SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation dataset for Uzbek language

Ulugbek Salaev*, Elmurod Kuriyozov†, Carlos Gómez-Rodríguez†
*Urgench State University, Department of Information Technologies
14, Kh.Alimdjan str, Urgench city, 220100, Uzbekistan
ulugbek0302@gmail.com

†Universidade da Coruña, CITIC
Grupo LYS, Depto. de Computación y Tecnologías de la Información
Facultade de Informática, Campus de Elviña, A Coruña 15071, Spain
{e.kuriyozov, carlos.gomez}@udc.es

Abstract

Semantic relatedness between words is one of the core concepts in natural language processing, thus making semantic evaluation an important task. In this paper, we present a semantic model evaluation dataset: SimRelUz - a collection of similarity and relatedness scores of word pairs for the low-resource Uzbek language. The dataset consists of more than a thousand pairs of words carefully selected based on their morphological features, occurrence frequency, semantic relation, as well as annotated by eleven native Uzbek speakers from different age groups and gender. We also paid attention to the problem of dealing with rare words and out-of-vocabulary words to thoroughly evaluate the robustness of semantic models.

Keywords: natural language processing, uzbek language, semantic evaluation, dataset, similarity, relatedness

1. Introduction

Having computational models that can measure the semantic relatedness and semantic similarity between concepts or words is an important fundamental task for many Natural Language Processing (NLP) applications, such as word sense disambiguation (Navigli, 2009; Agirre and Edmonds, 2007), thesauri, automatic dictionary generation (Mihalcea and Moldovan, 2001; Solovyev et al., 2020), as well as machine translation (Bahdanau et al., 2014; Brown et al., 1990). There are many language models that have been created that yield good quality semantic knowledge, yet their evaluation depends on gold standard datasets that have word/concept pairs scored by their semantic relations (such as synonymy, antonymy, meronymy, hypernymy, etc.), that come with cost due to their time-consuming context-generation process and high dependence on human annotators.

Many such datasets have been created so far for resource-rich languages (Hill et al., 2015; Finkelstein et al., 2001; Rubenstein and Goodenough, 1965). However, there is still a big gap of such datasets available for low-resource languages. Current work aims to fill that gap by providing, to our knowledge, the first semantic similarity and relatedness dataset for Uzbek language. In this paper, we describe all the steps we followed as a set of data collection and annotation guidelines, with the full statistics and results obtained. The main contributions of this paper are two-fold:

- Publicly available word pair semantic similarity and relatedness scoring web-based questionnaire software[1];
- Publicly available semantic evaluation dataset including both similarity and relatedness scores for the low-resource Uzbek language [2];

Furthermore, this paper also describes some important construction considerations about the dataset considering morphological and semantic attributes for a morphologically rich language, with their visualisations.

Uzbek language (native: *O'zbek tili*) is a member of the Eastern Turkic or Karluk branch of the Turkic language family, an official language of Uzbekistan, and also a second language in neighbouring Central-Asian countries. It has more than 30 million speakers inside Uzbekistan alone, and more than ten million elsewhere in Central Asian countries, Southern Russian Federation, as well as the North-Eastern part of China, making it the second most widely spoken language among Turkic languages (right after Turkish language)[3].

This paper has been organised as follows: It starts with a terminology section, explaining the basic definitions of terms used in the paper, then comes a related work section followed by a description of dataset creation and annotation process, moving onto some insights of the dataset, and in the end, authors describe their discussions, conclusions, as well as future work.

[1]Demo website: `https://simrel.urdu.uz`

[2]Both publicly available dataset and the source code of the web-application can be found here: `https://github.com/UlugbekSalaev/SimRelUz`.

[3]More information about Uzbek language: `https://en.wikipedia.org/wiki/Uzbek_language`

2. Terminology

In order to eliminate repetition, and to avoid confusion understanding the terms used in this paper, the terms similarity, relatedness, association, and distance may come with or without the prefix "semantic" interchangeably, but they are meant to mean the same respectively.

