Dependency Parsing of Turkish

Eryiğit, Nivre, & Oflazer(2008)

Anna Donohoe

Topics in NLP

31. 3. 2015

Outline

- Motivation
- Background
  - Turkish Morphology
  - Inflectional Groups
- Methodology
  - Treebank
  - Evaluation
- Models & Experiments
  - Probabilistic Dependency Parser
  - Classifier-Based Dependency Parser
- Results
- Conclusion
- References

Motivation

- Syntactic parsing of natural language has become more robust in the last few decades with data-driven and grammar-based methods
  - Many approaches only focus on constituency-based representations of English and a few other languages
  - Models and algorithms are often tailored to properties of specific languages or languages groups
- Eryiğit et al. demonstrate that free-constituent order and morphologically rich languages can be better analyzed using dependency-based representations and sublexical units

Dependency Parsing of Turkish

- Eryiğit et al. focus on Turkish, but view it as “representative of a class of languages that are very different from English and most other languages that have been studied in the parsing literature”
- Experiments investigate issues surrounding *morphology, lexicalization,* and *parsing methodology*
- Introduce two dependency parsing models, one probabilistic and one classifier-based that incorporates lexicalization

Turkish Morphology

- Turkish is a highly agglutinative, free constituent order language spoken by around 70 million people worldwide
- Because so much syntactic information is mediated by morphology in Turkish, it insufficient for a parser to only identify dependency relations between orthographic words
- For example…

OSMANLILAŞTIRAMAYABİLECEKLERİMİZDENMİŞSİNİZCESİNE

'Behaving as if you were of those whom we might consider not converting into an Ottoman'

49 letters, 13 morphemes…

OSMAN
+LI
+LAŞ
+TIR
+AMA
+YABİL
+ECEK
+LER
+İMİZ
+DEN
+MİŞ
+SİNİZ
+CESİNE

Another example

Bu okuldaki öğrencilerin en akıllısı şurada duran küçük kızdır

The school+at+this students-s' most intelligence+with+of there stand+ing little girl+is

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

Inflectional Groups (IGs)

- Eryiğit et al. build on previous work on Turkish morphology by splitting Turkish words into Inflectional Groups
- IGs express the root and derivational elements of a word, and are separated by Derivational Boundaries (DBs)
- IGs are are also annotated with POS and inflectional features

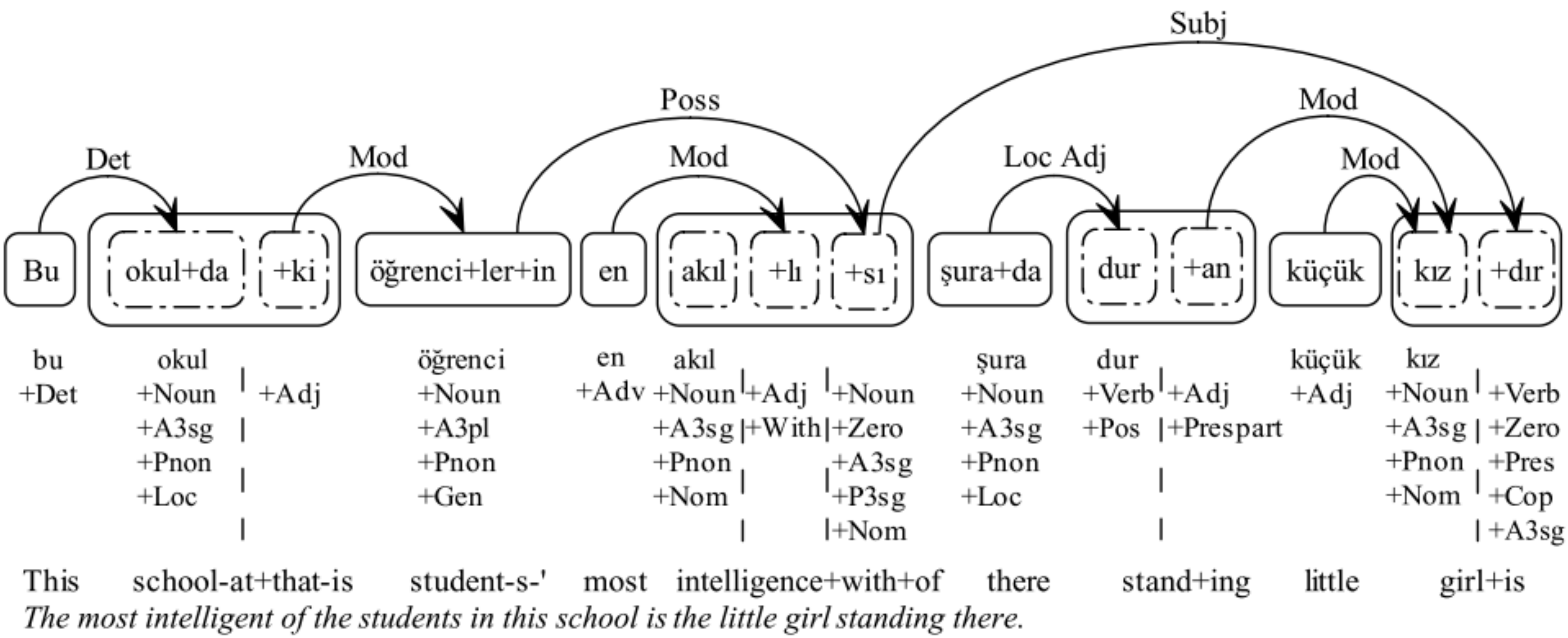
arabanızdaydı
('it was in your car')

