Generating a Concept Relation Network for Turkish Based on ConceptNet Using Translational Methods

Arif Sırrı Özçelik[1] and Tunga Güngör[1]

[1] Boğaziçi University, Computer Engineering, Istanbul, Turkey
arifsozcelik@gmail.com, gungort@boun.edu.tr

Abstract. ConceptNet is a large-scale network of concepts and relationships, based on various common sense knowledge bases. Turkish is a language that lacks similar resources for processing texts and extracting meaning. This study discusses various methods to create a Turkish ConceptNet using translational techniques based on English ConceptNet and explains the results herewith obtained. Multiple models were tested, using different knowledge sources and tools including WordNet, Wikipedia, and Google Translate. Results obtained from each model and the approaches to improve these results are discussed, while also explaining details, assumptions, and drawbacks relevant to each relation.

Keywords: Common sense knowledge base, ConceptNet, Turkish, Word sense disambiguation

1 Introduction

Common sense databases are resources used for extracting deeper semantic knowledge in texts. They express common sense knowledge in simple sentence forms. Examples are “birds have wings”, “the sun is very hot”, “candy tastes sweet”, and so on. When it comes to text processing, humans naturally and extensively use this source of knowledge in understanding and drawing conclusions. So, to actually capture the semantics of any given text, an apriori existence of this knowledge would help immensely.

ConceptNet [5] is a large-scale network of concepts and relations built initially on common sense databases. These assertions include examples like “pottery is made of clay” or “a cactus is capable of surviving with little water”. ConceptNet spans 12.5 million edges which represent 8.7 million assertions connecting 3.9 million multilingual nodes [10]. English is the most represented language by 11.5 million edges including at least one English concept.

ConceptNet defines a “Common Language” to be a language included in the network with a vocabulary size of at least 10 thousand terms. There are 68 languages considered to be a “Common Language” and Turkish is one of them. Turkish has a vocabulary size of around 66 thousand terms [11]. There are around 10.4 thousand assertions where both concepts are in Turkish.

Languages such as English, French, Portuguese have a good amount of ontological and common sense resources, but when it comes to Turkish there is an apparent need for similar sources. Turkish as a language lacks studies that either create common sense knowledge or somehow translate existing resources in other languages. In this work, we develop a concept relation network for Turkish by employing a method that translates from the ConceptNet resource. Our goal is to build an initial concept relation network that will benefit text processing systems which currently lack knowledge resources of this kind for the Turkish language.

2 Related Work

Studies with a similar aim of creating language resources for Turkish have been published in the past. Balkanet [2, 12] is a collective attempt to gradually create multilingual WordNet lexicons similar to WordNet [3] that spans Greek, Turkish, Romanian, Bulgarian, Czech, and Serbian. The work makes use of local monolingual WordNets if available, otherwise sources like dictionaries, corpora or language specific lexicons. The process then links each monolingual WordNet to an Inter-Lingual-Index that serves as a centralized index relating synsets among all languages.

Turkish WordNet [7] is a project lead by the Turkish team in the Balkanet project. The team started by translating base concepts into Turkish. Later on, a monolingual dictionary was used to extract synonyms, antonyms and hyponyms for these base concepts. In the second phase, the team gathered a “defining vocabulary” of most frequent words in the English language and compared these words to Turkish WordNet synsets. Missing terms were then used to extend the Turkish synset collection through hyperonym-hyponym relations.

In their study titled SentiTurkNet, Oflazer, Dehkharghani, Saygın and Yanıkoğlu [8] aimed to create a lexicon of polarity for words in Turkish similar to SentiWordNet for English that could be used by sentiment analysis methods. They used the Turkish version of WordNet [7] by semi-automatically assigning polarity values to create SentiTurkNet.

In their attempt to create a similar common sense list of assertions like what ConceptNet was built on, Özcan and Amasyalı [9] used an online game approach that would ask users to play a game and as a result generate common sense knowledge for Turkish. In this study, they look into a number of games previously implemented for English. They proposed using a game site called CSOYUN which they kept online for 4 years with 5 different games and reported that 57 thousand reliable concept relations were generated through these online games.