The term ***semantic similarity*** in general, stands for a sense of relatedness that is dependent on the amount of shared properties, thus the 'degree of synonymy'. Whereas the term ***semantic relatedness*** means a general sense of semantic proximity or semantic association, regardless of the causes of the connection humans can perceive. For instance *bus/train* are good examples of semantic similarity, where they share many properties, i.e. they are both means of transport, both consume similar sorts of energy, have engines to operate, etc. On the other hand, *teapot/cup* can be a good example of semantic relatedness, where they don't necessarily share common properties, but they are used in a similar context, since they both store tea, but teapot is for steeping tea in larger amounts, while a cup is for serving and drinking tea in smaller portions. Both above-mentioned examples can be used for semantic relatedness though, which means that semantic similarity is included inside semantic relatedness. Therefore, semantically similar things are, at the same time, semantically related, but the converse cannot be said to be the case in general.

3. Related Work

The first creation of a stand-alone semantic relation evaluation dataset dates back to the RG dataset (Rubenstein and Goodenough, 1965) , which was created for semantic similarity more than relatedness[4]. Although it was very small in size (limited to only 65 noun pairs), it clearly showed the scientific importance, so the research interest continued later with more datasets coming along. The FrameNet (Baker et al., 1998) dataset is a rich linguistic resource with morphological, as well as expert-annotated semantic information as well. Among the most important gold-standard semantic evaluation datasets, we can find the WordSim-353 (Finkelstein et al., 2001), MEN (Bruni et al., 2012), and SimLex-999 (Hill et al., 2015) datasets for English. WordSim-353[5] contains 353 noun pairs scored by multiple human annotators. Similar to SimLex-353, the MEN[6] dataset also is described as having similarity and relatedness distinctly, but the annotators only were asked to rate based on semantic relatedness. Later, introduction of the SimLex-999[7] dataset made it the state-of-the-art gold standard semantic relatedness evaluation source. Some popular datasets for other languages include the RG dataset's German translation (Gurevych, 2005), the database of paradigmatic semantic relation pairs for German (Scheible and Im Walde, 2014), and the Simlex-999's translation into three languages: Italian, German and Russian (Leviant and Reichart, 2015). The Multi-SimLex (Vulić et al., 2020) project includes datasets for 12 diverse languages, including both major languages (English, Russian, Chinese, etc.) and less-resourced ones (Welsh, Kiswahili). Multi-SimLex[8] was a project originated from Simlex-999, and was taken to another step by creating a larger and more comprehensive dataset. Linguistic databases such as VerbNet (Schuler, 2005) and WordNet (Miller, 1995; Fellbaum, 2010) together with their implementations for other languages also contain semantically rich information created by experts.

Since this is the first work of this kind for Uzbek language, the closest related work would be the related resources created for other Turkic languages, such as Turkish WordNets (Tufis et al., 2004; Bakay et al., 2021), and especially AnlamVer dataset (Ercan and Yıldız, 2018), where it contains both semantic similarity and relatedness scores annotated by many native speakers. Furthermore, the AnlamVer also shares useful knowledge of dataset design consideration when dealing with morphologially-rich and agglutinative languages.

Work on Uzbek language. Although there have been many papers published claiming that they have created NLP resources or developed some useful tools for Uzbek language, most of them, according to humble search results gathered by the authors, turned out to be "zigglebottom" papers (Pedersen, 2008). However, there are also many useful papers with publicly available resources, some of them are the first Uzbek morphological analyzer (Matlatipov and Vetulani, 2009), transliteration (Mansurov and Mansurov, 2021a), WordNet type synsets (Agostini et al., 2021), Uzbek stopwords dataset (Madatov et al., 2021), sentiment analysis (Rabbimov et al., 2020; Kuriyozov and Matlatipov, 2019), text classification (Rabbimov and Kobilov, 2020), and even a recent pretrained Uzbek language model based on the BERT architecture (Mansurov and Mansurov, 2021b). There is also a well established Finite State Transducer(FST) based morphological analyzer for Uzbek language with more than 60K lexemes in Apertium monolingual package[9].

