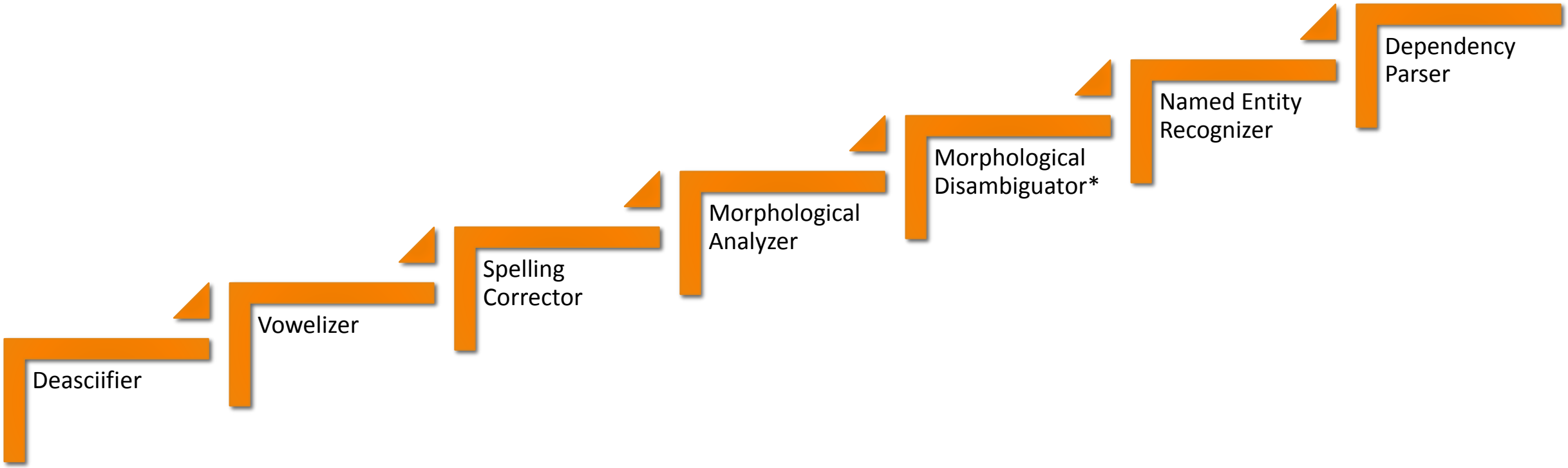
A review for

ITU NLP Web Service*

EMRAH BUDUR

* URL: http://tools.nlp.itu.edu.tr

7 main modules



* Publication in progress

URL: http://tools.nlp.itu.edu.tr

Deasciifier

Problem

Case 1

- kagit → kağıt

Case 2

- aci → açı? acı?

Case 3

- olumsuz → ölümsüz

URL: http://tools.nlp.itu.edu.tr

Deasciifier

Two stage method

URL: http://tools.nlp.itu.edu.tr

Deasciifier

Discriminative Sequence Classifier

Rüyamda evde olduğunu gördüm

Curr. Letter	Neig. Word(+1)	Curr. Word	Neig. Ch(-2)	Neig. Ch(-1)	Neig. Ch(+1)	Neig. Ch(+2)	Class Label
O	GOrdUm	OldUGUnU	-	-	l	d	**ö**
U	GOrdUm	OldUGUnU	l	d	G	U	**ü**
G	GOrdUm	OldUGUnU	d	U	U	n	**ğ**
U	GOrdUm	OldUGUnU	U	G	n	U	**ü**
U	GOrdUm	OldUGUnU	U	n	-	-	**ü**

Table 1: The capital letters denotes the character that needs to be labeled as either the ascii form or non-ascii form of it. Each row denotes a labeling step of the character given in the first column, which was also underlined in the column of *current word*. The expression $(\pm n)$ where $n \in \mathbb{Z}$ denotes the position of the feature relative to the current word or character.

Language Validator

- Number of all possible tokens: 2^5=32
- Valid tokens : **Olduğunu**, Öldüğünü

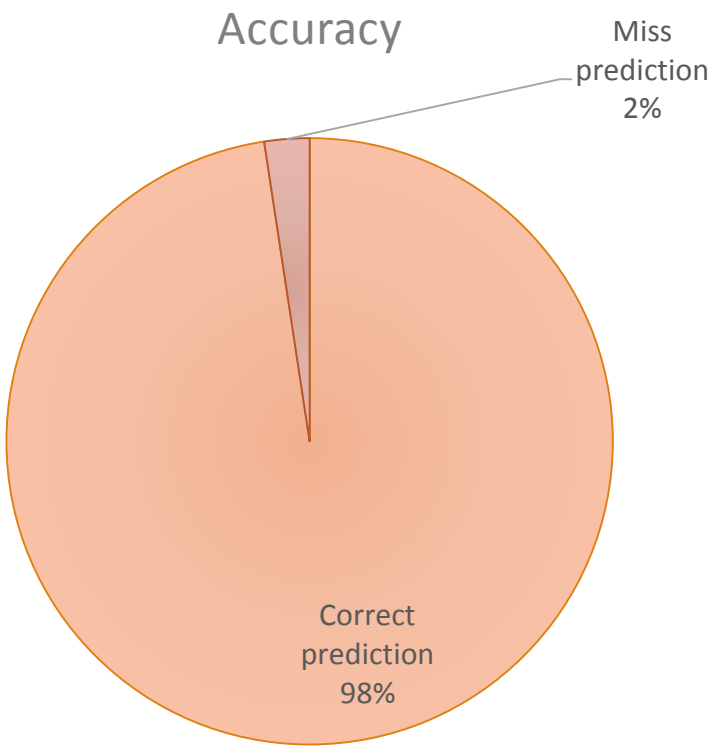
URL: http://tools.nlp.itu.edu.tr

Deasciifier

Result

System	Acc Overall	Acc Amb
Yuret [3]	95.93	91.05
Zemberek [4]	87.71	82.55
Proposed model [2]	**97.06**	**94.70**

