Creating Domain Dependent Turkish WordNet and SentiNet

Bilge Nas Arıcan
Starlang Yazılım Danışmanlık
bilge@starlangyazilim.com

Merve Özçelik
Starlang Yazılım Danışmanlık
merve@starlangyazilim.com

Deniz Baran Aslan
Starlang Yazılım Danışmanlık
deniz@starlangyazilim.com

Elif Sarmış
Starlang Yazılım Danışmanlık
elif@starlangyazilim.com

Selen Parlar
Starlang Yazılım Danışmanlık
selen.parlar@boun.edu.tr

Olcay Taner Yıldız
Özyeğin University
olcay.yildiz@ozyegin.edu.tr

Abstract

A WordNet is a thesaurus that has a structured list of words organized depending on their meanings. WordNet represents word senses, all meanings a single lemma may have, the relations between these senses, and their definitions. Another study within the domain of Natural Language Processing is sentiment analysis. With sentiment analysis, data sets can be scored according to the emotion they contain. In the sentiment analysis we did with the data we received on the Tourism WordNet, we performed a domain-specific sentiment analysis study by annotating the data. In this paper, we propose a method to facilitate Natural Language Processing tasks such as sentiment analysis performed in specific domains via creating a specific-domain subset of an original Turkish dictionary. As the preliminary study, we have created a WordNet for the tourism domain with 14,000 words and validated it on simple tasks.

1 Introduction

WordNet is a semantic network that represents semantic relations between different concepts by providing a graph consisting of nodes and links. A semantic network is a sine qua non of NLP applications which aim to integrate domain knowledge and lexical knowledge. To this end, since the primary purpose of using WordNet is obtaining the similarities and relations between words, WordNets have been employed in various fields of NLP such as word sense and root word disambiguation, information retrieval, machine translation, and sentiment analysis.

Sentiment analysis interprets and classifies the emotions in a data through natural language processing learning. It can be performed on a word, a sentence, or even a paragraph. With sentiment analysis, many data such as surveys, texts, customer comments and social media content can be analyzed. Especially in the business world, it has a very important place in understanding customers, so that products and services can be arranged to meet the needs.

Among many fields of NLP emplying WordNet, sentiment analysis, or opinion mining, refers to the study of people's opinions, sentiments, appraisals, attitudes, and emotions towards entities, which might include products, services, organizations, issues, individuals or events (Liu, 2015). Sentiment analysis primarily deals with opinions that express or imply positive, negative or neutral sentiments. In order to conduct such analyses, WordNets are of great importance since they provide data in an organized way, especially when the study relies on domain-specific data as in this study. Usually created for general usage, a WordNet can also be created and used for specific domains such as tourism, textile or technology, each of which may inherently contain different senses and relations of the same words. That is to say, depending on the WordNet one draws on, the output of sentiment analysis may change. To illustrate, being 'thick' has positive connotations for a carpet whereas it is often undesirable for smart phones. Therefore, conducting a sentiment analysis in a specific domain necessitates the creation of a domain-specific WordNet.

Prior to the creation of a WordNet, a lexicon with broad coverage should be created in the first place. However, there is no limit for the number of words in a lexicon for agglutinative languages like

Turkish. In addition to agglutination, polysemy, i.e., the coexistence of many possible meanings for a word creates hundreds of basic semantic inconsistencies, which indicates that covering all the words and their senses in a language is a highly demanding task. For instance, a Turkish lexicon carries more than 50,000 words; nevertheless, employing such a vast lexicon for a specific domain brings out ambiguous results since it leaves out the words that are not prevalent in daily usage but common in that specific domain.

To this end, this study aims to address the issue of utilizing an immense WordNet for a specific domain, namely tourism. The data for the study consists of online user reviews and preferences of a tourism company located in Turkey. Drawing on this data, we have initially corrected the misspelled words, put them into groups depending on their part of speech (noun, proper noun, adjective, verb and adverb), and finally tagged them based on the linguistic features of Turkish. These steps have provided us with a 14,000-word lexicon covering not only commonly used words but also domain-specific words. Compared to currently used Turkish dictionaries, this newly created dictionary has approximately four times fewer words, which is the reason why we draw on this dictionary while creating the domain-specific WordNet. As expected, the meanings of the words in this domain specific dictionary vary based on their area of usage. That is to say, the meanings of some words in general use acquire new meanings in the domain-specific dictionary, according to which we have arranged the hierarchy of the words.

