ÇUKUROVA UNIVERSITY

ENGINEERING AND ARCHITECTURE FACULTY

DEPARMENT OF COMPUTER ENGINEERING

GRADUATION THESIS

SUBJECT

Generation Of English-Turkish WordNet For Machine Translation

BY

2014555045 - Emine KISKANÇ

ADVISOR

Assoc.Prof. Umut ORHAN

June 2019

ADANA

ABSTRACTION

Machine Translation is one of the sub-branch of Natural Language Processing(NLP).Machine translation (MT) is fully automated software that can translate source content into target languages[1].

WordNet is a database of English words that are linked together by their semantic relationships. It is like a supercharged dictionary/thesaurus with a graph structure[2].

The aim of this study was to create an English-Turkish WordNet for Machine translation (MT). English WordNet[3]and Turkish-English dictionary with labels of words(127157 entries, 826K) [4] are used.

Key Terms: Machine Translation, WordNet, Graphs, Natural Language Processing.

LIST OF ABBREVIATIONS

PWN	:Princeton WordNet
NLP	:Natural Language Pcosessing
MT	:Machine Translation
WN	:WordNet
MLWN	:Multilingual WordNet
R.code	:Retionships code

LIST OF SHAPES

CONTENTS

INTRODUCTION

In spite of the increase of MT technologies, due to the lack of data (multilingual dictionaries, parallel corpus etc.), we cannot make translations with high accuracy rate. This study aims to create English-Turkish WordNet for use as source in English-Turkish machine translations. In section 1, the previously multilingual WN are mentioned. In section 2, the method and resources of the study are mentioned. In section 3, the results obtained are mentioned. In section 4, possible ambiguities were estimated when WN was used in machine translation and solutions were proposed. In the 5th setion, what can be done in the later studies is mentioned.

1.MULTILINGUAL WORDNETS

The most important factor for MT is the efficient dataset that can be used in translation. As it is known, the semantic relations of the words are important for MT. WordNets are used as the main source because it includes these semantic relations. Multilingual WordNet studies are important because there are basically two distinctions, including the source language and the target language. This section includes multilingual WordNet studies. In section 1.1 is mentioned EuroWordNet that includes several European languages.In section 1.2 is mentioned the Italian Wordnet which names MultiWordNet, created with the Wordnet of Princeton University.In section 1.3 is mentioned BalkaNet is included Balkan languages.

1.1.EuroWordNet

One of the most relevant endeavours has been the development of EuroWordNet, a project based on Wordnet structure whose ultimate purpose is to develop multilingual databases with Wordnets for several European languages. Each WN adopts an autonomous lexicalisation structure and all are interconnected through an interlinguistic index, for which relations have been added and modified and new levels identified in WN. For a multilingual description of EuroWordNet see[5,6].

1.2.MultiWordNet

MultiWordNet is a multilingual lexical database in which the Italian WordNet is aligned with Princeton WordNet 1.6[7,8].

The Italian synsets are created in correspondence with the Princeton WordNet synsets, whenever possible, and semantic relations are imported from the corresponding English synsets; i.e., we assume that if there are two synsets in PWN and a relation holding between them, the same relation holds between the corresponding synsets in Italian. While the project stresses the usefulness of a strict alignment between wordnets of different languages, the multilingual hierarchy implemented is able to represent true lexical idiosyncrasies between languages, such as lexical gaps and denotation differences.

The information contained in the database can be browsed through the MultiWordNet browser, which facilitates the comparison of the lexica of the aligned languages.

1.3.BalkaNet

Balkan WordNet aims to develop a multilingual dictionary database of WordNets for Balkan languages. The aim of Balkanet is to represent the semantic relations between words in each Balkan language and to link them together to develop a multilingual semantic network. Semantic relations will be classified in the independent WordNets according to a shared ontology. Then, all individual WordNets will be organized into a common database providing linking across them. Each of the WordNets will be structured along the same lines as the EuroWordNet through a WordNet Management System[9].

2.METHOD AND RESOURCES

In this section, resources are introduced (2.1.) and the method are mentioned(2.2.).

2.1.Resources

In this study basically two sources are used.One of these source is an English wordnet formed from the Wordnet of Princeton University, consisting of 206978 nodes ,1851093 relationships and 149229 different meaning words.The other is an English-Turkish dictionary has 127157 words. Each word in this dictionary has a format like that;

ENGLISH	TURKISH	LABEL
a little	azıcık	{a}
a little	azıcık	{adv}
a little	bir parça	{a}
a little	biraz	{adv}

Figure 1-Example of English-Turkish Dictionary Format

The Labels are:

Label	Represents	Label	Represents
{adv}	adverb	{k}	?
{a}	adjective	{n}	noun
{det}	determiner	{prep}	preposition
{id}	idiom	{v}	verb

