УДК 81’1:004.43
DOI: 10.36979/1694-500X-2024-24-2-99-102

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА КАК ОСОБЫЙ ИНСТРУМЕНТАРИЙ ОБРАБОТКИ ЯЗЫКОВЫХ ДАННЫХ

Г.Э. Жумалиева, С.А. Кайымова, Ж.Т. Асанова

Аннотация. Важной и перспективной задачей компьютерной лингвистики является создание лингвистических процессов для обеспечения связи с интеллектуальными автоматизированными информационными системами с использованием естественного языка или языка, близкого к естественному. Это автоматическое обнаружение и исправление ошибок при вводе текстов в ЭВМ при решении ряда прикладных задач: автоматический перевод текста с одного языка на другой, подключение компьютера к естественному языку, автоматическая классификация и индексация текстовых документов, их автоматическое суммирование, поиск документов в полнотекстовых базах данных.

Ключевые слова: компьютерная лингвистика; текстовая информация; машинное обучение; словари; чат-боты; голосовой перевод; машинный перевод; искусственный интеллект; естественный язык.

КОМПЬЮТЕРДИК ЛИНГВИСТИКА ТИЛДИК МААЛЫМАТТАРДЫ ИШТЕТҮҮНҮН АТАЙЫН КУРАЛЫ КАТАРЫ

Г.Э. Жумалиева, С.А. Кайымова, Ж.Т. Асанова

Аннотация. Компьютердик лингвистиканын маанилүү жана келечектүү милдети табигый тилди же табигый тилге жакын тилди колдонуп, интеллектуалдык автоматташтырылган маалымат системалары менен байланышты камсыз кылуу үчүн лингвистикалык процесстерди түзүү болуп эсептелет. Бир тилден экинчи тилге автоматтык түрдө бир катар колдонмо маселелерди чечүүдө компьютерге тексттерди киргизүүдө каталарды автоматтык түрдө табуу жана оңдоо тексттерди которууда, компьютерди табигый тилге туташтырууда, тексттик документтерди автоматтык түрдө классификациялоодо жана индекстөөдө, аларды автоматтык түрдө суммалоодо, толук тексттик маалымат базаларында документтерди издөөдө колдонулат.

Түйүндүү сөздөр: компьютердик лингвистика; тексттик маалымат; машина үйрөнүү; сөздүктөр; чат-боттор; үн котормолору; машиналык котормо; жасалма интеллект; табигый тил.

COMPUTER LINGUISTICS AS A SPECIAL TOOL FOR PROCESSING LANGUAGE DATA

G.E. Zhumalieva, S.A. Kaiymova, Zh.T. Asanova

Abstract. An important and promising task of computational linguistics is the creation of linguistic processes to ensure communication with intelligent automated information systems using natural language or a language close to natural. The issue of automatic detection and correction of errors when entering texts into a computer when solving a number of applied problems, automatically from one language to another. It is used when translating texts, connecting a computer to natural language, automatic classification and indexing of text documents, their automatic summation, searching for documents in full-text databases.

Keywords: computational linguistics; textual information; machine learning; dictionaries; chatbots; voice translation; machine translation; artificial intelligence; natural language.

Киришүү. XX кылымдын ортосунда компьютерлердин жаралышы жана алардын тез өнүгүшү, мурда элестетүүгө мүмкүн болбогон жаңы илимдердин пайда болушуна түрткү болду. Анын мүмкүнчүлүгүнө жараша бири-бири менен байланышы жок илимдердин кайчылашкан жеринде пайда болгон. Ошентип, биология жана инженерия кесилишинде бионик пайда болгон.