Table 2: The comparative results of the proposed model



URL: http://tools.nlp.itu.edu.tr

Vowelizer

Problem

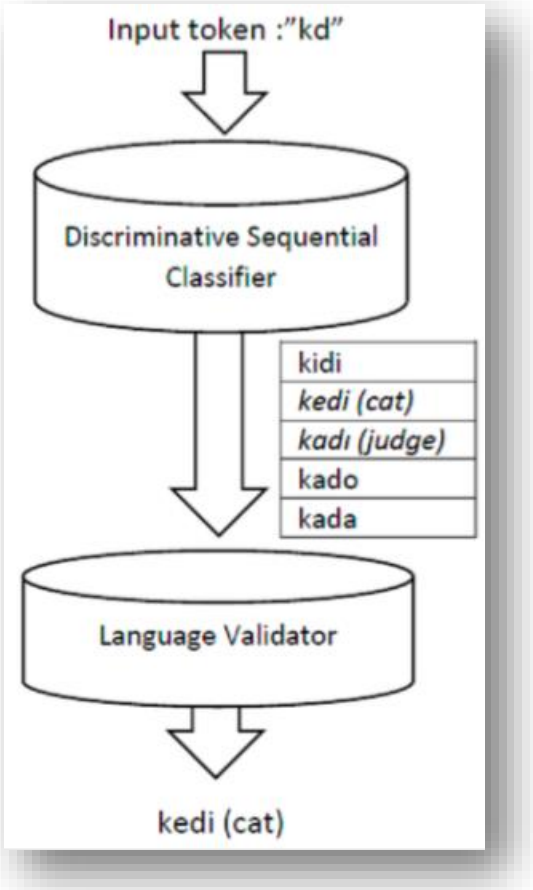
Examples

- slm→ selam
- Mrb → Merhaba

URL: http://tools.nlp.itu.edu.tr

Vowelizer

Example



URL: http://tools.nlp.itu.edu.tr

Vowelizer

Discriminative Sequence Classifier

Predicted Position	Curr. Word	Neig. Ch(-3)	Neig. Ch(-2)	Neig. Ch(-1)	Neig. Ch(+1)	Neig. Ch(+2)	Neig. Ch(+3)	Class Label
?s_l_m_	slm	-	-	-	s	l	m	-
_s?l_m_	slm	-	-	s	l	m	-	**e**
_s_l?m_	slm	-	s	l	m	-	-	**a**
_s_l_m?	slm	s	l	m	-	-	-	-

Table 3: Vowelization steps

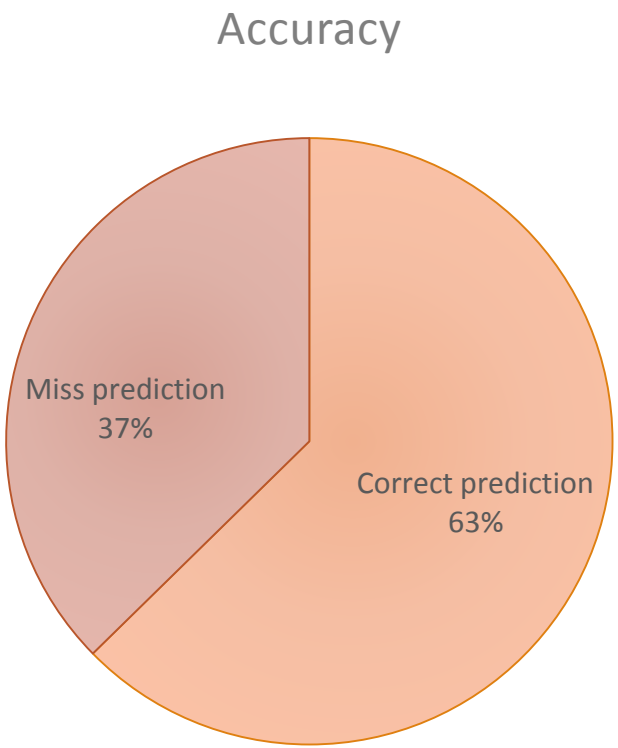
Language Validator

- Number of all possible tokens: 2^4=16
- Valid tokens : **Selam**, Salam, Sulama, Salım, Sılam

