An Overview of Resources Available for Turkish Natural Language Processing Applications

Tunga GÜNGÖR
Computer Engineering, Boğaziçi University
TurcLing 2016 Keynote Speaker
An Overview of Resources Available for Turkish Natural Language Processing Applications

Tunga Gungor
Boğaziçi University
Computer Engineering

Turkish WordNet

- Syntactic and structural quality was measured
- Coverage was measured using two sources
  - George Orwell's 1984 novel & a general-purpose corpus
  - Coverage of words was found to be 87%

Outline

- Turkish WordNet
- METU Turkish Corpus
- Turkish Treebank
- Boum Resources
- Kemik Group (YTU)
- ITU Turkish NLP Pipeline
- Turkish Labeled Text Corpus
- TS Corpus

- Other Resources
- Conclusions
- Paper List

Turkish WordNet

- Basic statistics (March 2004)

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synsets</td>
<td>11,628</td>
</tr>
<tr>
<td>Synset members</td>
<td>16,095</td>
</tr>
<tr>
<td>Average synset</td>
<td>1.38</td>
</tr>
<tr>
<td>Unlexicalized</td>
<td>987</td>
</tr>
<tr>
<td>Definitions</td>
<td>4,913</td>
</tr>
</tbody>
</table>

Turkish WordNet

- Part of BalkaNet project
  - WordNets in Central and Eastern European languages
- Linked to Princeton WordNet and Euro WordNet
- Methodology
  - Base concepts of Euro WordNet translated
  - Synonyms, antonyms and hyponyms extracted automatically from monolingual Turkish dictionary
  - Base concept set enlarged

Turkish WordNet

- Parts of speech

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>8,691</td>
<td>74.7%</td>
</tr>
<tr>
<td>Verbs</td>
<td>2,556</td>
<td>22.0%</td>
</tr>
<tr>
<td>Adjectives</td>
<td>361</td>
<td>3.3%</td>
</tr>
</tbody>
</table>
METU Turkish Corpus

Turkish Treebank

Subcorpus of METU Turkish Corpus
10,000 sentences with morphological and syntactic annotations
Syntactic relations:
Sentence Subject
Object Modifier
Modifier Determiner
Focus-Participle Question-Participle
Vocative Classifier
Direct-Adjective Attributive-Adjective
Locative-Adjective Instr-Adjective
Coordination Relation

Turkish Treebank

Annotation tool

This school-the is students' most intelligence of them standing little girl is.

The most intelligent of the students in this school is the little girl standing there.
Boun Resources

- Developed at Boğaziçi University
- Resources:
  - Stochastic morphological parser
  - Morphological disambiguator
  - Web corpus

Boun Resources

- Lexicon and morphotactics
  - Based on PC-KIMMO implementation (K. Ofazer)
  - Morphological features from Xerox implementation
  - 55K root word lexicon
  - 33 two-level rules hand-compiled to FSTs
  - Improved morphotactics
  - 11,000 words/sec

Boun Resources

- Morphological parser
  - Lexicon: Stores the root words and some morphological features
  - Morphotactics: (Morphosyntax) is the word grammar for the ordering of morphemes
  - Morphophonemics: Phonological rules describing morphological alternations
  - Stochasticized using the morphological disambiguator and web corpus

Boun Resources

- Lexicon and morphotactics
  - An example output for the word alını:
    - alını [Verb]+[Pos]+[Imp]+[A2sg]
    - alını [Noun]+[A3sg]+[Pron]
    - alını [Noun]+[A3sg]+[Pron]+[Verb+Pres+A3sg]
    - al [Noun]+[A3sg]+[Imp]+[P2sg]
    - al [Noun]+[A3sg]+[Imp]+[P2sg]+[Verb+Pres+A3sg]
    - al [Noun]+[A3sg]+[Imp]+[Gen]+[Verb+Pres+A3sg]
    - al [Verb]+[Pos]+[Imp]+[Verb+A2sg]
    - al [Adj]+[Noun]+[A3sg]+[Imp]+[P2sg]