In another study aiming to create a Turkish WordNet named KeNet,[1] Yıldız, Solak and Ehsani [13] start by extracting synonym candidates from an online dictionary for Turkish. Then they verify synonyms by manually annotating them and create a graph where nodes represent senses connected by synonymy relations. Finally, by looking at clusters they create synsets. They also mined Turkish Wikipedia for hypernym relations that increased the set of such relations obtained using only a dictionary.

[1] http://haydut.isikun.edu.tr/kenet.html

3 Methods

ConceptNet relations are assertions describing relations between two different concepts. They represent everyday common sense information. Some example assertions are:

- a bowl - *MadeOf* - steel
- an organism - *MadeOf* - cells
- chip - *PartOf* - computer
- edinburgh - *PartOf* - scotland
- brain - *UsedFor* - think
- breathing - *UsedFor* - meditating

We develop a method that translates similar assertions into Turkish. Before starting translating a relation, the following preparations and assumptions were made:

- Only English to English relations in ConceptNet were considered,
- Nodes on each side of the relation were preprocessed to remove initial stop words like "a", "an", "the", etc,
- Any translation of a concept to Turkish that fails is assumed to be a technical or domain specific term, so can be accepted as it is in Turkish,
- Except a few specific relations, all concepts were translated in their singular forms,
- Depending on relations certain Part of Speech (POS) categories (noun, adjective, verb, etc.) were used to filter senses while translating concepts,
- English terms were lemmatized using the Stanford Core NLP tool [6],
- Turkish terms were lemmatized using Zemberek [1],
- Crawlers were used to extract data from sites like Tureng,[2] Wiktionary,[3] Wordreference,[4] Wikipedia,[5] and Google Translate.[6]

ConceptNet includes 58 relations. Some of the relations were not included in this work because either there were no English to English examples in the relation (e.g. *TranslationOf*) or there were too few examples (e.g. *ParticipleOf* or *LocatedNear*).

Various models for translating English concepts into Turkish were tested, initially starting with using online bilingual dictionaries and then gradually introducing sources like WordNet, Wikipedia, Google Translate, and Google Search API.

Using only online dictionaries does not incorporate context and it is challenging to disambiguate between various translation candidates given that each example in a ConceptNet relation is short and lacks sufficient context.

[2] https://tureng.com/tr/turkce-ingilizce

[3] http://en.wiktionary.org

[4] https://www.wordreference.com/

[5] https://wikipedia.org/

[6] https://translate.google.com/

WordNet was used in an attempt to extend context best matching synsets. Synsets for one concept were filtered using the other concept by applying the Lesk [4] algorithm. An augmented version of the Lesk algorithm [4] was tested to enrich WordNet based contexts by adding hypernyms, hypernym ancestries, hypoynms, and part meronyms to the glosses of synsets for one concept before comparing them to synsets generated for the other concept.

As an example, given the concept “Laptop - *MadeOf* – Chip”, “Chip” is translated into Turkish in the sense of “French Fries”. The correct sense in WordNet includes terms like “Microchip”, “Silicon Chip”, and “Microprocessor Chip”. But “Laptop” reveals the terms “Laptop”, “Computer” and “Portable”. However, hypernym hierarchy for “Laptop” includes the terms “Microprocessor” and “Microcomputer". By using these hypernym terms it is possible to translate “Chip” in its correct sense in Turkish.