URL: http://tools.nlp.itu.edu.tr

Vowelizer

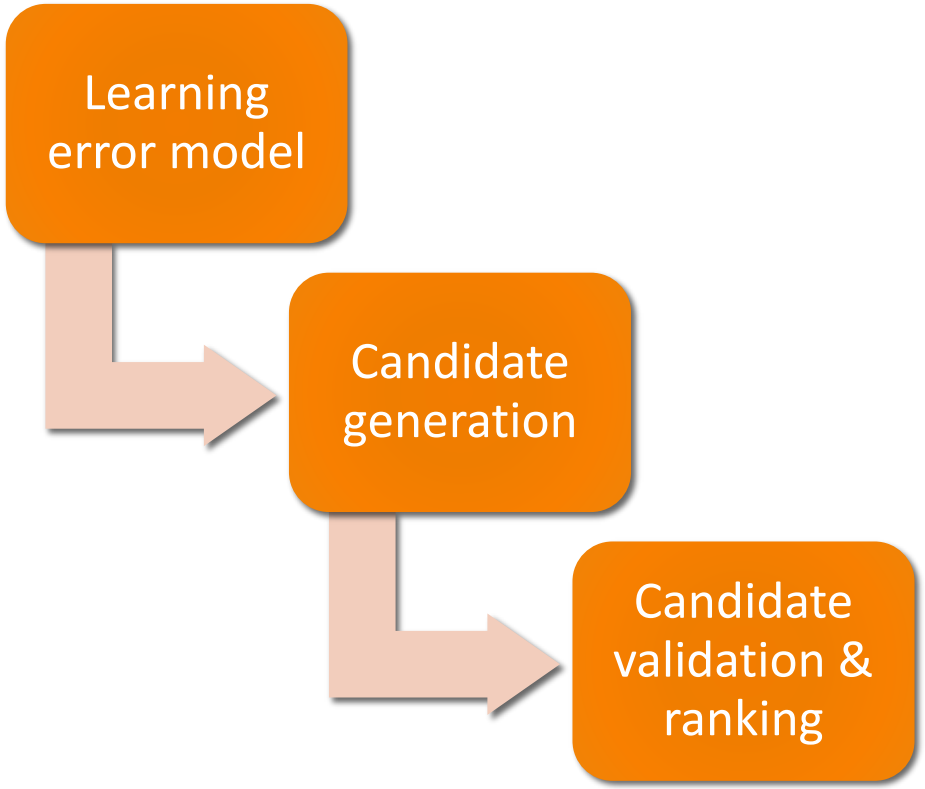
Result



URL: http://tools.nlp.itu.edu.tr

Spelling Corrector

Base model

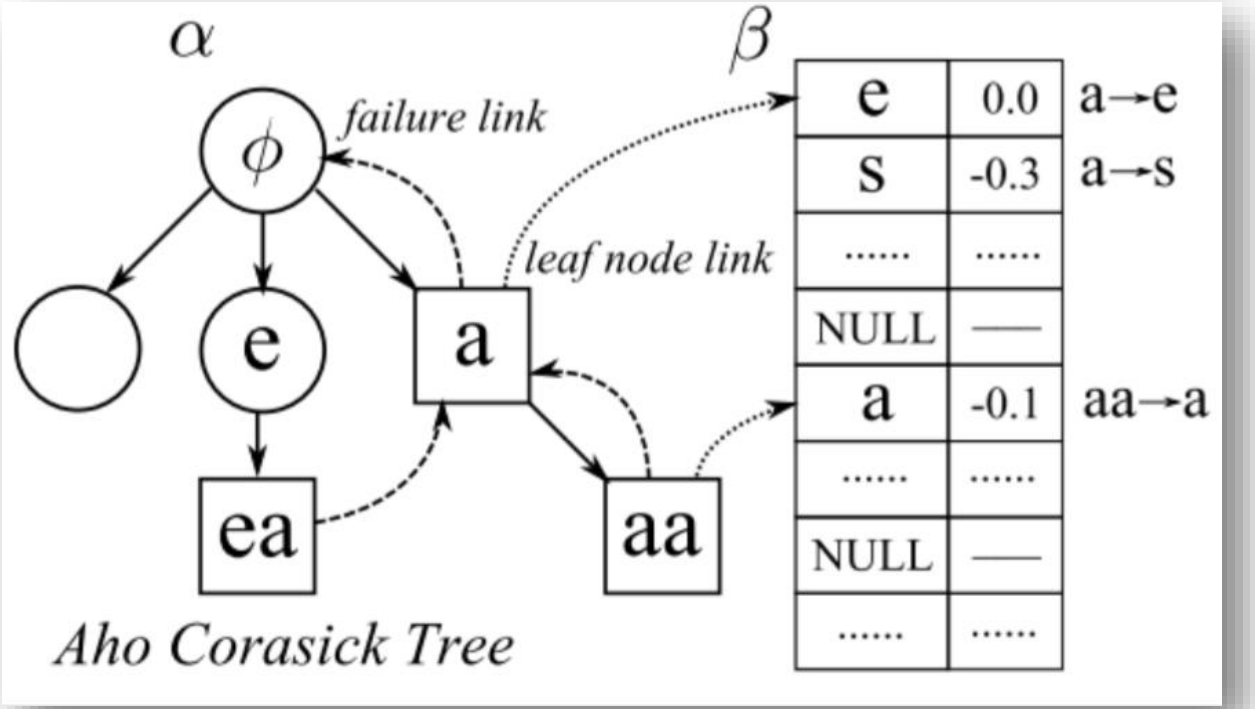


URL: http://tools.nlp.itu.edu.tr

Spelling Corrector

Aho-Corasick

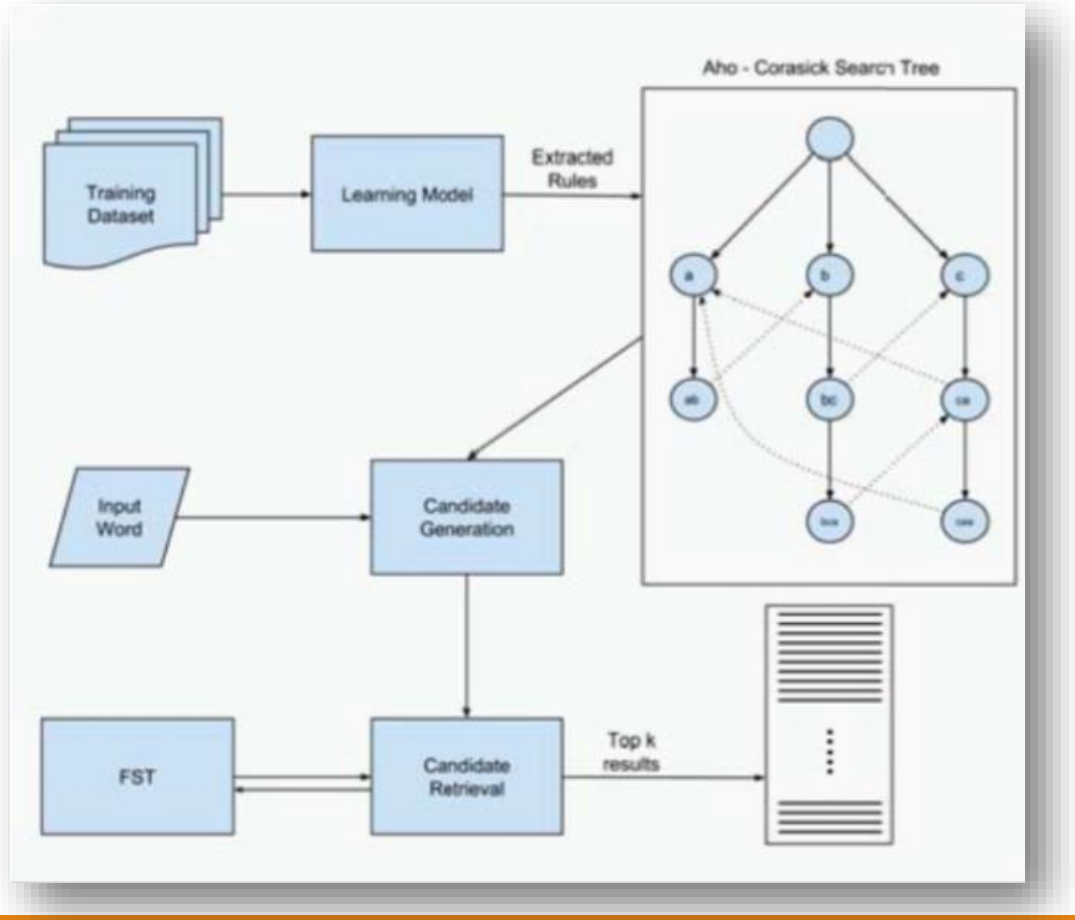
aacoustic → acoustic
~~ayoustic → acoustic~~



