Towards Turkish Abstract Meaning Representation

Zahra Azin
Informatics Institute
Istanbul Technical University
Istanbul, Turkey
azin18@itu.edu.tr

Gülşen Eryiğit
Department of Computer Engineering
Istanbul Technical University
Istanbul, Turkey
gulsen.cebiroglu@itu.edu.tr

Abstract

Using rooted, directed and labeled graphs, Abstract Meaning Representation (AMR) abstracts away from syntactic features such as word order and does not annotate every constituent in a sentence. AMR has been specified for English and was not supposed to be an Interlingua. However, several studies strived to overcome divergences in the annotations between English AMRs and those of their target languages by refining the annotation specification. Following this line of research, we have started to build the first Turkish AMR corpus by hand-annotating 100 sentences of the Turkish translation of the novel "*The Little Prince*" and comparing the results with the English AMRs available for the same corpus. The next step is to prepare the Turkish AMR annotation specification for training future annotators.

1 Introduction

For a long time, semantic annotation of natural language sentences was split into subtasks, i.e. there were independent semantic annotations for named entity recognition, semantic relations, temporal entities, etc. The ultimate goal of Abstract Meaning Representation (AMR) is to build a SemBank of English sentences paired with their whole-sentence logical meaning. To do this, one of the primary rules in AMR annotating sentences is to disregard many syntactic characteristics to unify the semantic annotations into a simple, readable SemBank (Banarescu et al., 2013).

According to the Abstract Meaning Representation specification, AMR is not an Interlingua. The assertion has attracted researchers' attention to sample AMR formalism on different languages. Several researches have been done to examine the compatibility of AMR framework with other languages such as Chinese and Czech (Xue et al., 2014; Hajic et al., 2014; Li et al., 2016). Other studies proposed methods to generate AMR annotations for languages with no gold standard dataset by implementing cross lingual and other rule based methods (Damonte and Cohen, 2017; Vanderwende et al., 2015).

In this work, we have manually annotated 100 sentences from the Turkish translation of the novel "*The Little Prince"* with AMRs to describe the differences between these annotations and their English counterparts. The next step is to prepare the Turkish AMR guideline based on the differences extracted in the previous phase for training future annotators who wish to construct the first Turkish AMR bank by hand-annotating 1562 sentences of "*The Little Prince"* for which the English AMR bank is available.

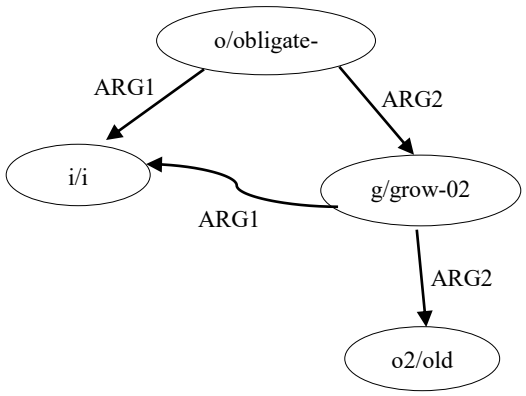
2 Abstract Meaning Representation

Abstract Meaning Representation is defined as a simple readable semantic representation of sentences with rooted, directional labeled graphs (Flanigan et al., 2014). The main goal was set to build a SemBank resembling the proposition bank which is independent and disregards syntactic idiosyncrasies.

The building blocks of AMR graphs are concepts represented in nodes and relations that hold among these concepts as the edges of the graph. Thus, instead of using syntactic features, AMR focuses on the relationships among concepts, some of which are extracted from PropBank and other words. Example 1 shows the English AMR for the sentence "I have had to grow old." The root of the graph is a reference to the sense *obligate-01* and is extracted from

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 43–47
Florence, Italy, July 28 - August 2, 2019.

PropbBank frames as the sentence contains the syntactic modal *had to*.



Example 1: The AMR annotation graph for the sentence " *I have had to grow old*."

