Automatic Propbank Generation for Turkish

Koray Ak
Department of Computer Engineering
Işık University
İstanbul, Turkey
koray.ak@isik.edu.tr

Olcay Taner Yıldız
Department of Computer Engineering
Işık University
İstanbul, Turkey
olcaytaner@isikun.edu.tr

Abstract

Semantic role labeling (SRL) is an important task for understanding natural languages, where the objective is to analyse propositions expressed by the verb and to identify each word that bears a semantic role. It provides an extensive dataset to enhance NLP applications such as information retrieval, machine translation, information extraction, and question answering. However, creating SRL models are difficult. Even in some languages, it is infeasible to create SRL models that have predicate-argument structure due to lack of linguistic resources. In this paper, we present our method to create an automatic Turkish PropBank by exploiting parallel data from the translated sentences of English PropBank. Experiments show that our method gives promising results.

1 Introduction

Semantic role labeling (SRL) is a well defined task that identifies semantic roles of the words in a sentence. Event characteristics and participants are simply identified by answering "Who did what to whom" questions. Having this semantic information facilitates NLP applications such as machine translation, information extraction, and question answering. After the development of statistical machine learning methods in the area of computational linguistics, learning complex linguistic knowledge has became feasible for NLP applications. Recent semantic resources specifically for SRL which provides input for developing statistical approaches are FrameNet (Fillmore et al., 2004), PropBank (Kingsbury and Palmer, 2002) (2003), (2005), (Bonial et al., 2014) and NomBank (2004). These resources enables us to understand language structure by providing a stable semantic representation.

Among these resources PropBank is a commonly used semantic resource which includes predicate - argument structure by stating the roles that each predicate can take along with the annotated corpora. It has been applied to more than 15 different languages. However, manually creating such semantic resource is labor-intensive, time-consuming and most importantly requires a professional linguistic perspective. Also limited linguistic data further blocks generating PropBank-like resources.

Various studies such as Zhuang and Zong (2010), Van der Plas et al. (2011) (2014), Kozhevnikov and Titov (2013), Akbik et al. (2015), which transfer semantic information using parallel corpus, are presented to cope with these problems. In this way, semantic information projected from a resource-rich language (English) to a language with inadequate resources and PropBank of the target language is automatically generated. Here the assumption is translated parallel sentences generally share same semantic information. Word and constituent based alignment techniques are widely used to construct mapping between source and target languages for annotation projection. Previous studies report translation divergences and language specific differences affect the quality of the projection. Filtering projections using learning methods is suggested to increase precision. In this paper, we present our study to create automatic Turkish PropBank using parallel sentences from English PropBank.

This paper is organized as follows: we first give brief information about English and Turkish PropBanks in Section 2. In Section 3, Studies for the automatic proposition bank generation are discussed. In the next section proposed methods are presented. First, we explain the annotation projection using parallel sentence trees. Then, we propose methods for aligning parallel sentence phrases not aligned with tree structure. Finally, in

Proceedings of Recent Advances in Natural Language Processing, pages 33–41,
Varna, Bulgaria, Sep 2–4 2019.

Section 5, we conclude with the results.

2 PropBank

2.1 English PropBank

PropBank is the bank of propositions where predicate-argument information of the corpora is annotated and semantic roles or arguments that each verb can take are posited. It is constituted on the Penn Treebank (Marcus et al., 1993) Wall Street Journal [WSJ]. The primary goal is to label syntactic elements in a sentence with specific argument roles to standardize labels for the similar arguments such as *the window* in John broke *the window* and *the window* broke. PropBank uses conceptual labels for arguments from Arg0 to Arg5. Only Arg0 and Arg1 indicate the same roles across different verbs where Arg0 means agent or causer and Arg1 is the patient or theme. The rest of the argument roles can vary across different verbs. They can be instrument, start point, end point, beneficiary, or attribute.

Moreover, PropBank uses ArgM's as modifier labels where the role is not specific to the verb group and generalizes over the corpora such as location, temporal, purpose, or cause etc. arguments. The first version of English PropBank, named as The Original PropBank, is constructed for only verbal predicates whereas the latest version includes all syntactic realizations of event and state semantics by focusing different expressions in form of nouns, adjectives and multi-word expressions to represent complete event relations within and across sentences.

