Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 85–89
Marseille, 20 June 2022

WordNet and Wikipedia Connection in Turkish WordNet KeNet

Merve Doğan♡, Ceren Oksal♡, Arife Betül Yenice♡
Fatih Beyhan♠, Reyyan Yeniterzi♠
Olcay Taner Yıldız♢
Starlang Yazılım Danışmanlık♡, Sabancı University♠, Özyeğin University♢
Istanbul, Turkey
{merve, ceren, arife}@starlangyazilim.com, olcay.yildiz@ozyegin.edu.tr

Abstract

This paper aims to present WordNet and Wikipedia connection by linking synsets from Turkish WordNet KeNet with Wikipedia and thus, provide a better machine-readable dictionary to create an NLP model with rich data. For this purpose, manual mapping between two resources is realized and 11,478 synsets are linked to Wikipedia. In addition to this, automatic linking approaches are utilized to analyze possible connection suggestions. Baseline Approach and ElasticSearch Based Approach help identify the potential human annotation errors and analyze the effectiveness of these approaches in linking. Adopting both manual and automatic mapping provides us with an encompassing resource of WordNet and Wikipedia connections.

Keywords: Wikipedia, WordNet, Turkish

1. Introduction

Words as the building blocks of any length and type of text, play a very important role in any Natural Language Processing task. These context dependent units can have different meanings and different types of relations between each other, which makes NLP tasks challenging. WordNet as a lexical database of these relations plays an important role in solving these linguistic challenges. WordNet consists of synonyms of synset members, making it a highly comprehensive dictionary that stores lexicographic information. In addition, semantic relations such as hypernyms and antonyms are captured by mapping through synsets.

In previous literature (Navigli and Ponzetto, 2012; Fernando and Stevenson, 2012; McCrae, 2018), one common way to enrich a WordNet is to connect it to another very detailed data resource which is Wikipedia. Wikipedia is a web-based encyclopedia which provides multilingual lexical knowledge by presenting specific concepts and named entities. Compared to WordNet which contains descriptions of words and some example usages, Wikipedia may contain much more detail regarding the corresponding concept. Combining the lexicographic knowledge of WordNet with the rich encyclopedic knowledge within Wikipedia will enable more comprehensive representation of words and therefore create a much more useful resource for the challenging NLP tasks.

This paper proposes to create this connection between Wikipedia and WordNet for the first time for Turkish language. KeNet (Bakay et al., 2021), which is WordNet for Turkish, has been mapped to Turkish Wikipedia. KeNet stores 76,757 synsets, which makes it the most comprehensive WordNet for Turkish. Not only does it have intralingual relations such as hypernym, derivational relatedness, and domain topic but it is also linked to Princeton WordNet (PWN) through interlingual relations. Turkish Wikipedia has almost 463,808 articles to date, and it is the 31st largest Wikipedia edition. Combining these two resources will be a significant contribution to Turkish NLP research.

In order to perform this important and yet challenging task, we initially started with manual annotations. After manually connecting more than 11000 synsets, we also applied some retrieval based approaches to analyze the effectiveness of these automatic approaches for future extentions and to help decreasing possible human annotation errors.

2. Literature Review

The previous studies have been shown to use automatic mapping between WordNet and Wikipedia. In this regard, one of the most important studies has been on BabelNet (Navigli and Ponzetto, 2012). In this study, a word-sense disambiguation algorithm has been used for the mapping. In this algorithm, they have used surrounding synsets and the article texts and thus, different contexts have been created for both WordNet and Wikipedia. The endeavor of mapping Wikipedia to WordNet via an automatic mapping has resulted in an F-measure of 82.7% with 81.2% Precision and this can be claimed to be a high-quality resource. Another important study (Fernando and Stevenson, 2012) has been conducted by the use of semantic similarity methods and the result has been an F-measure of 84.1%. However, the scale of this study has been small as it has involved only 200 words.

