TurkLang.2018

VI International Conference on Computer Processing of Turkic Languages

Tashkent, Uzbekistan
18-20 October, 2018

ТАШКЕНТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УЗБЕКСКОГО ЯЗЫКА И ЛИТЕРАТУРЫ ИМЕНИ АЛИШЕРА НАВОИ

АКАДЕМИЯ НАУК РЕСПУБЛИКИ ТАТАРСТАН

ИНСТИТУТ ПРИКЛАДНОЙ СЕМИОТИКИ

ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМЕНИ Л. Н. ГУМИЛЁВА

МИНИСТЕРСТВА ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН

НИИ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

VI МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ

«TURKLANG-2018»

(труды конференции)

ТАШКЕНТ
ИЗДАТЕЛЬСКО-ПОЛИГРАФИЧЕСКИЙ
ДОМ «NAVOIY UNIVERSITETI»
2018

For the words above TARAMOQ and QARAMOQ. m=6, t=0, $|S_1|=|S_2|=7$:

$$d_j = \frac{1}{3}\left(\frac{6}{7}+\frac{6}{7}+\frac{6}{6}\right) = 0.9047$$

l=0, p=0.1:

$$d_w = 0.9047 + \left(2 \cdot 0.1 \cdot (1 - 0.9047)\right) = 0.92376$$

There is some result of different test of program that is written in python programming language.

	DISTANCE OF DJARO	DISTANCE OF DJARO-WINKLER
BAHOR	0.8666666666666667	0.9066666666666667
NAHOR	0.7333333333333334	0.7333333333333334
XABAR	0.7	0.7
KITOB	0.8666666666666667	0.9066666666666667
KILOB	0.7333333333333334	0.7866666666666667

REFERENCE:

1. Choudhury, R., K. Kashyap, and N. Deb, 2016. A survey on the different approaches of context sensitive spell-checking. *International Journal of Engineering Science and Computing*, 6(6):6872–6873.
2. Fierman, W., 1992. *Language planning and national development: the Uzbek experience*. Berlin, Germany; New York, USA: Mouton de Gruyter.
3. Gupta, N. and P. Mathur, 2012. Spell checking techniques in NLP: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(12):217–221.
4. Li, X., J. Tracey, S. Grimes, and S. Strassel, 2016. Uzbek-English and Turkish-English Morpheme Alignment Corpora. In *Proceedings of LREC 2016*. pages 2925–2930.
5. Li, X., J. Tracey, S. Grimes, and S. Strassel, 2016. Uzbek-English and Turkish-English Morpheme Alignment Corpora. In Proceedings of LREC 2016. Pages 2925–2930.
6. Pirinen, T. and K. Linden, 2014. State-of-the-art in weighted finite-state spell-checking. In CICLing 2014: Computational Linguistics and Intelligent Text Processing, volume 8404 of LNCS. Berlin, Heidelberg: Springer.
7. Sayfullaev, A., 2016. Problems of rendering prepositions in translation (on the material of English and Uzbek languages). In *Scientific enquiry in the contemporary world: theoretical basics and innovative approach*. pages 91–96.
8. Singh, S.P., A. Kumar, L. Singh, M. Bhargava, K. Goyal, and B. Sharma, 2016. Frequency based spell checking and rule based grammar checking. In Proc. Of International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE.

АВТОЛИНГВИСТИЧЕСКИЕ СИСТЕМЫ ОБРАБОТКИ ТЕКСТОВ

М. Абжалова, Навоийский государственный горный институт,
Навои, Узбекистан, manzura_ok@mail.ru

Аннотация: *в статье обсуждается вопрос об автоматическом редактировании и автоматическом анализе систем текстов. А также говорилось о проводимых исследованиях по развитию узбекского литературного языка.*

Ключевые слова: *программное обеспечение, спелл-чекер, стемминг, лемматизация, токенизация, парсер.*

AUTOLINGUISTIC TEXT PROCESSING SYSTEMS

M. Abjalova, Navoi State Mining Institute,
Navoi, Uzbekistan, manzura_ok@mail.ru

Annotation: *in this article discussed about automatic editing and automatic analysis systems of texts. And also, it was spoken about conducted researches on development of the Uzbek literary language.*

Key words: *software, spell-checker, stemming, lemmatization, tokenization, parser.*

МАТНЛАРГА АВТО-ЛИНГВИСТИК ИШЛОВ БЕРИШ ТИЗИМЛАРИ

М. Абжалова, Навоий давлат кончилик институти,
Навоий, Ўзбекистон, manzura_ok@mail.ru

Аннотация: *мазкур мақолада матнларни автоматик таҳрир ва таҳлил қилувчи тизимлар борасида сўз юритилди. Шунингдек, мазкур масалада ўзбек адабий тили доирасида олиб борилаётган изланишлар баён этилди.*

Таянч сўзлар: *дастурий таъминот, спелл-чекер, стемминг, лемматизация, токенизация, парсер.*

Матнларни лингвистик таҳрир қилиш жараёнига азал-азалдан асосий филологик масала сифатида қаралган бўлиб, ушбу вазифа йиллар давомида инсон томонидан амалга ошириб келинмоқда.

Матнларга ишлов беришнинг автоматик таҳрир йўналиши XX аср ўрталарига келиб ривожланди. У матн муҳаррир дастурлари билан биргаликда янгича имкониятлар билан ривожланмоқда. Оддий муҳаррирлардан фарқи шундаки, унда таҳрир автоматик тарзда қисқа вақт ичида катта ҳажмли матнлар тез текширилиб, хатоларни самарали тўғрилаш каби имкониятлари бўлади.

Лингвистик таҳрирлаш (корректура) – турли кўринишдаги (илмий, бадиий, публицистик ва расмий услублардаги) матнларнинг орфографик, грамматик, стилистик ҳамда мантиқий қурилишини текшириш демакдир.