Figure 2-Labels In Dictionary

The English WordNet node format is as follows:

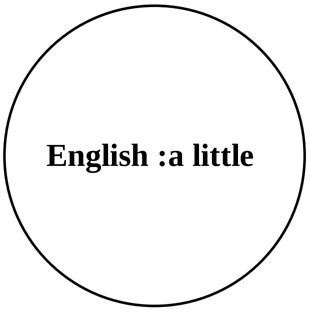
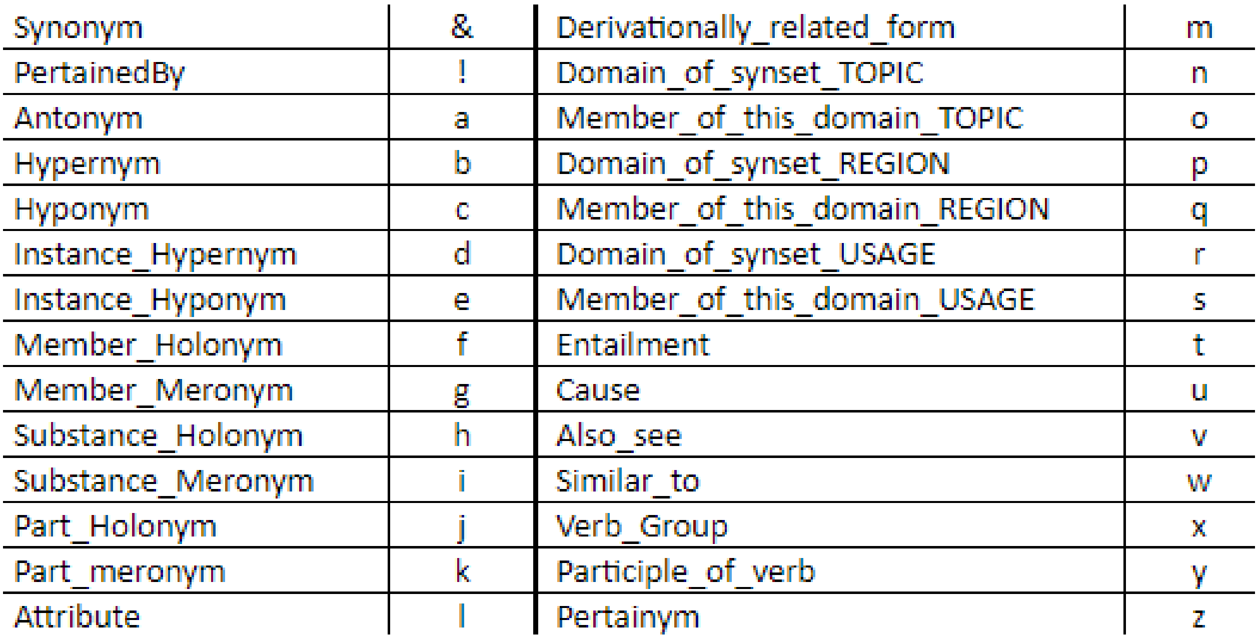


Figure 3-English WordNet Node Format

The English Wordnet relationship types and codes are as follows:

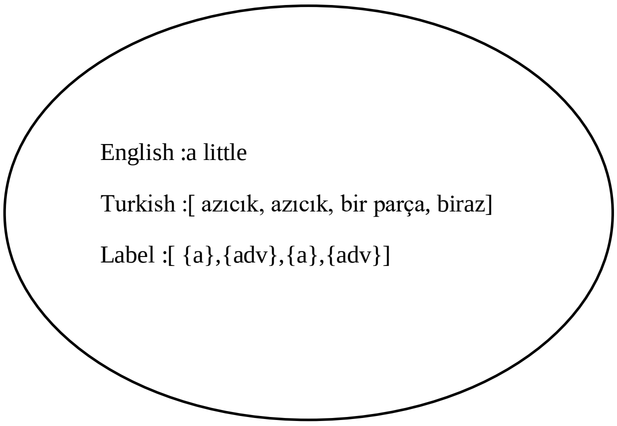


Relation type	R.code	Relation type	R.code
Synonym	&	Derivationally_related_form	m
PertainedBy	!	Domain_of_synset_TOPIC	n
Antonym	a	Member_of_this_domain_TOPIC	o
Hypernym	b	Domain_of_synset_REGION	p
Hyponym	c	Member_of_this_domain_REGION	q
Instance_Hypernym	d	Domain_of_synset_USAGE	r
Instance_Hyponym	e	Member_of_this_domain_USAGE	s
Member_Holonym	f	Entailment	t
Member_Meronym	g	Cause	u
Substance_Holonym	h	Also_see	v
Substance_Meronym	i	Similar_to	w
Part_Holonym	j	Verb_Group	x
Part_meronym	k	Participle_of_verb	y
Attribute	l	Pertainym	z

