Morphological Disambiguation for Turkish

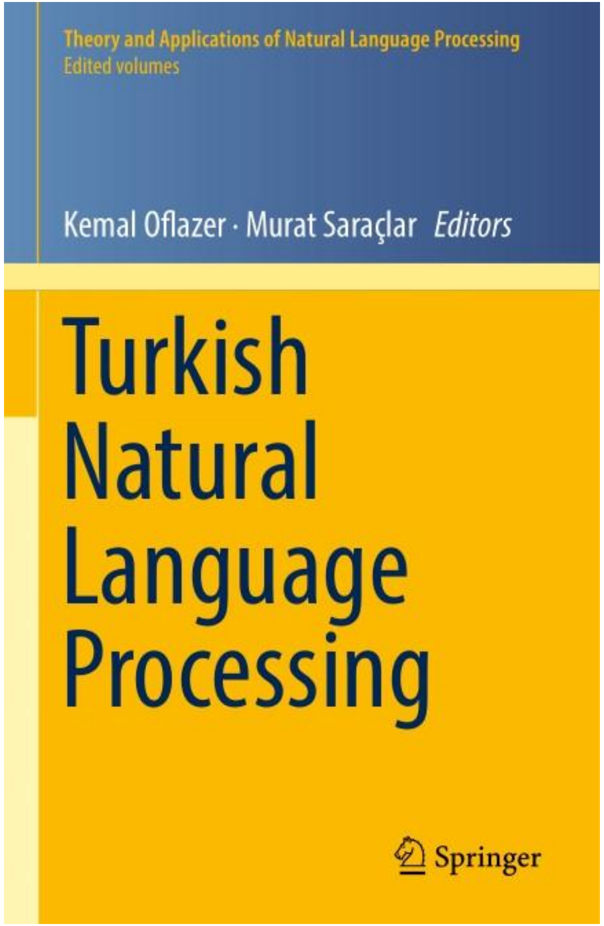
Selçuk Gülcan

Hakkani-Tür, Dilek Zeynep, et al. 2018, “Morphological Disambiguation for Turkish.” pp 53-67 in: Turkish Natural Language Processing. Springer, Cham.

Book

- Kemal Oflazer
- Murat Saraçlar

Oflazer, Kemal, and Murat Saraçlar, eds. Turkish Natural Language Processing. Springer, 2018.



Outline

- Turkish Language
- Morphological Ambiguity Problem
- Methods
- Datasets and Results

Turkish Language

- Free constituent order
- Consider words a, b, c
- All 6 permutation is valid:
  - a b c
  - a c b
  - b c a
  - ...

Turkish Language

- Ekin Çağla’yı gördü. (Ekin saw Çağla.)
- Çağla’yı Ekin gördü. (It was Ekin who saw Çağla.)
- Gördü Ekin Çağla’yı. (Ekin saw Çağla (but was not really supposed to see her.))
- Gördü Çağla’yı Ekin. (Ekin saw Çağla (and I was expecting that)
- Ekin gördü Çağla’yı. (It was Ekin who saw Çağla (but someone else could also have seen her.))
- Çağla’yı gördü Ekin. (Ekin saw Çağla (but he could have seen someone else.)

Turkish Language

- Turkish is an agglutinative language
- Morphemes attaches to a root word like "beads-on-a-string."
- yap+abil+ecek+se+k → if we will be able to do (it)

Problem Definition

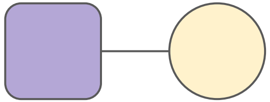
- Morphological Parsing: Dividing a word into its morphemes

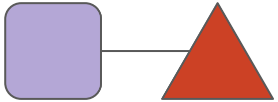
$$\underbrace{\underbrace{\underbrace{\underbrace{\underbrace{\underbrace{\text{ruhsat}}_{NOUN}+\text{lan}}_{VERB_{Acquire}}+\text{dır}}_{VERB_{Causative}}+\text{ıl}+\text{ama}}_{VERB_{Passive}}+\text{ma}+\text{sı}+\text{nda}}_{NOUN_{Infinitive}}+\text{ki}}_{ADJ_{Rel}}$$

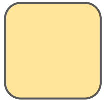
Turkish Language

- Root affects morpheme
  - Defter + ler
  - Kitap + lar
- Morpheme affects root
  - Tabak
  - Tabağ + ın