The final model to translate ConceptNet into Turkish used Google Translate, Google Search API, and Wikipedia. Fig. 1 shows the pseudocode of the algorithm.

```
FOR a relation type RELTYPE
FOR an instance of RELTYPE named RELATION
FOR each of the two concepts in RELATION named CONCEPT
 1. IF CONCEPT consists of more than 2 words, QUERY Google Translate
    for “CONCEPT” and RETURN the result as the correct translation.
    Otherwise proceed to next step.
 2. QUERY Google Translate as described in Fig. 2.
 3. LEMMATIZE all English terms in CONCEPT. LEMMATIZE all Turkish
    terms in TRANSLATION.
 4. QUERY Google Custom Search API for RELATION in Wikipedia
    specifically. COLLECT the first 10 ARTICLES.
 5. TRANSLATE and ALIGN sentences for Wikipedia ARTICLES as
    described in Fig. 3.
 6. ASSIGN scores to ARTICLE - TRANSLATION pairs as described in Fig.
    4. CHOOSE the highest scoring ARTICLE - TRANSLATION pair among
    all and RETURN TRANSLATION as the correct translation.
 7. REPEAT steps (1) through (7) for both concepts.
```

Fig. 1. Translation Model.

```
USING Google Translate:
  1. QUERY CONCEPT.
  2. COLLECT definitions, examples for source terms and the top 6 ranking
     TRANSLATIONS.
  3. IF there are no translations returned, default to Tureng.
     (a) IF Tureng does not return a TRANSLATION, RETURN CONCEPT.
     (b) Otherwise RETURN the first Tureng entry.
```

Fig. 2. Translation Model – Step 2.

```
For each Wikipedia ARTICLE:
  1. TRANSLATE an extract of the ARTICLE using Google Translate
     including the abstract.
  2. ALIGN English and translated Turkish sentences for both versions of the
     ARTICLE.
  3. For each aligned sentence, EN and TR of ARTICLE:
     (a) LEMMATIZE all English terms in EN.
     (b) LEMMATIZE all Turkish terms in TR.
```

Fig. 3. Translation Model – Step 5.

```
For each ARTICLE - TRANSLATION pair:
  1. For each lemmatized aligned sentence, EN and TR of ARTICLE:
     (a) COUNT how many times EN includes CONCEPT and TR includes
     TRANSLATION.
     (b) Assign the number of times both sentences had matches as the
     ALIGNED SENTENCE SCORE for TRANSLATION.
  2. SUM all ALIGNED SENTENCE SCORES to assign the ARTICLE -
     TRANSLATION pair a score.
```

Fig. 4. Translation Model – Step 6.

The proposed model uses Google Translate to generate a list of translations for each concept, searches for Wikipedia articles related to the example in the form “concept1 relatesTo concept2”, translates Wikipedia article extracts using Google Translate, and finally scores each translation candidate on the basis of matching terms within both English and Turkish extracts.

Google Search API was used to search for Wikipedia articles. Google Translate would be capable of translating articles into Turkish with a certain degree of success as article extracts will contain many terms, hence a large context.

Fig. 5 lays out the the algorithm of translating a concept in a flow chart.