2.2 PropBank Studies for Turkish

There have been different attempts to construct Turkish PropBank in the literature. Şahin (2016a; 2016b), Şahin and Adalı (2017) report semantic role annotation of arguments in the Turkish dependency treebank. They construct PropBank by using ITU-METU-Sabancı Treebank (IMST). In these studies, frame files of Turkish PropBank are constructed and extended by utilizing crowdsourcing. 20,060 semantic roles are annotated in 5,635 sentences. The size of the resource is stated as a drawback in the study. Recently, Ak et al. (2018) construct another Turkish Proposition Bank using translated sentences of English PropBank. So far, 9,560 of 17,000 translated sentences are annotated with semantic roles. Also, framesets are created for 1,330 verbs and 1,914 verb senses. These studies constitute a base for Turkish proposition bank, but their size is limited and construction of these proposition banks consumed a lot of time.

3 Automatic PropBank Generation Studies

PropBanks are also generated automatically for resource-scarce languages by using parallel corpus. In this section, proposition bank studies for automatic generation are presented. Zhuang and Zong (2010) proposed performing SRL on parallel corpus of different languages and merging the result via a joint inference model can improve SRL results for both input languages. In the study an English and Chinese parallel corpus is used. First each predicate is processed by monolingual SRL systems separately for producing argument candidates. After the candidates formed, a Joint Inference model selects the candidate that is reasonable to the both languages. Also, a log-linear model is formulated to evaluate the consistency. This approach increased F1 scores 1.52 and 1.74 respectively for Chinese and English.

Van der Plas et al. (2011) presents cross-lingual semantic transfer from English to French. English syntactic-semantic annotations were transferred using word alignments to French language. French semantic annotations gathered from the first step were then trained with a French joint syntactic-semantic parser along with the French syntactic annotations trained separately. Joint syntactic-semantic parser is used for learning the relation between semantic and syntactic structure of the target language and reduces the errors arising from the first step. This approach reaches 4% lower than the upper bound for predicates and 9% for arguments.

Kozhevnikov (2013) shows SRL model transfer from one language to another can be achieved by using shared feature representation. Shared feature representation for language pairs is constructed based on syntactic and lexical information. Afterwards, a semantic role labeling model is trained for source language and then used for the target language. As a result SRL model of the target language is generated. Process only requires a source language model and parallel data to construct target SRL model. Approach is applied for English, French, Czech and Chinese languages.

In the next study, Van der Plas (2014) improves the labeling results with respect to the previous work

(Van der Plas et al., 2011) by building separate models for arguments and predicates. Also, problems of transferring semantic annotations using parallel corpus is examined in the paper. Token-to-token basis annotation transfer, translation shifts, and alignment errors in the previous work is replaced with a global approach that aggregates information at corpus level. Instead of using English semantic annotations of roles and predicate together with French PoS tags to generate French semantic annotations, English annotations of predicates and roles used separately to generate one predicate and one role semantic annotations separately.

Akbik *et al.* propose a two stage approach (Akbik et al., 2015). In the first stage only filtered semantic annotation is projected. Since high confidence semantic labels projected, resulting target semantic labels will be high in precision and low in recall. In the next stage, completed target language sentences sampled and a classifier is trained to add new labels to boost recall and preserve precision. Proposed system is applied on 7 different languages from 3 different language family. These languages are Chinese, Arabic, French, German, Hindi, Russian, and Spanish.

4 Methods

Among the studies for Turkish proposition bank, Ak et al. (2018) is constructed on parallel English - Turkish sentences from the Original English PropBank. We have used the corpus provided in this study to automatically generate proposition bank.

4.1 Automatic Turkish PropBank Using Parallel Sentence Trees

Penn Treebank structure offers advantages for building fully tagged data set in accordance with syntactic labels, morphological labels and parallel sentences. We used this structure to add English PropBank labels for each word in the corpus. In this manner, we exploited this parallel dataset to transfer English PropBank annotations to an automatic Turkish PropBank.