[4]RG dataset: https://aclweb.org/aclwiki/RG-65_Test_Collection_(State_of_the_art)

[5]WordSim-353 datset: http://alfonseca.org/eng/research/wordsim353.html

[6]MEN dataset: https://staff.fnwi.uva.nl/e.bruni/MEN

[7]SimLex-999 dataset: https://fh295.github.io//simlex.html

[8]Multi-SimLex project and dataset: https://multisimlex.com

[9]https://github.com/apertium/apertium-uzb

4. Dataset Design and Methodology

The criterion for the construction of the dataset had to satisfy all the requirements available to make a high-quality semantic evaluation resource. So we followed the design choice and recommendations brought by authors of previous work (Finkelstein et al., 2001; Bruni et al., 2012; Hill et al., 2015; Ercan and Yıldız, 2018; Vulić et al., 2020), such as follows:

- `Clear definition`: The dataset must provide a clear definition of what semantic relation is supposed to be scored. So we decided to collect scores of both similarity and relatedness separately;
- `Language representativity`: The dataset should should be built considering diverse concepts of the language, such as parts of speech (i.e. verb, noun, adjective, ...), word formations (root, inflectional, or derivative), possible semantic relations (i.e. synonymy, antonymy, meronymy, ...), as well as the frequency range (i.e. frequent words, rare words, even out-of-vocabulary words);
- `Consistency and reliability`: Clear and precise scoring guidelines were provided to get consistent annotations from native speakers with different level of linguistic expertise.

More detailed information regarding each criteria are given below.

4.1. Design choice

For the design of the dataset we followed the AnlamVer project (Ercan and Yıldız, 2018), where instead of building two separate datasets for semantic similarity and relatedness, we decided to rate each word pair with two separate scores: one for similarity, and another for relatedness. This way, the resulting dataset was smaller in size, but richer in information. Moreover, this approach gave us an opportunity to visualize the dataset as a semantic relation space, using two scores as two dimensions, and creating a scatter plot. According to the methodology proposed by AnlamVer (Ercan and Yıldız, 2018) project, it is possible to predict the semantic relation of word pairs, by their location in the "Sim-Rel vector space", which is given in Figure 1.

4.2. Word candidates selection

Probably a relatively easy way to obtain candidate words with minimum work would be translating words from gold-standard resources available for rich-resource languages (i.e. Multi-Simlex (Vulić et al., 2020)). However, there have been various relevant problems that have been reported to be caused by the use of such translations, such as:

- Two synonym pairs from a source language being mapped to one word in target language (Both words in *car - automobile* pair in English would be mapped to a single *avtomobil* in Uzbek);
- A translation of a single word in a source language that makes it multiple words in a target one (the word *asylum* in English would be translated as *ruhiy kasalliklar shifoxonasi* in Uzbek);
- Loss in the similarity/relatedness scores due to other cross-lingual aspects of pairs, such as translation accuracy or semantic/grammatical/cultural differences, require human annotators to re-score, leaving the costly part to be done again.

Therefore, we decided to choose the candidate wordlist ourselves for better quality. The first thing to make was a comprehensive list of words in the language using a big language corpus. For the language corpus mentioned in this work, we used the Uzbek corpus from the CUNI corpora for Turkic languages (Baisa et al., 2012), which is, to our knowledge, the biggest Uzbek corpus collected with 18M tokens. To obtain their part-of-speech (POS) tags, we used the UzWordNET dataset (Agostini et al., 2021) (which contains very limited information of root words with their POS classes), and Apertium-Uzb monolingual data[10] (contains more than 60K of Uzbek root words with their POS tags). Then we extracted nouns, adjectives and verbs only (with descending order relatively, according to their frequencies in the corpus), following the custom of similar gold-standard semantic evaluation resources. Apart from only root forms of words, we also did manual selection of words with inflectional and derivational forms of words.

4.3. Frequency-based considerations

Considering the agglutinative nature of Uzbek language, creating the list of word frequencies in this language is not an easy task, since a single word can occur together with many different morphemes (either a single morpheme or a combination of many), making it difficult to obtain the actual count of occurrences of a single root-word. In this paper, we created a list of stems with their frequencies in Uzbek language using the biggest available Uzbek corpora (Baisa et al., 2012). Firstly, the CUNI corpus was tokenized into sentences, then all the sentences were fed to the Apertium morphological analyser tool for Uzbek language[11]. Then, all the parts except for the lemmas of the resulting output were removed, which allowed us to obtain a stem/root-word frequency list. Our priority was to include as many words with different frequencies as possible, so we used a technique similar