The necessary data for SentiNet, which is a domain-dependent resource for sentiment analysis, have been drawn from the Tourism WordNet we created. It should be said that the data used for sentiment analysis were matched with their counterpart in Turkish WordNet again after annotating. The Tourism WordNet and SentiNet data are linked to each other via senses. The synset IDs of SentiNet and Tourism Wordnet data are the same on both sides. In the annotating phase, care has been taken to annotate all data with more than one annotator and to ensure these annotators do not have information about each other's preferences. Although the line of objectivity is not possible for sentiment analysis markings, it is aimed to present a study that yields more successful results with these items that we pay attention to in the marking stages.

This paper is organized as follows: We first discuss the relevant literature on WordNets in Section 2. We explain how we generated the domain dependent WordNet and SentiNet in Sections 3 and 4. We provide details on the word-sense disambiguation task using our domain dependent wordnet in Section 5. The statistics and experimental results regarding our WordNet and SentiNet are given in Section 6. Lastly, we conclude in Section 7.

2 Literature Review

The first WordNet project is a lexical database for English, namely Princeton WordNet (PWN), which was initiated in 1995 by George Miller, (1995). Currently, the latest release of PWN, version 3.1, has 117,000 synsets, and 206,941 word-sense pairs. A more detailed history and description of PWN is given in (Fellbaum, 1998). Shortly after the release of PWN, WordNets for other languages have been constructed although their coverage is not as extensive as that of PWN, (Vossen, 1997), (Black et al., 2006). For Balkan languages, BalkaNet (Tufis et al., 2004) is the most comprehensive work up to date. For the Turkish WordNet part of BalkaNet (Bilgin et al., 2004), the researchers automatically extracted the synonyms, antonyms and hyponyms from a monolingual Turkish dictionary. The most comprehensive Turkish WordNet is KeNet, which has 80,000 synsets covering 110,000 word-sense pairs (Ehsani et al., 2018; Bakay et al., 2019b; Bakay et al., 2019a; Ozcelik et al., 2019; Bakay et al., 2020).

All this body of work mentioned above has been created and used for general purposes. However, the creation of a domain-specific WordNet is a more recent phenomenon, of which there are relatively few examples. ArchiWordNet is a WordNet created specifically for the architecture and construction domain drawing on Italian/English bilingual resources. Similarly, Jur-WordNet is another example of a domain-specific WordNet which was created as an extension for the legal domain of ItalWordNet by providing multilingual access to legal information sources. Specifically created to be used for software engineering tasks, SEthesaurus is a dictionary constructed based on informal discussions about programming on social platforms. By generating a WordNet specific to the tourism domain, we hope to contribute to this body of work, and provide inspiring ideas for future studies (Sagri et al., 2004; Bentivogli et al., 2003; Chen et al., 2019).

Regarding sentiment analysis, to the best of our knowledge, there have been no studies conducting a domain-specific sentiment analysis relying on a domain-specific WordNet. Therefore, it would be plausible to assert that we are presenting a pioneering study in this field.

3 Domain Dependent WordNet

3.1 Preprocessing

As stated in Section 1, the data used for this study consist of online customer reviews or customer preferences from the tourism domain. Since users usually prefer daily, informal language not paying attention to grammatical correctness but focusing mainly on the semantics, it is not feasible to perform further natural language processing based on the original input. Therefore, we employ the final version of the data following a preprocessing pipeline. The first step of this preprocessing is sentence splitting, where we divide paragraphs into sentences and each sentence into words, then perform case-folding to convert all the words to a particular case. Subsequently, we conduct the stemming process for which we only consider basic Turkish suffixes. For instance, we remove the plural suffix '*-lar*, *-ler*' ('-s, -es'), locative case suffix '*-de*, *-da*' (in, on, by), ablative case suffix '*-den*, *-dan*' (from, of), and dative case suffix '*-a*, *-e*' (to, towards). This stemming process provides us with tokens by unveiling distinct words.

Following the sentence splitting and stemming processes, the remaining single tokens need to be deasciified since not all tokens are spelled correctly by users. That is, we convert erroneously written Turkish characters into their correct forms. For instance, the word '*Türkçe*' (Turkish) which contains language-specific characters ('ü, ç') is mostly written by using English characters as '*Turkce*', which has no meaning in the lexicon. Moreover, if a word cannot be morphologically analyzed, after all, we interchange each letter with its closest neighbor. Provided that the resulting string still cannot be analyzed, we suggest the most similar word in the lexicon based on the Levenshtein distance between words. At the end of this preprocessing, we tokenize and retrieve the distinct words which are ready to be analyzed morphologically.