URL: http://tools.nlp.itu.edu.tr

Spelling Corrector

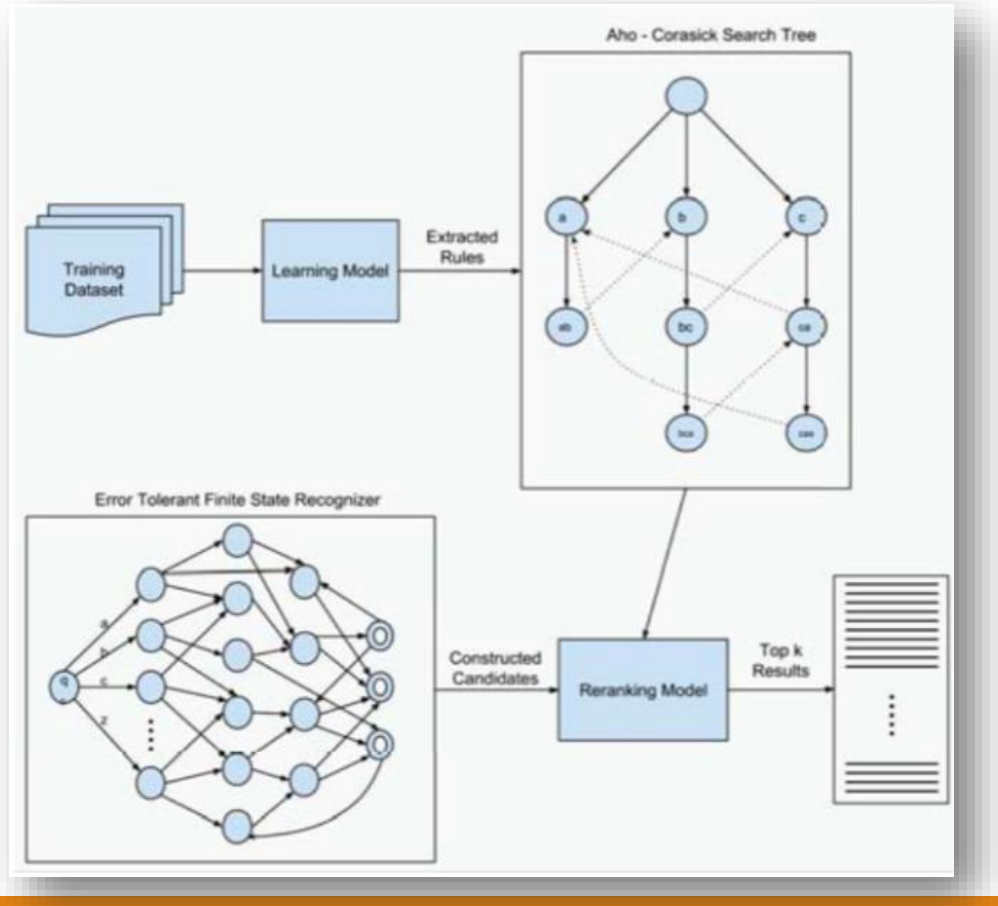
SC #1



URL: http://tools.nlp.itu.edu.tr

Spelling Corrector

SC #2



apelling → spelling

- Cost: -0.1

~~apelling →~~ ~~epelling~~

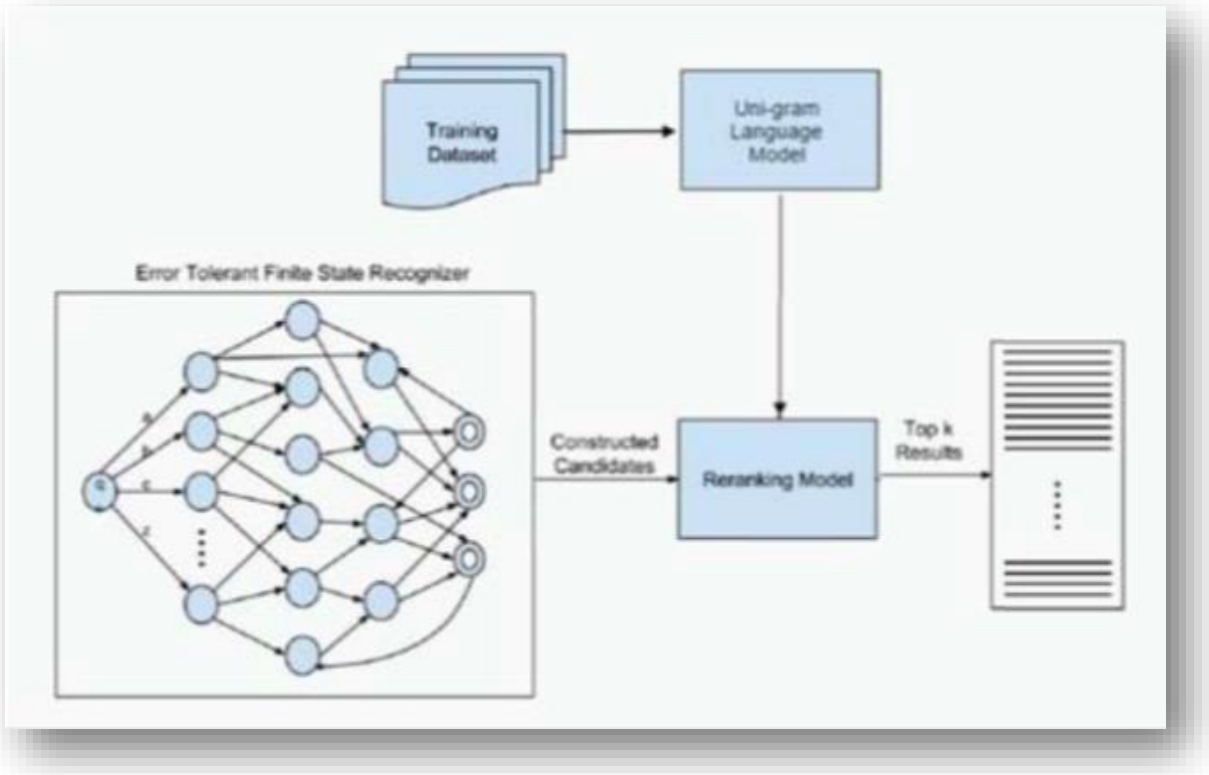
- Cost: 0.0

e	0.0	a→e
s	-0.3	a→s
.....	.....	
NULL	—	
a	-0.1	aa→a
.....	.....	
NULL	—	
.....	.....	

