

Using Morphosemantic Information in Construction of a Pilot Lexical Semantic Resource for Turkish

Gözde Gül Şahin
Department of Computer Engineering
Istanbul Technical University
Istanbul, 34469, Turkey
isguderg@itu.edu.tr;

Eşref Adalı
Department of Computer Engineering
Istanbul Technical University
Istanbul, 34469, Turkey
adali@itu.edu.tr;

ABSTRACT

Morphological units carry vast amount of semantic information for languages with rich inflectional and derivational morphology. In this paper we show how morphosemantic information available for morphologically rich languages can be used to reduce manual effort in creating semantic resources like PropBank and VerbNet; to increase performance of word sense disambiguation, semantic role labeling and related tasks. We test the consistency of these features in a pilot study for Turkish and show that; 1) Case markers are related with semantic roles and 2) Morphemes that change the valency of the verb follow a predictable pattern.

1 Introduction

In recent years considerable amount of research has been performed on extracting semantic information from sentences. Revealing such information is usually achieved by identifying the complements (arguments) of a predicate and assigning meaningful labels to them. Each label represents the argument's relation to its predicate and is referred to as a semantic role and this task is named as semantic role labeling (SRL). There exists some comprehensive semantically interpreted corpora such as FrameNet and PropBank. These corpora, annotated with semantic roles, help researchers to specify SRL as a task,

furthermore are used as training and test data for supervised machine learning methods [1]. These resources differ in type of semantic roles they use and type of additional information they provide.

FrameNet (FN) is a semantic network, built around the theory of semantic frames. This theory describes a type of event, relation, or entity with their participants which are called frame elements (FEs). All predicates in same semantic frame share one set of FEs. A sample sentence annotated with FrameNet, VerbNet and PropBank conventions respectively, is given in Ex.1. The predicate "buy" belongs to "Commerce buy", more generally "Commercial transaction" frame of FrameNet which contains "Buyer", "Goods" as core frame elements and "Seller" as a non-core frame element as in Ex. 1. FN also provides connections between semantic frames like inheritance, hierarchy and causativity. For example the frame "Commerce buy" is connected to "Importing" and "Shopping" frames with "used by" relation. Contrary to FN, VerbNet (VN) is a hierarchical verb lexicon, that contains categories of verbs based on Levin Verb classification.[2].

The predicate "buy" is contained in "get-13.5.1" class of VN, among with the verbs "pick", "reserve" and "book". Members of same verb class share same set of semantic

roles, referred to as thematic roles. In addition to thematic roles, verb classes are defined with different possible syntaxes for each class. One possible syntax for the class "get-13.5.1" is given in the second line of Ex. 1. Unlike FrameNet and VerbNet, PropBank (PB) [3] does not make use of a reference ontology like semantic frames or verb classes. Instead semantic roles are numbered from Arg0 to Arg5 for the core arguments.

[Jess]_{Buyer-Agent-Arg0} bought [a coat]_{Goods-Theme-Arg1} from [Abby]_{Seller-Source-Arg2}
Syntax: Agent V Theme {From} Source

Ex. 1

There doesn't exist a VerbNet, PropBank or a similar semantically interpretable resource for Turkish (except for WordNet [4]). Also, the only available morphologically and syntactically annotated treebank corpus: METU-Sabancı Dependency Treebank [5,6,7] has only about 5600 sentences, which has presumably a low coverage of Turkish verbs. VerbNet defines possible syntaxes for each class of verbs. However, due to free word order and excessive case marking system, syntactic information is already encoded with case markers in Turkish. Thus the structure of VerbNet does not fit well to the Turkish language. PropBank simplifies semantic roles, but defines neither relations between verbs nor all possible syntaxes for each verb. Moreover only Arg0 and Arg1 are associated with a specific semantic content, which reduces the consistency among labeled arguments. Due to lack of a large-scale treebank corpus, building a high coverage PropBank is currently not possible for Turkish. FrameNet defines richer relations between verbs, but the frame elements are extremely fine-grained and building such a comprehensive resource requires a great amount of manual work for which human resources are not currently available for Turkish.