Although the common strategy has been using automatic mapping to connect Wikipedia and WordNet, there is also a study in which manual mapping is adopted. With the aim of providing a gold standard for link discovery and creating richer, more usable resources for NLP, McCrae (McCrae, 2018) came

KeNet ID	Synset	Semantics	Wikipedia URL
TUR01-0301390	gentleman	A well-mannered man who can be a good friend	https://tr.wikipedia.org/wiki/Centilmen
TUR05-0800820	smiling	Slight laugh, smile	https://tr.wikipedia.org/wiki/Tebessüm
TUR03-2700020	green crescent society	Non-drinkers' association	https://tr.wikipedia.org/wiki/Yeşilay
TUR02-2200110	orient	East	https://tr.wikipedia.org/wiki/Doğu
TUR10-0256160	equilateral triangle	A triangle with three sides equal to each other	https://tr.wikipedia.org/wiki/Eşkenar_üçgen

Table 1: Example KeNet synsets with their unique IDs, semantic descriptions and connected Wikipedia links. Synset and Semantics are translated into English for convenience. Turkish correspondance of Synset column is as follows; *centilmen, gülümseme, yeşilay derneği, şark and eşkenar üçgen*, respectively.

up with mapping 7,742 instances between Princeton WordNet (PWN) and Wikipedia manually. These synsets in PWN are the instance hypernyms of 946 synsets in which it links a synset to an instance of a concept. The instance hypernyms of synsets that have been marked are named entities in the world. McCrae adopts the strategy to match the lemmas of WordNet entries to the titles of Wikipedia articles if it matches the title regardless of case before the first comma or parentheses or any page redirecting to this article. However, this results in significant ambiguity with approximately 21.6 candidates for each synset. Taking this into consideration, McCrae resorts to category mappings to determine the differences. Following this mapping, the links have been categorized as exact, broad, narrow, related and unnamed. This research stands out as the largest gold standard mapping for link discovery and an essential resource for NLP tasks.

In creating the connections between KeNet and Turkish Wikipedia, we use a combination of manual annotation with possible connection suggestions retrieved from automatic approaches.

3. KeNet and Turkish Wikipedia Linking

In this study, the initial connections have been created manually and then ElasticSearch[1] tool has been deployed to both analyze the effectiveness of automatic approaches and also to debug the manual annotations for any possible errors.

3.1. Manual Annotation

For the manual link creation process 47,169 synsets (all Nouns) from KeNet have been used. Linguistically informed human annotators manually iterated over these instances one by one and checked whether there is any Wikipedia page which describes the same concept. During this process, the meaning has been taken into consideration as semantics has been the focus.

The main focus has been on matching the article titles of Wikipedia with synsets and in addition to this, the content of Wikipedia has been checked to see whether it can be linked on the semantics level as well. The synsets of KeNet have been matched to the Wikipedia article if their meanings and the Wikipedia definitions correspond to each other. If the synset has been the subtitle of another Wikipedia article or when synset meaning has been given on that sub-title page, those synsets have not been linked. Therefore, one-to-one correspondence between KeNet and Wikipedia page has been paid attention and, in this respect, meaning component has been a crucial indicator.

Based on these manual mappings, 11,478 instances between KeNet and Wikipedia have been linked. Several example mappings are presented in Table 1. Each row in Table 1 corresponds to a synset with its unique KeNet ID and semantics as well as the manually mapped Wikipedia URL.

Almost 25% of the synsets have been mapped with this manual approach. Other synsets have not been matched due to a number of reasons. Firstly, many of these do not have any corresponding Wikipedia article. In this category, the metaphorical meanings of the synsets are quite common. For example, the synset "*ekmek parası*" which can be translated literally as "money for the bread" meaning "bread and butter" can not be found on Wikipedia and thus, there is no mapping.

Secondly, some of them appear as subtitles but we are only after the ones which are main titles. This has been done to get one-to-one correspondence between a KeNet entry and a Wikipedia main page, and with this in mind the subtitle matching have been ignored. For instance, the synset "*ağ*" which means "the web of a spider" is found as a subtitle of the main page "*örümcek*" (spider) and as a result, the mapping between these two cannot be realized.

Lastly, the content of the article does not match with the semantics of the synset. This has been encountered mostly with the words that have more than one synset and Wikipedia is able to provide generally one or two synsets for these types of words. As an example, there are two synsets for the word "*avcı*". One of the meanings is the animal who feeds on other animals by hunting and the other one is the name given to soldiers when they spread to combat. In the mapping process, the first synset is mapped to Wikipedia. On the other hand, the latter one cannot be mapped because there is not any correspondence on Wikipedia for this synset.

