FONDAZIONE
BRUNO KESSLER

Morphological Pre-processing for Turkish to English Statistical Machine Translation

Arianna Bisazza, Marcello Federico
FBK - Ricerca Scientifica e Tecnologica, Trento, Italy

Tokyo, Dec 1-2, 2009

FONDAZIONE
BRUNO KESSLER

Outline

- Turkish & SMT
- Morphological Segmentation
  - Preprocessing chain
  - Segmentation rules
- Lexical Approximation
- Experiments
- Future Work & Conclusions

Outline

- **Turkish & SMT**
- Morphological Segmentation
  - Preprocessing chain
  - Segmentation rules
- Lexical Approximation
- Experiments
- Future Work & Conclusions

FONDAZIONE
BRUNO KESSLER

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

FONDAZIONE BRUNO KESSLER

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

- **Agglutination**

  → large vocabulary, built by a wide range of suffix combinations

oda	[room]	= 'room'
odam	[room-my]	= 'my room'
odamda	[room-my-in]	= 'in my room'
odamdaydı	[room-my-in-was]	= 'was in my room'
odamdaydım	[room-my-in-was-I]	= 'I was in my room'

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

- **Agglutination**
  → large vocabulary, built by a wide range of suffix combinations

oda	[room]	= 'room'
$odam$	[room-my]	= 'my room'
$odamda$	[room-my-in]	= 'in my room'
$odamdaydı$	[room-my-in-was]	= 'was in my room'
$odamdaydım$	[room-my-in-was-I]	= 'I was in my room'

Some statistics on IWSLT09 training corpus :

	Tokens	Dict.size
TR	139,514	17,619
EN	182,627	8,345

OOV (devset2): 6.16%

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

- **Agglutination**
  → large vocabulary, built by a wide range of suffix combinations
- **Vowel harmony** and other phoneme alternation phenomena
  → systematic stem and suffix *allomorphy*

Ex. the suffix -*(I)m* = 'my':

saç+(I)m	→	*saç**ım***	'my hair'
el+(I)m	→	*el**im***	'my hand'
kol+(I)m	→	*kol**um***	'my arm'
göz+(I)m	→	*göz**üm***	'my eye'
kafa+(I)m	→	*kafa**m***	'my head'

FONDAZIONE BRUNO KESSLER

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

- **Agglutination**
  → large vocabulary, built by a wide range of suffix combinations
- **Vowel harmony** and other phoneme alternation phenomena
  → systematic stem and suffix *allomorphy*

Ex. the suffix -*(I)m* = 'my':

saç+(I)m	→	*saç**ım***	'my hair'
el+(I)m	→	*el**im***	'my hand'
kol+(I)m	→	*kol**um***	'my arm'
göz+(I)m	→	*göz**üm***	'my eye'
kafa+(I)m	→	*kafa**m***	'my head'

If splitted from words, suffixes undergo data sparseness
→ need to use a notation that factorizes

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

- **Agglutination**
  → large vocabulary, built by a wide range of suffix combinations
- **Vowel harmony** and other phoneme alternation phenomena
  → systematic stem and suffix *allomorphy*
- **Word order**
  → complex, long-span reorderings between TR and EN

Banyolu	*iki*	*kişilik*	*bir*	*oda*	*istiyorum.*
[bath-with]	[two]	[people-for]	[a]	[room]	[want-I]

‘I’d like a twin room with a bath please.’

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

- **Agglutination**
  → large vocabulary, built by a wide range of suffix combinations
- **Vowel harmony** and other phoneme alternation phenomena
  → systematic stem and suffix *allomorphy*
- **Word order**
  → complex, long-span reorderings between TR and EN

Importance of specific linguistic preprocessing:

FONDAZIONE BRUNO KESSLER

Turkish & SMT

Several linguistic features of Turkish can negatively affect an SMT system:

- **Agglutination**
  → large vocabulary, built by a wide range of suffix combinations
- **Vowel harmony** and other phoneme alternation phenomena
  → systematic stem and suffix *allomorphy*
- **Word order**
  → complex, long-span reorderings between TR and EN