In this paper, we discuss how the semantic information supplied by morphemes, named as morphosemantics, can be included in the construction of semantic resources for languages with less resources and rich morphologies, like Turkish. We try to show that we can decrease manual effort for building such banks and increase consistency and connectivity of the resource by exploiting derivational morphology of verbs; eliminate mapping costs by associating syntactic information with semantic roles and increase the performance of SRL and word sense disambiguation by directly using morphosemantic information supplied with inflectional morphology. Then, we perform a pilot study to build a lexical semantic resource that contains syntactic information as well as semantic information that is defined by semantic roles both in VerbNet and PropBank fashion, by exploiting morphological properties of Turkish language.

2 Morphosemantic Features

In morphologically rich languages, the meaning of a word is strongly determined by the morphemes that are attached to it. Some of these morphemes always add a predefined meaning while some differ, depending on the language. However, only regular features can be used for NLP tasks that require automatic semantic interpretation. Here, we determine two multilingual morphosemantic features: case markers and verb valency changing morphemes and analyze the regularity and usability of these features for Turkish.

2.1 Declension and Case Marking

Declension is a term used to express the inflection of nouns, pronouns, adjectives and articles for gender, number and case. It occurs in many languages such as Arabic, Basque,

Sanskrit, Finnish, Hungarian, Latin, Russian and Turkish. Even though the languages differ, the same case markers are used to express similar meanings with some variation. Relation between semantic roles and case markers can assist researchers in solving some of the challenging problems in natural language processing. In languages where case markers exist, these

- can be used as features for Semantic Role Labeling,
- can supply priori information for disambiguating word senses,
- can be used in language generation as such: Once the predicate and the sense is determined, the
- arguments can directly be inflected with the case markers associated with their roles.

2.2 Valency Changing Morphemes

The valency of a verb can be defined as the verb's ability to govern a particular number of arguments of a particular type. "In Turkish, verb stems govern relatively stable valency patterns or prototypical argument frames" as stated by [8]. Consider the root verb *giy* (to wear). One can derive new verbs from the root *giy* (to wear) such as *giy-in* (to get dressed), *giy-dir* (to dress someone) and *giy-il* (to be worn). These verbs are referred to as verb stems and these special suffixes are referred to as valency changing morphemes. By modeling the semantic role transformation from verb root to verb stem, we can automatically identify argument configuration of a new verb stem given the correct morphological analysis. By doing so, framing only the verb roots can guarantee to have frames of all verb stems derived from that root. This quickens the process of building a semantic resource, as well as automatizing and reducing the human

error. In this section we present a pilot study for some available valencies in Turkish language. For the sake of simplicity, instead of thematic roles, argument labeling in the PropBank fashion is used.

Reflexive

The reflexive suffix triggers the suppression of one of the arguments. In Fig. 1, observed argument shift is given.

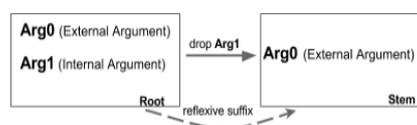


Figure 1: Argument transformation caused by reflexive suffix.

Reciprocal

Reciprocal verbs express actions done by more than one subject. The action may be done together or against each other. Reciprocal verbs may have a plural agent or two or more singular co-agents conjoined where one of them marked with COM case as shown in Fig 2. In both cases, the suppression of one of the arguments of the root verb is triggered. We have observed that the suppressed argument may be in different roles (patient, theme, stimulus, experiencer, co-patient), but usually appears as Arg1 and rarely as Arg2.

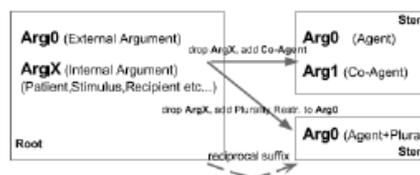


Figure 2: Argument transformation caused by reciprocal suffix.

Causative

Causative category is the most common valence-changing category among Bybee's [9]

world-wide sample of 50 languages. Contrary to other morphemes, causative morpheme introduces of a new argument called causer to the valence pattern. In most of the languages, only intransitive verbs are causativized. In this case, as shown in Fig. 3 the causee becomes the patient of the causation event. In other words, the central argument of the root verb, (Arg0 if exists, otherwise Arg1), is marked with ACC case and becomes an internal argument (usually Arg1) of the new causative verb. Some languages can have causatives from transitive verbs too, however the role and the mark of the causee may differ across languages. For the languages where the causee becomes an indirect object, like Turkish and Georgian, the central argument, Arg0 of the root verb, when transformed into a verb stem, receives the DAT case marker and serves as an indirect object (usually as Arg2), while Arg1 serves again as Arg1. This pattern for transitive verbs is given in Fig. 3.