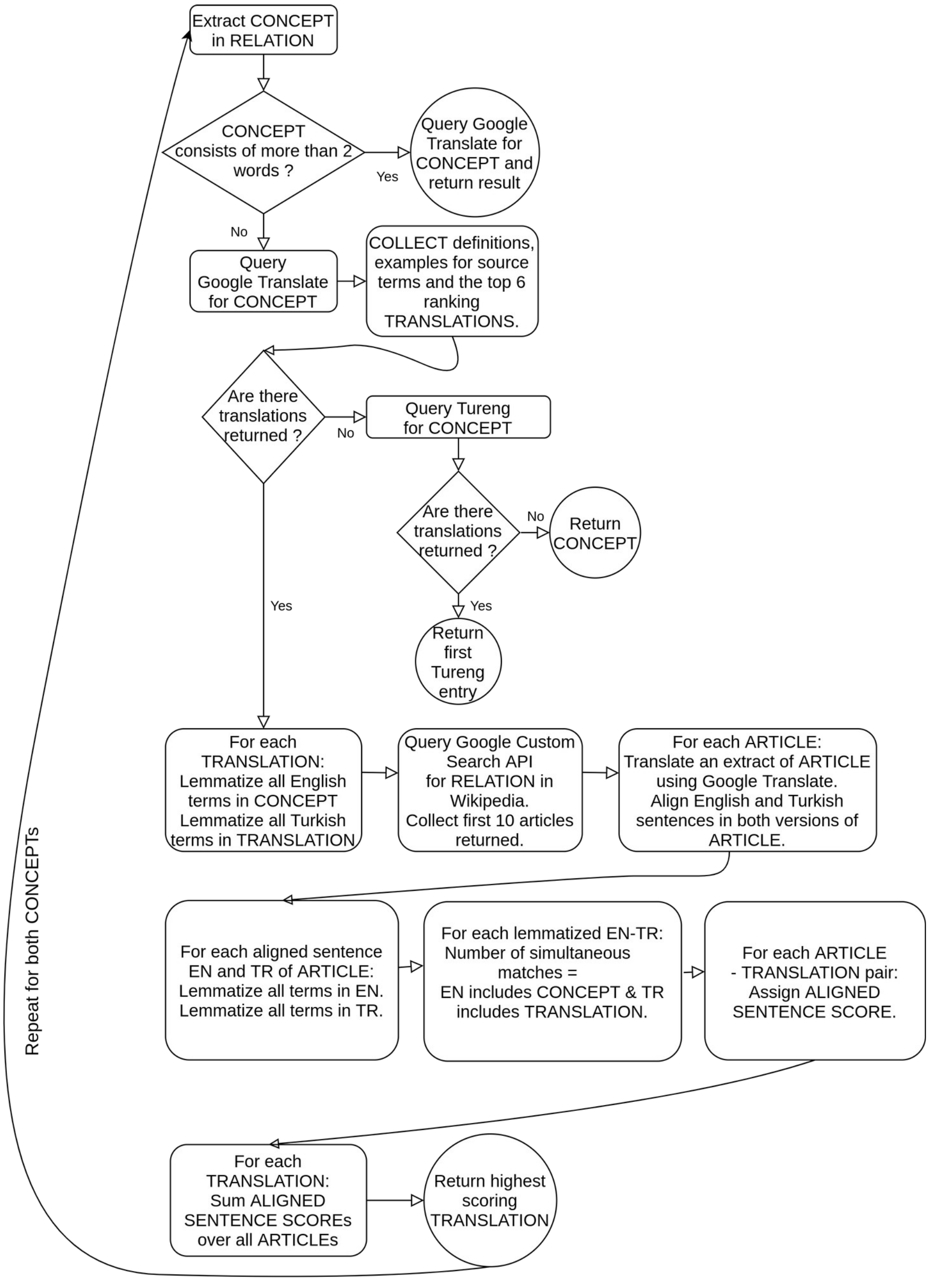
Extract CONCEPT in RELATION
CONCEPT consists of more than 2 words ?
Yes
Query Google Translate for CONCEPT and return result
No
Query Google Translate for CONCEPT
COLLECT definitions, examples for source terms and the top 6 ranking TRANSLATIONS.
Are there translations returned ?
No
Query Tureng for CONCEPT
Are there translations returned ?
No
Return CONCEPT
Yes
Return first Tureng entry
Yes
For each TRANSLATION: Lemmatize all English terms in CONCEPT Lemmatize all Turkish terms in TRANSLATION
Query Google Custom Search API for RELATION in Wikipedia. Collect first 10 articles returned.
For each ARTICLE: Translate an extract of ARTICLE using Google Translate. Align English and Turkish sentences in both versions of ARTICLE.
For each aligned sentence EN and TR of ARTICLE: Lemmatize all terms in EN. Lemmatize all terms in TR.
For each lemmatized EN-TR: Number of simultaneous matches = EN includes CONCEPT & TR includes TRANSLATION.
For each ARTICLE - TRANSLATION pair: Assign ALIGNED SENTENCE SCORE.
For each TRANSLATION: Sum ALIGNED SENTENCE SCOREs over all ARTICLEs
Return highest scoring TRANSLATION
Repeat for both CONCEPTs

Fig. 5. Translation Model – Flowchart.

4 Experiments and Discussion

Out of the 58 relations, 37 relations were translated. 10 relations were not included because they were small in size and 9 were not translated because they were huge and the resources of this study were limited. 2 of the relations, namely *influenced* and *influencedBy* were not translated because the majority of nodes for these relations were entity names, specifically famous people in various domains.