[1] https://github.com/elastic/elasticsearch

3.2. Automatic Approaches

In addition to the manual annotations, we also explored automatic approaches for both analyzing their effectiveness in linking and to double check for possible mistakes in manual annotations. In this paper we start our analysis with some classical ad-hoc retrieval and ranking approaches and leave the recent neural network based approaches for future work. Furthermore we use an exact match of the synset with the Wikipedia URL approach as our simple baseline.

In both of these approaches, the latest (1st of Jan 2022) Turkish Wikipedia dump [2], which consists of 463,808 Turkish wikipedia pages, is used.

3.2.1. Baseline Approach

Wikipedia websites have a URL base (*https://tr.wikipedia.org/wiki/*) which is followed by a unique page specific term or terms (similar to examples shown in Table 1). As a very simple baseline approach this base URL is concatenated with the synset from KeNet and checked whether there exists such an URL. If there is, then that Wikipedia link is connected to the corresponding synset. For example, for the word "*centilmen*" (gentleman) our baseline algorithm would suggest the page *https://tr.wikipedia.org/wiki/Centilmen*.

A portion of the synset entries has multiple terms and in these cases, the spaces between words are replaced with an underscore sign, as Wikipedia does. An example to such case is provided in Table 1 with "*eşkenar üçgen*" (equilateral triangle).

3.2.2. ElasticSearch based Approaches

In addition to the simple baseline, we approached the task as a search problem and utilized ElasticSearch (ES) to identify the possible connections.

463,808 Turkish Wikipedia articles were indexed. Unlike the simple baseline which only uses the URL, in here other more detailed parts of the Wikipedia pages are explored as well. The following two fields were created during indexing.

- *title*: Just the title of the Wikipedia page
- *all_text*: This is a concatenation of all the text in the title, text content, interwikies[3] and categories[4] of the Wikipedia page.

In addition to different fields of index, different retrieval mechanism were used as well. The *match* operator of the ElasticSearch retrieves documents with exact matches to at least one query term as its default behaviour (works like an OR operator). Additionally, the match operator can be used with an *AND* operator and in that case, it will retrieve only the pages which contain all the query terms. A more restricted version of this is the *match_phrase* operator which looks for documents with the exact query terms all in the same order (like a phrase) they were given in the query. These different exact match operators were analyzed.

Unlike *match* and *match_phrase*, the *fuzzy* search operator provides more flexibility in search by allowing retrieval of documents with possible typos or small variations of the query terms. Since the resources we are using are formal and well curated datasets, one may wonder whether *fuzzy* search is necessary at all. However, since the Turkish language has its own special characters such as *ü, ö, ğ, ç, ı*, fuzzy search may be useful in some cases.

Addition to aforementioned operators, we utilized the *bool* and *should* operators in order to create compound queries as well. The *bool* search with *should* inside, acts as an OR operator for a given set of queries being searched in different index fields.

While formulating the queries synset (SYN) field from the KeNet was used together with described query operators over described fields of index. The following experiments were conducted:

- **Exp1:** Using *match_phrase* query to search SYN in the *title* field
- **Exp2:** Using *match* query to search for SYN in the *title* field with the *AND* operator
- **Exp3:** Using *match* query to search for SYN in the *title* field with the *OR* operator
- **Exp4:** Using *fuzzy* query to search SYN in the *title* field
- **Exp5:** Using *match_phrase* query to search SYN in the *all_text* field
- **Exp6:** Using *match* query to search for SYN in the *all_text* field with the *AND* operator
- **Exp7:** Using *match* query to search for SYN in the *all_text* field with the *OR* operator
- **Exp8:** Using *fuzzy* query to search SYN in the *all_text* field
- **Exp9:** Using *bool* & *should* query operators to perform Exp2 and Exp6 together
- **Exp10:** Using *bool* & *should* query operators to perform Exp3 and Exp7 together
- **Exp11:** Using *bool* & *should* query operators to perform Exp3, Exp4, Exp7 and Exp8 altogether

[2] `https://dumps.wikimedia.org/trwiki/20220101/`

[3] Interwikies are the links to other Wikipedia pages. For instance, Wikipedia pages of Germany, France and Spain is in the interwikies section of the European Union's Wikipedia page, since they are mentioned within the contect of that page.