4.1.1 English PropBank Labels

Original English PropBank corpus (Palmer et al., 2004) is accessible through Linguistic Data Consortium (LDC). This resource is the initial version of the English PropBank and it only includes the relations with verbal predicates. In the newer versions adjective and noun relations are also annotated. Since we compare projection results with manually annotated corpus (Ak et al., 2018) which only contains verbal relations, we use the initial version of the English PropBank. We downloaded this dataset and imported annotations for the selected sentences. After this step 6,060 sentences among 9,558 were enhanced with the English annotations. Below in Figure 1, a sample sentence is presented. English annotations are inserted inside "englishPropbank" tags right after Turkish annotations which reside in "propbank" tags. Some of the words have only English annotation, because there is no word translated in the Turkish sentence for this node. As an example, "their" in Figure 1 has annotations in the englishPropbank tag but there is no equivalent translation in Turkish, presented as "*NONE*", so propbank tag does not exist. English tags have predicate information that annotation belongs to. "Müşterilerinin" (customers) in the same example has "ARG0$like_01#ARG1$think_01" in the englishPropbank tag which means there exists at least two words whose root is in verb form. Here the word is annotated with respect to "like" and "think" separately. We have separated multiple annotations with "#" sign and in each annotation predicate label and role is distinguished by "$" sign. In the Turkish annotation, WordNet id of the predicate was used instead of predicate label.

4.1.2 Transfering Annotations to Automatic Turkish PropBank Using Parallel Sentences

After importing English annotations, it is necessary to determine predicate(s) of the Turkish sentences. Morphological structures of the words are examined to detect predicate candidates. Words were morphologically and semantically analyzed in translated Penn TreeBank. We have used "morphologicalAnalysis" tag to check the morphological structure of the words. In Figure 1, sample morphological structure is displayed.

The word which has a verb root and verb according to last inflectional group is treated as the predicate of the sentence. Once we found a word suitable for these conditions, we gathered English PropBank annotation. If it is also labeled as predicate in English proposition bank, we got the predicate label, e.g. like_01, to find annotations with respect to this predicate. We searched for the found

Figure 1: Part of a sentence tree : English PropBank annotations reside in "englishPropBank" tags.

predicate label in the annotations and transfered annotations matching with the predicate label. If we could not find a predicate in Turkish sentence or the corresponding English label did not contain Predicate role annotation, we skipped to the next predicate candidate.

During the transfer, a mapping was needed due to the difference between English and Turkish (Ak et al., 2018) argument labeling. English PropBank corpus has "-" sign in ArgM's like ARGM-TMP and also some of the arguments from Arg1 to Arg5 are labeled with the prepositions such as ARG1-AT, ARG2-BY etc. We processed these differences and then transferred labels into the "propbank" tags. After analyzing Turkish sentences we found out some sentences have more than one predicate, so we continued to search for another predicate in the sentence and ran the same procedure for each predicate candidate.

4.1.3 Experiments

Annotations gathered from the English sentence were compared with the Turkish hand-annotated proposition bank (Ak et al., 2018). Comparisons were done at the word level by checking the annotations for each corpus. Among the 6,060 sentences enhanced with English PropBank roles, 848 sentences did not have a predicate in Turkish proposition bank. Therefore, in 5,212 sentences, 44,779 word annotations were compared. 31,813 annotations were transferred from English to Turkish. Results of the comparison are presented in Table 1. 19,373 words annotated with PropBank roles correctly . 6,441 annotations are incorrect, PropBank tags are different in both corpus. 5,999 annotations are undetermined, valid PropBank labels transferred from English annotations but no annotation exists in hand annotated proposition bank. Annotations to be compared is not valid so we did not include this set in the evaluation.

Transferred		**Untransferred**	
Correct	19,373	Not H.A.	8,837
Incorrect	6,441		
Undetermined	5,999	H.A.	4,129
Total	31,813	**Total**	12,966

Table 1: Results of the comparison between automatic proposition bank and hand annotated (H.A.) proposition bank.

When we remove undetermined 5,999 words in the comparison; 19,373 annotations from 25,814 annotations are correct, which gives us ~75% accuracy for transferred and comparable set. These 5,999 annotations may be hand-annotated and recompared for validity of the transferred annotations.

In Table 2, we present occurrences of erroneous annotation transfers. Only top ten occurrences are presented. Arg0-Arg1 transfers are the most occurred incorrect transfers 1,843 among 6,441 incorrect annotations. Second most occurred error is in Arg1-Arg2 labels. Errors in Arg0-Arg1 and Arg1-Arg2 labels forms ~44% of the transfer errors.

