik va leksik moslashtirish so‘z yoki iboralarni parallel yoki taqqoslanadigan korpuslar o‘rtasida moslashtirish uchun ishlatiladigan ikki usuldir. Ikkala usulning ham afzallik va kamchiliklari bor. Ulardan foydalanish tadqiqot maqsadiga bog‘liq bo‘lib statistik moslashtirishning ro‘li quyidagilarni o‘z ichiga oladi:

Parallel korpus: Statik aligner manba va maqsadli matnlardagi mos segmentlarni aniqlash orqali parallel korpuslarni moslashtirishga yordam beradi. Bu qiyosiy tahlil, tarjima tadqiqotlari va mashina tarjimasi uchun o‘quv ma’lumotlarini tayyorlash imkonini beradi.

Korpus izohi: Moslashtirish izohlarni matndagi tegishli so‘zlar yoki segmentlar bilan moslash orqali korpus izohini osonlashtiradi. Bu moslashtirish sintaktik, semantik yoki nutq tahlili kabi turli lingvistik tadqiqotlarda foydalidir.

Matn tahlili: Statik moslash matnni taqqoslash, kontrastli tahlil qilish va matnning turli jihatlarini aniqlashda yordam beradi. O‘xshash jihatlarni moslashtirish orqali tadqiqotchilar matnlar orasidagi farqlar yoki o‘xshashliklarni tahlil qilishlari mumkin.

Leksik asosda moslashtirish alohida so‘zlarni yoki leksik birliklarni parallel yoki taqqoslanadigan matnlarda moslashtirishga qaratilgan. Ushbu uslub so‘z darajasida yozishmalarni o‘rnatishga qaratilgan va birinchi navbatda lingvistik ma’lumotlarni leksik tahlil qilish va ajratib olish uchun ishlatiladi va quyidagi vazifalarni o‘z ichiga oladi:

Ikki tilli leksikografiya: Ikki tilli lug‘atlar yoki leksik resurslarni yaratish uchun ikki tilli leksikografiyada moslashtirish juda muhimdir. Parallel matnlardagi so‘zlar yoki iboralarni moslashtirish orqali leksikograflar aniq tarjimalarni o‘rnatishlari va to‘liq leksik ma’lumotlarni taqdim etishlari mumkin.

So‘z ma’nosini ajratish: Leksik moslashuv turli kontekstlarda ko‘p ma’noli so‘zlarning misollarini moslash orqali so‘z ma’nosini aniqlashga yordam beradi. Ushbu aligner so‘zning tegishli tarjimalari yoki moslashtirilgan matnlardagi qo‘llanishlari asosida mo‘ljallangan ma’nosini ajratishga yordam beradi.

Leksik o‘zgarishlarni tahlil qilish: leksik birliklarni turli matnlar bo‘ylab tekislash tadqiqotchilarga sinonimlar, birikmalar yoki idiomatik iboralar kabi leksik o‘zgarishlarni tahlil qilish imkonini beradi. Ushbu tahlil leksik tanlovlar, til o‘zgarishi va nutq shakllari haqida tushuncha berishi mumkin.

Xulosa qilib aytganda, statik asosli alignerlar qiyosiy tahlil va korpus annotatsiyasi uchun matnlarning kattaroq segmentlarini moslashga qaratilgan bo‘lsa, leksik asosidagi alignerlar alohida so‘zlar yoki leksik birliklarni leksikografik maqsadlarda moslash, so‘z ma’nosini aniqlash va leksik variatsiya tahlili bilan shug‘ullanadi. Ikkala usul ham korpus lingvistikasida korpuslardan lingvistik ma’lumotlarni o‘rganish va chiqarishga yordam beradi. Statistik moslashtirish katta hajmdagi ma’lumotlar bilan ishlashda, asosiy e’tiborni avtomatik qayta ishlashga qaratilayotganda foydali bo‘lsa, leksik moslashtirish muayyan so‘z yoki iboralarni semantik va sintaktik tahlil qilishga qaratilgan ishlarda mos keladi.

Foydalanilgan adabiyotlar ro‘yxati:

Brown P.F., Lai J.C., Mercer R.L. (1991). Aligning sentences in parallel corpora. Proceedings of the 29th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics. (169-176b).

Danielsson P., Ridings D. (1997, Fevral). Practical presentation of a “vanilla” aligner. TELRI Workshop in alignment and exploitation of texts,(88-93b).

Fung P., Church K.W. (1994, Avgust). K-vec: A new approach for aligning parallel texts. Proceedings of the 15th conference on Computational linguistics. Association for Computational Linguistics. (1096-1102b).

Fung P., McKeown K. (1994). Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping.

Gale W.A., Church K.W. (1993). A program for aligning sentences in bilingual corpora. Computational linguistics, *19*(1), (75-102b).

Gelbukh A., Sidorov G., Vera-Félix J. Á. (2006). A bilingual corpus of novels aligned at paragraph level. Advances in Natural Language Processing. Springer Berlin Heidelberg. (16-23 b).

Harris B. (1988). Bi-text, a new concept in translation theory. Language Monthly, 54, (8-10b).

Kit C., Webster J.J., Sin K.K., Pan H., Li H. (2004). Clause alignment for Hong Kong legal texts: A lexical-based approach. International Journal of Corpus Linguistics, (29-51b).

