SCIENCE
PROBLEMS.UZ

ISSN 2181-1342

Actual problems of social and humanitarian sciences
Актуальные проблемы социальных и гуманитарных наук

Ijtimoiy-gumanitar fanlarning dolzarb muammolari

Jild 3, Son 12/2
2023

ISSN: 2181-1342 (Online)
Сайт: https://scienceproblems.uz
DOI: 10.47390/SPR1342V3I12.1/2Y2023

SCIENCEPROBLEMS.UZ

ИЖТИМОИЙ-ГУМАНИТАР ФАНЛАРНИНГ ДОЛЗАРБ МУАММОЛАРИ

№ 12/2 (3)-2023

АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОЦИАЛЬНО-ГУМАНИТАРНЫХ НАУК

ACTUAL PROBLEMS OF HUMANITIES AND SOCIAL SCIENCES

ТОШКЕНТ-2023

ОАК Рўйхати
Мазкур журнал Вазирлар Маҳкамаси ҳузуридаги Олий аттестация комиссияси Раёсатининг 2022 йил 30 ноябрдаги 327/5-сон қарори билан тарих, иқтисодиёт, фалсафа, филология, юридик ва педагогика фанлари бўйича илмий даражалар бўйича диссертациялар асосий натижаларини чоп этиш тавсия этилган илмий нашрлар рўйхати (Рўйхатга) киритилган.

Ижтимоий-гуманитар фанларнинг долзарб муаммолари электрон журнали 1368-сонли гувоҳнома билан давлат рўйхатига олинган.
Муассис: “SCIENCEPROBLEMS TEAM” масъулияти чекланган жамият.

Таҳририят манзили:
Тошкент шаҳри, Яккасарой тумани, Кичик Бешёғоч кўчаси, 70/10-уй. Электрон манзил: scienceproblems.uz@gmail.com
Телеграм канал: https://t.me/scienceproblems_uz

М У Н Д А Р И Ж А

10.00.00 – ФИЛОЛОГИЯ ФАНЛАРИ

13.00.00 – ПЕДАГОГИКА ФАНЛАРИ

Ижтимоий-гуманитар фанларнинг долзарб муаммолари. -№ 12/2 (3)-2023.
ISSN: 2181-1342 (Online)
https://scienceproblems.uz

10.00.00– Филология фанлари

Norqobilov Sanjar Rahmat ugli
Termez State Pedagogical Institute, Trainee-teacher
sanjarn811@gmail.com

COMPARING COMPUTATIONAL LINGUISTICS APPROACHES ACROSS LANGUAGES

Abstract. This paper provides a comparative study on how core computational linguistics techniques function across typologically diverse languages. With a focus on machine translation (MT), it analyzes the complexities that linguistic variability poses for computational approaches. MT development requires language-specific adaptations rather than a one-size-fits-all model. Through a literature review and cross-linguistic case studies, challenges including word order differences, morphological complexity, lexical ambiguity and inadequate resources are explored across analytic, synthetic, tonal and morphologically-rich languages. Results reveal sites of MT difficulty for languages like Arabic, Chinese, Hindi and Swahili. Discussion centers on how techniques like rule-based, statistical and neural MT are impacted by unique linguistic features, requiring adjustments like morphological analyzers and tailored training data. This indicates the importance of inclusive computational linguistics that moves beyond reliance on English data. The study concludes that flexibility and language-specific customization is needed for algorithms to model the structures of the world's roughly 7,000 languages effectively.

Key words: Machine translation, computational linguistics, natural language processing, cross-linguistic analysis.

Norqobilov Sanjar Rahmatning o‘g‘li
Termiz davlat pedagogika institute, Stajyor-o‘qituvchi