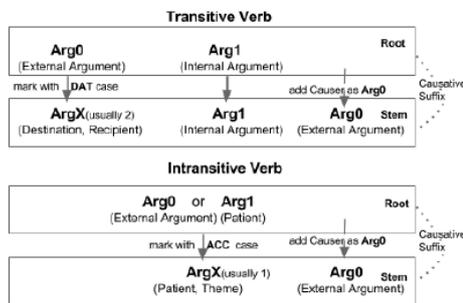


Figure 3: Argument transformation caused by causative suffix.

3 Methodology

We have performed a feasibility study for using morphosemantic features in building a lexical semantic resource for Turkish. As discussed in Section 3.2, we assume we can automatically frame a verb (e.g. saklan(reflexive)) that is derived with a regular valency changing morpheme (e.g. n), if the

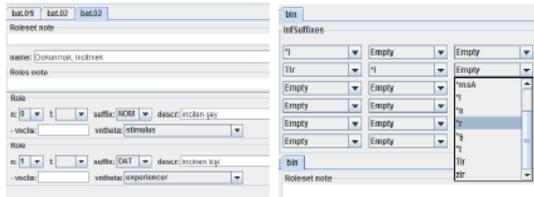
argument configuration of the root verb (e.g. sakla) is known. Hence, we have only framed root verbs. We have framed 233 root verbs and 452 verb senses. We have calculated the total number of valence changing morphemes as 425. This means 425 verbs can be automatically framed by applying the valency patterns to 233 root verbs. In this analysis we have only considered one sense of the verb since there may be cases where valency changing morpheme can not be applied to another sense of the verb. This can not be automatically determined. Moreover, a verb stem may have multiple senses. In that case automatically extracted argument transformation may be wrong, because the verb stem may have a completely different meaning.

Turkish is not among rich languages by means of computational resources as discussed before. Turkish Language Association (TDK) is a trustworthy source for lexical datasets and dictionaries. To run this pilot study, we have used the list of Turkish root verbs provided by TDK and the TNC corpus⁴. The interface built for searching the TNC corpus gives the possibility to see all sentences that were built with the verb the user is searching for [10]. The senses of the verbs and case marking of their arguments are decided by manually investigating the sentences appear in search results of the TNC corpus. Then, the arguments of the predicates are labeled with VerbNet thematic roles and PropBank argument numbers, by checking the English equivalent of Turkish verb sense. This process is repeated for all verb senses.

For framing purposes, we have adjusted an already available open source software, cornerstone [11]. To supply case marking information of the argument, a drop down menu containing six possible case markers in Turkish is added as shown in Fig 4a. Finally, another drop down menu that contains all possible suffixes that a Turkish verb can have is added, shown in Fig 4b. Theoretically, the

number of possible derivations may be infinite for some Turkish verbs, due to its rich generative property.

However, practically the average number of inflectional groups in a word is less than two. TDK provides a lexicon for widely used verb stems derived from root verbs by a valency changing morpheme. To avoid framing a nonexisting verb, we have used a simple interface shown in Fig 4b to enter only the stems given by TDK. An example with the Turkish verb "bin" (to ride) is given in Fig 4b. The first line defines that one can generate a stem "bin-il" (to be ridden by someone) from the root "bin" by using the suffix "I". Similarly, second line illustrates a two layer derivational morphology, which can be interpreted as producing two verbs: "bin-dir" (cause someone to ride something) and "bindir-il" (to be caused by someone to ride something).



(a) Case marker info given in suffix list (b) Verb derivational info as a drop down menu

Figure 4: Cornerstone Software Adjusted for Turkish

4 Experiments and Results

In Table 1, number of co-occurrences of each thematic role with each case marker are given. Since in PropBank only Arg0 and Arg1 have a certain semantic interpretation, we have used VerbNet thematic roles in our analysis. Some roles look highly related with a case marker, while some look arbitrary. Results can be interpreted in two ways: 1) If the semantic roles are known and case marker information is needed, Agent will be marked with NOM, Destination with DAT, Source with ABL and

Recipient with DAT case with more than 0.98 probability, furthermore Patient and Theme can be restricted to NOM or ACC cases; 2) If case markers are known and semantic role information is needed, only restrictions and prior probabilities can be provided. Highest probabilities occur with COM-instrument, LOC-location, DAT-destination, ACC-Theme and NOM-Agent pairs. We have applied our proposed argument transformation on verbs with different valencies, and compared the argument configurations of the roots and stems.