McEnery A., Xiao R. Z. (2008). Paralell and comparable corpora: what are they up to? Incorporating Corpora: Translation and the Linguist. Translating Europe. Clevendon: Multilingual Matters.

Macklovitch E., Hannan M.L. (1998). Line ‘em up: advances in alignment technology and their impact on translation support tools. Machine Translation, 13(1), (41-57b).

Meyers A., Kosaka M., Grishman R. (1998, Oktyabr). A multilingual procedure for dictionary-based sentence alignment. Conference of the Association for Machine Translation in the Americas. Springer Berlin Heidelberg. (187-198b).

Moore R. C. (2005, Iyun). Association-based bilingual word alignment. Proceedings of the ACL Workshop on Building and Using Parallel Texts. Association for Computational Linguistics. (1-8b).

Sennrich R., Volk M. (2010, Noyabr). MT-based sentence alignment for OCRgenerated parallel texts. The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010*),* Denver, Colorado.

Sennrich R., Volk M. (2011, May). Iterative, MT-based sentence alignment of parallel texts. 18th Nordic Conference of Computational Linguistics, NODALIDA

Simões A. (2004). Parallel corpora word alignment and applications (master thesis). Universidade do Minho, Braga.

Natural Language Toolkit (NLTK): https://www.nltk.org/

British National Corpus (BNC): https://www.natcorp.ox.ac.uk/

Text Encoding Initiative (TEI): https://tei-c.org/

METHODS OF TEXT ALIGNMENT IN MACHINE TRANSLATION

Shahlo Hamroyeva[1]
Noila Matyakubova[2]

Abstract

The text alignment is animportant process of different Machine Translation systems. This task consists in identifying correspondences between words, sentences or paragraphs of a source text and their translation (parallel corpus). There are two main approaches to perform parallel corpus alignment: the statistical-based methods and lexical-based methods. This article provides a comprehensive overview of widely used statistical-based methods and lexical-based methods in parallel corpora and the related scientific works done in the field.

Keywords: *Text alignment, machine translation, statistical methods, lexical methods, natural language processing.*

References:

Brown P.F., Lai J.C., Mercer R.L. (1991). Aligning sentences in parallel corpora. Proceedings of the 29th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics. (169-176b).

Danielsson P., Ridings D. (1997, Fevral). Practical presentation of a "vanilla" aligner. TELRI Workshop in alignment and exploitation of texts,(88-93b).

Fung P., Church K.W. (1994, Avgust). K-vec: A new approach for aligning parallel texts. Proceedings of the 15th conference on Computational linguistics. Association for Computational Linguistics. (1096-1102b).

Fung P., McKeown K. (1994). Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping.

Gale W.A., Church K.W. (1993). A program for aligning sentences in bilingual corpora. Computational linguistics, *19*(1), (75-102b).

Gelbukh A., Sidorov G., Vera-Félix J. Á. (2006). A bilingual corpus of novels aligned at paragraph level. Advances in Natural Language Processing. Springer Berlin Heidelberg. (16-23 b).

[1]*Hamroyeva Shahlo Mirdjanovna* – doctor of philological sciences, associate professor Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
Email: shaxlo.xamrayeva@navoiy-uni.uz
ORCID: 0000-0002-5429-4708
[2]*Matyakubova Noila Shakirjanovna* – PhD student of Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: matyakubovanoila@navoiy-uni.uz
ORCID: 0009-0009-3154-723X

Harris B. (1988). Bi-text, a new concept in translation theory. Language Monthly, 54, (8-10b).

Kit C., Webster J.J., Sin K.K., Pan H., Li H. (2004). Clause alignment for Hong Kong legal texts: A lexical-based approach. International Journal of Corpus Linguistics, (29-51b).

McEnery A., Xiao R. Z. (2008). Paralell and comparable corpora: what are they up to? Incorporating Corpora: Translation and the Linguist. Translating Europe. Clevendon: Multilingual Matters.

Macklovitch E., Hannan M.L. (1998). Line 'em up: advances in alignment technology and their impact on translation support tools. Machine Translation, 13(1), (41-57b).

Meyers A., Kosaka M., Grishman R. (1998, Oktyabr). A multilingual procedure for dictionary-based sentence alignment. Conference of the Association for Machine Translation in the Americas. Springer Berlin Heidelberg. (187-198b).

Moore R. C. (2005, Iyun). Association-based bilingual word alignment. Proceedings of the ACL Workshop on Building and Using Parallel Texts. Association for Computational Linguistics. (1-8b).

Sennrich R., Volk M. (2010, Noyabr). MT-based sentence alignment for OCR generated parallel texts. The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010*),* Denver, Colorado.

Sennrich R., Volk M. (2011, May). Iterative, MT-based sentence alignment of parallel texts. 18th Nordic Conference of Computational Linguistics, NODALIDA

Simões A. (2004). Parallel corpora word alignment and applications (master thesis). Universidade do Minho, Braga.

Natural Language Toolkit (NLTK): https://www.nltk.org/

British National Corpus (BNC): https://www.natcorp.ox.ac.uk/

Text Encoding Initiative (TEI): https://tei-c.org/