Figure 4-English-Turkish WordNet Relationships

2.2.Method

In used WordNet nodes and relation types are defined. The purpose of this study was to find the Turkish equivalents and labels of the English words in the nodes from the dictionary and add them to the nodes as new properties. The target node format is as follows:



English :a little

Turkish :[azıcık, azıcık, bir parça, biraz]

Label :[{a},{adv},{a},{adv}]

Figure 5-The English-Turkish WordNet node format

2.2.1.Step 1

In order to reach the node format in Figure-5, firstly univocal Turkish words were added to the English wordnet.For example:

zoo	hayvanat bahçesi	{n}
zucchini	kabak	{n}
abduct	kaçırmak	{v}

In dictionary, There are 18144 words which are non-homonym. 13684 of them were placed in WordNet.

2.2.2.Step 2

The second step was to add the Turkish meanings and tags of words with more than or equal two homonym words to WordNet as a list.Like that for “a little”:

Turkish :[azıcık, azıcık, bir parça, biraz]

Some examples more than or equal two homonym words:

(1)

a priori	muhtemel	{a}
a priori	muhtemel	{adv}
a priori	olasi	{a}

(2)

abc	alfabe	{n}
abc	ilkeler	{n}

(3)

absurd	abes	{a}
absurd	anlamsiz	{a}
absurd	mantiksiz	{a}
absurd	olanaksiz	{a}
absurd	saçma	{a}
absurd	ipe sapa gelmez	{a}

(4)

a little	azicik	{a}
a little	azicik	{adv}
a little	bir parça	{a}
a little	biraz	{adv}

If “a little “ represents (a) word,for (a), the mean of Turkish (a) specified in the form of a.Turkish[0] and the label of (a) specified in the form of a.Label[0]. So, the table in Figure-1 is updated for (a) as follows:

a.English = “a little”	
Turkish	Label
a.Turkish[0]=” azıcık”	a.Label[0]=” {a}”
a.Turkish[1]=” azıcık”	a.Label[1]=” {adv}”
a.Turkish[2]=” bir parça”	a.Label[2]=” {a}”
a.Turkish[3]=”biraz”	a.Label[2]=” {adv}”

Figure 6-Represent Table Of Figure-1

3.RESULTS

In this section, the results of the study were evaluated. And the statistics of placed words are presented.There are 206978 nodes in English WordNet. 74804 words were placed at the end of the process. Also, 52353 words could not placed.The results are presented in the following table:

Number of Homonym	Dictionary	Added(in WN)	Percentage(%)
1	18.144	13.684	75,4
2	17.230	10.136	58,8
>= 3	91.783	50.984	55,5

Figure 7-Results and Statistics

WordNet	Added (%)	Dictionary	Added(%)
206978 (nodes)	36,1	127157(words)	58,8

Figure 8-Total Statistics

Some words cannot be placed because the words in the dictionary and wordnet are not enough overlapped.

4.POSSIBLE AMBIGUITIES AND SOLUTIONS

The most important feature of machine translation is that it is not intended to be subject to any supervision during translation. So it is wanted to install automatic systems. This situation brings some problems. This section consists of semantic uncertainties due to the dictionaries used in machine translation and solutions to be proposed to solve them.

4.1 Some example of ambiguities

The fact that the words are firm can cause various ambiguities. In English words, we often encounter homonym problem in the pronunciation.For instance , ”two”-”to”,”ate”-”eight”,”you’re”-”your”,etc[10].

In English, homonym words is a problem for speech recognition while in Turkish, homonym words is a problem for text processing because of Turkish as a language read as written. Therefore, this section will focus on the Turkish homonym words. The ambiguities created by the homonym words are illustrated below.

Lemma 1: Homonym words can be obtained by subtracting some words from some words or with making attachments and shooting attachments[11].

Gül:

1. Çiçek (flower),

2. Imperative of “gülmek(laugh)” is a verb

Kır:

1. kırsal alan(rural area),

2. imperative of “kırmak(break)”,

3. beyaz (grey-heired)

Yazma:

1. baş örtüsü(scarf),

2.negative imperative of "yazmak(write)",

3. yazma işi(writing job)

Geç[13]:

1.beklenen zamandan sonra(late)
2. Imperative of "geçmek(pass,go)"

Yüz:

1.Çehre, surat, sima (face, visage)

2. doksan dokuzdan sonra gelen sayı (hundred)

3.Imperative of "Yüzmek(swim) "

4.yüzmek,derisini çıkarmak, soymak (shave)

Lemma 2: There are some words which are totally same written words.Like in the bellow sentence:

"Çay kenarında çay içmek zevklidir."[11].There is two meanings "çay" in this sentence.

