A Multilayer Annotated Corpus for Turkish

Olcay Taner Yıldız Koray Ak Gökhan Ercan Ozan Topsakal Cengiz Asmazoğlu

Department of Computer Engineering, Işık University, İstanbul, Turkey

ICNLSP 2018

Outline

Motivation

- Two central tasks: semantic analysis, syntactic analysis.
- Most of the NLP studies focus on analytic languages like English.
- An acute need for these kinds of resources, especially for non-English languages.
- We present the first multilayer annotated corpus for Turkish.

Morphological Disambiguation

- Each word consist of one or more morphemes, the smallest linguistic unit having a particular meaning or grammatical function.
- Turkish: agglutinative language, derivational and inflectional suffixes attached to the roots.
- Morphological disambiguation is the process of identifying the correct morphological analysis of a word.

sorun + NOUN + A3SG + PNON + ACC (the problem)
sorun + NOUN + A3SG + P3SG + NOM (her/his problem)
soru + NOUN + A3SG + P2SG + ACC (your question)

Named Entity Tagging

- Anything that is denoted by a proper name, i. e., for instance, a person, a location, or an organization.

Tag	**Sample Categories**
PERSON	people, characters
ORGANIZATION	companies, teams
LOCATION	regions, mountains, seas
TIME	time expressions
MONEY	monetarial expressions

[$_{ORG}$ Türk Hava Yolları] bu [$_{TIME}$ Pazartesi'den] itibaren [$_{LOC}$ İstanbul] [$_{LOC}$ Ankara] hattı için indirimli satışlarını [$_{MONEY}$ 90 TL'den] başlatacağını açıkladı.
[$_{ORG}$ Turkish Airlines] announced that from this [$_{TIME}$ Monday] on it will start its discounted fares of [$_{MONEY}$ 90TL] for [$_{LOC}$ İstanbul] [$_{LOC}$ Ankara] route.

Shallow Parsing

- Shallow parsing is the process of identifying the flat, non-overlapping parts of a sentence.

Tag	**Question**	**Parts**
ÖZNE	Who, What	Subject
ZARF TÜMLECİ	When, How, Why	Adverbial Clause
DOLAYLI TÜMLEÇ	Where, To/From whom	Adverbial Clause
NESNE	What, Whom	Object
YÜKLEM	Predicate	

[*OZNE* Türk Hava Yolları] [*ZARF TUMLECI* Salı günü] [*NESNE* yeni indirimli fiyatlarını] [*YUKLEM* açıkladı]
[*SUBJECT* Turkish Airlines] [*PREDICATE* announced] [*OBJECT* new discounted fares] [*ADVERBIAL CLAUSE* on Tuesday]

Word Sense Disambiguation

- The task of choosing the correct sense for a word is called word sense disambiguation.
- In a **lexical sample** task, a small selected set of target words is chosen, along with a set of senses for each target word.
- In an **all-words** task, systems are given entire sentences and a lexicon with the set of senses for each word in each sentence.

Sense	**Definition**
yüz[1] (hundred)	The number coming after ninety nine
yüz[2] (swim)	move or float in water
yüz[3] (face)	face, visage, countenance

Semantic Role Labeling

- Semantic Role Labeling is a well-defined task where the objective is to analyze propositions expressed by the verb.

Tag	**Meaning**	**Tag**	**Meaning**
Arg0	Agent or Causer	ArgM-EXT	Extent
Arg1	Patient or Theme	ArgM-LOC	Locatives
Arg2	Instrument, start point, end point		
ArgM-CAU	Cause	ArgM-MNR	Manner
ArgM-DIS	Discourse	ArgM-ADV	Adverbials
ArgM-DIR	Directionals	ArgM-PNC	Purpose
ArgM-TMP	Temporals		

[$_{ARG0}$ Türk Hava Yolları] [$_{ARG1}$ indirimli satışlarını] [$_{ARGM-TMP}$ bu Pazartesi] [$_{PREDICATE}$ açıkladı].
[$_{ARG0}$ Turkish Airlines] [$_{PREDICATE}$ announced] [$_{ARG1}$ its discounted fares] [$_{ARGM-TMP}$ this Monday].