Problem Definition

- ev + in (your) house 

- ev + in of the house 

- evin wheat grain 

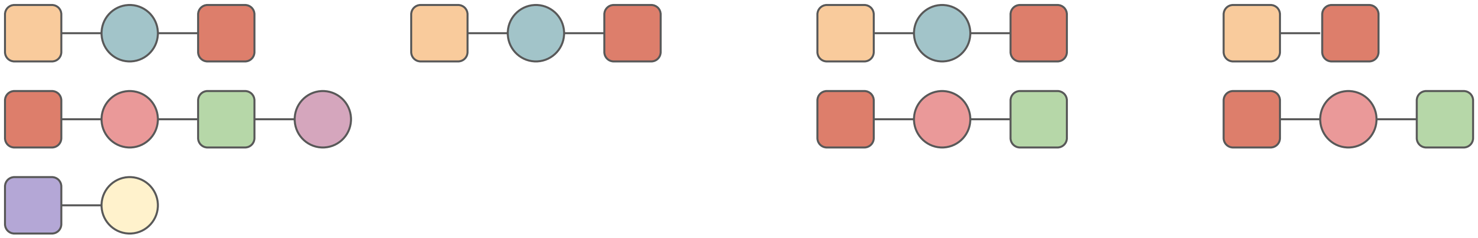
Problem Definition

- Morphological disambiguation is the task of determining the contextually correct morphological parses of tokens in a sentence.
- Ambiguity quite common: Each word has 2 different morphological interpretation on average.

Problem Definition

Sentence : $\text{Word}_1 + \text{Word}_2 + \text{Word}_3 + \text{Word}_4$

Parse :



3 x 1 x 2 x 2

12 Possible Candidate Parses, which one is correct?

Approaches

- Rule based methods
- Statistical methods
    - Hidden Markov Model (HMM)
    - Averaged Perceptron Algorithm

Rule Based Methods

- Manually written constraints
- Need an expert
- No need for data

Oflazer K, Kuruöz İ (1994) Tagging and morphological disambiguation of Turkish text. In: Proceedings of ANLP, Stuttgart, pp 144–149

Hidden Markov Model

- Generative Model

$$\hat{T} = \underset{T}{\operatorname{argmax}}\, P(T|W) = \underset{T}{\operatorname{argmax}}\, P(T) \times P(W|T)$$

- Markov Assumption:

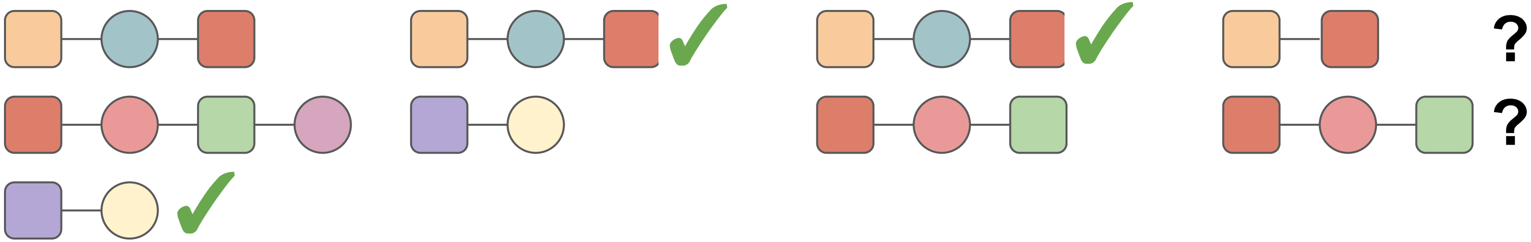
$$\hat{T} = \underset{T}{\operatorname{argmax}} \prod_{i=1}^{n} P(t_i|t_{i-2}, t_{i-1}) \times P(w_i|t_i)$$

Hakkani-Tür DZ, Oflazer K, Tür G (2002) Statistical morphological disambiguation for agglutinative languages. Comput Hum 36(4):381–410

Hidden Markov Model

Sentence : $\text{Word}_1 + \text{Word}_2 + \text{Word}_3 + \text{Word}_4$

Parse :

