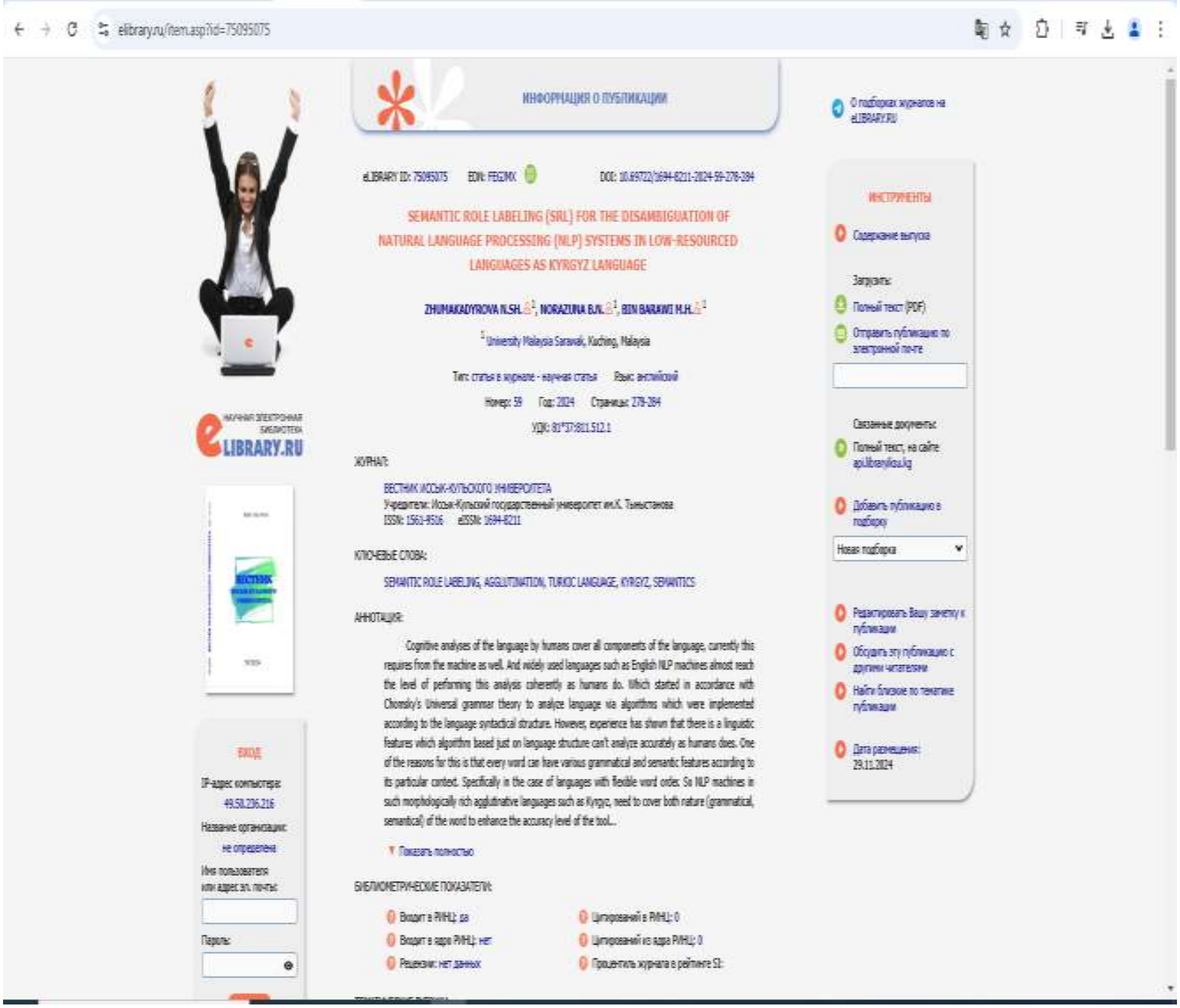

elibrary.ru/item.asp?id=75095075

ИНФОРМАЦИЯ О ПУБЛИКАЦИИ

eLIBRARY ID: 75095075 EDN: FEGZMX DOI: 10.69722/1694-8211-2024-59-278-284

SEMANTIC ROLE LABELING (SRL) FOR THE DISAMBIGUATION OF NATURAL LANGUAGE PROCESSING (NLP) SYSTEMS IN LOW-RESOURCED LANGUAGES AS KYRGYZ LANGUAGE