Ал эми компьютердик лингвистика илими атын бир нече жолу өзгөрткөн илим катары, адегенде математикалык лингвистика, андан кийин структуралык лингвистика жана XX кылымдын акыркы жылдарында гана компьютердик лингвистика тил илими деп аталып калды. Акыры, анын компаниясы анын артына бекем орноду. Жаңы илимдин пайда болушуна эки себеп болгон. Биринчиден, лингвистика боюнча изилдөөчүлөр азыркы так илимдер жана биринчи кезекте математика тил илимине жетишпеген тактыкка ээ болууга жардам берет деп үмүт кылышкан. Компьютердин пайда болушу бул үмүттөрдү бекемдеди, анткени компьютерлер "тез иштеген арифмометрлер" гана эмес, тексттер менен иштөөнү автоматташтыруунун кубаттуу куралы экендиги эң башынан эле көптөгөн лингвисттерге түшүнүктүү болгон. Тексттерди статистикалык иштетүү, ар түрдүү сөздүк жана лексикалык файл шкафтарын жүргүзүү сыяктуу көптөгөн эмгекти көп талап кылуучу процесстерди автоматташтыруу мүмкүн болду. Экинчиден, компьютерлердин пайда болушу менен даярдыгы жок колдонуучулардын алар менен баарлашуу маселеси дароо эле пайда болду. Албетте, мындай колдонуучулар үчүн эң жакшы форма тааныш табигый тил болушу мүмкүн. Бирок мындай өз ара аракеттенүүнү уюштуруу үчүн эң оболу адамдардын ортосундагы баарлашуу процессинде табигый тилдин колдонулушунун мыйзамдарын жана өзгөчөлүктөрүн түшүнүү керек [1].

Компьютердик лингвистиканын өнүгүүсүнө серептеме

Компьютердик лингвистика – табигый тилди компьютерде моделдеп иштетүүчү колдонмо лингвистиканын бир тармагы. Тилди компьютерде моделдөө жагынан информатика, жасалма интеллект, программалар теориясы өңдүү кибернетикалык багыттар менен тогошот. Компьютердик лингвистика табигый тилде берилген маалыматты автоматтык түрдө иштетүү маселелерин чечүүгө байланышкан билимдердин тармагы, башкача айтканда, аны колдонмо маселелерин: тексттерди компьютерге киргизүүдө, каталарды автоматтык түрдө аныктоо жана оңдоо, оозеки кепти автоматтык түрдө анализдөө жана синтездөө, тексттерди бир тилден экинчи тилге автоматтык түрдө которуу, тексттик документтерди автоматтык түрдө классификациялоо жана индекстөө, аларга автоматтык шилтеме берүү жана толук тексттик маалымат базаларында документтерди издөө, тил менен башка чөйрөлөрдүн байланышы сыяктуу максаттарга багытталган. Жалпысынан, компьютер аркылуу ишке ашырылган тил илиминдеги бардык багыттар компьютердик лингвистиканын объектиси болуп эсептелет. 1950-жылдары атактуу америкалык лингвист, публицист жана философ Ноам Хомскийдин табигый тилдин түзүлүшүн формалдаштыруу боюнча изилдөөсүнөн, ошондой эле машиналык котормодогу сыноо эксперименттеринен жана табигый тилди түшүнүү үчүн биринчи AI программаларын түзүп баштоо зарылдыгы айтылган. Кээ бир өлкөлөрдө бир тилден экинчи тилге тексттерди машинанын жардамы менен которуунун эксперименталдык жана өндүрүштүк системдер, атап айтканда, Россия, США, Германия, Франция сыяктуу өлкөлөрдө табигый тилдин компьютер менен байланышы боюнча бир катар эксперименталдык системдер түзүлүп, терминологиялык маалымат базаларын, тезуарустарды, эки тилдүү жана көп тилдүү машина сөздүктөрүн түзүү боюнча иш чаралары жүргүзүлүүдө, кепти автоматтык талдоо жана синтездөө системдер курулууда. Ал эми Россия, США, Япония өлкөлөрүндө табигый тилдердин моделдердин түзүү тармагында изилдөөлөр XX кылымдын экинчи жарымынан баштап жүргүзүлүп жатат [2].