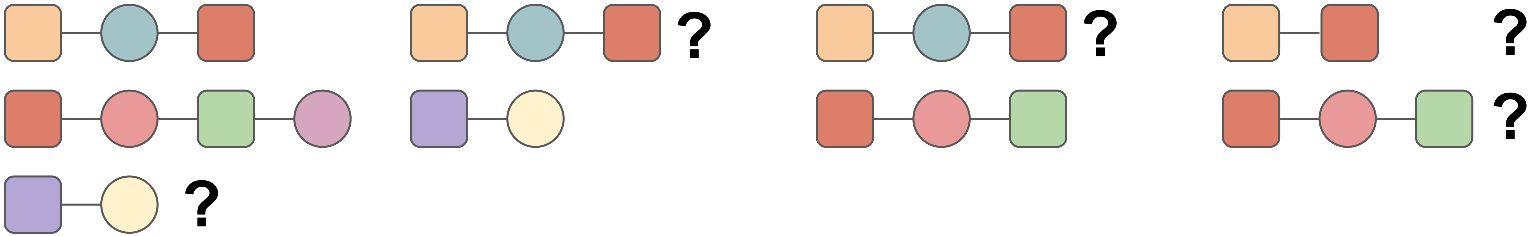


The correct parse of word 4 depends on correct parse of word 3 and word 2

Hidden Markov Model

Sentence : $\text{Word}_1 + \text{Word}_2 + \text{Word}_3 + \text{Word}_4$

Parse :



- Correct states are hidden
- We have to guess them from observations
- Observations = Candidate Parses

Averaged Perceptron Algorithm

- Neural network with one layer
- Handcrafted features

$$P(T|W) = \frac{e^{\mathbf{\Phi}(W,T)\cdot\overline{\alpha}}}{\sum_{T'\in\mathbf{GEN}(W)} e^{\mathbf{\Phi}(W,T')\cdot\overline{\alpha}}}.$$

Sak H, Güngör T, Saraçlar M (2011) Resources for Turkish morphological processing. LangResour Eval 45(2):249–26

Averaged Perceptron Algorithm

Gloss	Feature
Morphological parse trigram	(1) $t_{i-2}t_{i-1}t_i$
Morphological parse bigram	(2) $t_{i-2}t_i$ and (3) $t_{i-1}t_i$
Morphological parse unigram	(4) t_i
Morpheme tag with previous tag	(5) $t_{i-1}m_i$
Morpheme tag with second to previous tag	(6) $t_{i-2}m_i$
Root trigram	(7) $r_{i-2}r_{i-1}r_i$
Root bigram	(8) $r_{i-2}r_i$ and (9) $r_{i-1}r_i$
Root unigram	(10) r_i
Morpheme tag trigram	(11) $m_{i-2}m_{i-1}m_i$
Morpheme tag bigram	(12) $m_{i-2}m_i$ and (13) $m_{i-1}m_i$
Morpheme tag unigram	(14) m_i
Individual morpheme tags	(15) $m_{i,j}$ for $j = 1 \ldots n_i$
Individual morpheme tags with position	(16) $jm_{i,j}$ for $j = 1 \ldots n_i$
Number of morpheme tags	(17) n_i

Datasets and Results

- METU dataset: 5635 sentences, 56 K words
- ITU dataset: 300 sentences, 3.7 words
- Training set: 650 K unambiguous tokens & 32 K disambiguated tokens

Disambiguator	Manual test	METU-Sabancı Treebank	ITU validation set
Hakkani-Tür et al. (2002)	95.48%	–	–
Yuret and Türe (2006)	95.82%	78.76%	87.67%
Sak et al. (2011)	96.45%	78.23%	87.84%

References

- Hakkani-Tür, Dilek Zeynep, et al. 2018, “Morphological Disambiguation for Turkish.” pp 53-67 in: Turkish Natural Language Processing. Springer, Cham.
- Oflazer K, Kuruöz İ (1994) Tagging and morphological disambiguation of Turkish text. In: Proceedings of ANLP, Stuttgart, pp 144–149
- Hakkani-Tür DZ, Oflazer K, Tür G (2002) Statistical morphological disambiguation for agglutinative languages. Comput Hum 36(4):381–410
- Sak H, Güngör T, Saraçlar M (2011) Resources for Turkish morphological processing. LangResour Eval 45(2):249–26
- Yuret D, Türe F (2006) Learning morphological disambiguation rules for Turkish. In: Proceedings of NAACL-HLT, New York, NY, pp 328–334