On the other hand, when we look at the all word results, 12,966 roles were not transferred. If we take these untransferred instances as incorrect; 19,373 annotations out of 38,780 annotation are true and the accuracy drops to ~50%. However, 8,837 of untransferred annotation are not an-

Different Arguments	# of Occurrence
ARG0-ARG1	1,843
ARG1-ARG2	961
ARG2-ARGMEXT	462
ARG1-PREDICATE	255
ARG0-ARG2	229
ARG4-ARGMEXT	226
ARG1-ARGMPNC	220
ARG1-ARGMMNR	186
ARG1-ARGMTMP	160
ARG1-ARGMLOC	148

Table 2: Counts of different argument annotations between transferred annotations and hand annotations.

notated in the hand-annotated corpus. Only 4,129 are valid PropBank arguments. In this respect, if we count only valid arguments for untransferred annotations, accuracy is ~65%.

4.2 Automatic Turkish PropBank Using Parallel Sentence Phrases

In the previous method, annotation projection using parallel sentence trees is discussed. However, finding such a resource in a special format is difficult especially if you are working with a resource-scarce language. Most of the time creating a formatted parallel resource like tree structured sentences complicates translation procedure. In this section, automatic generation with translated sentences without tree structure will be examined.

4.2.1 Phrase Sentence Structure

For the phrase sentences, English sentences retranslated without tree structure. Prior the annotation projection, linguists in the team annotated phrase sentences and populated "propbank" and "shallowParse" tags so that we check the correctness after the annotation transfer. 6,511 sentences among 9,557 phrase sentences have predicate according to hand annotations for newly translated sentences. However, only 5,259 sentences have English PropBank annotation, so we take this set to transfer annotations. As you remember, the same number in the previous section was 5,212. Here translation and annotation differences change the processed sentence count.

Tag structure of Penn Treebank is preserved to simplify morphologic and semantic analysis requirements during the annotation transfer. In Figure 2, sample phrase sentence can be seen. Unlikely Figure 1, syntactic tags which indicate tree structure are not included. We used original tree formatted English sentence to extract English propbank annotations. However, since the target sentence do not have tree structure definition we used other word alignment methods to determine annotation projection.

4.2.2 Semantic Alignment Using WordNet

In order to transfer annotations, first we tried to match predicates of English sentence and Turkish translation. Again we utilize "morphologicalAnalysis" tags to determine predicate candidates in the phrase sentence. Words which have a verb root and verb according to last inflectional group is treated as the predicate candidates of the sentence. Once we found all the words ensuring these conditions, we gathered all English PropBank annotation labels which are tagged as "Predicate" in 'englishPropbank" tag. To align predicates in different languages, we tried to exploit WordNet's (Ehsani et al., 2018) interlingual mapping capabilities. For each predicate in English sentence we find Turkish translation by searching English synset id in the WordNet. English synset id is located in englishSemantics tags as in the sample in Figure 1. If there exists any translation in the WordNet, we take Turkish synset id and search it in the predicate candidates found for phrase sentence. Whenever translation found, we align predicates and try to transfer annotation with respect to aligned English label. For annotation transfer of other arguments we again align words using WordNet's interlingual mapping. An example WordNet record is presented in Figure 3.

First results gathered with only WordNet mapping were very low. True annotation count is 2,195 among 29,168 annotations tagged manually which yields 7.53%. However, transferred false annotation count is only 342. System heavily relies on semantic annotations for both English and Turkish words where some of the words failed to have semantic annotation. We look deeper into dataset provided by Ak et al. (2018), 11,006 English words do not have semantic annotation so we failed to match these words with Turkish counterparts.

Some words are not annotated semantically such as, proper nouns, time, date, numbers, ordinal numbers, percentiles, fractional numbers, number intervals, and reel numbers. Most of these words are same in Turkish translation so we matched English and Turkish words by string match. For example if a sentence contains proper

Figure 2: Part of a phrase sentence : Translated words in Turkish tags. Helper tags gives additional information for each word.

```
<SYNSET>
    <ID>TUR10-0682580</ID>
    <SYNONYM>
        <LITERAL>sevmek<SENSE>5</SENSE></LITERAL>
    </SYNONYM>
    <POS>v</POS>
    <ILR>ENG31-01779085-v<TYPE>SYNONYM</TYPE></ILR>
    <ILR>ENG31-01779456-v<TYPE>SYNONYM</TYPE></ILR>
    <ILR>ENG31-01780873-v<TYPE>SYNONYM</TYPE></ILR>
    <ILR>ENG31-01781131-v<TYPE>SYNONYM</TYPE></ILR>
    <DEF>Yerini, şartlarını uygun bulmak</DEF>
    <EXAMPLE>Bu ağaç nemli ortamı sever.</EXAMPLE>
</SYNSET>
```