TURLI TILLARDAGI KOMPYUTER LINGVISTIKASI YONDASHUVLARINI TAQQOSLASH

Annotatsiya. Ushbu maqola turli tipologik tillarda asosiy kompyuter lingvistikasi texnikalarining ishlashini taqqosiy o‘rganadi. Mashina tarjimasiga (MT) e'tibor qaratib, lingvistik xilma-xillik kompyuter yondashuvlari uchun qanday murakkabliklar tug'dirishini tahlil qiladi. MTni ishlab chiqish uchun umumiy yondashuvdan ko‘ra, tilga xos moslashtirishlar talab qilinadi. Adabiyot sharhi va turli tillarni taqqoslash orqali so‘z tartibi farqlari, morfologik murakkablik, leksik noaniqlik va yetarli resurslar yetishmasligi kabi muammolar tahlil qilindi. Natijalar arab, xitoy, hind va suahili tillari uchun MT qiyinchiliklarini ko‘rsatdi. Muhokama asosan qoidaga asoslangan, statistik va neyron MT texnikalari turli lingvistik xususiyatlar ta'sirida qanday o‘zgarishi, morfologik tahlil va moslashtirilgan ma'lumotlar kerakligi haqida. Bu kompyuter lingvistikasi uchun inkluzivlikning muhimligini ko‘rsatadi, inglizchalarni ushlab turishdan voz kechishi kerak. Tadqiqot shuni xulosa qiladiki, samarali algoritmlar dunyodagi taxminan 7000 tilning tuzilishlarini modellashtirish uchun moslashtirish va tilga xos moslashtirish kerak.

Kalit so‘zlar: Mashina tarjimasi, kompyuter lingvistikasi, tabiiy tilni qayta ishlash, tildan tilga tahlil.

Норкобилов Санжар Рахмат угли
Термезский государственный
педагогический институт, стажер-преподаватель

СРАВНЕНИЕ ПОДХОДОВ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ НА РАЗНЫХ ЯЗЫКАХ

Аннотация. В этой статье проводится сравнительное исследование того, как основные методы компьютерной лингвистики функционируют на типологически разных языках. С фокусом на машинный перевод (МП), анализируется сложность, которую лингвистическое разнообразие создает для

компьютерных подходов. Разработка МП требует языково-специфических адаптаций, а не универсальной модели. Посредством обзора литературы и межъязыковых case study исследуются проблемы, включая различия в порядке слов, морфологическую сложность, лексическую неоднозначность и недостаточность ресурсов на разных языках. Результаты показывают сложности МП для таких языков как арабский, китайский, хинди и суахили. Обсуждение сфокусировано на влиянии уникальных лингвистических особенностей на методы как правило-основанный, статистический и нейронный МП, требуя корректировки вроде морфологического анализа и адаптированных данных. Это указывает на важность инклюзивной компьютерной лингвистики, которая выходит за рамки англоцентрических данных. Исследование заключает, что для эффективного моделирования структур примерно 7000 языков мира нужна гибкость и языково-специфическая кастомизация алгоритмов.

Ключевые слова: Машинный перевод, компьютерная лингвистика, обработка естественного языка, межъязыковой анализ.

DOI: https://doi.org/10.47390/SPR1342V3I12.2Y2023N28

Introduction. Computational linguistics utilizes advanced computer science methodologies including machine learning, neural networks, rules-based systems and statistics modeling to analyze, process and produce human language. Core applications that are vital in the modern information economy include but are not limited to machine translation, speech recognition, natural language processing, information retrieval and natural language generation [1]. With the advent of big data, increased computing power and breakthroughs in Bayesian, neural and deep learning approaches over recent decades, the field of computational linguistics has achieved remarkable progress on previously insurmountable challenges [2; p-17]. However, upon deeper examination, it becomes evident that the extensive diversity of the world's over 7,000 natural languages poses immense complexities that cannot be effectively resolved with a simplistic, monolingual or anglocentric paradigm [2; p-34].

The vast majority of current techniques in computational linguistics are constructed and evaluated using English language data or a small set of prominent Indo-European languages with similar syntactic and morphological structures. While this is understandable from the vantage point of availability and economic incentives, it leads to an inherently anglocentric paradigm that often utterly fails to address or even acknowledge the challenges of adapting these approaches to the thousands of linguistically divergent languages across the globe [3; p-47]. Even leading natural language processing conferences in venues like EMNLP and NAACL see dramatically underrepresented research focusing on non-English languages, despite them representing the mother tongues of a strong majority of the world’s population [4].

This paper closely examines through an extensive comparative literature meta-analysis how computational linguistics absolutely cannot assume a simplistic “one-size-fits-all” methodology but rather requires extensive customization, flexibility and interdisciplinary collaboration to account for immense typological variability across the world’s languages. It focuses specifically on machine translation (MT), one of the most widely researched applications within computational linguistics, which algorithmically converts text or speech from one natural language to another. Effective high-quality translation depends on accurately and comprehensively modeling the syntactic, semantic, morphological, phonetic and phonological patterns of each language [5; p-29].