Importance of specific linguistic preprocessing:

→ *reduction of data sparseness (dict. size from 17.6K to 10.4K)*

→ *decrease of OOV rate by more than half*

→ *improvement of 5 points BLEU*

Outline

- Turkish & SMT
- **Morphological Segmentation**
  - Preprocessing chain
  - Segmentation rules
- Lexical Approximation
- Experiments
- Future Work & Conclusions

FONDAZIONE
BRUNO KESSLER

Morphological Segmentation

Idea: selectively splitting or removing suffixes from the words

FONDAZIONE BRUNO KESSLER

Morphological Segmentation

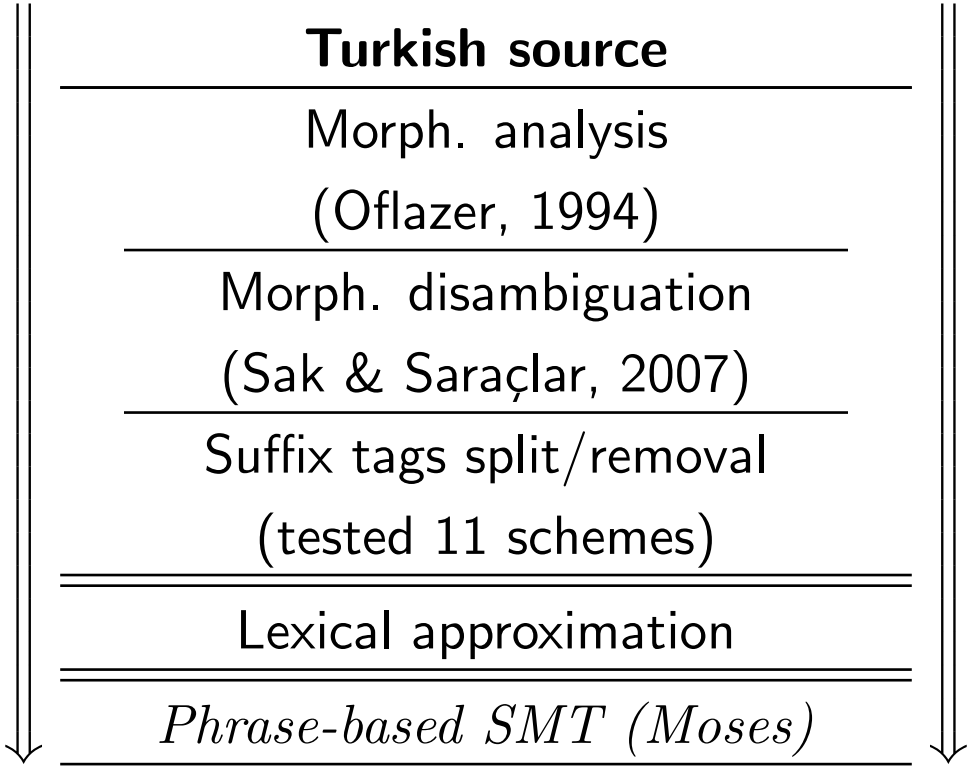
Idea: selectively splitting or removing suffixes from the words

Already explored by:

- Habash & Sadat, 2006 [1] on an Arabic-English task
  - similar method: comparison of segmentation schemes
  - different language: Arabic affixation less rich
- Oflazer & Durgar El-Kahlout, 2007 [2] on an English-Turkish task
  - similar preprocessing chain
  - translating *into* a morphologically rich language

FONDAZIONE BRUNO KESSLER

Preprocessing chain



FONDAZIONE BRUNO KESSLER

Preprocessing chain

1. **Morphological analysis** (Oflazer, 1994 [3])

'Are there any tours of famous stars' homes?'

Ünlü yıldızların <u>evine</u> turlar var mı ?