ZHUMAKADYROVA N.SH.[1], NORAZUNA B.N.[1], BIN BARAWI M.H.[1]

[1] University Malaysia Sarawak, Kuching, Malaysia

Тип: статья в журнале - научная статья Язык: английский

Номер: 59 Год: 2024 Страницы: 278-284

УДК: 81'37:811.512.1

ЖУРНАЛ:

ВЕСТНИК ИССЫК-КУЛЬСКОГО УНИВЕРСИТЕТА
Учредители: Иссык-Кульский государственный университет им.К. Тыныстанова
ISSN: 1561-9516 eISSN: 1694-8211

КЛЮЧЕВЫЕ СЛОВА:

SEMANTIC ROLE LABELING, AGGLUTINATION, TURKIC LANGUAGE, KYRGYZ, SEMANTICS

АННОТАЦИЯ:

Cognitive analyses of the language by humans cover all components of the language, currently this requires from the machine as well. And widely used languages such as English NLP machines almost reach the level of performing this analysis coherently as humans do. Which started in accordance with Chomsky's Universal grammar theory to analyze language via algorithms which were implemented according to the language syntactical structure. However, experience has shown that there is a linguistic features which algorithm based just on language structure can't analyze accurately as humans does. One of the reasons for this is that every word can have various grammatical and semantic features according to its particular context. Specifically in the case of languages with flexible word order. So NLP machines in such morphologically rich agglutinative languages such as Kyrgyz, need to cover both nature (grammatical, semantical) of the word to enhance the accuracy level of the tool...

▼ Показать полностью

БИБЛИОМЕТРИЧЕСКИЕ ПОКАЗАТЕЛИ:

- Входит в РИНЦ: да
- Входит в ядро РИНЦ: нет
- Рецензии: нет данных
- Цитирований в РИНЦ: 0
- Цитирований из ядра РИНЦ: 0
- Процентиль журнала в рейтинге SI:

О подборках журналов на eLIBRARY.RU

ИНСТРУМЕНТЫ

- Содержание выпуска

Загрузить:

- Полный текст (PDF)
- Отправить публикацию по электронной почте

Связанные документы:

- Полный текст, на сайте: api.library.iksu.kg

- Добавить публикацию в подборку

Новая подборка

- Редактировать Вашу заметку к публикации
- Обсудить эту публикацию с другими читателями
- Найти близкие по тематике публикации

- Дата размещения: 29.11.2024

ВХОД

IP-адрес компьютера: 49.50.236.216

Название организации: не определена

Имя пользователя или адрес эл. почты:

Пароль:

SEMANTIC ROLE LABELING (SRL) FOR THE DISAMBIGUATION OF NATURAL LANGUAGE PROCESSING (NLP) SYSTEMS IN LOW-RESOURCED LANGUAGES AS KYRGYZ LANGUAGE.

Zhumakadyrova N. Sh., MA., zhumakadyrova7@gmail.com

ORCID:

UNIMAS University Malaysia Sarawak, Kuching

Dr. Norazuna bt Norahim, Assoc., prof., nazuna@unimas.my

ORCID: 0009-0005-0456-430X

UNIMAS University Malaysia Sarawak, Kuching

ORCID: 0000-0003-4014-313X

UNIMAS University Malaysia Sarawak, Kuching

Annotation: Cognitive analyses of the language by humans cover all components of the language, currently this requires from the machine as well. And widely used languages such as English NLP machines almost reach the level of performing this analysis coherently as humans do. Which started in accordance with Chomsky's Universal grammar theory to analyze language via algorithms which were implemented according to the language syntactical structure. However, experience has shown that there is a linguistic features which algorithm based just on language structure can't analyze accurately as humans does. One of the reasons for this is that every word can have various grammatical and semantic features according to its particular context. Specifically in the case of languages with flexible word order. So NLP machines in such morphologically rich agglutinative languages such as Kyrgyz, need to cover both nature (grammatical, semantical) of the word to enhance the accuracy level of the tool. Thus this work will analyze the word's semantic meaning beside grammatical features such as word correlation (who did what to whom, when and where). **Thus this paper will investigate the challenges of implementing Semantic Role Labeling (SRL) in the context of the Kyrgyz language**, a low-resourced language, by examining its unique linguistic properties and the limitations of existing NLP tools. With **identification and analysis of the specific challenges faced by the SRL model** in handling ambiguous or complex sentences in the Kyrgyz language, providing insights into areas for future improvements.