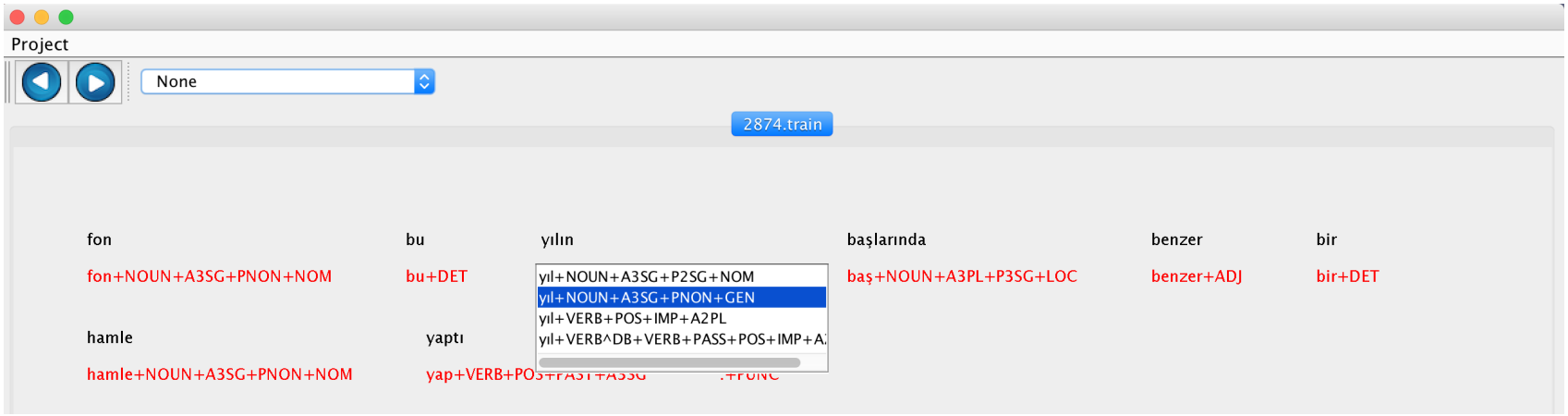
Annotation Setup

- The original data for our corpus is drawn from the Penn-Treebank corpus.
- Selected sentences from this Penn-Treebank corpus containing less than 15 words are translated into Turkish.
- The corpus currently contains 9,600 sentences.
- We worked with six annotators, all undergraduate students of Işık University.
- Video guidelines for all editors for annotation were prepared based on guidelines provided by linguists.