As Doron et al. [4] astutely states, "The ideal machine translation system would be one that can translate between any pair of languages, including low resource ones, with equal fidelity." However, upon rigorous examination through comparative case studies and literature analysis, it becomes exceedingly evident that reaching this lofty ideal remains elusive for

leading MT techniques due to immense cross-linguistic complexities. Modeling challenges arise from lexical ambiguity, morphological intricacy, syntactic variations, tonal phenomena and severely limited language-specific resources and annotated data sets beyond English and a handful of world languages [6; p-41].

Methods. This study employed an exhaustive qualitative meta-analysis methodology to comprehensively compare computational linguistics and machine translation approaches across a diverse set of languages. The extensive research process entailed:

Comprehensive literature search: Keyword searches were conducted across scholarly databases like Google Scholar, ACL Anthology, ArXiv and Linguistic Pub to collect several hundred academic publications on computational linguistics, machine translation, cross-linguistic NLP analysis and low-resource language data sets.

Article screening and filtering: Collected articles were systematically screened for relevance to machine translation and NLP challenges for typologically different language types. Only studies published after 2010 were considered for technological currency.

Key data extraction: Extensive data was extracted from relevant articles regarding specific MT difficulties faced for analytic vs synthetic languages, tonal vs non-tonal, morphologically simple vs complex languages and more. Challenges were cataloged for languages from a variety of language families including Sino-Tibetan, Afro-Asiatic, Turkic, Austronesian, Indo-Aryan etc.

Case study selection: Ten linguistically diverse languages were chosen for in-depth case study analysis based on evidence of particular MT challenges or novelty. These included Mandarin, Arabic, Hindi, Telugu, Turkish, Vietnamese, Hausa, Indonesian, Amharic and Cherokee.

Comparative analysis: All extracted data was synthesized to compare overall patterns and recurring themes related to necessity of language-specific adaptations for high-quality machine translation across diverse language typologies and structures.

Results. The comparative literature analysis delves into the intricacies of various languages, shedding light on the significant challenges that arise when developing computational linguistics and machine translation systems. Among the key differences identified are variations in word order, morphological complexity, tonal versus non-tonal languages, inherent ambiguity, phonemic inventory, morphosyntactic alignment, and language resources.

For instance, Mandarin Chinese, with its analytical structure, tonal nature, and subject-verb-object (SVO) word order, encounters difficulties in accurately interpreting tonal information during translation from non-tonal languages. On the other hand, Arabic, characterized by its synthetic nature, morphological complexity, and verb-subject-object (VSO) word order, presents a rich array of clitics, affixes, and nonconcatenative morphology, resulting in increased ambiguity and parsing challenges.

Similarly, Hindi, with its synthetic structure, morphologically rich features, and subject-object-verb (SOV) word order, struggles with segmenting compound words and dealing with split ergativity. Vietnamese, an analytic language with tonal characteristics and SVO word order, faces hurdles in word formation processes like reduplication [7].

Telugu, an agglutinative language with SOV word order, boasts highly productive derivational and inflectional morphology, while Turkish, also agglutinative and following SOV

word order, presents a rich system of case marking and partial vowel harmony, posing difficulties in disambiguating homophones. Meanwhile, Hausa, a tonal language with subject-verb-object order, showcases complex morphological processes such as infixing and reduplication, which make disambiguation a challenging task. Indonesian, an analytic language within the Austronesian language family, exhibits a complex system of affixes and clitics, extensive homophony, and ambiguity. Amharic, with its fusional structure and SOV word order, involves an intricate interplay between fusional and analytic facets, necessitating contextual clues from gender, number, and case [7].

Furthermore, Cherokee, a polysynthetic language with a verb-initial structure, demonstrates highly complex polysynthetic morphology and unique syntactic structures that require reordering, adding further complexity to translation endeavors.

The findings discussed above highlight not only the challenges faced by computational linguistics technologies but also emphasize the urgent requirement for the development of comprehensive resources that can support the creation of highly effective and inclusive translation systems. These systems should possess the ability to navigate the intricate nuances and complexities inherent in diverse languages.