Key words: Semantic role labeling, agglutination, turkic language, Kyrgyz, semantics

Dr. Mohamad Hardyman bin Barawi

Assoc., prof., bmhardyman@unimas.my

1. Introduction

In each language, a word has semantic meaning as well as grammatical meaning [11]. Whereas humans' cognitive acquisition can comprehend both automatically, natural language processing (NLP) machines require such tools as SRL for comprehensive language processing [12]. Specifically the needs for this approach increase in cases of agglutinative languages where depending on the context, words can have multiple semantic meanings. Hence the research will be focused on analyzing challenges for SRL tools for agglutinative Kyrgyz language, based on the Kyrgyz corpus dataset. The history of development of the language has undergone several historical and linguistic stages.The early history is closely connected with settlement patterns of the Kyrgyz peoples in Central Asia [2]. Historically, the Kyrgyz language was primarily an oral language, coming through generations by oral traditional folklore and storytelling. During the medieval Islam was spread through Central Asia, hence Arabic script and vocabulary influenced the Kyrgyz language as well. In 14th to 19th centuries, the Chagatai Turkic language had a considerable influence on the written form of the language due the political situation in the region. In the early 20th century, the Cyrillic script was adopted into the Kyrgyz language in 1940-1950. This process aimed at standardizing the written form of Kyrgyz and aligning it with other Turkic languages within the Soviet Union [8]. Currently with the project of Universal dependencies Kyrgyz language aims to be preserved from loss by collecting corpus data. Hence in this paper we will be analyzing flexible word order from agglutinative language nature, to observe challenges for the parsing system to understand semantic features of the word in a context [1]. As well as errors that may occur due to the subjective interpretation of NLP tools [13].

The rest of this paper is organised as follows: In Section 2, we provide an overview of related work in the field of text data analysis, covering previous approaches to SRL for other agglutinative languages and recent advances of models for natural language processing tasks in Turkish language. Section 3 We look at linguistic examples in details, including the key stages of SRL-based text analysing in the target language. In Section 4, we discuss the analysis results and evaluate the current stage of the tool from linguistic part. Finally, in Section 5, we conclude the paper and discuss future directions for research in text SRL.

2. Literature Review

As we know Language is an interdependent system in which the meaning of any given term depends on the other terms' random presence which the human brain automatically analyzes, including all aspects of NL. Every linguistic aspect must be included in NLP as a cognitive process of analyzing and presenting results[4]. One linguistic unit for more than one meaning[7]. Since language is the big structure (tool) to code and decode information [3, pp. 73–75], the disambiguation of morpho-semantic ambiguities by human depends from the style, mood and way of handling this tool by human being, so the concept of ambiguity is psycholinguistics specialty of the communication process as well as disambiguation process of this concept.

So, the POS tagging tool is needed for the disambiguation process of some cognitive gaps for the machine to comprehend natural language by covering the grammatical features of the term.

With the part of speech tagging tool (POS), we can cover grammatical features of the word. However, as is well known, a single word can have multiple semantic meanings depending on the context. As a result, SRL improves machine understanding of natural language (NL) by defining grammatical features of words under specific word sequences[5,] by identification of grammatical features of word under particular context [10] this allows more accurate interpretation of the NLP applications.

Specifically in the case of morphologically rich and low resourced Turkic languages *(Language family is commonly appears in: Kyrgyzstan, Uzbekistan, Kazakhstan, Turkmenistan, Turkey, China, South Siberia, Lithuania, different areas of Balkan, Cyprus, Azerbaijan)* such as Kyrgyz implementation of SRL in NLP tasks may propose solutions to enhance comprehension of the machine. So according to [6] the algorithm will be able to define changes in the segmentation of morphemes. Beside POS tagging methods which can identify grammatical features can make errors with ambiguous words of the target language because NL interpretation by human is a process more than we see and definition of the language by machine requires the coverage of multiple levels and features. Where SRL defines the semantic feature of the word within context together with POS tagging grammatical definition leads to the comprehensive review from NLP tool. Since semantic meaning relies on the syntactic feature it increases the accuracy of the tool. According to [4] quick disambiguation is possible with SRL.