1.tea

2.watercourse

So, the translation of this sentence must be :

"Tea is a pleasure to drink alongside of watercourse"

Kara:

1.siyah(black)

2.yeryüzü,toprak parçası (land)

Arı[13]:

1.bal yapan böcek(bee)
2. katıksız saf(pure)

Arz[13]:
1.sunma(supply)
2.yeryüzü yer(earth)

Sakin[13]:
1.durgun sessiz(quiet)
2.bir yerde oturan(habitant,resident)

Zar[13]:
1.ince deri ya da kabuki(skin, integument)
2.küp biçiminde üzerinde sayılar bulunan oyun aracı(die)

As you can seen, there are quite a lot of ambiguities about the translation –particularly- from Turkish into English. Suggestion a way to overcome these ambiguities have been proposed in below:

4.2.Suggestion

The location of the words in the sentence is important. The words in the sentence in Turkish are sorted by SOV (Subject-Object-Verb) structure. According to this structure, the subject must be primarily the object in the middle and the predicate must be at the end of the sentence.In Turkish, predicates are usually verbs. For this reason in the ambiguities encountered during translation, the meaning of the word can be determined by looking at the word's label and position in the sentence.For example:

Yüz:

1.Çehre, surat, sima (face, visage)	{n}
2. doksan dokuzdan sonra gelen sayı (hundred)	{n}
3.Imperative of "Yüzmek(swim) "	{v}
4.yüzmek,derisini çıkarmak, soymak (shave)	{v}

➔ O denizde yüzüyordu.
Step-1: If word of "yüzüyordu" is lemmatized, lemma of the word is "yüz".
Step-2:"yüzüyordu" is at end of the sentence and it is predicate.
Step-3:So, it should be a verb.
Step-4:There are two means of word(yüz) with {v} label.
a).yüzmek (swim)
b). yüzmek,derisini çıkarmak, soymak (shave)
Step 5: There are two ways:
1.Either we take first meaning or frequently used meaning.
2.or we should look other words in sentence

If we choose second way;

Since data is stored relationally as wordnet, translation of words can be guess by choosing the closest relational path to the in the sentence by looking at the other word in the sentence.

In this sentence,there are two others words except predicate("-yüzüyordu"):

a).O (lemma is "o(he/she/it)")

b).denizde (lemma is "deniz(sea)")

And the predicate has two meaning form of verb("-yüzmek"):

1.yüzmek (swim)

2.yüzmek ,derisini çıkarmak, soymak(shave)

s: the number of other words in the sentence

t : other meanings of the predicate

$$f(x) = \sum_{i=1}^{sxt} x = \sum_{t=1}^{t} \sum_{s=1}^{s} (\text{shortest\_path}(\text{s}, \text{t}))$$

Note that, the ∑ (summation) is used to indicate the loop.

The results in f(x) which are x, the minimum should be chosen.

There may be disadvantages of this method. The number of words in the sentence may be increased or a word can have more than 2 meanings. These reasons can increase the number of dual you need to make comparisons. Therefore the cost can be increased.

5. WHAT CAN BE DONE

As mentioned in section 1, multi-lingual wordnet studies are available. Some studies have also been made for Turkish but multi-lingual wordnet is not prepared at the desired level since multilingual dictionaries do not contain enough words or have without class labels. Methods can be developed for automatic generation of multilingual data sets with labeled data for use in the further multilingual wordnet studies.

Translation platforms with multiple supervisors can be created for Turkish-English or other languages. Supervisors on these platforms:

1. Can tag words

2. Can indicate what the meaning of words is meant

3. Or can translate sentences completely.

These corrections made by the supervisors can be automatically generated by an English-Turkish multilingual wordnet using as training data.

Also, every supervisor should see every corrections because of wrong tagging data.

REFERENCES

[1] https://www.gala-global.org/what-machine-translation

[2] https://stevenloria.com/wordnet-tutorial/

[3] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[4] http://www.fen.bilkent.edu.tr/~aykutlu/sozluk.txt

[5] Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities, 32(2–3) (1998) 73–89.

[6]. Green, Rebecca, Pearl, L., Dorr, B.J., and Resnik, P.: Mapping lexical entries in verbs database to WordNet senses. Proceedings of the 39þ Annual Meeting of the Association for Computational Linguistics (ACL 2001), Toulouse, France, July 9–11 (2001).

[7]. https://wordnet.princeton.edu/

[8]. http://multiwordnet.fbk.eu/english/home.php

[9]. http://www.dblab.upatras.gr/balkanet/

[10] https://www.fluentu.com/blog/english-tur/ingilizce-sestes-sozcukler/

[11] http://sestes-kelime-ornekleri.nedir.org/

[12] https://www.dilbilgisi.net/konular/es-sesli-kelimeler/

[13]https://www.turkdilbilgisi.com/sozcukte-anlam/es-sesli-kelimeler-ornekler-sozlugu.html