Figure 3: Sample WordNet record found by searching "ENG31-01781131-v", English synset id, from the sentence in Figure 1.

noun "Dow Jones", the same string also exists in the Turkish translation too. However, it may take additional suffixes, so we only check whether English words starts with Turkish root word. Also, translational differences are encountered like decimal separator in English is "." where some Turkish translations "," is used. We replace this differences by looking whether the first morphological tag is "NUM". After these tunings, we rerun the procedure and get 2,680 true and 531 false annotations which increases true annotations to 9.19%. Another problem is erroneous semantic annotations. If English and Turkish semantic annotation is not right, alignment is not possible. Even in the best scenario where both word is annotated, if WordNet mapping is incomplete, an alignment can not be established.

As an alternative we decided to reinforce annotation transfer by using constituent boundaries identified with shallowParse tags by our linguist team mates. Example of shallowParse tags can be seen in Figure 2. Prior to the annotation transfer, phrase sentences are annotated for constituent boundaries which can be used to group argument roles in the sentence. After transferring annotations with respect to semantic annotations, we run another method over phrase sentences which calculates maximum argument types for each constituent and tags any untagged word with the calculated max argument role within the constituent boundary. This procedure further enhance true annotations to 4,255 but also increase false annotations to 1,202. After constituent boundary calculation, correct annotation transfer percent is increased to ~14.59%. In Figure 4 annotation of the sentence 7076.train is presented. Untagged words in "Özne" and "Zarf Tümleci" constituent boundaries are tagged with the found argument role within the boundary. Note that, we did not use the constituent types but we use boundaries of the constituents.

4.2.3 Word Alignment Using IBM Alignment Models

Word alignment through semantic relation requires fair semantic annotation for both languages and also sufficient semantic mapping between languages. We search different word alignment methods between English and Turkish sentences. IBM

(1) [The less-rigorous Senate version]$_{\text{ARG0}}$ [would]$_{\text{ARGM-MOD}}$ [defer]$_{\text{Predicate}}$ [the deductibility]$_{\text{ARG1}}$ for [roughly five years.]$_{\text{ARG2-for}}$

(2) [Daha az sıkı türden bir Senato versiyonu]$_{\text{Özne - Subject}}$ [aşağı yukarı beş yıl için]$_{\text{Zarf Tümleci - Adverbial Clause}}$ [düşülebilirliği]$_{\text{Nesne - Object}}$ [ertelerdi.]$_{\text{Yüklem - Predicate}}$

(3) [Daha az sıkı türden bir]$_{\text{NONE}}$ [Senato]$_{\text{ARG0}}$ [versiyonu]$_{\text{NONE}}$ [aşağı yukarı]$_{\text{ARG2}}$ [beş]$_{\text{NONE}}$ [yıl]$_{\text{ARG2}}$ [için düşülebilirliği]$_{\text{NONE}}$ [ertelerdi.]$_{\text{PREDICATE}}$

(4) [Daha az sıkı türden bir Senato versiyonu]$_{\text{ARG0}}$ [aşağı yukarı beş yıl için]$_{\text{ARG2}}$ [düşülebilirliği]$_{\text{NONE}}$ [ertelerdi.]$_{\text{PREDICATE}}$

Figure 4: Annotation reinforced with respect to constituent boundaries: (1) English sentence (2) constituent boundaries identified with shallow-Parse tags for sentence in 7076.train, (3) Argument roles for the same sentence after annotation transfer, (4) Argument roles for the same sentence after reinforce method.

alignment models offer solution to our word alignment problem. IBM Models are mainly used for statistical machine translation to train a translation model and an alignment model. IBM Model 1 (Brown et al., 1993) is the primary word alignment model offered by IBM. It is widely used for solving word alignments while working with parallel corpora. It is a generative probabilistic model that calculates probabilities for each word alignment from source sentence to target sentence. It takes a corpus of paired sentences from two languages as training data. These paired sentences are possible translation of the sentences from source language to target. With this training corpus, parameters of the model estimated using EM (expectation maximization). IBM Model 2 has an additional model for alignment and introduce alignment distortion parameters. We decided to use IBM model 1 & 2 to establish word alignments instead of WordNet's interligual mapping. We input sentence pairs and gather alignment probabilities for each English word to Turkish equivalent. 244,024 word pairs are taken as output where for each English word, 10 most probable Turkish words are listed. Alignment probabilities for word "Reserve" is presented in Table 3 and 4 for IBM Model 1 and 2 respectively.