Table 1: Example words from the Tourism WordNet

Word	**Instance Hypernym**
Sicily	Island
Metrogarden	Mall
Nestle	Food brand
Izmir	City
Mimarova	Neighborhood
Merlin	Hotel
Italy	Country

3.2 Dictionary

Relying on the words and comments from the online system of a tourism company, a dictionary is prepared for the creation of the WordNet by three Turkish native speakers, who specialize in Turkish linguistics. This makes sure that the dictionary reflects the most commonly used words in the domain such as meals, hotel names, holiday items, etc. Based on their part of speech, these words are tagged as a proper noun, noun, verb, adjective or adverb, which determines the area of usage for each word. In addition to these main categories, some words receive extra labels such as vowel harmony tags while verbs are re-grouped based on their grammatical features.

In addition, we have created a set of "misspelling data" consisting of the misspelled words, which contain 120,000 entries. In this way, we have identified the words that are most frequently misspelled by users so that these words can be automatically corrected for future studies.

3.3 WordNet

In a WordNet, which plays a crucial role in NLP, words are first grouped based on their part of speech under the categories of proper nouns, nouns, verbs, adjectives, and adverbs, after which the words in each category are clustered depending on their semantic relations. In our Tourism Dictionary, there are three major part of speech categories (See Table 1, which are proper noun, noun, and adjective.

Following the categorization of the words, each category is exclusively studied on its own. The words in the noun category are organized depending on several semantic relations, namely synonym, antonym, member holonym, substance holonym, part holonym, domain topic, and attribute. Regarding the proper noun category, we have paid attention to the areas that the words belong to; therefore, all proper nouns that do not have a particular importance are grouped under the same category, the majority of which consists of hotel names. However, given names and surnames have been removed from the data. Finally, the adverb category has been dismissed from the scope of this study due to the small number of words in that category.

4 Doman Dependent SentiNet

Since sentiment analysis focuses on whether entities are positive, negative or neutral, the words in our tourism corpus have been labeled as positive, negative or neutral by three annotators, who are native speakers of Turkish. Following the first labeling process, the words labeled as positive and negative have been subjected to a second labeling process, marked as strong or weak since the degree of positivity or negativity may vary as in the difference between the words ”güzel” (beautiful) and ”harika” (excellent). This allows a more precise analysis of the positive or negative value that the word adds to the sentence. Furthermore, we have paid attention to both the dictionary meaning of the word and the way it is used in daily life in this specific domain. In cases where a word was labeled differently by the annotators, we have relied on the opinion of the majority.

Following the labeling process, we have found that the majority of the words are neutral while the ratio of negative words is higher than positive words. Moreover, we have found that the weak positive and weak negative tags are more prevalent than the strong positive and strong negative. In addition, the automated analysis of the sentences are accelerated since the positive, negative and neutral values of the words can be better processed by the algorithm. Therefore, we believe that the automatic analysis of the words will be much easier and faster.

5 Usage of WordNet in Semantic Annotation: All-Words Sense Annotation

The study has been conducted on a 20,000-sentence corpus created using data from the tourism domain. The words and their definitions have been drawn from the Tourism WordNet. Two interfaces have been created to employ in the semantic annotation process, which consisted of two steps. The sentences were processed by the annotators after each word was subjected to morphological analysis and matched with its equivalent in the Turkish WordNet. Four annotators worked simultaneously in the first step using the interface that displays each sentence individually. As can be seen in Figure 1, each word can be annotated individually, and the buttons at the top are used to navigate the corpus. When a word is clicked on, a list of every possible definition is displayed. The annotators chose the appropriate definition manually. Punctuation marks were annotated automatically. The annotators also made use of the ”annotate each occurrence of the word with the same definition” feature, making the process semiautomatic and increasing efficiency. This feature annotates all occurrences of the selected word in the corpus with the same definition from the list. Through this feature, words that happen to only have a single definition, in general or in this specific domain, have been annotated more easily. Sentences that produced errors in the morphological analysis phase were corrected manually using the same analyzer. Each word was annotated primarily using the definitions in the Tourism WordNet. The definitions in the Turkish WordNet were made use of where the Tourism WordNet was not sufficient.