[10] `https://github.com/apertium/apertium-uzb`

[11] Although we have used the CLI version of the Apertium morphoological analyzer, it also can be accessed on the web to check its features: `https://turkic.apertium.org/index.eng.html?choice=uzb#analyzation`

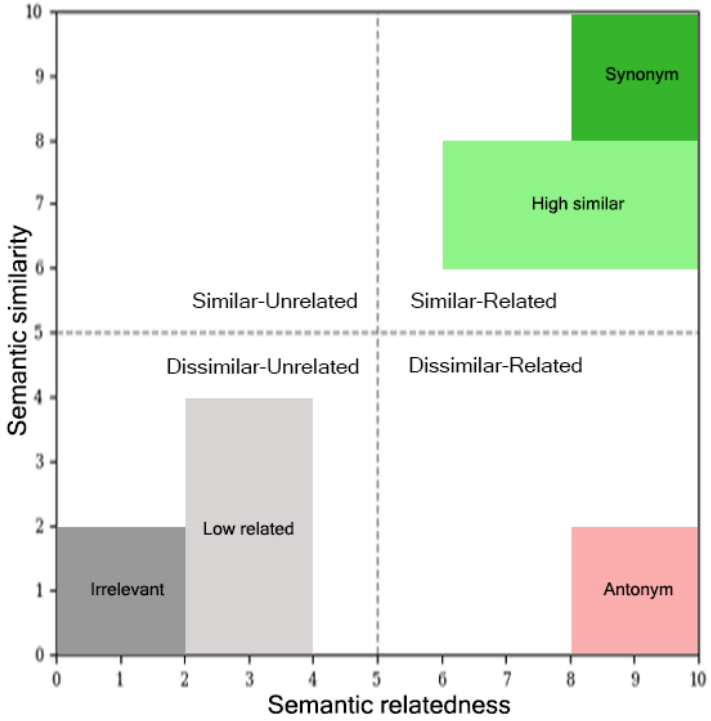

Figure 1: Semantic relation vector-space (proposed by AnlamVer project).

to the one issued by the RareWords dataset (Luong et al., 2013) - grouping words by their frequencies, dividing into three groups labeled as *low*, *medium*, *high* with [2,5],[6,49],[50+] count ranges respectively.

4.4. Rare and OOV words

Furthermore, to make the dataset useful for checking the robustness of the semantic models, considering less-frequent words, even words that do not exist in the language dictionary but might appear in the context due to some morphological (surface words), syntactical (typo), or phonetical (homophones) reasons is also an important aspect. Thus, the words where their root form does not appear more than 3 times in the corpus were grouped as rare words, and their representatives were manually selected for the word list.

Considering the rich morphological aspect of Uzbek language, like other Turkic languages, there is a high inflection and derivation rate, where words are made in an agglutinative way: by combining stem and one or more morphemes (as prefix or suffixes). Hence, there is a high chance that a word may be grammatically wrong, but was created following surface-word creation rules (of which almost an unlimited number can be created). So we chose the following two most common out-of-vocabulary word cases, which are formally incorrect, but considered as acceptable forms for native speakers, and added some examples to the dataset:

- `Stem-morpheme ambiguation`: It is a frequent case in Uzbek where stem and morpheme are combined directly, skipping the slight changes to fit them. E.g. *yaxshiliq* instead of *yaxshilik* (goodness), *qamoqga* instead of *qamoqqa* (to jail);
- `Phonetic ambiguation`: Two letters in Uzbek alphabet: “x” and “h” are phonetically so close to each-other, it is hard to identify them when used in a context, so people frequently mistake one for another when writing. E.g. *pahta* instead of *paxta* (cotton), *shaxzoda* instead of *shahzoda* (prince).

In total, 128 examples from both rare and OOV words with diverse POS types and word forms were added to the dataset.

After going through all the above mentioned steps and considerations, we gathered 1963 unique words to construct pairs. All their distribution among ford types, word forms, as well as word frequencies are given in Table 1.