After gathering alignment data, we transfer annotations to phrase sentences from English PropBank labels in the tree structured sentences. All

English Word	Turkish Word	Probability
Reserve	Reserve	0.72270917
Reserve	Rezerv	0.15328414
Reserve	mevduat	0.03056293
Reserve	Bankası'nın	0.02731664
Reserve	kuruluşlarındaki	0.01375332
Reserve	komisyonları	0.01375332
Reserve	Bankasının	0.00611259
Reserve	kuruluşlarında	0.00458444
Reserve	komisyon	0.00458444
Reserve	Federe	0.00458444

Table 3: Word alignment probabilities for English word "Reserve" calculated by IBM Model 1.

English Word	Turkish Word	Probability
Reserve	Reserve	0.67700755
Reserve	Rezerv	0.14360766
Reserve	Federe	0.06154614
Reserve	Bankası	0.05265972
Reserve	tasarruf	0.03072182
Reserve	kuruluşlarına	0.02117394
Reserve	üzerindeki	0.01111856
Reserve	bu	0.00212005
Reserve	kurumlarına	0.00004452
Reserve	Merkez	0.00000002

Table 4: Word alignment probabilities for English word "Reserve" calculated by IBM Model 2.

words tagged with "PREDICATE" tag in English sentence are stored into a map which includes predicate label from the "englishPropbank" tag *e.g.* "like_01" and English word from the "english" tag *e.g.* "like". Then we search alignments for each found English predicate. Here we observed that aligned Turkish words may not occur in the phrase sentence as they found in the alignment table. Words may include additional suffixes, so we use Finite State Machine(FSM) morphological analyzer available in our NLP Toolkit of Ak et al. (2018) to extract roots of the aligned Turkish words. Since we have several possible morphological parse for each aligned word, we created an array for possible roots. In parallel, we found predicate candidates from the phrase sentence as we stated in the previous methods. Then we tried to match aligned words and possible roots with the found predicate candidates. If there exists a predicate candidate that matches with the aligned word or one of its roots in the array, we tagged the candidate as "PREDICATE" and update map as predicate label and synset id of Turkish predicate.

After finishing predicate discovery, we transfer annotations for found predicates. To do that we look for the annotations with respect to the predicate labels in the map. For each record in map we

took the predicate label and corresponding Turkish synset id. When we found an annotation with this predicate label, first we extract the argument and try to find aligned word for the processed English word. For the alignment again we find the most probable word from the table and use FSM morphological analyzer to extract possible roots. Then for each word we search Turkish sentence to match words with aligned word or possible roots extracted. If matched Turkish words do not have argument annotation, we transfer argument with the synset id found in the map record.

As we discuss in the previous annotation transfer procedure 4.2.2, some of the English words such as proper nouns, time, date, numbers, ordinal numbers, percentiles, fractional numbers, number intervals, and reel numbers stay same or take additional suffixes in Turkish translation. So we include the same method used for matching these words. In a case words are not aligned with the information from alignment table, and a valid annotation present in English word, we search exact string match or any word starts with the root of English word in the Turkish sentence.

We run our procedure with IBM Model 1 & 2 separately. We add reinforce step previously used in Section 4.2.2. Unlikely previous attempts, after examining language structure we decided to add rules to tag any untagged words after annotation transfer. We observed argument types affect noun inflections, for some argument types the last word in constituent boundary is taking certain suffixes. So first we find untagged word and select the last word in its constituent boundary. Since we run reinforce step beforehand, only untagged constituents exists in the sentence. In this respect, we set the following rules to determine argument annotation for untransfered words;

- For nouns and proper nouns:
  - Have no suffix then ARG0
  - Last morpheme tag is "ACCUSATIVE" (-(y)H, -nH) or "DATIVE" (-(y)A, -nA) then ARG1
  - Last morpheme tag is "LOCATIVE" (-DA, -nDA) or "ABLATIVE" (-DAn, -nDAn) then ARGMLOC
  - Last morpheme tag is "INSTRUMENTAL" (-(y)lA) then ARG2
- For all word types
  - Morphological parse contains date, time then ARGMTMP
  - Morphological parse contains cardinal number, fraction, percent, range, real number, ordinal number then ARGMEXT

We use these rules to tag any untagged word. After applying these rules annotation transfer result is as shown in Table 5 and 6. Results show that rules applied slightly change the correct annotations. For model 1 rules output much more correct annotation than the incorrect ones whereas in model 2 the number of correct and incorrect annotations gathered are nearly same. However, precision for model 1 is improved to 59.44% and for model 2 precision become 59.86%.