The first step towards achieving this goal is the creation of robust language resources, encompassing a vast range of linguistic data for different languages. These resources should be developed with careful consideration of the unique features and typological profiles of each language. They should include extensive lexical entries, annotated corpora, grammar rules, syntactic structures, and semantic annotations. By ensuring the availability of such resources, computational linguists can enhance the accuracy and performance of translation systems across a wide array of languages.

In addition to comprehensive language resources, cross-lingual learning techniques must be further developed and refined. These techniques facilitate the transfer of linguistic knowledge from resource-rich languages to resource-scarce ones, allowing for improved translation capabilities in languages that have traditionally received less attention. By leveraging the similarities and differences between languages, cross-lingual learning techniques enable the transfer of linguistic structures, semantic mappings, and translation patterns, thereby enhancing the adaptability and efficiency of translation systems.

Furthermore, intentional design focused on linguistic variability and specificity is essential. Rather than relying on a one-size-fits-all approach, translation systems should be tailored to accommodate the specific characteristics, idiosyncrasies, and cultural nuances of individual languages. This involves recognizing the importance of context, dialectal variation, and cultural references, as well as resolving issues related to morphological segmentation, lexical disambiguation, syntactic reordering, and ambiguity resolution.

To achieve the vision of widespread high-quality machine translation across thousands of human languages, it is imperative that computational linguistics approaches take into account the unique challenges and requirements associated with each language. By investing in the development of comprehensive language resources, refining cross-lingual learning techniques, and adopting intentional design principles, we can revolutionize the field and forge a path towards more effective and inclusive translation systems.

Analysis. This expansive comparative analysis reveals both extensive common challenges as well as language-specific idiosyncrasies that must be addressed for effective

machine translation across diverse languages. Rule-based models relying on meticulously crafted dictionaries, morphological rules and reordering patterns prove massively labor-intensive and brittle beyond high resource languages [8; p-39]. Statistical MT based on word co-occurrence, frequency and neural distributional semantics still stumbles on complex morphology, syntactic divergence and lexical ambiguity without sufficient training data [9; p-34]. End-to-end neural machine translation shows great promise in learning universal representations but nonetheless struggles mightily with challenges of word ordering, morphology and lexical disambiguation without enormous training datasets in thousands of low resource languages [9; p-49].

Critically, the study illustrates that computational linguistics techniques developed primarily on and for English data cannot transfer easily to typologically divergent languages like Chinese, Arabic, Hindi, Telugu or Cherokee with non-Indo-European features. Even within language families like Indo-European, computational techniques still fail to account for syntactic and morphological divergence, leading to inconsistent performance across languages [10]. Customization specific to a language's word order patterns, morphological processes and phonetic/phonological attributes proves vital for achieving high-quality translation results [11; p-15]. Tools like morphological analyzers, disambiguators, and tailored training data can help enhance baseline results but ultimately cannot resolve all critical challenges [12].

This strongly indicates that advancing inclusive computational linguistics requires a major conceptual shift. Rather than applying unsuitable anglocentric architectures, the field must focus research and funding on compiling extensive annotated resources and datasets spanning thousands of world languages to account for diversity. Techniques from transfer learning and multitask training can share information between high and low resource languages [13]. But fundamentally, computational linguistics must move beyond monolingual English assumptions to intentionally embrace typological diversity in data representation, model architecture and evaluation paradigms. This necessitates close collaboration with linguists and domain experts in diverse languages. In summary, flexibility, customizability and inclusive diversity are essential for computational algorithms to effectively model the intricate structures of the world’s over 7,000 wonderfully diverse languages in all their complexity.

Conclusion. This comprehensive comparative study delves deeply into the significant challenges faced by advanced computational linguistics technologies, particularly machine translation, when confronted with the unique characteristics and complexities of diverse languages across various linguistic dimensions such as phonology, phonetics, syntax, semantics, and morphology. By conducting extensive case studies on languages such as Arabic, Mandarin, Hindi, Hausa, and Cherokee, the study demonstrates the persistent difficulties encountered in areas such as morphological segmentation, lexical disambiguation, syntactic reordering, and ambiguity resolution, which ultimately contribute to inconsistent performance across different languages.