Boun Resources

- Lexicon and morphotactics
  - Implemented as a single finite-state transducer
  - An example for Turkish nominal inflection:
    ![Example Diagram]
  - Output side is input to two-level rules transducer
  - Special symbols are used for phonological alternations
  - Compose lexicon/morphotactics transducer with the morphophonemics transducer

Boun Resources

- Morphological disambiguator
  - Based on averaged perceptron algorithm
    - alını [Verb]+[Pos]+[Imp]+[A2sg]
    - alını [Noun]+[A3sg]+[Pron]
    - alını [Noun]+[A3sg]+[Pron]+[Verb+Pres+A3sg]
    - al [Noun]+[A3sg]+[Imp]+[P2sg]
    - al [Noun]+[A3sg]+[Imp]+[P2sg]+[Verb+Pres+A3sg]
    - al [Noun]+[A3sg]+[Imp]+[Gen]+[Verb+Pres+A3sg]
    - al [Verb]+[Pos]+[Imp]+[Verb+A2sg]
    - al [Verb]+[Pron]+[Pos]+[Imp]+[A2sg]
    - al [Adj]+[Noun]+[A3sg]+[Imp]+[P2sg]
  - 97.8% success rate
Morphological disambiguator

<table>
<thead>
<tr>
<th>Gloss</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological parse tag</td>
<td>(1) L_{i} L_{j} L_{k}</td>
</tr>
<tr>
<td>Morphological parse tag</td>
<td>(2) L_{i} L_{j} (3) L_{i} L_{j}</td>
</tr>
<tr>
<td>Morphological parse tag</td>
<td>(4) L_{i}</td>
</tr>
<tr>
<td>Morpheme tag with previous tag</td>
<td>(5) L_{i} L_{j} m_{k}</td>
</tr>
<tr>
<td>Morpheme tag with second to previous tag</td>
<td>(6) L_{i} L_{j} m_{k} n_{l}</td>
</tr>
<tr>
<td>Root tag</td>
<td>(7) L_{i} L_{j} L_{k}</td>
</tr>
<tr>
<td>Root tag</td>
<td>(8) L_{i} L_{j} L_{k} (9) L_{i} L_{j} L_{k}</td>
</tr>
<tr>
<td>Root unigram</td>
<td>(10) L_{i}</td>
</tr>
<tr>
<td>Morpheme tag unigram</td>
<td>(11) m_{i} m_{j} m_{k}</td>
</tr>
<tr>
<td>Morpheme tag unigram</td>
<td>(12) m_{i} m_{j} (13) m_{i} m_{j}</td>
</tr>
<tr>
<td>Individual morpheme tags</td>
<td>(14) m_{i}</td>
</tr>
<tr>
<td>Individual morpheme tags</td>
<td>(15) m_{i} m_{j} for j = 1, ..., n</td>
</tr>
<tr>
<td>Individual morpheme tags with position</td>
<td>(16) m_{i} m_{j} for j = 1, ..., n</td>
</tr>
<tr>
<td>Number of morpheme tags</td>
<td>(17) n</td>
</tr>
</tbody>
</table>

Web Corpus

Coverage statistics

<table>
<thead>
<tr>
<th>Vocabulary size</th>
<th>Coverage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1K</td>
<td>50.5</td>
</tr>
<tr>
<td>5K</td>
<td>67.0</td>
</tr>
<tr>
<td>10K</td>
<td>73.0</td>
</tr>
<tr>
<td>20K</td>
<td>81.5</td>
</tr>
<tr>
<td>50K</td>
<td>89.0</td>
</tr>
<tr>
<td>100K</td>
<td>93.2</td>
</tr>
<tr>
<td>200K</td>
<td>94.7</td>
</tr>
<tr>
<td>400K</td>
<td>96.2</td>
</tr>
<tr>
<td>1M</td>
<td>98.9</td>
</tr>
<tr>
<td>3M</td>
<td>99.2</td>
</tr>
<tr>
<td>4M</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Web Corpus

Formed of two subcorpora:
- NewsCor: Milliyet, Nhmsnbnc, Radikal
- GenCor: General sampling of web pages

About 500M tokens

Cleaned using morphological parser and heuristics

Converted to XCES XML format

Web Corpus

Type statistics

Web Corpus

Stem and lexical ending statistics
Kemik Group (YTU)