URL: http://tools.nlp.itu.edu.tr

Spelling Corrector

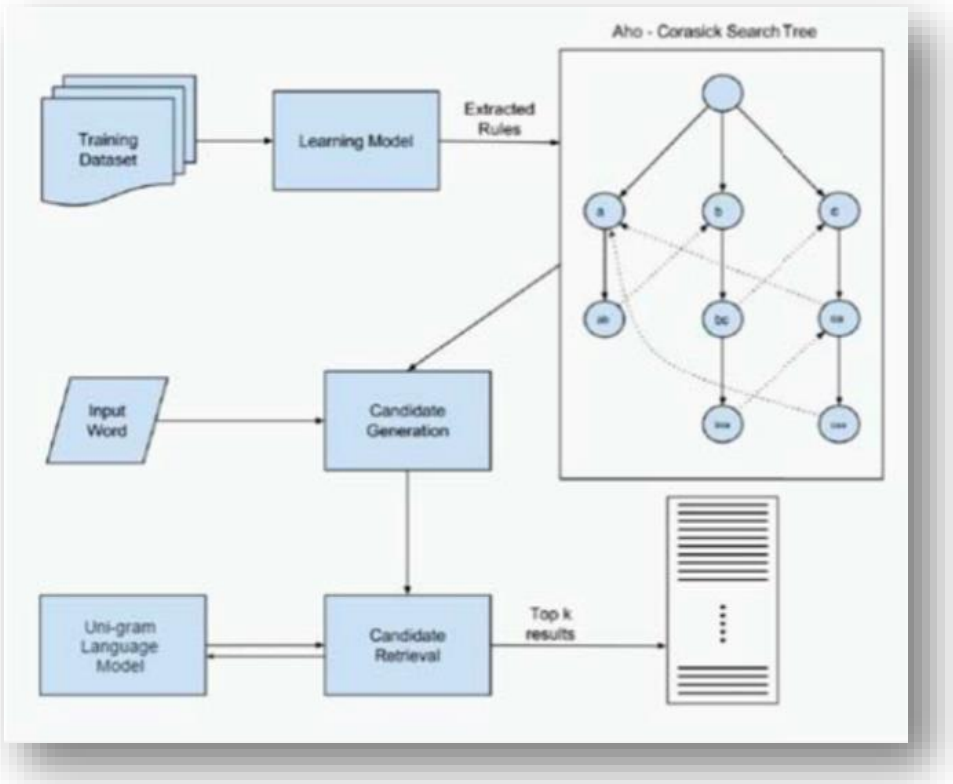
SC #3



URL: http://tools.nlp.itu.edu.tr

Spelling Corrector

SC #4



URL: http://tools.nlp.itu.edu.tr

Spelling Corrector

Results

	Accuracy
ETFSR	49.0%
Zemberek	61.4%
MsWord	66.3%
SC1	68.6%
SC2	67.8%
SC3	78.7%
SC4	80.7%

Table 4: Benchmarking with existing solutions

URL: http://tools.nlp.itu.edu.tr

Morphological Analyzer

Problem

Example: kalem

- kalem → N(kalem) ⇒ pencil
- kale-m → N(kale) + 1PS-POSS(m) ⇒ my castle

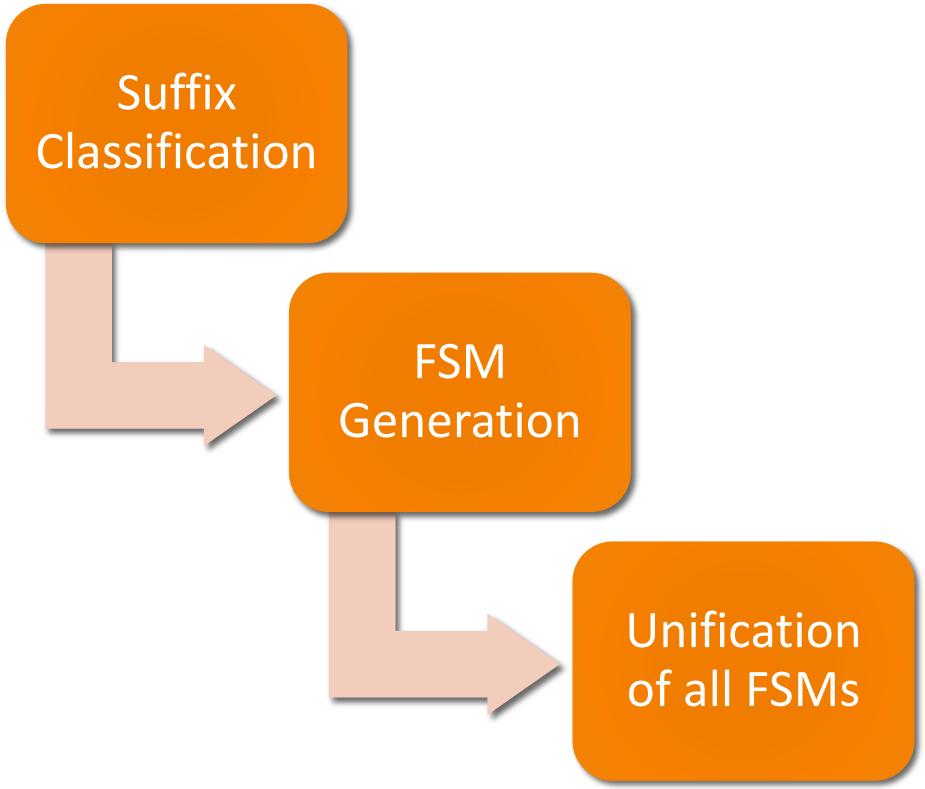
Example (kalemler)

- kalem → N(kalem) + PLUR(ler) ⇒ pencils
- ~~kale-m → N(kale) + 1PS-POSS(m) + PLUR(ler) ⇒ ?my castles~~

URL: http://tools.nlp.itu.edu.tr

Morphological Analyzer

Base model



URL: http://tools.nlp.itu.edu.tr

Morphological Analyzer

Step 1: Suffix Classification

Class #	Class	Type
1	Nominal verb suffixes	Inflectional
2	Derivational suffixes	Derivational
3	Noun suffixes	Inflectional
4	Tense & person verb suffixes	Inflectional
5	Verb suffixes	Inflectional