At the end of the first step, there were still words without annotations. The second step was an effort to fill in these gaps and check the results manually. A different interface displaying all sentences simultaneously was used in this step. The words were arranged alphabetically, and grouped based on their sentences. In this way, the words were compared to one another in different contexts, and their definitions were decided on by reviewing the entire corpus. The missing annotations were completed based on the existing ones. Two annotators worked on this step in cooperation in order to ensure consistency between their annotations.

In the annotation process, an optional automatic annotation function was also employed. This function automatically matches the words with only one definition in the dictionary with that one definition without asking the annotator. Afterwards, these were verified by the annotators and corrected when necessary. The semantic annotation interface can also detect multi-word expressions, which allows the annotation of words that come together to form a single unit of meaning.

Table 2: An example of positive marking

Annotation	ID	Word	Definition
p	TUR10-0318100	güzel (beautiful)	Göze ve kulağa hoş gelen, hayranlık uyandıran (Pleasing to the eye, admirable)

Table 3: An example of neutral marking

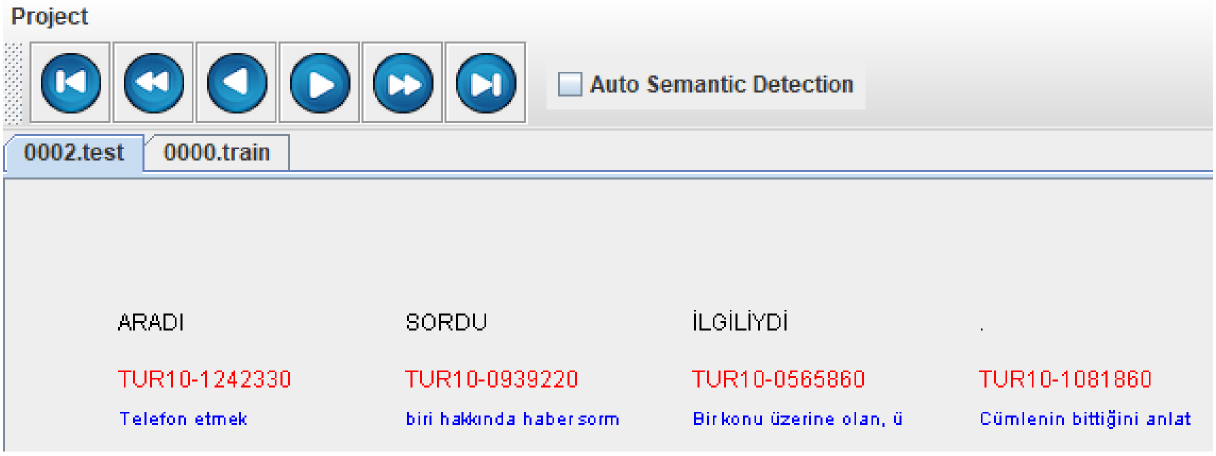
Annotation	ID	Word	Definition
o	TUR10-0016080	ahşap (wood)	Ağaçtan, tahtadan yapılmış (Made of wood)

Table 4: An example of negative marking

Annotation	ID	Word	Definition
n	TUR10-0335560	Çığlık (scream)	Acı, ince ve keskin ses, feryat (Painful, subtle and sharp sound, howl.)

Table 5: Markings in the second stage of a positive sample

Ann.1	Ann.2	Ann.3	ID	Word	Definition
s	s	s	TUR10-0318100	Güzel (beautiful)	Göze ve kulağa hoş gelen, hayranlık uyandıran (Pleasing to the eye, admirable)
w	s	w	TUR10-0246270	Empati (empathy)	Aynı duyguları paylaşma (Sharing the same emotions)
w	w	w	TUR10-0421970	Hesaplı (economic)	Az masraflı, kazançlı, hesaplı, iktisadi (Low-cost, profitable, affordable, economic)



ARADI	SORDU	İLGİLİYDİ	.
TUR10-1242330	TUR10-0939220	TUR10-0565860	TUR10-1081860
Telefon etmek	biri hakkında haber sorm	Bir konu üzerine olan, ü	Cümlenin bittiğini anlat

Figure 1: Interface used in the first phase

Turkish has a great volume of two-word verbal expressions (e.g. ”kabul etmek”, to accept; ”memnun kalmak”, to be satisfied”), which is reflected in the tourism corpus. The senses that do not show up when these words occur by themselves are included in the list of possible definitions if they appear consecutively in the right order, which the annotators chose manually.