Datasets

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Explanation</th>
<th>Download</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 Turkish Tweets</td>
<td>200,000 Turkish Tweets</td>
<td>2000tweets.zip</td>
<td>0.1 MB</td>
</tr>
<tr>
<td>1500 Columns</td>
<td>Columns from 1500 different Columns in raw and aiff formats</td>
<td>1500columns.zip</td>
<td>7 MB</td>
</tr>
</tbody>
</table>

Kemik Group (YTU)

Dataset: 2000 parsed Turkish sentences

Manually tagged with syntactic labels

<table>
<thead>
<tr>
<th>Label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zaman</td>
<td>Zaman Tümearı (Time Complement)</td>
</tr>
<tr>
<td>DT</td>
<td>Dolaylı Tümearı (Indirect Object)</td>
</tr>
<tr>
<td>İle</td>
<td>Birliklik (With)</td>
</tr>
<tr>
<td>Y</td>
<td>Yüklem (Verb)</td>
</tr>
<tr>
<td>Öne</td>
<td>Öne (Subject)</td>
</tr>
<tr>
<td>Nasıl</td>
<td>Nasıl (How)</td>
</tr>
<tr>
<td>Baz</td>
<td>Baz (Indefinite Object)</td>
</tr>
<tr>
<td>Bil</td>
<td>Bil (Definite Object)</td>
</tr>
<tr>
<td>O</td>
<td>On (Punctuation)</td>
</tr>
<tr>
<td>İkin</td>
<td>İkin (Cause)</td>
</tr>
<tr>
<td>Kadar</td>
<td>Kadar (Until)</td>
</tr>
</tbody>
</table>

Kemik Group (YTU)

Dataset: 2000 parsed Turkish sentences

Example sentence:

Ağır    Zaman
çay    DT
bağışında    DT
arkadaşıyla    İle
bulsuçak    Y

Kemik Group (YTU)

Dataset: 2000 parsed Turkish sentences

Example sentence (with morphological labels):

Ağır  000...000  Zaman
çay  010...000  DT
bağışında  010...000  DT
arkadaşıyla  010...000  İle
bulsuçak  100...000  Y

.  000...000  O
Kemik Group (YTU)

Dataset: 2000 parsed Turkish sentences

- 113 morphological labels

Dataset: 2M Turkish tweets

- In the form: date – user – tweet
- Collected during July-August 2012 period

Kemik Group (YTU)

Dataset: 2500 Columns

- 50 columnists
- 50 texts for each
- Collected during 2012-2013 period

Dataset: Word meanings

- An online dictionary containing 10.5K words and their meanings

Kemik Group (YTU)

Software and Tools

- TextRank v3.1: A program that automatically extracts various features of Turkish texts and saves in .aff format.
- Author Recognition Setup: Author Recognition is done using Artificial Neural Networks. Also, for use in Win ARFF format files is created.
- Document Similarity: A program that calculates Fugum and saldile statistics either from Turkish dictionary of spelling or any given Turkish text.
- Author Detection: An Author Detection program which utilizes Artificial Immune System and Anti-Corruption Optimization Methods.
ITU Turkish NLP Pipeline

Preprocessing

- A cascaded approach for text normalization
- Two stages:
  - Ill-formed word detection
  - Using morphological analyzer and abbreviation list
  - Candidate word generation
  - Rule-based and machine learning
  - Seven normalization layers

Different layers of normalization

- User interface via three channels:
  - Web interface
  - Uploading a file
  - Using Web APIs

- Ill-formed word detection: 86%
- Candidate generation: 71%
ITU Turkish NLP Pipeline

Preprocessing

- Vowel and diacritic restoration
- Two stage model:
  - Discriminative sequence classifier
    - Uses CRF to produce the most probable words
    - Views as a character level sequence labeling task
  - Language validator
    - Uses morphological analyzer to select the most probable valid word among the n-best sequences

Dependency parser

- State-of-the-art syntactic parser for Turkish
- Dependencies based on inflectional groups rather than words