`ev+Noun+A3sg+P2sg+Dat`	[to your house]
`ev+Noun+A3sg+P3sg+Dat`	[to his/her/its house]
`evin+Noun+A3sg+Pnon+Dat`	[to the kernel]

FONDAZIONE BRUNO KESSLER

Preprocessing chain

1. **Morphological analysis**
2. **Morphological disambiguation in context** (Sak & Saraçlar, 2007 [4])

'Are there any tours of famous stars' homes?'

Ünlü yıldızların evine turlar var mı ?

	`ev+Noun+A3sg+P2sg+Dat`	[to your house]
`->`	`ev+Noun+A3sg+P3sg+Dat`	**[to his/her/its house]**
	`evin+Noun+A3sg+Pnon+Dat`	[to the kernel]

FONDAZIONE BRUNO KESSLER

Preprocessing chain

1. **Morphological analysis**
2. **Morphological disambiguation in context** (Sak & Saraçlar, 2007 [4])

'Are there any tours of famous stars' homes?'
Ünlü yıldızların evine turlar var mı ?

	`ev+Noun+A3sg+P2sg+Dat`	[to your house]
`->`	`ev+Noun+A3sg+P3sg+Dat`	**[to his/her/its house]**
	`evin+Noun+A3sg+Pnon+Dat`	[to the kernel]

Note: some tags encode implicit features (i.e. with no surface form)

We use feature tags to:
→ abstract from suffix allomorphy
→ deal with non-ambiguous symbols
→ make more readable rules

FONDAZIONE BRUNO KESSLER

Preprocessing chain

1. **Morphological analysis**
2. **Morphological disambiguation in context**
3. **Rules for splitting/removal of suffix tags**
   - rules based on feature tags $\longrightarrow$ simple regular expressions
   - 11 segmentation schemes developed and tested
   - mainly focus on nominal, but also some verbal inflection

FONDAZIONE
BRUNO KESSLER

Preprocessing chain

1. **Morphological analysis**
2. **Morphological disambiguation in context**
3. **Rules for splitting/removal of suffix tags**

FONDAZIONE
BRUNO KESSLER

Segmentation rules

Idea: Split off tags expected to have English counterpart, remove others.
When decision is not straightforward → experiment

FONDAZIONE BRUNO KESSLER

Segmentation rules

Idea: Split off tags expected to have English counterpart, remove others.
When decision is not straightforward → experiment

- **Nominal case**
  - split off if expected to have an English counterpart:
    - dative (oda/ya) ≈ 'to'
    - ablative (oda/dan) ≈ 'from'
    - locative (oda/da) ≈ 'in'
    - instrumental (oda/yla) ≈ 'with/by'
  - removed otherwise:
    - nominative $(oda\text{-})$
  - doubtful cases:
    - accusative $(oda/yı)$ (≈ 'the') ⇒ removed
    - genitive $(oda/nın)$ (≈ 'of/'s') ⇒ removed

Segmentation rules

Idea: Split off tags expected to have English counterpart, remove others.
When decision is not straightforward $\rightarrow$ experiment

- **Nominal case**
- **Possessive**
  - split off if expected to have an English counterpart:
    - 1st and 2nd sing. $(oda/m,\ oda/n)$ $\approx$ 'my', 'your'
    - 1st, 2nd and 3rd plur. $(oda/mız,\ oda/nız,\ oda/ları)$ $\approx$ 'our', 'your', 'their'
  - removed otherwise:
    - no_possessive $(oda\text{-})$
  - doubtful cases:
    - 3rd sing. $(oda/sı)$ ($\approx$ 'his/her') $\Rightarrow$ removed

FONDAZIONE BRUNO KESSLER

Segmentation rules

Idea: Split off tags expected to have English counterpart, remove others.
When decision is not straightforward → experiment

- **Nominal case**
- **Possessive**
- **Copula 'to be'**
  - always split off. Example:

- *oda temiz/dir*	litt. [room clean-is]	'the room **is** clean'
- *oda temiz/di*	litt. [room clean-was]	'the room **was** clean'

Segmentation rules

Idea: Split off tags expected to have English counterpart, remove others.
When decision is not straightforward → experiment