[4] Categories section of a Wikipedia page is used in order to gather articles under the common topics. For instance, Wikipedia pages of Germany, France and Spain have *Countries in Europe* category in their categories section.

Experiment	Compound	IndexField	ESQueryType	S@1	S@5	S@10	Ave. # Pages
Baseline	-	-	-	47.60	-	-	-
Exp1	No	title	match_phrase	46.28	50.47	51.65	3.28
Exp2			match (AND)	46.70	51.02	52.20	3.33
Exp3			match (OR)	63.30	78.67	81.42	7.35
Exp4			fuzzy	35.88	41.96	43.61	4.56
Exp5	No	all_text	match_phrase	17.96	34.84	40.63	6.08
Exp6			match (AND)	23.17	42.93	49.03	7.03
Exp7			match (OR)	29.94	57.53	66.47	9.59
Exp8			fuzzy	11.69	23.38	28.32	5.14
Exp9	Yes	title all_text	match (AND)	51.25	60.65	63.05	7.07
Exp10	Yes	title all_text	match (OR)	**68.11**	83.10	86.51	9.62
Exp11	Yes	title	match (OR) fuzzy	66.37	**85.15**	**88.79**	9.87
		all_text	match (OR) fuzzy				

Table 2: Evaluation results of simple baseline and ElasticSearch with different experiments. *Bool* and *should* query operators were used in order to build the compound queries.

3.2.3. Evaluation and Results

The trec_eval[5], the standard evaluation tool of the TREC community, was used to evaluate the automatically generated candidates. Unlike other ad-hoc retrieval tasks, our dataset is designed to have a single relevant page (the Wikipedia page) rather than a list of possible relevant pages. Hence, instead of precision or recall, we used *Success@1* (S@1), *Success@5* (S@5) and *Success@10* (S@10) evaluation metrics. Given a list of candidate pages ordered based on their retrieval scores, S@N evaluation metric would return 1 in case the correct page is in the top N candidate pages.

The results of all the experiments are presented in Table 2. The first column displays the experiment ID and the next three columns detail whether the query is a compound query, the Wikipedia field used for indexing and the ElasticSearch query type in order. In addition to the Success@N scores, the average number of retrieved pages are shown in the last column. This number is specifically important because these retrieved pages are manually checked that will affect the size of the pool of pages to be assessed.

According to Table 2, our simple *baseline* is not so bad at all. It correctly identified almost half of the connected pages. The *Exp1* and *Exp2* are the most similar experiments to this *baseline* as these also searched for the whole synset in the title of the page. Overall these restricted queries return approximately 3-4 Wikipedia pages which is really efficient but with cost of missing relevant pages.

Match with the *OR* operator (*Exp3*) performed much better across all S@N metrics. In our analysis we observed that in some nominal compounds the second element which is possessed noun may be missing in Wikipedia or in the synset. For example, we have "*yeşilay derneği*" as one of our synsets and there is only "*yeşilay*" entry on Wikipedia. So, this case which had been missed with previous experiments was caught with *Exp3*. Of course this more relaxed search comes with a larger pool of around 7-8 pages per query.

Using *fuzzy* query (*Exp3*) did not help at all and returned the lowest scores so far. Also using all text within the Wikipedia (*Exp4-Exp8*) instead of the title did not provide any improvement in any aspects, as we got lower S@N scores and higher average number of retrieved pages.

In addition to simple one field searching queries, more complicated compound queries are tried as well to see the effects of combining information from different fields. Both *Exp9* and *Exp10* returned improvements over the individual experiments *Exp2* and *Exp3* respectively. With *Exp9* the average number of retrieved pages increased more than twice compared to *Exp2*. At this point *Exp3* is still better than *Exp9* with a slightly larger pool. Therefore we did not continue working on *match* with *AND* operators. Instead we continued with *Exp10* and tried extending it with *fuzzy* cases as well. Even though adding *fuzzy* (in *Exp11*) lowered the S@1 scores, it still returned the highest S@5 and S@10 scores so far. The correct Wikipedia page is retrieved within top 5 documents 85% of the time.