URL: http://tools.nlp.itu.edu.tr

Morphological Analyzer

Step 2: FSM Generation

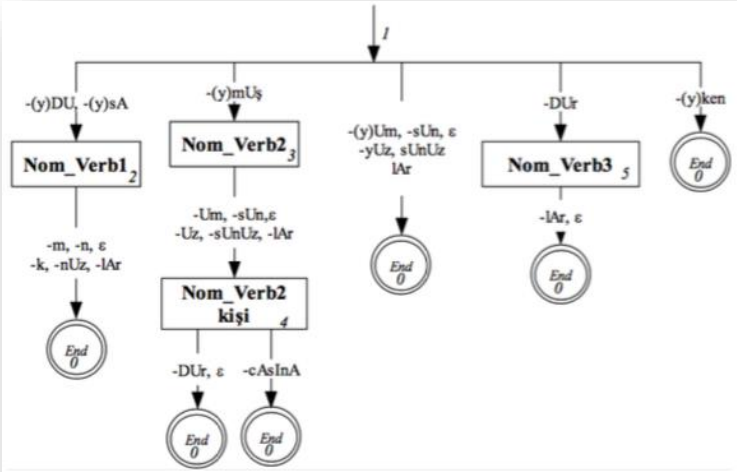
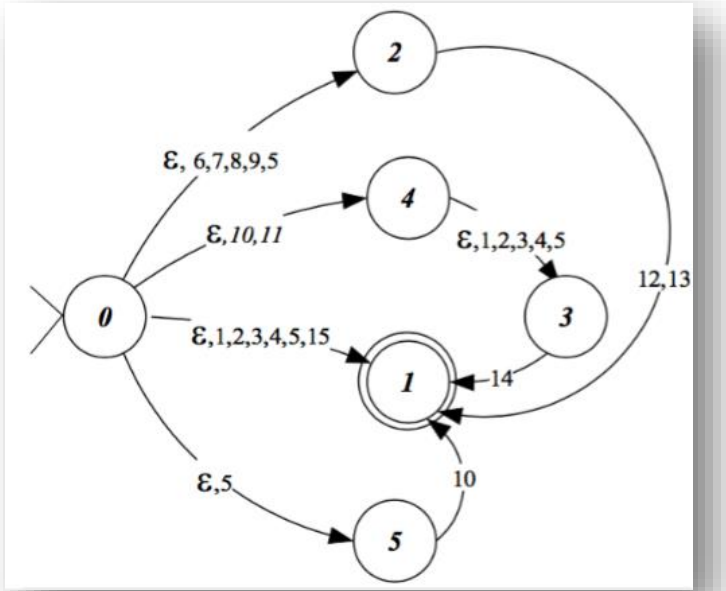
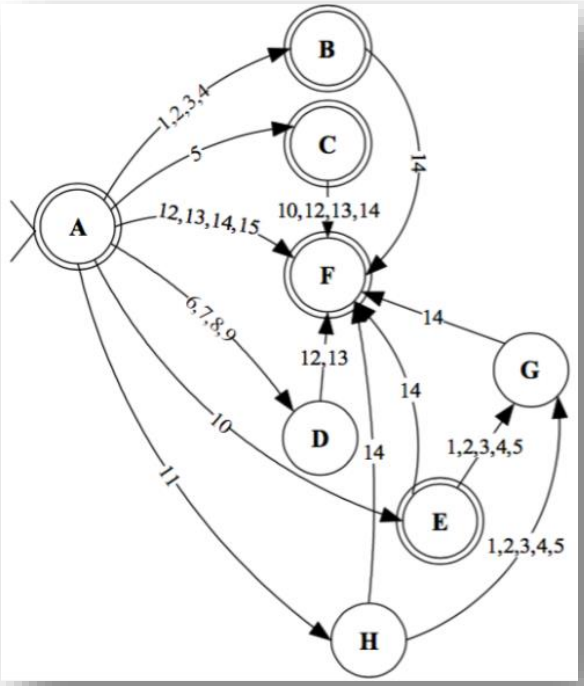


Figure 7: Nominal Verb Suffix - Left to right FSM [6]

#	Suffix	#	Suffix	#	Suffix
1	–(y)Um	6	–m	11	–cAsInA
2	–sUn	7	–n	12	–(y)DU
3	–(y)Uz	8	–k	13	–(y)sA
4	–sUnUz	9	–nUz	14	–(y)mUş
5	–lAr	10	–DUr	15	–(y)ken

Table 6: Nominal Verb Suffixes





URL: http://tools.nlp.itu.edu.tr

Morphological Analyzer

Step 3: Unification of all FSMs

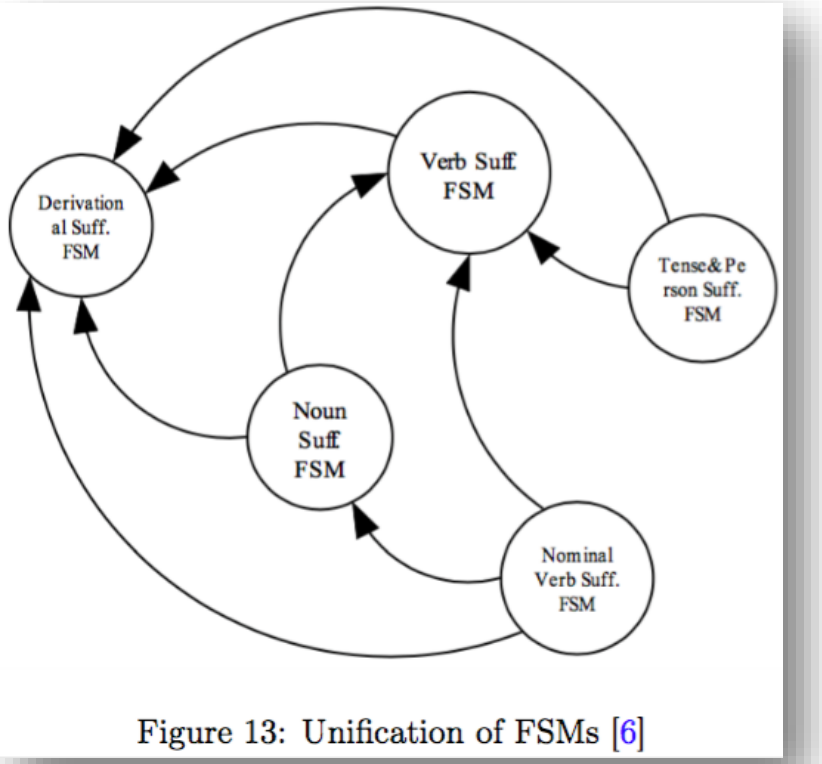
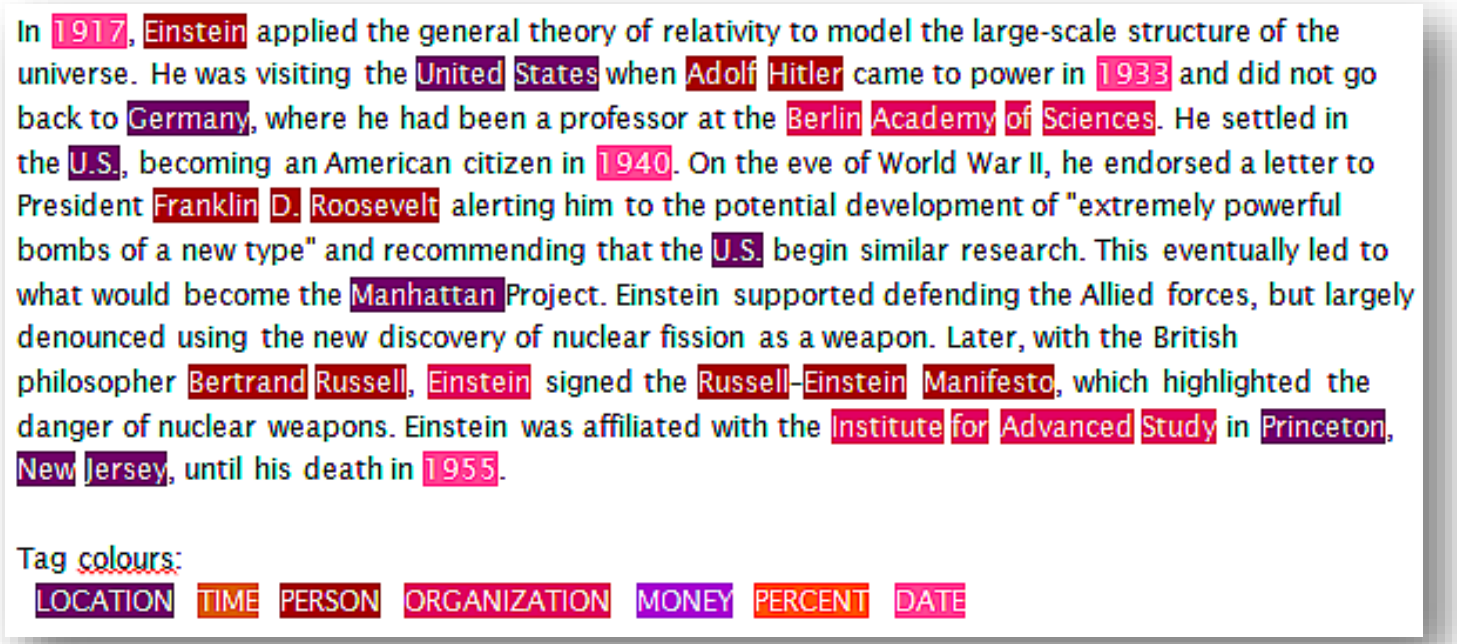