AMR does not annotate every single word in the sentence since its goal is to represent the analysis of a sentence in predicative and conceptual levels. Furthermore, AMR does not represent inflectional morphology for syntactic categories like tense which results in the same meaning representation of similar sentences with different wordings or word order. For example, the two sentences "*The boss has decided to fire the employee."* and "*This is the boss decision to fire the employee.*" have same AMR annotations.

3 AMR Resources

Inspired by the UNL project[1], a freely downloadable annotated corpus of the novel "*The Little Prince"* containing 1562 sentences has been released by the project initiators[2]. The purpose was to release a corpus so that other researchers could compare their annotated sentences based on the same text. There is another annotated corpus, Bio AMR, freely available on the same website which contains cancer-related articles including about 1000 sentences. Moreover, Abstract Meaning Representation release 2.0 which contains more than 39,260 annotated sentences was developed by the Linguistic Data Consortium (LDC), SDL/language Weaver, Inc., The University of Colorado, and the University of Southern California and is distributed via the LDC catalog.

4 AMR Parsing

The ultimate goal of semantic formalisms such as Abstract Meaning Representation in natural language processing is to automatically map natural language strings to their meaning representations. In an AMR parsing system, we work on graphs which have their own characteristics specified by AMR formalism. These properties like reentrancy in which a single concept participates in multiple relations or the possibility to represent a sentence with different word orders by a single AMR make the parsing phase challenging. On the bright side, similar to dependency trees, AMR has a graph structure in which nodes contain concepts and edges represent linguistic relationships.

Several AMR parsing algorithms have been proposed so far (Wang et al., 2015; Vanderwende et al., 2015; Welch et al., 2018; Damonte et al., 2016; Damonte and Cohen, 2016) among which JAMR is the first open-source automatic parser published by the project initiators[3]. It works based on a two-stage algorithm in which concepts and then relations are identified using statistical methods. On the other hand, the transition-based method which transforms the dependency tree to an AMR graph seems promising because of its use of available dependency trees for different languages (Wang et al., 2015).

Sometimes, in natural language processing, due to limited resources or lack of NLP tools, researchers seek to discover methods to get the most out of it. Cross-lingual Abstract Meaning Representation parsing (Damonte and Cohen, 2017) for which we do not require a standard gold data seems to overcome the structural differences between English and a target language in AMR annotation process using "annotation projection" method. The parser works based on annotation projection from English to a target language and has been trained for Italian, German, Chinese, and Spanish.

[1] http://www.unlweb.net/unlweb/
[2] https://amr.isi.edu/download.html
[3] https://github.com/jflanigan/JAMR

Building a semantically hand-annotated corpus like an AMR bank is an arduous time-consuming task. However, annotating a small amount of data manually results in achieving an understanding of the formalism, in the first place, and facilitating the evaluation of AMR parsers. The annotated AMR corpus of this study can be utilized in evaluating future Turkish AMR parsers.

5 Turkish AMR

As AMR is not an interlingua, several studies have examined the differences between AMR annotations of sentences in languages like Chinese and Czech with English AMR annotations (Xue et al., 2014; Hajic et al., 2014; Li et al., 2016) so far and some have introduced cross-lingual and rule based methods to generate AMR graphs for languages other than English (Damonte and Cohen, 2015; Vanderwende et al., 2015). However, none of them had ever tackled an agglutinative language in which there is a possibility to derive and inflect words by cascading suffixes indefinitely.

One of the main challenges in developing language models for morphologically rich languages with productive derivational morphology like Hungarian, Finnish, and modern Turkish is the number of word forms that can be derived from a root. According to Turkish Language Association (TDK)[4], 759 root verbs exist in Turkish. Moreover, 2380 verbs are derived from nouns and 2944 verbs from verbs. Thus, there is almost no limit on suffixes a verb can take which results in tens of possible word formations.

Another challenge in Turkish processing is its free word order that allows sentence constituents to move freely at different phrase levels. One should note that as the word order changes, some pragmatic characteristics such as focus and topics change as well. This property of Turkish might lead to several challenges such as the need for collecting as much data as possible to cover all possible word orders.