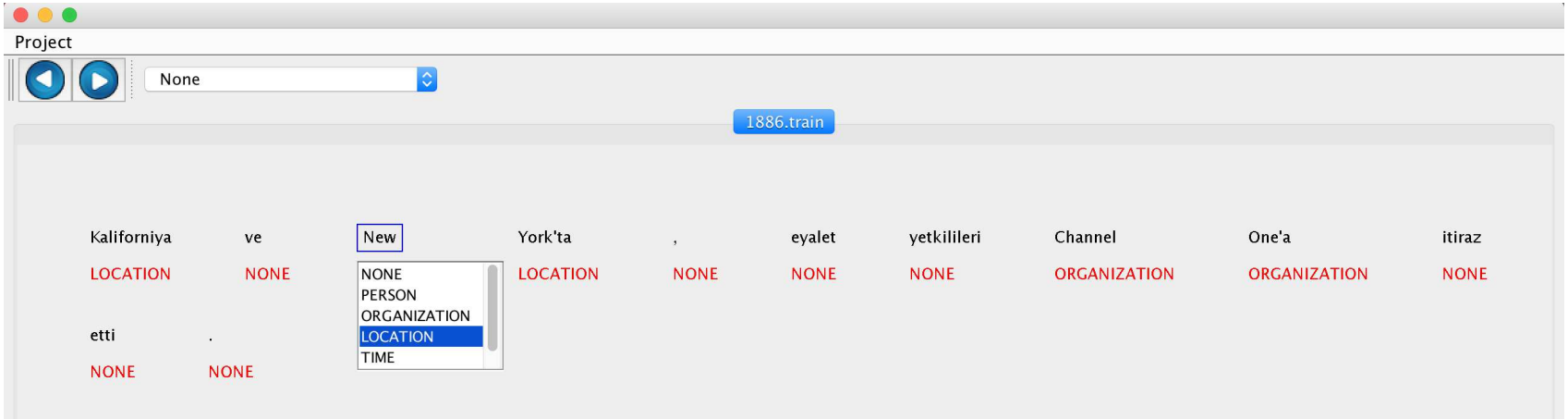
Data format

{turkish=yatırımcılar}
{analysis=yatırımcı+NOUN+A3PL+PNON+NOM}
{semantics=0841060}{namedEntity=NONE}
{shallowParse=ÖZNE}{propbank=ARG0:0006410}

Interface (I)