Figure 13: Unification of FSMs [6]

URL: http://tools.nlp.itu.edu.tr

Named Entity Recognition

Problem

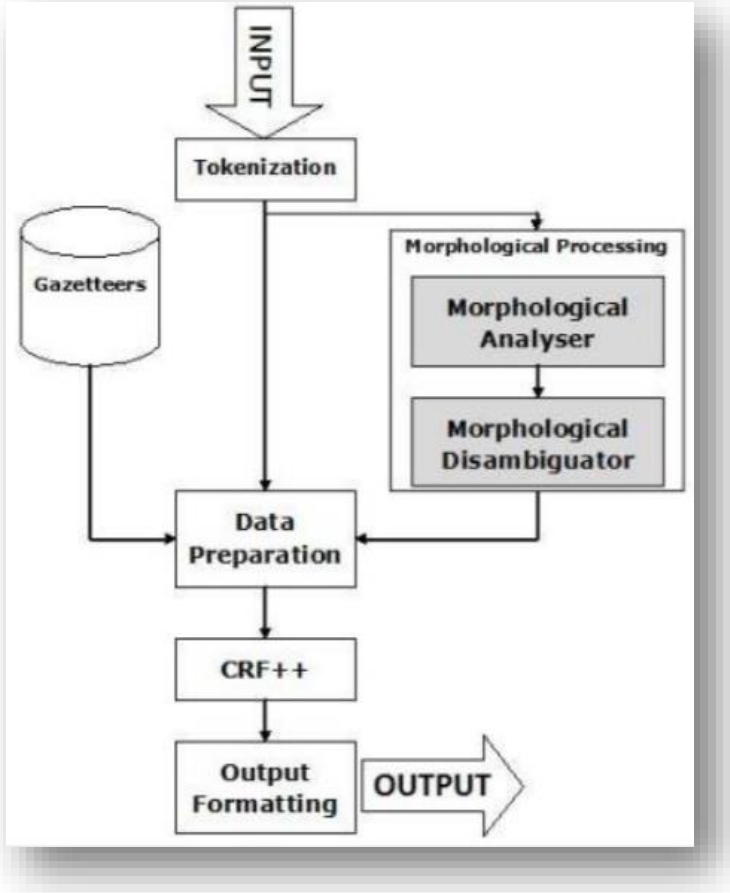


Adopted from http://europeana-newspapers.eu

URL: http://tools.nlp.itu.edu.tr

Named Entity Recognition

Method



URL: http://tools.nlp.itu.edu.tr

Named Entity Recognition

Results

Related work	Best Result	Ev.Metr.	Domain
Ozkaya and Diri [11]	84.24	n/a	E-mail texts
Kucuk and Yazici [12]	90.13	OTHER	General news
Tur et al. [13]	91.56	MUC	General news
Bayraktar and Temizel [14]	81.97	MUC	Financial Texts
Seker et al. [15]	94.59	MUC	General news
Tatar and Cicekli [16]	91.08	CoNLL	Terrorism news
Yeniterzi [17]	88.94	CoNLL	General news
Seker et al. [15]	91.94	CoNLL	General news

Table 10: Comparative results of related works on NER

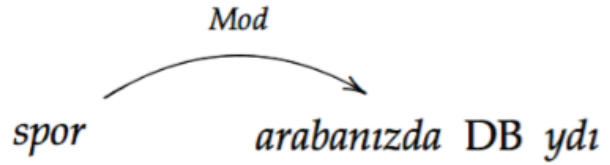
URL: http://tools.nlp.itu.edu.tr

Dependency Parser

Problem

arabanızdaydı
('it was in your car')

arabanızda	DB	*ydı*
araba+Noun+A3sg+P2pl+Loc	DB	+Verb+Zero+Past+A3sg
IG_1		IG_2
'in your car'		'it was'