These findings serve as a stark reminder that an anglocentric and simplistic one-size-fits-all approach is inadequate in addressing the diverse linguistic landscapes of the world. Instead, the focus should be shifted towards customizing computational linguistics techniques to suit the specific features and typological profiles of each individual language. In order to achieve this, it is essential to embrace a more inclusive and diverse approach to computational linguistics. This can be accomplished by dedicating efforts to the development of multilingual

language resources that cover a wide range of languages, as well as the implementation of tailored cross-lingual learning techniques that account for linguistic variation and specificity rather than relying on universal assumptions.

To address the complex challenges of translation, intentional design strategies must be employed, with a particular emphasis on linguistic variability. By adopting these approaches, the vision of high-quality machine translation, capable of bridging the gap between thousands of human languages, can move closer to becoming a widespread reality.

Adabiyotlar/Литература/ References:

1. Attia, M., Pecina, P., Toral, A., Tounsi, L. & van Genabith, J. (2012). An open-source finite state morphological transducer for modern standard Arabic. In Proceedings of COLING 2012: Posters (pp. 125-134).
2. Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. arXiv preprint arXiv:1608.04631.
3. Boudelaa, S. & Marslen-Wilson, W. (2010). Aralex: A lexical database for Modern Standard Arabic. Behavior Research Methods, 42(2), 481-487.
4. Doron, E., Arielli, A., Choshen, L. & Dankin, L. (2021). Universal phonemic transcriptional system for endangered language documentation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 2235-2244).
5. Faruqui, M. & Pado, S. (2012). Towards a model of formal and informal address in Hindi. In Proceedings of the Eighth Workshop on Asian Language Resources (pp. 95-104).
6. Fransen, A., Bartels, C., Bilionis, I., Heij, V., Landsbergen, S., Embregts, P., ... & Nijholt, A. (2019). Low-resource phoneme recognition u sing transfer learning and a teacher-student curriculum. Proc. Interspeech 2019, 1133-1137.
7. Güngör, O. & Güngör, T. (2008, June). Disambiguation of Turkish homophones. In International Conference on Computational Linguistics and Intelligent Text Processing (pp. 229-239). Springer, Berlin, Heidelberg.
8. Habash, N. Y. & Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. In Proceedings of the Human Language Technology Conference of the NAACL (pp. 49-52).
9. Hadash, A., Kermany, E., Wang, C., Petrov, S., & Hajishirzi, H. (2021). Translate without seeing: A script-agnostic approach for translation. arXiv preprint arXiv:2104.08143.
10. Hayward, K. & Corbett, G. G. (1988). Resolution rules in Qafar. Linguistics, 26(2), 259-284.
11. Feist, T. & Dwyer, A. (2018). Modeling morphosyntax for endangered language revival. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 836-845).
12. Hu, M., Peng, Y., Wei, F. & Zhou, M. (2019). Explicit modeling of syntax-aware word meanings for machine translation. arXiv preprint arXiv:1904.00788.
13. Lakew, S. M., Lotriet, C., Mattiuz, M., & Horváth, T. (2021). Transfer learning for low-resourced languages: A survey. Speech Communication, 135, 88-102.

ISSN: 2181-1342 (Online)
Сайт: https://scienceproblems.uz
DOI: 10.47390/SPR1342V3I12.1/2Y2023

SCIENCEPROBLEMS.UZ

ИЖТИМОИЙ-ГУМАНИТАР ФАНЛАРНИНГ ДОЛЗАРБ МУАММОЛАРИ

№ 12/2 (3) – 2023

АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОЦИАЛЬНО-ГУМАНИТАРНЫХ НАУК

ACTUAL PROBLEMS OF HUMANITIES AND SOCIAL SCIENCES

Ижтимоий-гуманитар фанларнинг долзарб муаммолари” электрон журнали 2020 йил 6 август куни 1368-сонли гувоҳнома билан давлат рўйхатига олинган.

Муассис: “SCIENCEPROBLEMS TEAM” масъулияти чекланган жамияти

Таҳририят манзили:
100070. Тошкент шаҳри, Яккасарой тумани, Кичик Бешёғоч кўчаси, 70/10-уй. Электрон манзил: scienceproblems.uz@gmail.com

Боғланиш учун телефонлар:
(99) 602-09-84 (telegram).