4. Evolving Datasets

Both KeNet and Wikipedia are evolving resources. As time passes new synsets are introduced to KeNet. Similarly new Wikipedia pages can be created or the existing ones can be updated (a change in the title also affects the URL of the page) or even deleted. Therefore keeping track of these resources and updating the

[5] `https://github.com/usnistgov/trec_eval`

connections between them is necessary. This continuous update or extension process will be easier to handle with the help of these automatic tools. With these tools this time consuming manual process becomes both efficient and user-friendly. The aforementioned potential updates occurred even in the creation phase of this dataset. Initially we started the annotation phase with the latest Wikipedia dump of that time. Later on as we moved to the automatic linking approaches, we started working with another version (again latest of that time; 1st of Jan 2022) of Wikipedia dump. Between these different dumps of Wikipedia we have seen that around 100 Wikipedia URLs, which were assigned as labels to our synsets, were not in the Wikipedia dump that we started using recently. However, when we tried to open these links, Wikipedia redirected us to new pages which are the updated version of the requested pages. For instance, the Wikipedia page for *Mersingiller*, which is a type of a flower family, were labelled as *https://tr.wikipedia.org/wiki/Mersingiller*, however the updated version of the same page has *https://tr.wikipedia.org/wiki/Myrtaceae* as its URL. Overall the automatic retrieval process helped the annotators to catch these changes and update the connections accordingly.

In addition to helping with the updates, the automatic approaches even help with finding the missing connections and therefore extending the connections lists. After manually mapping almost 25% of the synsets, there were 35583 synsets which were not mapped to any Wikipedia page, yet. Even our simple baseline experiments showed that almost half of our dataset was mapped correctly only with concatenation of the synset and the Wikipedia URL base. We utilized our simple baseline to create candidate URLs for the unlinked 35583 synsets. Among the candidate URLs which were generated by the baseline algorithm, there were only 2961 URLs that existed in the Wikipedia dump. An annotator manually checked these 2961 URLs to validate whether there are any missed connections. 83 URLs were identified as missing in the original dataset which were included in the final version of our dataset. As expected these are the results of human errors which exist in almost all annotated data collections. This error frequency being low is also a good indication that our initial manual annotations are in good quality.

5. Conclusion

In this paper, we have presented the connections between KeNet and Wikipedia for Turkish language. The fact that it is possible to find different parts of speech such as nouns, verbs, adjectives and adverbs in a WordNet, only nouns are found in Wikipedia. In this regard, the combination of two comprehensive resources bears fruitful results for future usages in NLP tasks because of their complemantary nature. By combining lexicographic knowledge of WordNet with rich encyclopedic knowledge of Wikipedia, we have been able to map synset instances between those two resources. Both manual mapping and automatic approaches of this linking have made possible to reach an exact match of synset with the Wikipedia page. While mapping manually have been great tool for matching process, automatic approaches consisting of classical ad-hoc retrieval and ranking approaches have helped to see how successful manual mapping has been and enabled us to retrieve the possible connections and thus, double-check also the synsets that haven't been matched. Thus, Wikipedia and WordNet connection that has been shown is crucial for machine-readable dictionary for future NLP tasks.

6. Bibliographical References

Bakay, Ö., Ergelen, Ö., Sarmış, E., Yıldırım, S., Arıcan, B. N., Kocabalcıoğlu, A., Özçelik, M., Sanıyar, E., Kuyrukçu, O., Avar, B., and Yıldız, O. T. (2021). Turkish WordNet KeNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 166–174, University of South Africa (UNISA), January. Global Wordnet Association.

Fernando, S. and Stevenson, M. (2012). Mapping WordNet synsets to Wikipedia articles. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 590–596, Istanbul, Turkey, May. European Language Resources Association (ELRA).

McCrae, J. P. (2018). Mapping wordnet instances to wikipedia. In *Proceedings of the 9th Global WordNet Conference*. Zenodo, January.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.