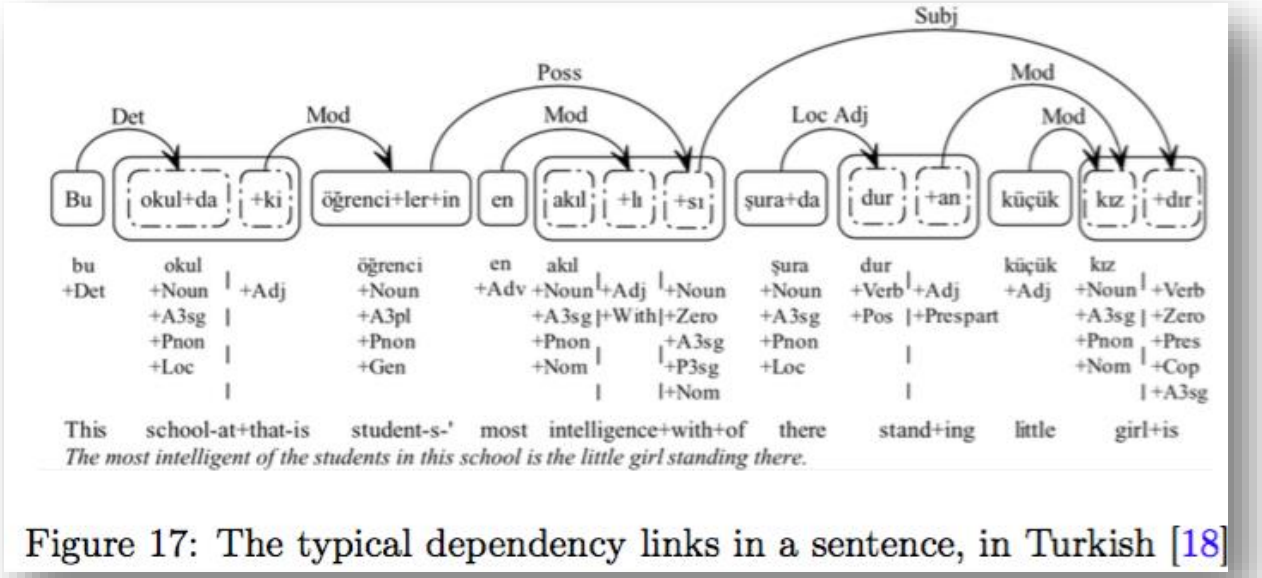


Figure 17: The typical dependency links in a sentence, in Turkish [18]

URL: http://tools.nlp.itu.edu.tr

Dependency Parser

Word-based vs Inflectional Group (IG) based models

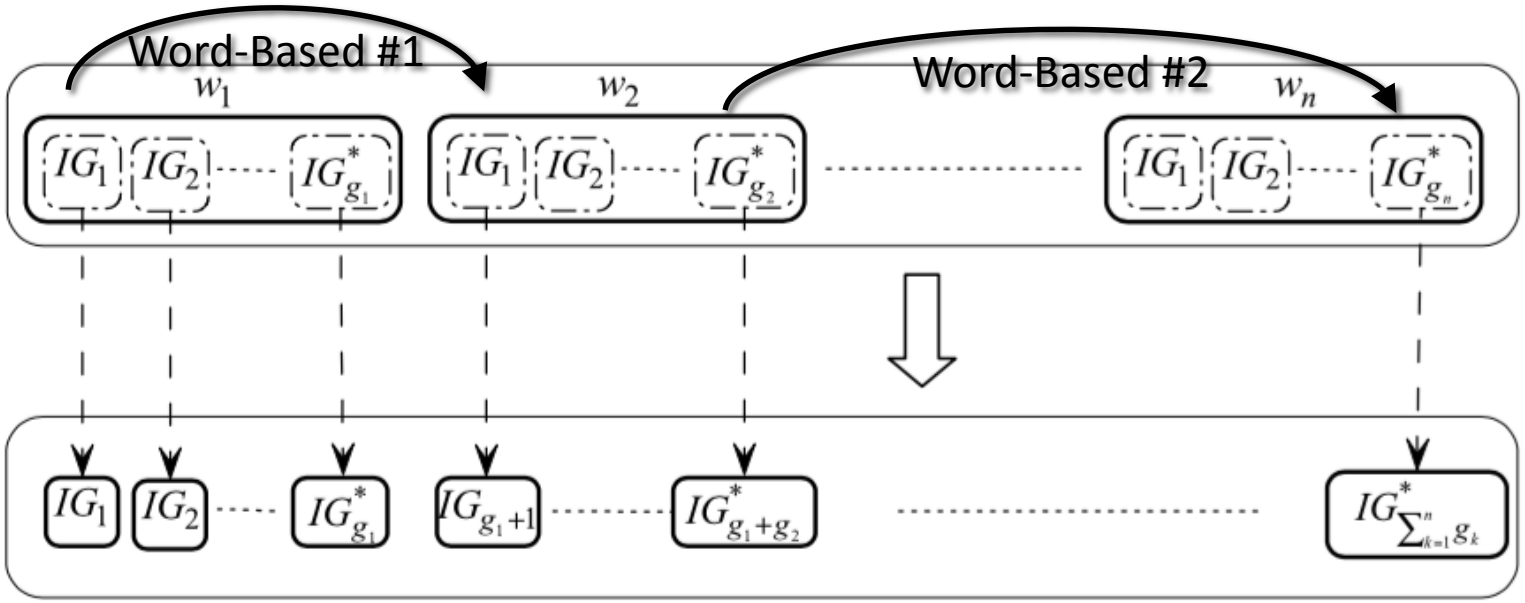


Figure 2
Mapping from word-based to IG-based representation of a sentence.

URL: http://tools.nlp.itu.edu.tr

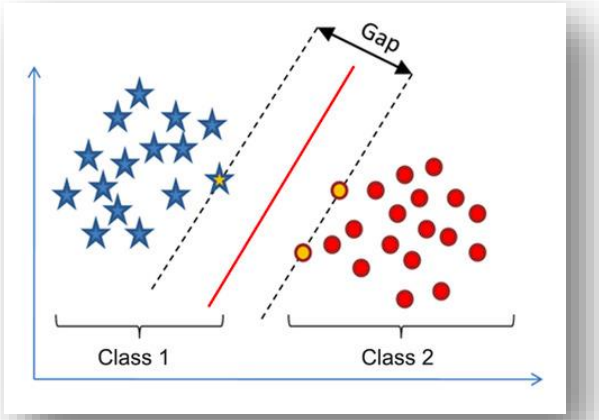
Dependency Parser

Probabilistic Parser

$$T^* = \arg\max_T P(T|S) = \arg\max_T \prod_{i=1}^{n-1} P(dep(u_i, u_{H(i)})|S)$$

Classification-based Parser

- Features
  - Root form
  - POS tag
  - Inflectional features



Adopted from http://diggdata.in

URL: http://tools.nlp.itu.edu.tr

Dependency Parser

Results

Parsing Model (baselines)	ASU	WWU
Attach-to-next (first IG)	56.0	63.3
Attach-to-next (last IG)	54.1	63.3
Rule-based	70.5	79.3

Table 11: The experimentation results of baseline parsers

Parsing Model (parameters)	AS_U	WW_U
Word-based model #1	68.1±0.4	77.1±0.7
Word-based model #2	68.3±0.3	77.6±0.5
IG-based model	72.1±0.3	79.0±0.7

Table 12: The experimentation results of probabilistic parser

Parsing Model (baselines)	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

Table 13: Experimental results of classifier based model

URL: http://tools.nlp.itu.edu.tr

Thanks

URL: http://tools.nlp.itu.edu.tr