arabanızda DB *ydı*

$$\underbrace{\textit{araba}\text{+Noun+A3sg+P2pl+Loc}}_{IG_1} \quad \text{DB} \quad \underbrace{\text{+Verb+Zero+Past+A3sg}}_{IG_2}$$

'in your car' 'it was'

Dependency Tree with IGs



= word boundaries = IG boundaries + = morpheme boundaries

Treebank & Evaluation

- Turkish Treebank, a small subset of the Metu Turkish Corpus
  - A balanced corpus of 5,000+ sentences; words are represented with IG-based gold-standard morphological representations and dependency links between IGs
- Evaluated on entire treebank using 10-fold cross-validation
- Results reported as mean scores of the cross-validation, with standard error taken into account
- Evaluation Metrics
  - **Unlabeled Attachment Score (AS_U)**– proportion of IGs that are attached to the correct head
  - **Labeled Attachment Score (AS_L)**– proportion of IGs that are both attached to the correct head *and* labeled correctly

Parsing Models I
Probabilistic Dependency Parser

- Data-driven, statistical parser that uses a conditional probabilistic model

- Assigns a probability to each candidate dependency link based on frequency of similar dependencies in the training set

- $$T^* = \underset{T}{\operatorname{argmax}}\, P(T|S) = \underset{T}{\operatorname{argmax}} \prod_{i=1}^{n-1} P(dep\,(u_i, u_{\mathcal{H}(i)})\,|\,S) \tag{1}$$

- $$P(dep\,(u_i, u_{\mathcal{H}(i)})\,|\,S) \approx P(link(u_i, u_{\mathcal{H}(i)})\,|\,\Phi_i\;\Phi_{\mathcal{H}(i)}) \tag{2}$$