For the first step, we have started hand-annotating the Turkish translation of "*The Little Prince"* aligning to its English AMR annotation to find out divergences and at the same time developing the very first Turkish AMR specification based on both English AMR guideline and differences between the two languages. The sentences were annotated by a non-Turkish linguist who aligned the English sentences with their literary translation in Turkish and created the AMR graphs using the Online AMR Editor[5]. Final annotations were proofread by a Turkish speaker.

We annotated 100 sentences and came up with following observations. First, a small number of sentences have exactly the same AMR structure as their English translation. An example is shown in figure 1. As it is illustrated in the textual form of the annotation, which is in the form of PENMAN notation (Matthiessen and Bateman, 1991), concepts and relations are aligned, although objects of the two sentences are different.

```
(t / talk-01
  :ARG0 (i / i)
  :ARG1 (a / and
        :op1 (b / bridge)
        :op2 (g / golf)
        :op3 (p / politics)
        :op4 (n2 / necktie))
  :ARG2 (h / he))
```

```
(k / konuşmak
    :ARG0 (b / ben)
    :ARG1 (v / ve
          :op1 (b2 / briç)
          :op2 (g / golf)
          :op3 (p / politika)
          :op4 (b3 / boyun-bağı))
    :ARG2 (o / onlar))
```

Figure 1: Textual forms of AMR annotations for the sentence "*I would talk to him about bridge, and golf, and politics, and neckties*." and its Turkish translation ("*Onlarla/them-with briç/bridge, golf/golf, politika/politics ve/and boyun bağları/neckties hakkında/about konuştum/I talked*.")

Second, most of the AMR annotations' divergences were due to different word choices in

[4] www.tdk.gov.tr

[5] https://www.isi.edu/cgi-bin/div3/mt/amr-editor/login-gen-v1.7.cgi

translating the text. Third, Turkish seems to be more expressive as suffixes add nuances to the words such as possession markers and intensifiers. Figure 2 shows AMR annotations for two sentences from the parallel corpus where ARG0 of *live-01* in English has been changed to a non-core role, *:poss*, which shows possession in Turkish. Although there was the possibility to ignore the possession marker and list the arguments of the predicate, *yaşamak* (to live), like its English counterpart, we chose to leave it as it is to highlight the differences between English and Turkish as an agglutinative language in AMR annotation.

Another important characteristic of Turkish is that unlike English, there are many light verbs and multiword expressions. In English AMR, we simply remove light verb constructions and use onto-notes predicate frames to deal with verb-particle combinations. However, due to the highly productive nature of Turkish and its idiosyncratic features, we need to be more cautious dealing with multiword expressions and light verb constructions. Figure 3 shows the inclination of Turkish toward productivity by duplicating the adjective, *uzun* (long), to be used as an adverb.

In our future study, we will also investigate how morphosemantic features like case markers might help specifying relations between concepts in Turkish and whether adding these properties to the AMR annotation structure may help achieving more accurate results.

```
(s / small
  :degree (v / very)
  :domain (e / everything)
  :location (l2 / live-01
          :ARG0 (i / i)))

(k / küçük
     :degree (x / çok)
     :domain (x2 / şey
          :mod (h / her))
     :location (y / yaşamak
          :poss (b / ben)))
```

Figure 2: Textual forms of AMR annotations for the sentence "*Where I live, everything is very small*." and its Turkish translation ("Benim/my yaşadığım/where live-I yerde/place-in her/every şey/thing çok/very küçük/small.")