4.5. Word pairs selection

Choosing word pairs randomly and scoring them would require the dataset to be huge in size, taking a very long time to annotate, so we tried to provide best quality semantic evaluation dataset with a limited number of word pairs by pre-establishing common semantic relations, such as synonymity, antonymity, hypernymity, and meronymity. This way the dataset would achieve a diverse distribution of scores, rather than filled up with very low scores due to most words not being related. Thus, we selected common semantic relation categories, namely synonyms, antonyms, meronyms and

Word classes		Word forms		Word frequencies	
Nouns	1154	Root form	995	High frequency	1136
Verbs	351	Infelctional	423	Medium frequency	448
Adjectives	457	Derivational	544	Low frequency & OOV	378
Total number of unique words: 1962					

Table 1: Distribution of words by different word types, word forms, and word frequencies.

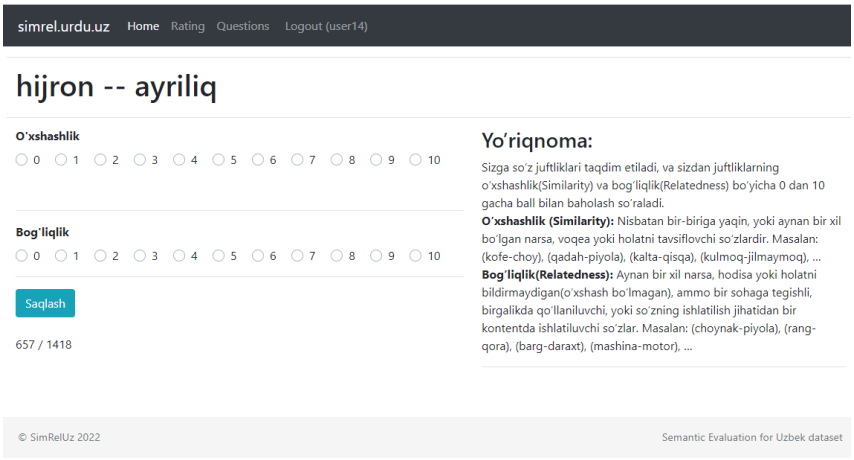

Figure 2: User interface of web-based annotation app.

hypernyms, and manually combined words from the word candidates list, tagging the pairs by a category where they most likely fit. Furthermore, we added word pairs by random allocation, which we named this category of pairs "irrelevant" (not in the sense of irrelevant pairs but in the sense of the magnitude of their semantic similarity and relatedness, as they are more likely to have very low scores on both sides).

Overall, 1418 word pairs were selected for the annotation, Table 2 shows the number of word pairs for each individual category.

Category	# of word pairs
Synonyms	639
Antonyms	239
Hypernyms	220
Meronyms	193
Irrelevant/Random	127
Total	**1418**

Table 2: Distribution of word pairs by their pre-established semantic relations.

5. Annotation process

For the annotation process, we have created a web-based survey application where each annotator is given a unique username and password, where they can access the website and rate given word pairs with two separate scores at once. General user interface of the annotation page can be seen in Figure 2.

In total, eleven annotators (including two authors), who are native Uzbek speakers with different linguistic background, from different age groups and genders, have participated at the annotation, rating each pair once, with two scores (one for similarity, and the other for relatedness) from 0 to 10. Based on a statistical analysis from (Snow et al., 2008), more than ten annotators for a semantic evaluation are reliable enough. In the end, there were eleven scores of similarity and the same amount for relatedness for each word pair, and we took their averages as the final scores. Figure 3 shows the distribution of age and gender between annotators.

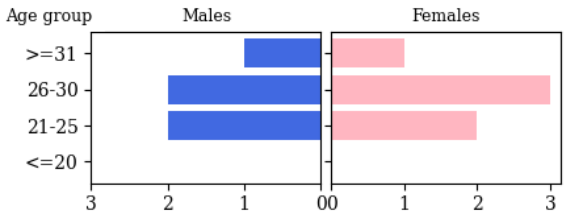

Figure 3: Distribution of annotators based on gender and age-groups.