The following results assume a translation to be correct only if it makes sense in Turkish. Some translations are grammatically slightly incorrect or are seemingly correct but semantically not sensible either in the source example or translation. Some of the grammatical errors are caused by the tools used and others are results of source examples with poor quality.

Nearly all relations have examples that do not make much sense in English. There are also many examples which are hard to translate. Some examples are actually asymmetrically divided long sentences. They do not conform to the assumption that there are two concepts on both sides of a relation that are isolated units of meaning. These are considered to be incorrect.

Some of these unexpected examples are:

- difference between an entranceway and a patio door: patio door - *MadeOf* - glass
- pizza usually - *MadeOf* - tomato sauce, cheese and crust
- stabbing to death may - *SymbolOf* – dead domination to some person
- graph - *MadeOf* - set of vertices and a set of edge

Examples above do not actually satisfy the assumption made in this study that each side of a relation consists of simple isolated concepts.

Relations like *NotHasProperty*, *adjectivePertainsTo*, *adverbPertainsTo* and *NotCapableOf* did not seem to perform well under the model proposed. This is because either they contain many hard to translate or noisy examples or are very domain specific and it is hard to translate without a domain specific resource.

Some hard to translate examples for these relations are:

- accidents can happen to someone who - *NotHasProperty* - careful
- fenestral - *adjectivePertainsTo* - fenestra
- ravishingly - *adverbPertainsTo* - ravish
- television - *NotCapableOf* - need to be watered

For relations like *mainInterest*, *MemberOf*, and *notableIdea*, the concept to be translated can be very domain specific. Some examples are:

- martin heidegger - *notableIdea* - desein
- bomarea - *MemberOf* - amaryllidaceae

Examples where either concept consists of only stop words were discarded (not translated). Some examples of this type are:

- almost - *NotCapableOf* - count except in horseshoes
- they - *HasA* - nice smell
- it - *HasProperty* - dirty or clean

Table 1. Results for all relations.

Relation	Size	Coverage	Accuracy
SymbolOf	166	165	84%
DesireOf	280	275	83%
Entails	408	404	79%
NotHasA	409	390	61%
NotIsA	478	402	73%
CreatedBy	503	499	74%
Attribute	639	624	85%
notableIdea	908	908	72%
NotHasProperty	1144	1085	72%
MadeOf	2198	2177	82%
mainInterest	2764	2764	85%
adverbPertainsTo	2880	2841	61%
NotCapableOf	2915	2440	60%
HasLastSubevent	3065	3063	66%
adjectivePertainsTo	3313	3297	56%
HasFirstSubevent	4208	4202	62%
NotDesires	4280	4239	71%
Desires	5062	4870	74%
CausesDesire	5176	5158	69%
DefinedAs	6406	6179	61%
HasContext	8851	8615	71%
HasA	9762	9283	62%
ReceivesAction	10429	10090	61%
SimilarTo	11061	10679	74%
MemberOf	12190	12052	55%
PartOf	14151	13791	65%
spokenIn	15590	15427	53%
MotivatedByGoal	15960	15605	68%
Causes	18355	18143	55%
languageFamily	19713	19504	60%
HasProperty	19823	18615	67%
HasPrerequisite	24545	24155	69%
Antonym	26551	24478	71%
Field	26732	26450	83%
HasSubevent	26911	26602	62%
knownFor	27519	27224	75%
UsedFor	46522	45381	64%

Random samples of size 150 were selected and evaluated by the authors. Estimated accuracies were also based on annotations done by the authors.

Table 1 lists results after applying the proposed model to 37 relations. Size is the number of examples in a relation, Coverage is the number of these examples that were processed, and Accuracy is the relative frequency of successful translations obtained in randomly selected samples of size 150.

Higher scoring relations seem to have a tendency to contain shorter concepts while mid or lower scoring relations are more spread out. This is consistent with the assumption made throughout this study assuming ConceptNet consisted of short concepts on both sides of a relation. Concepts that are longer in size were not considered to be disambiguated and Google Translate or Tureng results were accepted instead.