Компьютердик лингвистика 1954-жылы январь айында Джорджтаун университетинде пайда болгон деп айтууга болот, анткени Америка Кошмо Штаттарында алгачкылардан болуп аталган университетте машина которуу боюнча дүйнөдө биринчи коомдук эксперимент өткөрүлгөн. Инженерлер 60тан ашык сүйлөмдөрдү орус тилинен англис тилине толук автоматтык режимде которууга жетишишкен. 1980-жылдардын аягында Интернеттин өнүгүшү менен электрондук формадагы тексттердин көлөмү кескин көбөйдү, бул маалыматты издөөдө сандан сапаттык секирикке алып келди [3].

Компьютердик лингвистика – негизинен тил маалыматтарын иштетүүчү негизги каражат.

Тексттерди автоматтык иштетүү – компьютердин жардамы менен текстти жасалма же табигый тилге которуу процесси. Автоматташтырылган оңдоо иштери, текстти трансформациялоо компьютердин эсинде текстке оңдоолорду жана толуктоолорду киргизүүдөн турат. Текстти форматтоо темаларды бөлүп көрсөтүүдөн керектүү форматтагы саптарды жана баракчаларды түзүүдөн аны компьютердик басып чыгаруу түзүлүштөрүндө кайра чыгаруу үчүн тексттин бөлүмдөрү менен бөлүп көрсөтүүдөн жана долбооорлордон турат. Компьютердик лингвистикасы өзгөчө колдонмо дисциплина катары биринчи кезекте өзүнүн куралдары менен айырмаланат. Бүгүнкү күндө компьютердик лингвисттер табигый тилди иштетүү программаларын, текстти таануу жана кеп таануу куралдарын, котормо системаларын, текст редакторлорун, тил үйрөнүү материалдарын, үн жардамчыларын, акылдуу чат-ботторду жана башкаларды иштеп чыгышат. Текстти интеллектуалдык автоматтык иштетүүгө болгон муктаждык негизинен эки себеп менен пайда болот, алардын экөө тең өндүрүлгөн тексттердин көлөмүнө байланыштуу [4].

Компьютердик лингвистиканы иштетүү объектиси табигый тилдеги тексттер болгондуктан, анын өнүгүшүн жалпы тил илими жаатындагы базалык билимсиз элестетүү мүмкүн эмес. Ошол эле учурда биринчи машина үйрөнүү алгоритмдери жана статистикалык машина которуу системалары түзүлүп, 2010-жылдан баштап тилди иштетүү тармагындагы терең үйрөнүү алгоритмдери өнүгө баштады. Ошондон бери компьютердик лингвистика илиминин маселелерин чечүү үчүн көптөгөн өнүгүүлөр пайда болуп, иштелип чыгууда [5].

Компьютердик лингвистиканын алдында турган негизги маселелеринин эң маанилүүлөрүнө төмөнкүлөр кирет:

1) машиналык сөздүктөрдүн лингвистикалык иштетүү жана түзүү жана аларды автоматташтыруу;
2) тексттерди компьютерге киргизүүдөгү каталарды табуу жана оңдоо процессин автоматташтыруу;
3) маалыматтык суроо-талаптарды жана документтерди автоматтык түрдө индекстөө;
4) автоматтык түрдө документтерди классификациялоо жана жалпылоо;
5) маалыматтарды бир тилдик жана көп тилдик маалымат базалардан издөө процесстерин лингвистикалык камсыздоо;
6) тексттерди бир тилден экинчи тилге машинанын жардамы менен которуу;
7) колдонуучулардын автоматташтырылган интеллектуалдуу маалымат системдери менен байланышты камсыз кылган лингвистикалык процесстерди куруу;
8) фактографиялык маалыматтарды формалдаштырылбаган тексттерден бөлүп алуу.

Тексттик маалыматтарды автоматтык түрдө иштетүүчү системдердин дээрлик баарында машиналык сөздүктөр алардын ажырагыс бир бөлүгү болуп саналат. Алар туруктуу илимий техникалык түшүнүктөрдү билдирүүчү сөз же сөз айкалыш сөздүктөрү болушу мүмкүн. Сөздүктөрдү түзүүдө тексттердин лексикалык түзүүнүн жогорку даражада чагылдырууга аракет кылат.