IBM Model 1 + Reinforce + Rules			
Transferred		**Untransferred**	
Correct	17,340	Not H.A.	1,151
Incorrect	9,664	H.A.	2,170
Undetermined	14,384		
Total	41,388	**Total**	3,321

Table 5: Results for IBM Model 1 alignment.

IBM Model 2 + Reinforce + Rules			
Transfered		**Untransfered**	
Correct	17,464	Not H.A.	1,078
Incorrect	9,635	H.A.	2,075
Undetermined	14,457		
Total	41,556	**Total**	3,153

Table 6: Results for IBM Model 2 alignment.

5 Conclusion

We proposed methods to generate automatic Turkish proposition bank by transferring cross-language semantic information. Using the parallelism with English proposition bank gives us an opportunity to create a proposition bank in a short time with less effort. We currently have 64% accuracy with the hand-annotated proposition bank (Ak et al., 2018) for parallel sentence trees. When we consider only transferred annotations, accuracy is rising to ~75%. We also present annotation projection to phrase sentences using WordNet and IBM alignment models. WordNet alignment heavily relies on semantic annotations, correct annotations transferred after this method is ~14.59%. However, 4,255 correct argument roles are transferred among 5,457 arguments which means 79% of the transferred roles are correct. To increase annotation transfer for phrase sentences, we have also proposed alignment with IBM Model 1 and 2. Both models yields ~60% correct annotations. Annotations transferred with these methods can provide a basis for proposition bank creation in resource-scarce languages. Annotations may then

be checked quickly by the annotators and proposition bank reach the final state.

References

Meyers A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*. Association for Computational Linguistics, Boston, Massachusetts, USA, pages 24–31.

K. Ak, O. T. Yıldız, V. Esgel, and C. Toprak. 2018. Construction of a Turkish proposition bank. *Turkish Journal of Electrical Engineering and Computer Science* 26:570 – 581.

A. Akbik, L. Chiticariu, M. Danilevsky, Y. Li, S. Vaithyanathan, and H. Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. In *ACL (1)*. The Association for Computer Linguistics, pages 397–407.

C. Bonial, J. Bonn, K. Conger, J. D. Hwang, and M. Palmer. 2014. Propbank: Semantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.* 19(2):263–311.

G. G. Şahin. 2016a. Framing of verbs for turkish propbank. In *TurCLing 2016 in conj. with 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016)*.

G. G. Şahin. 2016b. Verb sense annotation for turkish propbank via crowdsourcing. In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016)*.

G. G. Şahin and E. Adalı. 2017. Annotation of semantic roles for the Turkish proposition bank. *Language Resources and Evaluation* .

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

C. J. Fillmore, J. Ruppenhofer, and Collin F. Baker. 2004. *FrameNet and Representing the Link between Semantic and Syntactic Relations*, Institute of Linguistics, Academia Sinica, Taipei, pages 19–62. Language and Linguistics Monographs Series B.

P. Kingsbury and M. Palmer. 2002. From treebank to propbank. In *LREC*. European Language Resources Association.

P. Kingsbury and M. Palmer. 2003. Propbank: The next level of treebank. In *Proceedings of Treebanks and Lexical Theories*. Växjö, Sweden.

M. Kozhevnikov and I. Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *ACL (1)*. The Association for Computer Linguistics, pages 1190–1200.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics* 19(2):313–330.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.* 31(1):71–106.

M. Palmer, P. Kingsbury, O. Babko-Malaya, S. Cotton, and B. Snyder. 2004. Proposition bank i. Philadelphia: Linguistic Data Consortium. LDC2004T14.

L. Van der Plas, M. Apidianaki, and C. Chen. 2014. Global methods for cross-lingual semantic role and predicate labelling. In *COLING*. ACL, pages 1279–1290.

L. Van der Plas, P. Merlo, and J. Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *ACL (Short Papers)*. The Association for Computer Linguistics, pages 299–304.

T. Zhuang and C. Zong. 2010. Joint inference for bilingual semantic role labeling. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '10, pages 304–314.