$$P(u_i \text{ links to some head } dist(i, H(i)) \text{ away}\,|\,\Phi_i)$$

Parsing Model II
Classifier-Based Dependency Parser

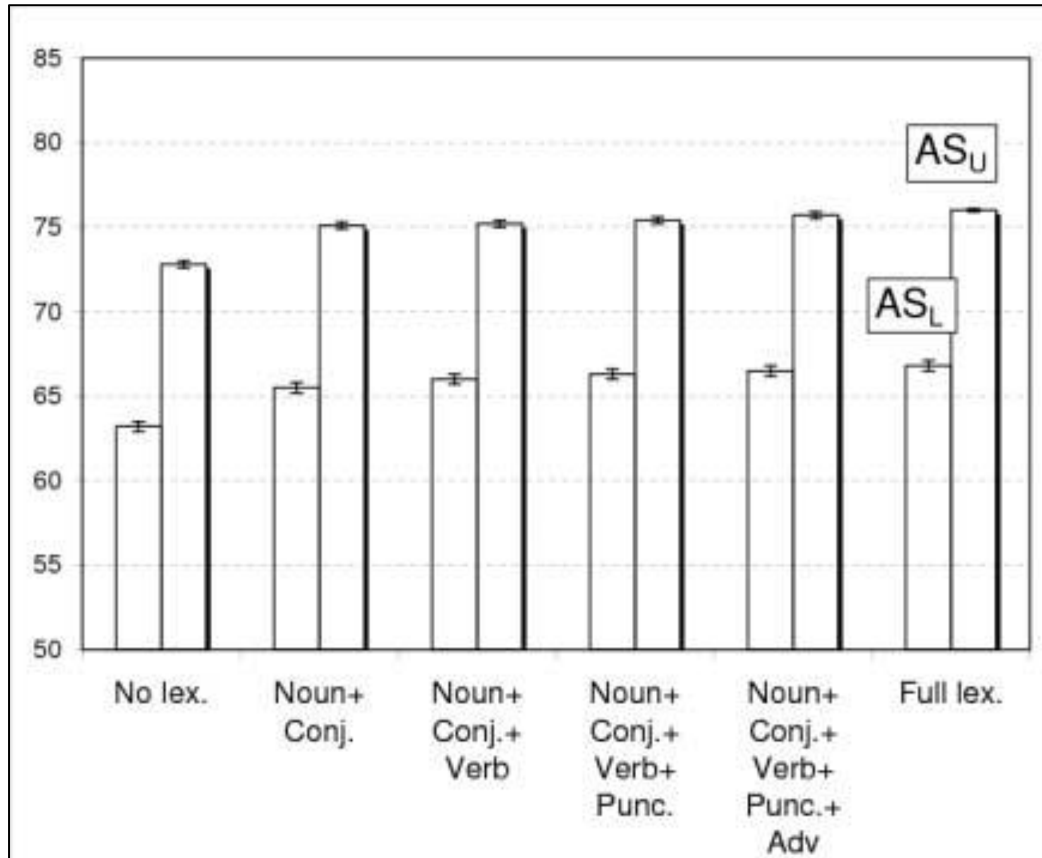
- Data-driven, deterministic classifier-based parser using discriminative learning

- Linear-time algorithm that derives a labeled dependency graph in one pass, with partially processed tokens stored in a stack and remaining input tokens stored in a list

- Types of Parsing Actions
  - **Shift**: Push the next token onto the sack
  - **Left-Arc$_r$**: Add a dependency arc from the next token to the top token (*r*), then pop the stack
  - **Right-Arc$_r$**: Add a dependency arc from the top token to the next token(*r*), then replace next token with the top token at head of input list

Parsing Model II
Lexicalization

- The classifier-based parser incorporates various levels lexicalization
- Lexicalization can improve parsing accuracy under this model because, unlike the probabilistic model, it is less sensitive to sparse data
  - Unlabeled scores are higher than labeled scores



Probabilistic Dependency Parser Results

Parsing Model	AS_U	AS_L
Word-based model	67.1±0.3	57.8±0.3
IG-based model	70.6±0.2	60.9±0.3

- The IG-based model outperformed the word-based model in terms of both Unlabeled and Labeled Attachment Score
  - IG-based model considers IG and word relations and head words
  - Word-based model ignores within-word dependencies and labels

Classifier-Based Dependency Parser Results

CoNLL-X shared task results on Turkish (taken from Table 5 in Buchholz and Marsi [2006]).