	NOM	ACC	DAT	LOC	ABL	COM	Total	Explanation
Agent	318	0	1	0	0	0	319	Human or an animate subject that controls or initiates the action.
Patient	36	34	0	0	0	0	70	Participants that undergo a state of change.
Theme	101	117	14	0	7	1	240	Participants in a location or experience a change of location
Beneficiary	1	2	5	0	0	0	8	Entity that benefits negatively or positively from the action.
Location	0	0	2	6	0	0	8	Place or path
Destination	1	0	66	0	0	0	67	End point or direction towards which the motion is directed.
Source	0	0	0	0	29	0	29	Start point of the motion.
Experiencer	13	5	4	0	0	0	22	Usually used for subjects of verbs of perception or psychology.
Stimulus	8	2	4	0	2	0	16	Objects that cause some response from Experiencer.
Instrument	0	0	0	0	0	10	10	Objects that come in contact with an object and cause a change.
Recipient	0	1	13	0	0	0	14	Animate or organization target of transfer.
Time	1	0	2	2	0	0	5	Time.
Topic	0	1	3	0	2	0	6	Theme of communication verbs.
Total	479	162	114	8	40	11	814	

Table 1. Results of Semantic Roles – Case Marking

	#Intransitive	#Transitive	#Hold	#!Hold	Total
Reflexive	0	20	20	0	20
Reciprocal	8	18	26	0	26
Causative	26	11	37	0	37

Table 2. Results of argument transformation

In Table 2, rows represent the valency changes applied to verb root, where Intransitive column contains the number of intransitive verbs that the pattern is applied to, and Transitive similarly. The #Hold column shows the number of root verbs for which the proposed patterns hold, and #!Hold shows the number of times the pattern can not be observed. Reflexive pattern can only be applied to transitive verbs, while others can be applied to both. Experiments are done for reflexive, reciprocal and causative forms. Our preliminary results on a small set of root verbs show that proposed argument transformation can be seen as a regular transformation.

6 Acknowledgements

We thank Gülşen Eryiğit for insightful comments and suggestions that helped us improve this work.

7 References

- [1] **Ana-Maria Giuglea and Alessandro Moschitti.** 2006. Semantic Role Labeling via FrameNet, VerbNet and PropBank. Proceedings of the 21st International Conference on Computational Linguistics, pp. 929-936. 2006.
- [2] **K. Schuler** 2006. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon PhD diss., University of Pennsylvania
- [3] **Martha Palmer, P Kingsbury, and D Gildea.** 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1):71–106
- [4] **Orhan Bilgin, Ozlem Çetinoğlu and Kemal Oflazer.** 2004. Building a wordnet for Turkish. Romanian Journal of Information Science and Technology, 7.1-2 (2004): 163-172.
- [5] **Gülşen Eryiğit, Tugay Ilbay, Ozan A. Can.** 2011. Multiword Expressions in Statistical Dependency Parsing. In Proceedings of the Workshop on Statistical Parsing of Morphologically-Rich Languages SPRML at IWPT, Dublin.
- [6] **Kemal Oflazer, Bilge Say, Dilek Z. Hakkani-Tür, Gökhan Tür.** 2003. Building a Turkish Treebank. Invited chapter in Building and Exploiting Syntactically annotated Corpora, Anne Abeille Editor, Kluwer Academic Publishers
- [7] **Nart B. Atalay, Kemal Oflazer, Bilge Say.** 2003. The Annotation Process in the Turkish Treebank. In Proceedings of the EACL Workshop on Linguistically Interpreted Corpora - LINC, Budapest, Hungary
- [8] **Geoffrey Haig.** 1998. Relative Constructions in Turkish. Otto Harrassowitz Verlag., ISBN 3447040041, (1998)
- [9] **Joan L. Bybee.** 1985. Morphology: A Study of the Relation between Meaning and Form. Typological Studies in Language 9 Amsterdam, Philadelphia: Benjamins
- [10] **Yeşim Aksan, Mustafa Aksan** 2012. Construction of the Turkish National Corpus (TNC). (LREC 2012). Istanbul.
- [11] **Jinho D. Choi, Claire Bonial, and Martha Palmer.** 2010. Propbank Frameset Annotation Guidelines Using a Dedicated Editor, Cornerstone. (LREC 10), Valletta, Malta