Компьютердик лингвистикада түзүлгөн жана колдонулган колдонмо линвистикалык каражаттарды шарттуу түрдө эки бөлүккө бөлсө болот: декларативдик жана процедуралык. Декларативдик бөлүккө тил жана кеп бирдиги, сөздүктөр, тексттер жана ар кандай түрдөгү грамматикалык таблицалар кирет.

Процедуралык бөлүккө тексттерди жана грамматикалык таблицалардын, тилдин жана кеп бирдиктерин башкаруу каражаттары кирет.

Компьютердик лингвистиканын колдонмо маселелерин чечүүдөгү ийгилик баарынан мурда декларативдик каражаттардын компьютердин эс тутумуна толук жана так берилишинен көз каранды. Азыркы күндө бул жаатгагы маселелер керектүү деңгээлде чечилбегендиктен бардык өнүккөн мамлекеттерде компьютердик лингвистика тармагындагы көптөгөн иш чаралар жүргүзүлүп жатат. Ошентсе да, компьютердик лингвистика тармагындагы олуттуу илимий жана практикалык жетишкендиктерди белгилей кетсе болот [6].

Корутунду. Акыркы учурда колдонмо лингвистиканын негизги өркүндөп жаткан багыты катары компьютердик лингвистика эсептелет. Колдонмо лингвистиканын бул чөйрөсү белгилүү

бир шарттарда, кырдаалдарда, көйгөйлүү аймактарда жана башкаларда тилдин иштешин моделдөө үчүн маалыматтарды уюштуруу жана иштетүү үчүн компьютердик каражаттарды – программаларды, компьютердик технологияларды колдонууга багытталган. Компьютердик лингвистиканын колдонуу чөйрөлөрүнө жана багыттарына төмөнкүлөр кирет: корпустук лингвистика электрондук текст корпусун түзүү жана пайдалануу, электрондук сөздүктөрдү, тезаурус, онтологияларды түзүү, тексттерди автоматтык которуу, автоматтык фактыларды алуу, текстти терендеп иликтөө, автореферaттоо. Бул функция, мисалы, билимдин башкаруу системаларын куруу, эксперттик системаларды кароо, суроолорго жооп берүү системаларын түзүү ж.б.у.с.

Поступила: 14.11.23; рецензирована: 28.11.23;
принята: 01.12.23.

Адабияттар

1. *Асанов У.А.* “Кыргызстан” улуттук энциклопедиясы / У.А. Асанов. Бишкек: Мамлекеттик тил жана энциклопедия борбору, 2012. 832 б.
2. *Чеповский А.* Неразрешимая проблема компьютерной лингвистики / А. Чеповский // Компьютера. М., 2002. № 30. С. 3–7.
3. *Ермаков А.Е.* Синтаксический разбор в системах статистического анализа текста / А.Е. Ермаков, В.В. Плешко // Информационные технологии. М.: Мир ПК, 2002. № 7.
4. *Жумалиева Г.Э.* Кыргыз тилинин компьютердик лингвистикасынын негиздери / Т. Садыков, Г.Э. Жумалиева, М.Ж. Түмөнбаева, Б.Ш. Шаршембаев. Бишкек: Имак Офсет, 2015. 400 б.
5. *Белоногов Г.Г.* Компьютерная лингвистика и перспективные информационные технологии: теория и практика построения систем автомат. обработки текстовой информации / Г.Г. Белоногов, Ю.П. Калинин, А.А. Хорошилов. М.: Информац.-издат. агентство “Русский мир”, 2004. 246 с.
6. Компьютерная лингвистика / А.Н. Баранов // Большая российская энциклопедия: в 35 т. / гл. ред. Ю.С. Осипов. М.: Большая российская энциклопедия, 2004–2017.