6 Results

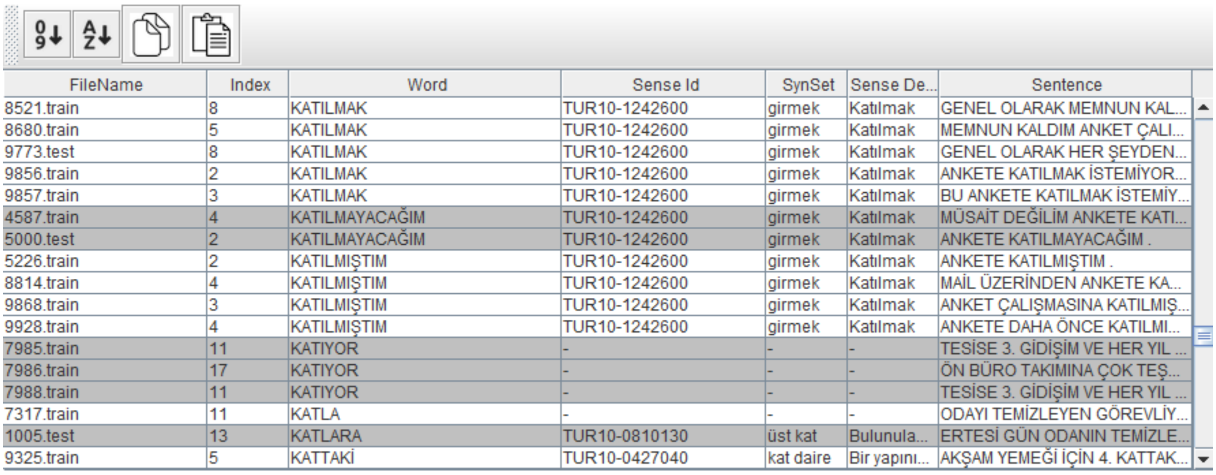
6.1 Statistics About WordNet and SentiNet

Designing WordNets and dictionaries entails working with a huge body of data, which is the reason why this study relies on a large amount of data from online user reviews and user preferences from the tourism domain. Before the study, for the tourism domain, we created a lexicon of 14,000 entries by using the words extracted from the most common 20,000 reviews by users, e.g., the customers of a holiday resort, or a tour, an example of which can be seen in Table 7.

Generally, users do not pay attention to the conventions of standard grammar or spelling while typing their comments in online surveys. Therefore, we have enacted several pre-processing steps described in Section 3 in order to retrieve the cor-

Table 6: Markings in the second stage of a negative sample

Ann.1	Ann.2	Ann.3	ID	Word	Definition
s	s	s	TUR10-0827940	Yalan (lie)	Aldatmak amacıyla bilerek ve gerçeğe aykırı olarak söylenen söz (A word that is not true)
s	w	s	TUR10-0600400	Öksüz (orphan)	Anası veya hem anası hem babası ölmüş olan çocuk (Child whose mother or father has died)
w	w	w	TUR10-0201160	Dezavantaj (disadvantage)	Avantajlı olmama durumu (A disadvantaged situation.)