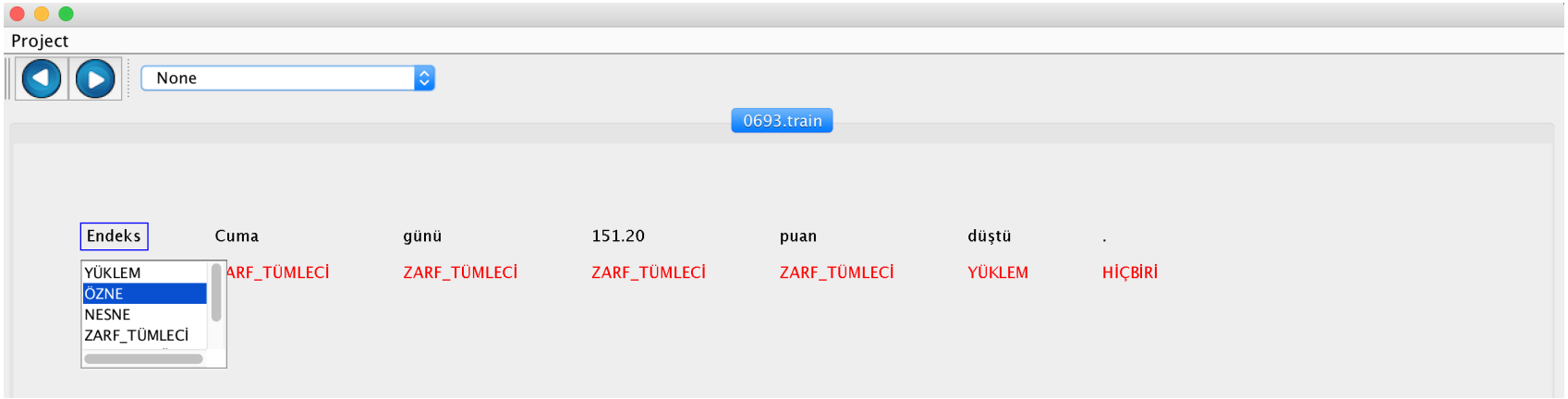


(a) Morphological disambiguation interface

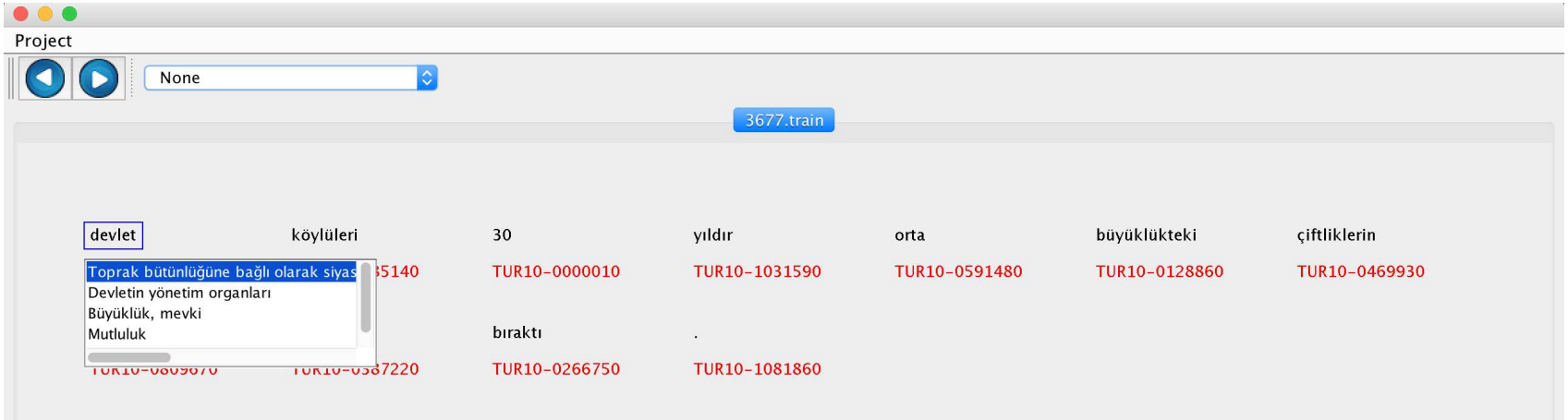


(b) Named entity annotation interface

Interface (II)