Teams	AS_U	AS_L
Nivre et al. (2006)	75.8	65.7
Johansson and Nugues (2006)	73.6	63.4
McDonald, Lerman, and Pereira (2006)	74.7	63.2
Corston-Oliver and Aue (2006)	73.1	61.7
Cheng, Asahara, and Matsumoto (2006)	74.5	61.2
Chang, Do, and Roth (2006)	73.2	60.5
Yüret (2006)	71.5	60.3
Riedel, Çakıcı, and Meza-Ruiz (2006)	74.1	58.6
Carreras, Surdeanu, and Marquez (2006)	70.1	58.1
Wu, Lee, and Yang (2006)	69.3	55.1
Shimizu (2006)	68.8	54.2
Bick (2006)	65.5	53.9
Canisius et al. (2006)	64.2	51.1
Schiehlen and Spranger (2006)	61.6	49.8
Dreyer, Smith, and Smith (2006)	60.5	46.1
Liu et al. (2006)	56.9	41.7
Attardi (2006)	65.3	37.8

- The authors’ Unlabeled Attachment Score of 75.8 is the highest reported accuracy for parsing the Turkish Treebank

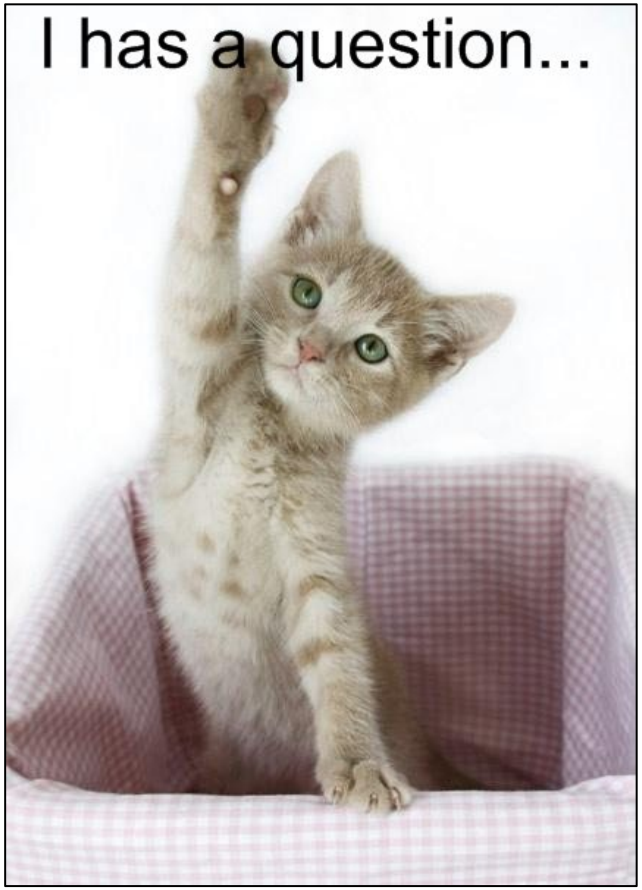
Conclusion

- Using sublexical parsing units (IGs) substantially improves parsing accuracy for Turkish
- Parsing of Turkish (and by extension, other morphologically rich and flexible constituent order languages) benefits from incorporating dependency relations
- Future work
  - Extend the existing system to cover other languages
  - Incorporate non-projective dependency structures (crossing arcs) into the classifier-based parsing model

References

- Eryiğit, G., Nivre, J., & Oflazer, K. (2008). Dependency parsing of Turkish. *Computational Linguistics*, 34(3), 357-389.
- Hwa, R., Resnik, P., & Weinberg, A. (2005). Breaking the resource bottleneck for multilingual parsing. *Maryland University College Park Institute for Advanced Computer Studies*.
- Nivre, J. (2005). Dependency grammar and dependency parsing. *MSI report*, 5133(1959), 1-32.
- Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and linguistic computing*, 9(2), 137-148.
- Tsarfaty, R., Seddah, D., Kübler, S., & Nivre, J. (2013). Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1), 15-22.

Questions?



- Thanks for your attention!