FileName	Index	Word	Sense Id	SynSet	Sense De...	Sentence
8521.train	8	KATILMAK	TUR10-1242600	girmek	Katılmak	GENEL OLARAK MEMNUN KAL...
8680.train	5	KATILMAK	TUR10-1242600	girmek	Katılmak	MEMNUN KALDIM ANKET ÇALI...
9773.test	8	KATILMAK	TUR10-1242600	girmek	Katılmak	GENEL OLARAK HER ŞEYDEN...
9856.train	2	KATILMAK	TUR10-1242600	girmek	Katılmak	ANKETE KATILMAK İSTEMİYOR...
9857.train	3	KATILMAK	TUR10-1242600	girmek	Katılmak	BU ANKETE KATILMAK İSTEMİY...
4587.train	4	KATILMAYACAĞIM	TUR10-1242600	girmek	Katılmak	MÜSAİT DEĞİLİM ANKETE KATI...
5000.test	2	KATILMAYACAĞIM	TUR10-1242600	girmek	Katılmak	ANKETE KATILMAYACAĞIM .
5226.train	2	KATILMIŞTIM	TUR10-1242600	girmek	Katılmak	ANKETE KATILMIŞTIM .
8814.train	4	KATILMIŞTIM	TUR10-1242600	girmek	Katılmak	MAİL ÜZERİNDEN ANKETE KA...
9868.train	3	KATILMIŞTIM	TUR10-1242600	girmek	Katılmak	ANKET ÇALIŞMASINA KATILMIŞ...
9928.train	4	KATILMIŞTIM	TUR10-1242600	girmek	Katılmak	ANKETE DAHA ÖNCE KATILMI...
7985.train	11	KATIYOR	-	-	-	TESİSE 3. GİDİŞİM VE HER YIL ...
7986.train	17	KATIYOR	-	-	-	ÖN BÜRO TAKIMINA ÇOK TEŞ...
7988.train	11	KATIYOR	-	-	-	TESİSE 3. GİDİŞİM VE HER YIL ...
7317.train	11	KATLA	-	-	-	ODAYI TEMİZLEYEN GÖREVLİY...
1005.test	13	KATLARA	TUR10-0810130	üst kat	Bulunula...	ERTESİ GÜN ODANIN TEMİZLE...
9325.train	5	KATTAKİ	TUR10-0427040	kat daire	Bir yapını...	AKŞAM YEMEĞİ İÇİN 4. KATTAK...

Figure 2: Interface used in the second phase

Table 7: A review sample from the Tourism Domain

AİLE OTELİ OLARAK TAVSİYE EDERİM .
I recommend the hotel as a family hotel.
HERŞEY GÜZELDI .
Everything was good.
ÇOCUKLU AILELERE ÖNERIRIM .
I recommend the hotel to families with children.
DENIZI ÇOK GÜZELDI .
The sea was nice.

rected sentences for the annotations. For instance, the sentence *"HERŞEY GÜZELDI ."* is not orthographically correct since the lemma *ŞEY* (thing) should be written separately from the previous word *HER* (every) according to the standard orthographic conventions of Turkish. Moreover, since there is a *capital i* (İ) in the Turkish alphabet, the *I* should be corrected as *İ*. In this case, the correct form of this sentence would be *"HER ŞEY GÜZELDİ ."*.

Following the pre-processing of the data, we manually assign POS tags to each word in order to perform morphological analysis. For instance, the word *"Samsun"*, which is a city in Northern Turkey, is a proper name and its tag is represented as "IS_OA" in the dictionary. Similarly, the word *"ev"* (house) is a common noun and it is represented as "CL_ISIM" in the dictionary. Table 8 shows the percentages of the four most frequently used POS tags in the Tourism and Turkish dictionaries, which are IS_OA (proper name), CL_ISIM (common name), CL_FIIL (verb), and IS_ADJ (adjective) respectively.

Table 8: Percentage of frequently used POS tags of 2 dictionaries.

	Tourism	Turkish
PROPER NOUN	32.44	36.87
NOUN	45.92	53.07
VERB	8.42	8.35
ADJECTIVE	13.53	7.38

Since users or customers generally use the daily language in texts, the Tourism Dictionary has a lot of words in common with the Turkish Dictionary, which accounts for the result that 70.5% of the Tourism Dictionary is identical to the Turkish Dictionary. Table 9 shows the percentage of the POS tags of the intersecting words in the Tourism and Turkish dictionaries.

Table 9: Percentage of frequently used POS tags of common words in Tourism-Turkish dictionaries.

	Tourism-Turkish
PROPER NOUN	28.41
NOUN	51.27
VERB	9.19
ADJECTIVE	12.64

Table 10: The percentages of the top 5 hypernym relations in the Tourism WordNet

Otel (Hotel)	42.74
İlçe (District)	4.17
Ülke (Country)	2.23
Şehir (Town)	1.90
İl (City)	1.61

Table 11: Percentages of analyzed sentences and words with different sizes of tourism dictionaries and a Turkish Dictionary.

Dictionary	**Size**	**Sentence**	**Word**
Tourism	5,000	98.52	99.66
Tourism	10,000	98.93	99.75
Tourism	14,000	98.92	99.75
Turkish	51,552	95.97	99.07

Furthermore, we have extracted the hypernym relation, i.e., the hierarchy of word-senses from WordNet to obtain a more precise picture of the data. Table 10 shows the top 5 hypernyms in the tourism domain. As expected, the tourism dictionary predominantly consists of hotel names under the word hotel.