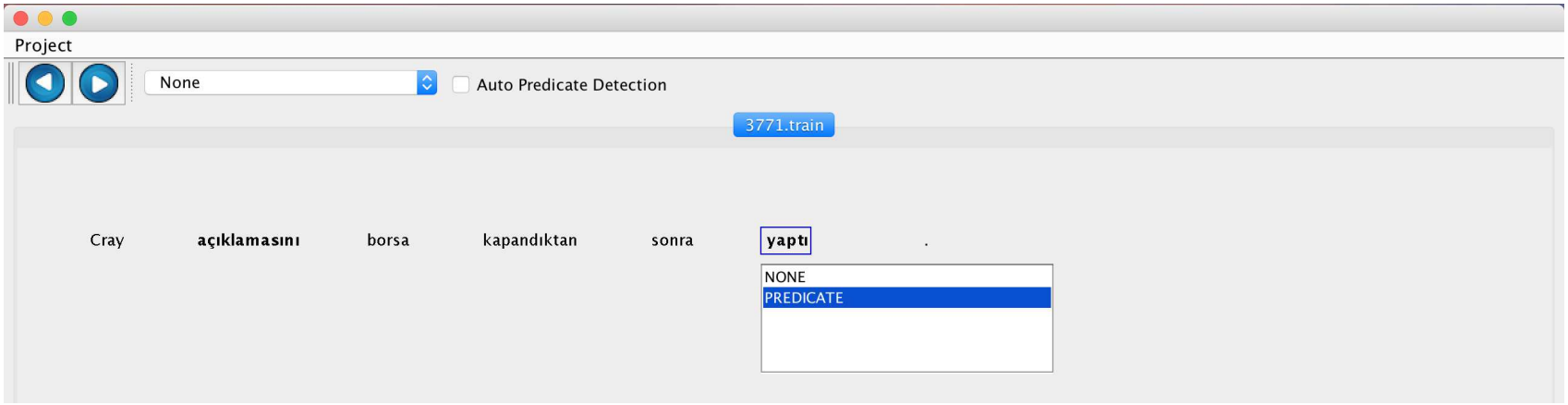


(c) Shallow parse annotation interface

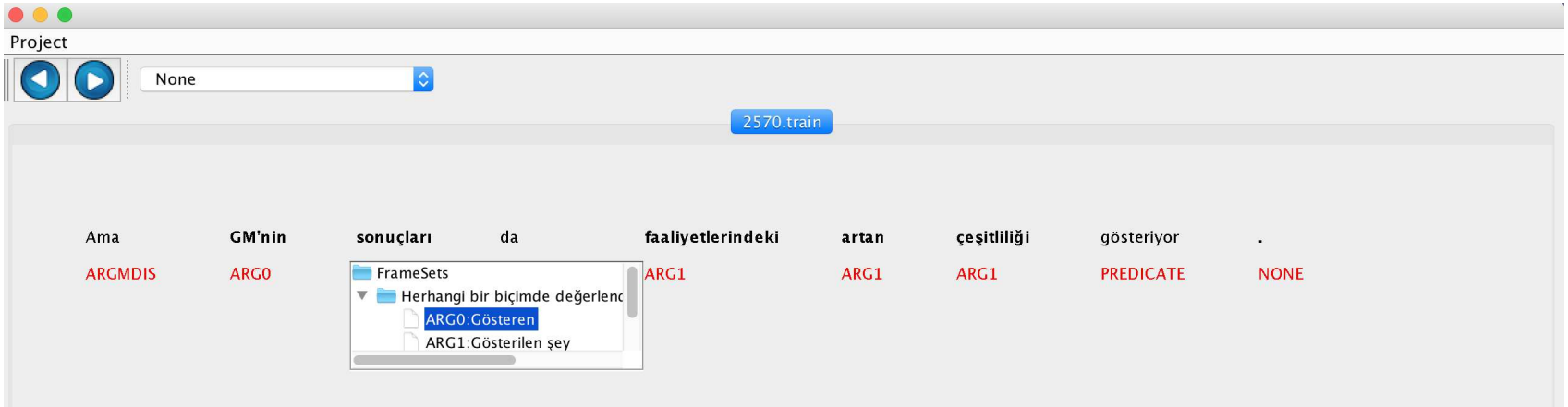


(d) Sense annotation interface

Interface (III)



(e) Predicate selection interface



(f) Semantic role labeling interface

Inter-annotator Agreement

Layer	Agreement	E. Agreement	C. Kappa
NER	0.975	0.167	0.969
Shallow Parse	0.79	0.167	0.748
Sense Annotation	0.785	0.545	0.527

10 most-frequent root words and POS tags of the root words

Root word	Count	Root word	Count
olmak (be)	1421	hisse (share, stock)	380
etmek (do)	796	dolar (dollar)	373
bay (mr.)	476	milyon (million)	362
yapmak (do)	391	artmak (increase)	347
şirket (company, firm)	385	gelmek (come)	326

Word Type	Count	Word Type	Count
Noun	34,433	Number	4,240
Punctuation	13,896	Adverb	2,994
Verb	11,964	Pronoun	1,049
Adjective	6,643		

Distribution of the named entity tags

Tag	**Count**	**Percentage**
ORGANIZATION	4,418	5.05
PERSON	2,612	2.99
MONEY	2,240	2.56
LOCATION	1,303	1.49
TIME	1,194	1.37
NONE	75,682	86.54
Total	87,449	100.00

Distribution of the shallow parse tags

Tag	**Count**	**Percentage**
NESNE	15,768	20.02
ÖZNE	13,996	17.77
ZARF TÜMLECİ	13,003	16.51
YÜKLEM	11,607	14.74
DOLAYLI TÜMLEÇ	7,252	9.21
HİÇBİRİ	17,134	21.75
Total	78,760	100.00

Distribution of the semantic roles

Tag	Count	Tag	Count
ARG0	716	ARGM-LOC	74
ARG1	1,066	ARGM-EXT	67
ARG2	55	ARGM-DIS	41
ARGM-MNR	157	ARGM-ADV	24
ARGM-TMP	103	ARGM-CAU	13

Questions?