6. Results

The resulting dataset is composed of 1418 word pairs from different word types (nouns, adjectives and verbs), different word forms (root, inflectional, derivational), with different frequencies (high, mid, low frequencies, rare and OOV words), and with diverse pre-established semantic relations (synonym, antonym, meronym, hypernym, not related). All the pairs have two scores, one for semantic similarity, while the other

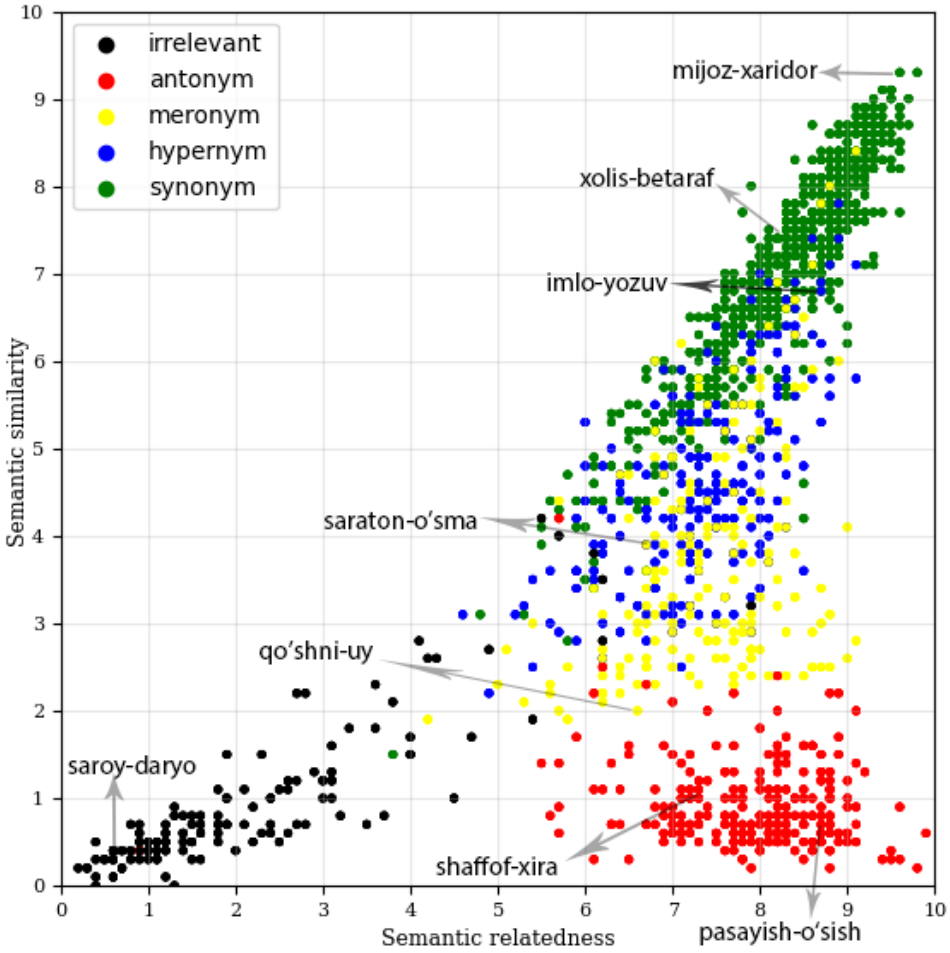

Figure 4: Visualisation of the created dataset in a Sim-Rel vector space.

is for semantic relatedness. No field in the dataset was left empty (as was requested from annotators in the guidelines, even for the OOV cases), and the average pairwise inter-annotator agreement scores (apia) were computed for both semantic similarity and relatedness separately, where we achieved 0.71 and 0.69 apia scores for semantic similarity and relatedness respectively, meaning that although we have scored less than AnlamVer dataset (0.75), it still performed better than most semantic evaluation datasets (SimLex=0.67, MEN=0.68). The resulting dataset can be plotted into the Sim-Rel vector space as shown in Figure 4.

Discussions. As can be seen from the scatter plot of the dataset in a vector space (Figure 4), it can be concluded that average scores of word pairs visually correlate to our pre-established relation types, since they are scattered mostly inside and around the determined areas in the vector-space. Irrelevant and random pairs can be easily detected from the plot, that it has no much overlap with other types. It is also worth mentioning that none of the word pair is in the Similar-Unrelated (top-left quarter of the vector-space) part of the plot, confirming its reliability, since a word cannot be similar, but not related at once. There is a big overlap between hypernym, meronym, and partially synonym pairs, as expected, as they share similar score ranges. Handling OOV words by annotators has also met our expectations, where they treated them as regular words and scored accordingly.