- **Nominal case**
- **Possessive**
- **Copula 'to be'**
- **Verb person**
  - split off person suffixes from finite verb forms and copula. Example:
    - $gidiyor/um$ litt. [go-I] 'I go'
    - $gidiyor/sun$ litt. [go-you] 'you go'

Segmentation rules

Idea: Split off tags expected to have English counterpart, remove others.
When decision is not straightforward → experiment

- **Nominal case**
- **Possessive**
- **Copula 'to be'**
- **Verb person**

Example: 'I was in my room'

$odamdaydım$	→	$oda\ /m/da/ydı/m$
[room-my-in-was-I]		[room] [my] [in] [was] [I]

Segmentation rules

Idea: Split off tags expected to have English counterpart, remove others.
When decision is not straightforward → experiment

- **Nominal case**
- **Possessive**
- **Copula 'to be'**
- **Verb person**

Example: 'I was in my room'

odamdaydım	→	*oda /m/da/ydı/m*
[room-my-in-was-I]		[room] [my] [in] [was] [I]

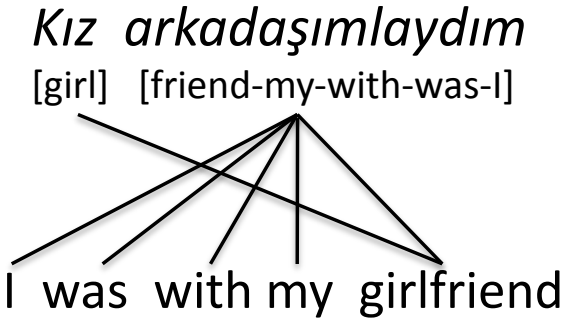
`oda+Noun+A3sg/+P1sg/+Loc/^DB+Verb+Zero+Past/+A1sg`

lemma poss. case copula v.pers

FONDAZIONE BRUNO KESSLER

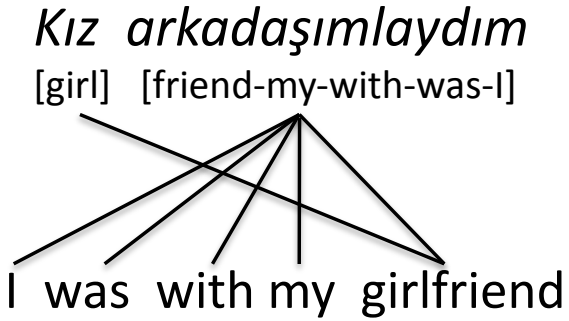
Looking into the alignments

Before segmentation:



Looking into the alignments

Before segmentation:



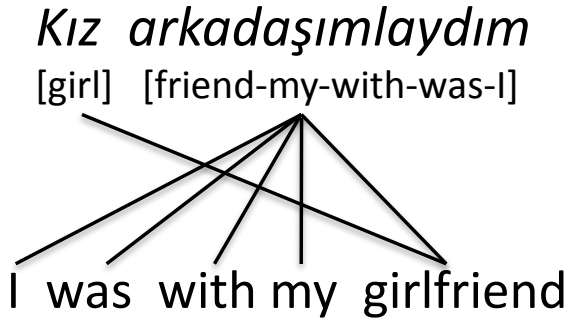
After segmentation:

Kız arkadaş/ım/ la/ ydı/ m

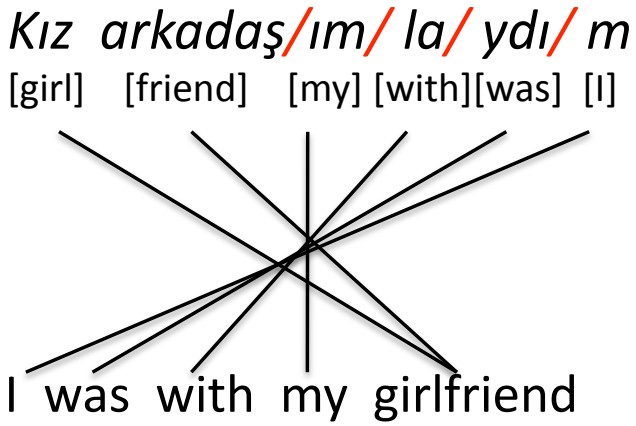
[girl] [friend] [my] [with][was] [I]