Бугунги кунда матнларни таҳрир қилувчи ёхуд унинг таҳлилини ҳам амалга оширувчи автоматик лингво-тизимлар яратилган бўлиб, уларнинг ишлаш принциплари янада такомиллаштирилмоқда. Қуйида шундай тизимлар борасида сўз юритилди.

Имлони текшириш тизими (спелл-чекер ингл. spell checker) – компьютер дастури бўлиб, киритилган матннинг орфографик текширувини — таҳрирни амалга оширади. Аниқланган имло хатолари махсус тарзда белгиланади, яъни хато ёзилган лексеманинг тагига чизилади. Кўп ҳолларда матн терувчига имловий хатоларга ишора қилишдан ташқари дастур махсус эслатмаси сифатида сўзнинг тўғри ёзилиш вариантларини ҳам таклиф қилади. Шунингдек, матнга қандай тузатиш киритиш мумкинлигига изоҳлар ҳам берилади.

Биринчи матн териш текшируви тизимлари 1970-йилларнинг охирларида фойдаланила бошланган. Жоржтаун университетининг олти нафар тилшуносидан иборат гуруҳи IBM компанияси учун биринчи тизимни ишлаб чиқди. CP/M ва TRS-80 шахсий компьютерларига 1980 йилда ушбу тизим киритилган, 1981 йилда IBM PCда тизимнинг биринчи пакетлари пайдо бўлди.

Орфографик текширув (Speller) – матнни орфографик жиҳатдан тўлиқ текширувчи модул: унинг қулайлиги шундаки, келтирилган кўрсатмалар орқали матнни

киритувчи дастурнинг лингвистик таъминотига янги сўзлар, сўзшаклларни киритиб автоматик луғатни шу заҳоти бойитиши мумкин.

Лемматизация (lemmatization) – бу сўзшаклларини унинг луғатдаги оддий шакли – леммага келтириш жараёни[11].

Лемматизация сўзларнинг лексиконидан фойдаланиб, уларнинг морфологик таҳririни амалга оширувчи аниқ жараён бўлиб, лемматизация жараёнида фақат флекцияга учраган аффикслар учирилади ва лемма деб аталмиш таъминотдаги сўзнинг асосий ёки луғат шаклига қайтарилади.

Стемминг (stemming) – бу киритилган сўзнинг асоси (ўзак)ини топиш жараёни. Бунда топилган сўз асоси морфологияда қабул қилинган сўз ўзагига мос келиши талаб қилинмайди. Боиси дастур таъминотига «тайёр қолипли сўзлар» асос сифатида алоҳида категория қилиб киритилиш эҳтимоли юқори бўлади. Масалан, ҳуқуқ, тадқиқ каби лексемаларга эгалик қўшимчаси қўшилганда лексема сўнгидаги қ ундоши ўзгаришга учрамагани боис, бундай истисноли лексемалар «тайёр қолипли сўзлар» категориясига ҳуқуқи, тадқиқи каби киритилади ва улар стеммингда асос деб қабул қилинади.

Стемминг, одатда, тахминий жараён дейилади. Сабаби, стеммингда сўзшакллари аффикслари охиридан бошлаб лингвистик таъминотга киритилган асос шаклга қадар кесиб келинади. Кўп ҳолларда бу ҳолат ўзини оқлаган.

Стемминг бир асосдан юзага келган сўзшаклларидаги белгилар билан ишлайди, лемматизация эса бир лемманинг флектив (аффикс қўшилиши натижасида ўзгаришга учраган шакл) шаклини эътиборга олади.

«Токенизация» инглиз. tokenizing сўзидан олинган бўлиб, информатиклар томонидан киритилган атама ҳисобланади ва тилшуносликда лексик таҳлил бирикмаси билан тушунтирилади. Токенизация – бу киритилаётган белгилар кетма-кетлигини муайян гуруҳларга, яъни лексемаларга ажратиб, таҳлил қилиш жараёни. Ушбу жараён чиқишда «токен» («сўзлар билан гуруҳланган ҳарфлар») деб аталмиш идентификация қилинган кетма-кетликни олиш мақсадида амалга оширилади.

Парсер (инглиз. parser; parse – таҳлил қилиш) ёхуд синтактик анализатор — дастур қисми ёки фақат синтактик таҳлилни амалга ошириш учун яратилган махсус дастур. Кириш маълумотларини (матн) муайян форматга келтириш орқали таҳлил қилади.

Ўзбек тилидаги матнларни автоматик таҳрир қилиш дастурини яратишда, албатта, тилшунослар томонидан яратиладиган лингвистик маълумотлар базаси муҳим аҳамиятга эга. Бу борада кўп босқичли ўзбек тилининг таҳрир ва таҳлил дастурининг лингвистик таъминоти устида иш олиб борилмоқда. Матнларни автоматик равишда таҳрир қилувчи дастурларнинг лингвистик модулларининг яратилиши мукаммал дастурларнинг ишлаб чиқилишига замин яратади, бу эса ўзбек адабий тилидаги матларнинг саводли ёзилиши ва дунё тиллари қаторидан ўрин олишига хизмат қилади.

АДАБИЁТЛАР:

1. Русский орфографический словарь: ок. 200 000 слов / РАН. Ин-т рус. яз. им. В. В. Виноградова / Под ред. В. В. Лопатина, О. Е. Ивановой. — Изд. 4-е, испр. и доп. — М.: АСТ-Пресс Книга, 2012. — С. 709. — (Фундаментал. словари рус. яз.). — ISBN 978-5-462-01272-3.

[11] Атамаларга айнан изоҳ беришда https://ru.wikipedia.org/wiki/ сайтидаги маълумотлардан фойдаланилди.