3. Methodology

This work conducts qualitative analyses with corpus based descriptive and comparative methods. And will work on the output Stanza NLP tool and based on Turkic- Kyrgyz language corpora data. By comparing POS tagging and SRL comprehension for the sentence analyses which include ambiguous words. The sentences will be chosen by language corpus frequency formatting, after the human request for ambiguous words.

Sentence 1

Мунун кесепетинен улам көлгө ***кир*** *суулар кошулуп жатат.(As a result,* ***dirty*** *water is being poured into the lake).*

Кир (laundry, come in, dirty) is a Kyrgyz ambiguous word that might be Noun, Verb and Adjective depending on the context. In this sentence fulfilling the role of ADJ but relying just on the grammatical feature will lead to the misinterpretation of the machine. Since POS tagging was able to catch only one function of this word.

Sentence	**POS**	**SRL**
Мунун		*(error)*
кесепетинен		
улам		
көлгө	*(error)*	*Predicate: көлгө*
кир	*(error)*	
суулар		
кошулуп		Predicate: кошулуп

жатат		Predicate: жатат

Table #1 Stanza POS and SRL output for the Kyrgyz sentence

Sentence 2

***<u>Кир</u>**, кире гой, - деп маңдайындагы отургучту көрсөттү Толгонай. (**<u>Come in</u>**, come in, - Tolgonay pointed to the chair in front of him).*

Кир (laundry, come in, dirty) In this sentence fulfilling the role of Verb but machine misinterpreted it as Noun for the 1st word again covering just the first meaning of the word. However, the SRL tool adds extra meaning to the word by defining it as a word with extra meaning to the Noun.

Sentence	**POS**	**SRL**
Кир	*(error)*	*Кир, Relation: nmod*
кире *гой*		*Predicate: кире*
деп		*Predicate: деп*
маңдайындагы		*маңдайындагы, Relation: obl*
отургучту		*отургучту, Relation: nmod*
көрсөттү		*Predicate: көрсөттү,*
Толгонай		*Predicate: Толгонай,*

Table #2 *Stanza POS and SRL output for the Kyrgyz sentence*

Sentence 3

*Жакында эле үч **<u>барак</u>** толтура суроолор жазылган кагаз келди. (A **<u>paper</u>** for three pages with full of questions arrived recently).*

<u>Барак</u> (piece of paper, type of the house) is a Kyrgyz ambiguous word that might be Noun and Adj depending on the context. In this sentence the word playing the role of Noun, but the system gives wrong grammatical feature. Since SRL data didn't cover this word, the interpretation will be with errors.

Sentence	**POS**	**SRL**
Жакында	*VERB*	Argument: Жакында, Relation: obl
эле	*NOUN*	
үч	***NOUN***	

барак	*(error)*	
толтура	*VERB*	
суроолор	*VERB*	Argument: суроолор, Relation: nmod
жазылган	*ADV*	Argument: жазылган, Relation: advcl
кагаз	*NOUN*	Argument: кагаз, Relation: obj
келди		Predicate: келди,

Table #3 *Stanza POS and SRL output for the Kyrgyz sentence*

Sentence 4

Биз ***барак*** *үйдө жашачубуз. (We lived in* ***барак (type of the house)*** *a house).*

Барак (piece of paper, type of the house) in this sentence the word playing role of the ADJ, and POS tagging defined correct grammatical feature which makes us to assume this is most frequent grammatical feature, since even if the system didn’t define the SRL meaning, the interpretation was without errors.