ITU Turkish NLP Pipeline

Preprocessing

- Spell correction
- Finding words in a dictionary most similar to the misspelled word
- A probabilistic model is trained
- A set of rules for transformation are defined

Dependency parser

- **masa** + Noun + A3sg + Pron + Nom + DB + Adj + With

  - **stem**
  - **features**
  - **inflectional group (IG)**
  - **derivational boundary**
  - **IG**

- IGs can independently act as heads or modifiers in syntactic dependencies

Named entity recognizer

- Entity types:
  - person | date | money
  - location | time | percentage
  - organization

- Uses CRF as the probabilistic model
- Raw tags & IOB tags
- Uses morphological processor and gazetteers
- Morphological, lexical, and gazetteer features

Tokenizer
**TS Corpus**

- A project that aims at building Turkish corpora, NLP tools, and linguistic datasets.
- Started in 2011, first release published on March 2012

---

**TS Corpus – Main corpus**

- Corpus can be queried online

---

**TS Corpus**

- General purpose corpora
  - TS Corpus
  - Wikipedia
  - Tweets
  - Idioms and proverbs
- Specialized corpora
  - Abstracts
  - Newspaper corpus
  - Gezi corpus
  - Constitutions

---

**TS Corpus – Main corpus**

- Corpus can be queried online
  - Simple query mod
  - CQP syntax

<table>
<thead>
<tr>
<th>Attributes in this corpus:</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
</tr>
<tr>
<td>Lemma</td>
</tr>
<tr>
<td>Morph</td>
</tr>
<tr>
<td>PoS/Tag</td>
</tr>
</tbody>
</table>

- Attributes in this corpus:
  - <text> Text
  - <text_id> Text ID

---

**TS Corpus – Main corpus**

- Based on Boun Web Corpus
- 491M tokens and 4.9M word types
- Words with POS tags and morphological annotations
- Based on CWB (IMS Open Corpus Workbench) CQP (Corpus Query Processor) structure
  - Queries can be written using regular expressions or CQP commands
TS Corpus – Main corpus

- Collected from Turkish Wikipedia pages, beginning from July 2013
- A corpus with diverse topics
- 215K pages
- 45M tokens
- 1.7M word types
- Preprocessed and cleaned
  - Auto-generated empty entries eliminated
  - External URL, image, table and other non-text contents deleted
  - POS-tagged and morphologically annotated

TS TweetS Corpus

- Data collected from Twitter
- 1M tweets
- 12.5M tokens
- POS-tagged and morphologically annotated

TS Wikipedia Corpus

- Academic text from various disciplines
- Domains:
  - Social sciences
  - Physical sciences
- Genres:
  - Humanities & arts
  - Medicine
  - Natural sciences
  - Politics, law & education
  - Social sciences
  - Technology & engineering
- Useful for text genre classification

TS Abstract Corpus

- Based on Turkish Labeled Text Corpus
- 32 disciplines
- 1M tokens
- 6.2K abstracts
TS Gezi Corpus
- Collected from Turkish and foreign press, during Taksim Gezi Park protests
- 9 sources (BBC, Cumhuriyet, France 24, Hürriyet, NY Times, Spiegel, Telegraph, Washington Post, Zaman)
- 2,968 articles
- 1.1M tokens
- Specialized corpus
- Covers a period of 1 month
- Can be queried wrt source, language, genre, period

TS Constitution Corpus
- Collection of three constitutions (1924, 1961, 1982)
- 32M tokens
- 5.3K word types
- Grouped wrt constitution and article number
- Can be queried wrt constitution year

TS Idioms and Proverbs Corpus
- 10K idioms and proverbs
- 27.3K tokens
- Query returns the full idiom or proverb

Other Resources
- Zemberek
  [https://github.com/ahmetsaa/zemberek-nlp](https://github.com/ahmetsaa/zemberek-nlp)
- TRmorph and other tools
  [https://github.com/coltekin/](https://github.com/coltekin/)
- Morphological disambiguator (D. Yuret)

Conclusions
- Currently available NLP resources for Turkish language
  - Semantic relations knowledge base
  - Treebank
  - Morphological analysis and disambiguation
  - Named entity recognition
  - Text normalizer
  - Syntactic analysis
  - Various corpora