6.2 Morphological Analysis Tests

We have created a domain-dependent dictionary and WordNet using the dataset described in Section 6.1, and performed some analyses with the newly created domain-specific dictionary WordNet, and the general Turkish Dictionary. In order to validate our lexicon, we have tested it on tourism datasets and compared the results with that of the general Turkish Dictionary on the same datasets.

Table 11 shows the results of two analyses, a sentence-based and a word-based analysis, for three different sizes of tourism dictionaries and a Turkish dictionary. For the sentence-based analysis, we check the Tourism Dictionary's ability to correctly perform a morphological analysis of 20,000 sentences. For the word-based analysis, we check the accuracy of the performance of a morphological analysis on each of the 93,483 words separately. It can been observed that there is a 2.55% improvement in the sentence-based analysis, and the results of the word-based analysis are also similar. Nevertheless, after the dictionary size reaches 10,000 entries, no sufficient improvement is observed.

Table 12: Morphological analyses of size 1 using different dictionaries

	% of Morphological Analyses
Tourism	61.05
Turkish	54.11

Table 13: The 20 topmost annotated synsets and their counts

Id	SynSet	Count
TUR10-1081860	.	19,995
TOU01-1010440	çok	3,016
TUR10-0388960	iyi	2,529
TUR10-0105580	bir	1,981
TOU01-1063690	memnun kalmak	1,929
TUR10-0000000	(özel isim)	1,759
TUR10-0624490	personel	1,557
TUR10-0318110	güzel	1,396
TUR10-0513570	yemek	1,330
TUR10-0495010	tesis	1,247
TUR10-0816400	ve	1,221
TUR10-0346660	hizmet	1,042
TUR10-0593590	otel	1,014
TUR10-1121820	puan vermek	1,010
TUR10-0318100	güzel	957
TUR10-0097260	bey	924
TUR10-0582130	oda	915
TUR10-0187890	değil	769
TUR10-0473520	konum	740
TUR10-0565860	ilgili	708

Having multiple morphological analyses for a word introduces an ambiguity problem. With our approach, we aim to address this ambiguity issue by diminishing the dictionary size. To do so, we include only the domain-related senses of words, and discard the rest. To test its performance, we count the number of the words that have only one possible morphological analysis. This leads to a 7% improvement in the tourism domain as shown in Table 12. Thus, it is plausible to assert that reducing the dictionary size is an effective method to solve the disambiguation problem.

6.3 Semantic Annotation Statistics

Following the processing of 20,000 sentences, 93,653 words were annotated semantically, during which a total of 1,849 senses were used. While only 111 of these were from the Tourism WordNet, the remaining 1,737 were from the Turkish WordNet. As for the words, while 8,455 were annotated with senses from the Tourism WordNet, the remaining 85,186 were annotated from the Turkish WordNet. The results showed that 4,788 entries among the 13,555 in the Tourism WordNet were specific to the tourism domain whereas the remaining 8,767 were from the Turkish WordNet.

As can be seen in Table 13, function words such as ”değil (not), bir (a), ve (and)” are highly frequent, which is an expected case regardless of domain. However, the domain effects are observable through content words such as ”personel” (staff), ”tesis” (facility), ”hizmet” (service) and ”otel” (hotel)”, which make up a significant portion of the corpus. As the data is comprised of customer reviews, adjectives such as ”iyi (good), güzel (good / pretty)” are also highly frequent. Furthermore, due to the inclusion of punctuation, the full stop at the end of each sentence appears as the most frequent ”word”. Other frequent words that are not listed in Table 13 include evaluative adjectives such as ”yeterli (sufficient), kötü (bad)” and of course the comma. Finally, another anticipated result is the frequent occurrence of proper names such as the names of hotels and hotel staff.

As mentioned previously, multi-word expressions were also included in the annotation process. Table 14 shows that the majority of these were expressions such as ”memnun kalmak” (to be satisfied) or ”puan vermek” (to give points), frequently used in customer reviews. The inclusion of multi-word expressions were not limited to two-word expressions; thus, the occurrence of three and even four-word expressions was also frequent.