Sentence	***POS***	***SRL***
Биз		
барак	*(correct)*	
үйдө		Argument: үйдө, Relation: nmod
жашачубуз		Predicate: жашачубуз,

Table #4 Stanza POS and SRL output for the Kyrgyz sentence

4. Analysis

Morphologically rich Turkic languages where most information is expressed via word formation rather than with syntactic [9]. Hence the parsing process is challenging as we can notice from the examples above. Hence it is necessary for the more accurate results to implement SRL which provides additional meaning to the words in the context. Then interpretation of the word by machine will be more accurate. As in the *sentence 1* first step with POS defined the ambiguous word *кир* with error, additional SRL gave additional semantic feature highlighting relation with object character. However, SRL for low resourced languages may make errors as shown in the *sentence 2* due additional meaning and need to expand the data number and enhance the accuracy level.

From the examples 3,4 we can assume that the POS tagging tool is coherent to define the most frequent feature of the word. However, SRL tools need to be developed in terms of the data gap for the low resourced Kyrgyz language. And it can be useful since the tool was helpful in some examples via giving extra characteristics for the decoding.

5. Conclusion

This paper was addressed to the low-resourced language as Kyrgyz language to define semantic complexities for NLP machines due to the ambiguous characteristics. And gave characteristic of the errors that are made by the system within initiation of the way how humans disambiguate words using a large pool of latent semantic factors and connections between senses. Beside the paper focuses on linguistic suggestions for semantic factors interpretation to become coherent under the disambiguation process.

And future works will be addressed for the disambiguation process of semantic ambiguities for the low-resourced Kyrgyz language. Since all the resources employed in this work may apply to several languages of Turkic family , the direction will be convenient for the adaptation to other languages. And will lead to the large future projects in this field.

Then we may achieve high results for Kyrgyz language NLP field.

References

1. Altıntaş, M., & Tantuğ, A. C. (2022). *Boosting Dependency Parsing Performance by Incorporating Additional Features for Agglutinative Languages.*
2. Bahry, S. (2018). Towards “Mapping” a Complex Language Ecology: The Case of Central Asia. In *Handbook of the Changing World Language Map* (pp. 1–39). Springer International Publishing. https://doi.org/10.1007/978-3-319-73400-2_4-1
3. Budiyono, S., Pranawa, E., Yuwono, S. E., Widya, U., & Klaten, D. (2021). *Seminar Nasional Riset Linguistik dan Pengajaran Bahasa (SENARILIP V)*. http://ojs.pnb.ac.id/index.php/Proceedings/73
4. Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. In *IEEE Computational Intelligence Magazine* (Vol. 9, Issue 2, pp. 48–57). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MCI.2014.2307227
5. Fei, H., Zhang, M., Li, B., & Ji, D. (2021). *End-to-end Semantic Role Labeling with Neural Transition-based Model*. www.aaai.org
6. Li, Z., Li, X., Sheng, J., & Slamu, W. (2020). AgglutiFiT: Efficient Low-Resource Agglutinative Language Model Fine-Tuning. *IEEE Access*, *8*, 148489–148499. https://doi.org/10.1109/ACCESS.2020.3015854
7. Rodd, J. (n.d.). *Settling into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access Word Meaning Access: The Challenge of Lexical Ambiguity*. www.jennirodd.com
8. Shermatova, F. (n.d.). *Construction of Youth Identity via Language in Kyrgyzstan: A Study of Russian and Kyrgyz*. www.academiccanvas.highereduhry.com
9. Tsarfaty, R., Bareket, D., Klein, S., & Seker, A. (2020). *From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)?* http://arxiv.org/abs/2005.01330

10. ВОПРОСЫ ЯДЕРНОГО ПРЕДЛОЖЕНИЯ И ИХ АНАЛИЗ В ПРЕДЛОЖЕНИЯХ КЫРГЫЗСКОГО ЯЗЫКА Ийсаева А.Д. В сборнике: EUROPEAN RESEARCH. сборник статей победителей VIII международной научно-практической конференции. 2017. С. 222-224
11. Zhaparov, Sheraly., Sydykova, T., & Sydykov, Akmatali. (2008). *Azyrky kyrgyz tilinin leksikalogii͡asy : lekt͡sii͡alyk sabaktardyn materialdary*. Teknik.
12. Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. In *Engineering* (Vol. 6, Issue 3, pp. 275–290). Elsevier Ltd. https://doi.org/10.1016/j.eng.2019.12.014