

A Novel Approach to Morphological Disambiguation for Turkish

Onur Görgün, Olcay Taner Yıldız

Department of Computer Engineering
Işık University
Istanbul, Turkey

ISCIS 2011

Outline

- 1 Morphology
- 2 Morphological Disambiguation
- 3 Proposed Approach
- 4 Conclusions

Morphology

• Turkish

- Agglutinative, inflectionally rich language.
- Syntactical information is expressed by suffixation (morphology).

Gelmeyeceğim. (**I will not come**)

gel + me + (y)ecek + im
| | | |
come not will I

Morphological Parse

gel + **Verb** + **Pos** + **^DB** + **Adj** + **FutPart** + **Pnon** (*Surface Form*)

Root Morph. Derivational
 Feature Boundary
 Inflectional Group



root + **IG₁** + **^DB** + **IG₂** + ... + **IG_n** (*General Form*)

- Turkish → 126 morph. feature, theoretically unlimited number of tags.
- Generated by morphological parsers → [Oflazer94, Sak, et al.08].

Turkish Morphology

otur → sit

oturt → make X sit

oturttur → have Y make X sit

oturtturt → have Z have Y make X sit

oturtturtur → have T have Z have Y make X sit

oturtturturt → have W ...

- Theoretically unlimited # of word forms → unlimited # of possible tags.
- Data sparseness!

Morphological Disambiguation

- Selection of accurate morphological parse among all possible parses
- Agglutinative languages → Turkish
 - \simeq 40% ambiguity rate, \simeq 4 parses for ambiguous tokens
- Parses for word "gelecek":

gelecek + Noun + A3sg + Pnon + Nom

gelecek + Adj

gel + Verb + Pos + Fut + A3sg

gel + Verb + Pos + ^DB + Adj + FutPart + Pnon

Related Work

- Two main approaches:
 - Rule-based approaches
 - Greedy Prepend Algorithm (GPA) [Yüret, et al.06]
 - Statistical approaches
 - Baseline Trigram-based Model [Hakkani-Tür,et al.02]

Problem Definition

- Find the correct morphological parse from N possible parses excluding the root word.
- Each distinct parse corresponds to one class
- Parses for word "gelecek" :

gelecek + Noun + A3sg + Pnon + Nom

gelecek + Adj

gel + Verb + Pos + Fut + A3sg

gel + Verb + Pos + ^DB + Adj + FutPart + Pnon



Class 1: Noun + A3sg + Pnon + Nom

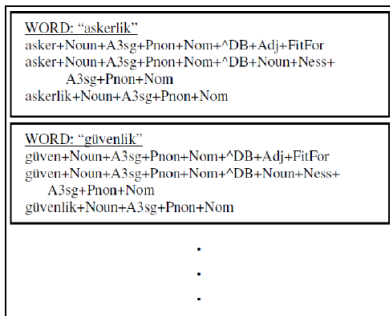
Class 2: Adj

Class 3: Verb + Pos + Fut + A3sg

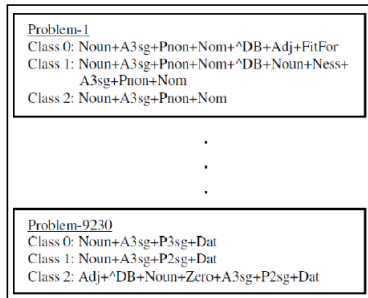
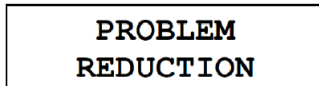
Class 4: Verb + Pos + ^DB + Adj + FutPart + Pnon

- Total of 399223 problems for 1M size ambiguous data

Problem Reduction



Initial problem set of 399223 problems



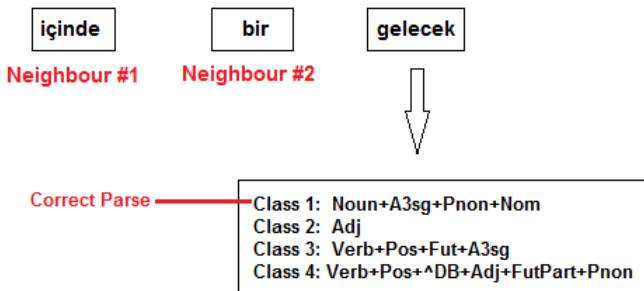
Problem Reduction

- Distribution of problems with respect to the # of instances

Number of Instances	Number of Problems
1 – 10	7213
11 – 100	1617
101 – 1000	427
1001 – 10000	60
10001 – 100000	3

Training Data

- 3-word window → ambiguous word and 2 preceding neighbours.
- Training instance → existence of morp. features in all parses of neighbour words.
- 126x2 feature + class info.



Experimental Setup-Data

- Training set → 1M tagged disambiguated tokens (including end of sentence, title and document markers). 50673 sentences.
- Test set → 958 tokens, and 42 sentences. 379 tokens are morphologically ambiguous.
- Test set problem definitions are subset of training phase problem set!!

Experimental Results

Method	Acc.(%)	Method	Acc.(%)
NaiveBayes	93.83	Logistic Regression	94.67
Conjunctive Rule	66.25	SVM	94.98
k -NN($k=10$)	95.40	J48 Tree	95.61
LWL	94.67	Baseline Trigram-Based Model	95.48
J48 Tree(no pruning)	95.09	Greedy Prepend Algorithm	95.82
KStar	94.36	Perceptron(23 Features)	96.28
NNge	90.49		

Conclusions

- Define MD problem as multiple classification problems.
- J48 is the best performer among 10 classifiers → slightly better than "Baseline Trigram-Based Model".
- Performance improvement → expanding the feature set or linear/non-linear feature extraction mechanism