As shown in Table 15, the majority of the sentences in the corpus have a length of three to six words, while there are also sentences longer than 10 words, which make up a minority. At the end of the two-step process, approximately 100.000 words have been annotated, and a significant portion of these annotations have been observed to be a small set of frequently repeated expressions. Most of these frequent expressions have been annotated semi-automatically. Therefore, the words that took the longest time to annotate were the least frequent ones, occurring once or twice in the entire corpus.

Table 14: The 20 topmost annotated multi-word synsets and their counts

Id	SynSet	Count
TOU01-1063690	memnun kalmak	926
TUR10-1121820	puan vermek	404
TUR10-1154960	tavsiye etmek	321
TUR10-1181550	tercih etmek	187
TUR10-0728240	güler yüzlü	154
TUR10-0893550	yardımcı olmak	113
TUR10-1160460	teşekkür etmek	102
TUR10-0181700	damak tadı	51
TUR10-0847620	yeme içme	47
TOU01-1063820	aile oteli	45
TUR10-0839560	sağ olsun	43
TUR10-0227360	haberdar olmak	42
TUR10-1199410	bilgi vermek	34
TOU01-1041440	aqua park	30
TUR10-0089100	hoşuna gitmek	26
TOU01-1063770	çocuk dostu	24
TUR10-0004240	açık büfe	20
TUR10-0019600	dört dörtlük	19
TUR10-0565860	ilgi alaka	17
TUR10-0084000	her zaman	16

Table 15: Number of words in a sentence and their occurrences

# of Words	# of Occurences
2	824
3	4,475
4	6,761
5	4,584
6	1,632
7	601
8	341
9	157
10	134

7 Conclusion

Overall, we have created a domain-specific lexicon with user reviews and preferences from the tourism domain. Based on this newly created lexicon, we have designed a novel WordNet, and employed it for domain-specific sentiment analysis. By doing so, we have managed to mitigate the disambiguation problem for this specific domain. Finally, we have improved the performance of sentence-based morphological analysis by approximately 7% in the tourism domain.

References

Ozge Bakay, Ozlem Ergelen, and Olcay Taner Yildiz. 2019a. Integrating Turkish WordNet KeNet to Princeton WordNet: The case of one-to-many correspondences. In *Innovations in Intelligent Systems and Applications*.

Ozge Bakay, Ozlem Ergelen, and Olcay Taner Yildiz. 2019b. Problems caused by semantic drift in wordnet synset construction. In *International Conference on Computer Science and Engineering*.

O. Bakay, O. Ergelen, E. Sarmis, S. Yildirim, A. Kocabalcioglu, B. N. Arican, M. Ozcelik, E. Saniyar, O. Kuyrukcu, B. Avar, and O. T. Yıldız. 2020. Turkish WordNet KeNet. In *Proceedings of GWC 2020*.

L. Bentivogli, A. Bocco, and E. Pianta. 2003. Archiwordnet: Integrating wordnet with domain-specific knowledge.

O. Bilgin, O. Cetinoglu, and K. Oflazer. 2004. Building a wordnet for Turkish. *Romanian Journal of Information Science*, 7:163–172.

W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum. 2006. Introducing the Arabic wordnet project. In *International Wordnet Conference*, pages 295–300. Masaryck University, Brno, Czeck Republic.

X. Chen, C. Chen, D. Zhang, and Z. Xing. 2019. Sethesaurus: Wordnet in software engineering. *IEEE Transactions on Software Engineering*, pages 1–1.

R. Ehsani, E. Solak, and O.T. Yildiz. 2018. Constructing a wordnet for Turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3):24.

C. Fellbaum. 1998. *Wordnet: an electronic lexical database. Cambridge*. MIT Press, MA, USA.

G.A. Miller. 1995. Wordnet: a lexical database for English. *ACM Communications*, 38:39–41.

Riza Ozcelik, Selen Parlar, Ozge Bakay, Ozlem Ergelen, and Olcay Taner Yildiz. 2019. User interface for Turkish word network KeNet. In *Signal Processing and Communication Applications Conference*.

Maria Teresa Sagri, Daniela Tiscornia, and Francesca Bertagna. 2004. Jur-wordnet.

D. Tufis, D. Cristea, and S. Stamou. 2004. Balkanet: Aims, methods, results and per- spectives. a general overview. *Romanian Journal of Information Science*, 7:9–43.

V. Vossen. 1997. Eurowordnet: a multilingual database for information retrieval. In *DELOS workshop on Cross-language Information Retrieval*. Vrije Universiteit, Amsterdam, Czech Republic.