7. Conclusion

In this paper, we presented SimRelUz, a novel semantic evaluation dataset for the low-resource Uzbek language, with semantic similarity and relatedness scores for 1418 word pairs, which were selected based on their morphological classes, word-forms, frequencies, also including rare and out-of-vocabulary words for better evaluation of semantic language models. This kind of dataset is a useful resource to be used for evaluation of computational semantic analysis systems that will be created in the future for Uzbek, in simpler words, for formal analysis of meaning in language models. Moreover, we have also presented an open-source web-based semantic evaluation tool designed for multiple-user annotation. Our future work includes intrinsic and extrinsic analysis of created dataset, also creating big WordNet-type knowledge-base for Uzbek language.

8. Acknowledgements

This work has received funding from ERDF/MICINN-AEI (SCANNER-UDC, PID2020-113230RB-C21), from Xunta de Galicia (ED431C 2020/11), and from Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (ERDF - Galicia 2014-2020 Program), by grant ED431G 2019/01. Elmurod Kuriyozov was funded for his PhD by El-Yurt-Umidi Foundation under the Cabinet of Ministers of the Republic of Uzbekistan. The authors would also like to thank the NLP team of Urgench State University for their tremendous help with the web hosting, and annotation.

9. Bibliographical References

Agirre, E. and Edmonds, P. (2007). *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., and Mamasaidov, M. (2021). Uzwordnet: A lexical-semantic database for the uzbek language. In *Proceedings of the 11th Global Wordnet conference*, pages 8–19.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baisa, V., Suchomel, V., et al. (2012). Large corpora for turkic languages and unsupervised morphological analysis. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA)*.

Bakay, Ö., Ergelen, Ö., Sarmış, E., Yıldırım, S., Arıcan, B. N., Kocabalcıoğlu, A., Özçelik, M., Sanıyar, E., Kuyrukçu, O., Avar, B., et al. (2021). Turkish wordnet kenet. In *Proceedings of the 11th global wordnet conference*, pages 166–174.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

Ercan, G. and Yıldız, O. T. (2018). Anlamver: Semantic model evaluation dataset for turkish-word similarity and relatedness. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3819–3836.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *International conference on natural language processing*, pages 767–778. Springer.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Kuriyozov, E. and Matlatipov, S. (2019). Building a new sentiment analysis dataset for uzbek language and creating baseline models. In *Multidisciplinary Digital Publishing Institute Proceedings*, volume 21, page 37.

Leviant, I. and Reichart, R. (2015). Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR, abs/1508.00106*.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the seventeenth conference on computational natural language learning*, pages 104–113.

Madatov, K., Bekchanov, S., and Vičič, J. (2021). Lists of uzbek stopwords.

Mansurov, B. and Mansurov, A. (2021a). Uzbek cyrillic-latin-cyrillic machine transliteration. *arXiv preprint arXiv:2101.05162*.

Mansurov, B. and Mansurov, A. (2021b). Uzbert: pretraining a bert model for uzbek. *arXiv preprint arXiv:2108.09814*.

Matlatipov, G. and Vetulani, Z. (2009). Representation of uzbek morphology in prolog. In *Aspects of Natural Language Processing*, pages 83–110. Springer.

Mihalcea, R. and Moldovan, D. I. (2001). Automatic generation of a coarse grained wordnet.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Pedersen, T. (2008). Last words: Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Rabbimov, I. and Kobilov, S. (2020). Multi-class text classification of uzbek news articles using machine learning. In *Journal of Physics: Conference Series*, volume 1546, page 012097. IOP Publishing.

Rabbimov, I., Mporas, I., Simaki, V., and Kobilov, S. (2020). Investigating the effect of emoji in opinion classification of uzbek movie review comments. In

International Conference on Speech and Computer, pages 435–445. Springer.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Scheible, S. and Im Walde, S. S. (2014). A database of paradigmatic semantic relation pairs for german nouns, verbs, and adjectives. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 111–119.

Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Solovyev, V., Bochkarev, V., and Khristoforov, S. (2020). Generation of a dictionary of abstract/concrete words by a multilayer neural network. In *Journal of Physics: Conference Series*, volume 1680, page 012046. IOP Publishing.

Tufis, D., Cristea, D., and Stamou, S. (2004). Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., et al. (2020). Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.