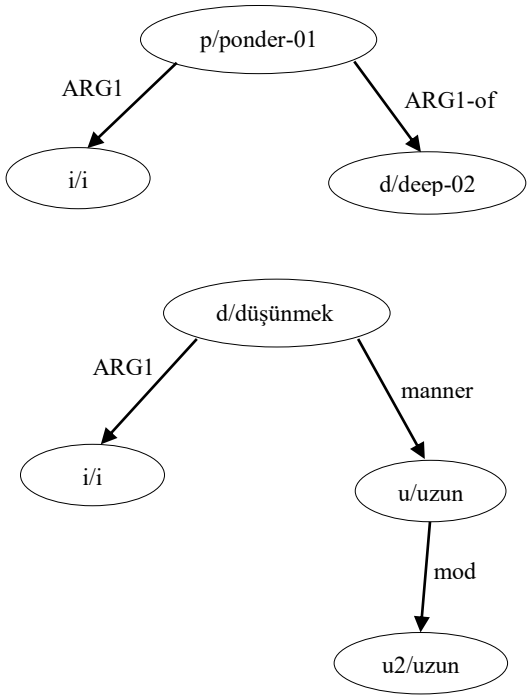


Figure 3: The AMR annotation graph for the sentence *"I pondered deeply"* which is translated as ("*uzun uzun/long long düşündüm/ thought-I*.")

6 Future Work

We have started the Turkish AMR project by annotating the first 100 hundred sentences of our parallel corpus, "*The Little Prince"*, and analyzing the divergences between our annotations and English AMR annotations. Currently, we are developing an AMR annotation guideline to construct the first Turkish Abstract Meaning Representation standard gold data. Finally, based on Turkish language peculiarities, we are going to create a transition-based parser to generate Turkish AMRs, which will be the first AMR parser for an agglutinative language.

Acknowledgments

We would like to thank Tuğba Pamay for her invaluable insights and discussions. We also want to thank Valerio Basile and anonymous reviewers for their comments and suggestions.

References

Atalay, N. B., Oflazer, K., & Say, B. 2003. The annotation process in the Turkish treebank. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 1003*. http://aclweb.org/anthology/W03-2405

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., ... & Schneider, N. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp.178-186). http://aclweb.org/anthology/W13-2322

Damonte, M., Cohen, S. B., & Satta, G. 2016. An incremental parser for abstract meaning representation. *arXiv preprint arXiv:1608.06111*. http://aclweb.org/anthology/E17-1051

Damonte, M., & Cohen, S. B. 2017. Cross-lingual abstract meaning representation parsing. *arXiv preprint arXiv:1704.04539*. https://doi.org/ 10.18653/v1/N18-1104

Flanigan, J., Thomson, S., Carbonell, J., Dyer, C., & Smith, N. A. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1426-1436). https://doi.org/ 10.3115/v1/P14-1134

Hajic, J., Bojar, O., & Uresova, Z. 2014. Comparing Czech and English AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing* (pp. 55-64). https://doi.org/ 10.3115/v1/W14-5808

Li, B., Wen, Y., Weiguang, Q. U., Bu, L., & Xue, N. 2016. Annotating the little prince with chinese amrs. In *Proceedings of the 10th linguistic annotation workshop held in conjunction with acl 2016 (law-x 2016)* (pp. 7-15). https://doi.org/ 10.18653/v1/W16-1702

Mathiessen, C. M., & Bateman, J. 1991. Text Generation and Systemic-Functional Linguistics. *London: Pinter*.

Şahin, G. G. 2016. Verb sense annotation for Turkish propbank via crowdsourcing. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 496-506). Springer, Cham.

Vanderwende, L., Menezes, A., & Quirk, C. 2015. An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Demonstrations* (pp. 26-30). https://doi.org/ 10.3115/v1/N15-3006

Wang, C., Xue, N., & Pradhan, S. 2015. *Boosting transition-based AMR parsing with refined actions and auxiliary analyzers.* In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Vol. 2, pp. 857-862). https://doi.org/ 10.3115/v1/P15-2141

Wang, C., Xue, N., & Pradhan, S. 2015. A transition-based algorithm for amr parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*(pp. 366-375). https://doi.org/ 10.3115/v1/N15-1040

Welch, C., Kummerfeld, J. K., Feng, S., & Mihalcea, R. 2018. World Knowledge for Abstract Meaning Representation Parsing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. http://aclweb.org/anthology/L18-1492

Xue, N., Bojar, O., Hajic, J., Palmer, M., Uresova, Z., & Zhang, X. 2014. *Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech*. In LREC (Vol. 14, pp. 1765-1772). http://www.lrecconf.org/proceedings/lrec2014/pdf/384_Paper.pdf