Table 2. Estimated accuracies vs. concept lengths.

Relation	**Combined Concept Length**				
	2	**3**	**4**	**5**	**6**
mainInterest	100%	82%	83%	96%	75%
SymbolOf	89%	83%	72%	100%	40%
DesireOf	88%	77%	93%	82%	
F’eld	81%	83%	90%	100%	
MadeOf	88%	88%	65%	83%	25%
Entails	82%	81%	100%		
knownFor	77%	80%	66%	95%	
CreatedBy	89%	71%	83%	44%	40%
Desires	86%	67%	60%	80%	78%
SimilarTo	79%	56%			

Table 2 shows how successful translations are distributed when examples are grouped by combined concept lengths, for the 10 top scoring relations. A combined concept length is the sum of the number of words in each concept. As an example, 65% of *MadeOf* relations having a combined concept length of 4, were accurately translated. In most of the relations there seems to be a decrease of performance as the concepts become larger. For sizes larger than 4, at least one of the concepts are translated directly through Google Translate, which explains the increase in performance. The relation field is the only exception to this observation and this is mainly caused by technical or domain specific terms in concepts.

5 Conclusion

Building resources for Turkish like WordNet, ConceptNet or other common sense knowledge bases manually is time and resource consuming. Instead, attempting to translate resources from other languages is more feasible.

The work described throughout this study attempts to translate as many examples of ConceptNet relations as possible from English into Turkish by making use of different existing tools. The goal was to create a network of everyday knowledge for Turkish, a language that lacks a proper common sense knowledge base. Looking at the results, it could be said that the method used to translate performed slightly better with relatively small examples, consisting of simple nodes.

Future work could integrate KeNet [13] and possibly cross lingual WordNet links into the algorithm. It is also possible to improve incorrect translations through feedback implementations or manual corrections.

References

1. Akın, A.A.: Zemberek-NLP, https://github.com/ahmetaa/zemberek-nlp (2019).
2. Cristeau, D., Tufis, D.I., Stamou, S.: BalkaNet: aims, methods, results and perspectives. a general overview. Romanian Journal of Information Science and Technology, Vol.7, 9-43 (2004).
3. Fellbaum, C.: WordNet and wordnets. In: Encyclopedia of Language and Linguistics. Elsevier, Vol.3, 665-670 (2006).
4. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from a ice cream cone. In: Proceedings of SIGDOC, pp.24-26 (1986).
5. Liu, H., Singh, P.: ConceptNet: a practical commonsense reasoning tool-kit. BT Technology Journal, Vol.22 (2004).
6. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford coreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp.55-60 (2014).
7. Oflazer, K., Bilgin, O., Çetinoğlu, Ö.: Building a wordnet for Turkish. Romanian Journal of Information Science and Technology, Vol.7, 163-172 (2004).
8. Oflazer, K., Dehkharghani, R., Saygın, Y., Yanıkoğlu, B.: SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. Language Resources and Evaluation, Vol.50, 667-685 (2016).
9. Özcan, S., Amasyalı, M.F.: Turkish commonsense database (CSDB) and csoyun (a game with a purpose). Sigma Journal of Engineering and Natural Sciences, Vol.32, 116-127 (2014).
10. Speer, R., Havasi, C.: Representing general relational knowledge in conceptNet 5. Language Resources and Evaluation (2012).
11. Speer, R.: ConceptNet5 languages, https://github.com/commonsense/conceptnet5/wiki (2016).
12. Stamou, S., Azer, O., Pala, K., Christoudoulakis, D.: “BalkaNet: a multilingual semantic network for Balkan languages. 1st International Wordnet Conference, pp.12-14, India (2002).
13. Yıldız, O.T., Solak, E., Ehsani, R.: Constructing a wordNet for Turkish using manual and automatic annotation. ACM Transactions on Asian and Low-Resource Language Information Processing, Vol.17(3) (2018).