Looking into the alignments

Before segmentation:



After segmentation:



Outline

- Turkish & SMT
- Morphological Segmentation
  - Preprocessing chain
  - Segmentation rules
- **Lexical Approximation**
- Experiments
- Future Work & Conclusions

FONDAZIONE
BRUNO KESSLER

Lexical Approximation

Lexical Approximation

Idea: replace OOV words in the test with morphologically similar words of training

Cf. previous IWSLT's works on Arabic:

- Mermer & al., 2007 [5]
- Shen & al., 2008 [6]

FONDAZIONE
BRUNO KESSLER

Lexical Approximation

Idea: replace OOV words in the test with morphologically similar words of training

- Possible replacers → known words sharing the same lemma
- Similarity function → priority to words sharing more contiguous tags
- Deterministic choice of 1-best candidate

FONDAZIONE BRUNO KESSLER

Lexical Approximation

Idea: replace OOV words in the test with morphologically similar words of training

- Possible replacers → known words sharing the same lemma
- Similarity function → priority to words sharing more contiguous tags
- Deterministic choice of 1-best candidate

Word	Gloss	Preprocessed (MS11)	Score
çıkışlar	**exits**	**çık+Verb+Pos^DB+Noun+Inf3+A3pl**	
çıkış	**exit**	**çık+Verb+Pos^DB+Noun+Inf3+A3sg**	**93**
çıkma	going out	çık+Verb+Pos^DB+Noun+Inf2+A3sg	66
çıkacak	will go out	çık+Verb+Pos^DB+Noun+FutPart+A3sg	66
çıkan	who goes out	çık+Verb+Pos^DB+Adj+PresPart	44
çıkıyor	is going out	çık+Verb+Pos+Prog1	27
çıkmıyor	isn't going out	çık+Verb+Neg+Prog1	0
çıkarır	takes out	çık+Verb^DB+Verb+Caus+Pos+Aor	-15

FONDAZIONE
BRUNO KESSLER

Outline

- Turkish & SMT
- Morphological Segmentation
  - Preprocessing chain
  - Segmentation rules
- Lexical Approximation
- **Experiments**
- Future Work & Conclusions

FONDAZIONE
BRUNO KESSLER

Experiments

FONDAZIONE BRUNO KESSLER

Experiments

	Training set		SMT (on devset2)			
Preprocessing	Tokens	Dict.	%OOV	%BLEU	%WER	%PER
baseline	139,514	17,619	6.16	52.26	37.75	29.95
MS2 (case)	151,410	14,343	4.35	53.89	37.21	28.51
MS6 (case,poss)	156,390	12,009	3.49	54.10	37.29	28.19
MS7 (case,poss,cop)	157,927	11,519	3.18	55.05	37.73	27.67
MS11 (case,poss,cop,v.pers)	168,135	10,450	2.54	56.23	36.59	26.37

FONDAZIONE BRUNO KESSLER

Experiments

Preprocessing		Training set		SMT (on devset2)			
		Tokens	Dict.	%OOV	%BLEU	%WER	%PER
baseline		139,514	17,619	6.16	52.26	37.75	29.95
MS2	(case)	151,410	14,343	4.35	53.89	37.21	28.51
MS6	(case,poss)	156,390	12,009	3.49	54.10	37.29	28.19
MS7	(case,poss,cop)	157,927	11,519	3.18	55.05	37.73	27.67
MS11	(case,poss,cop,v.pers)	168,135	10,450	2.54	56.23	36.59	26.37

- segmentation minimizes differences in word granularity between TR and EN
- reduces dictionary size and data sparseness

FONDAZIONE BRUNO KESSLER

Experiments

Preprocessing		Training set		SMT (on devset2)			
		Tokens	Dict.	%OOV	%BLEU	%WER	%PER
baseline		139,514	17,619	6.16	52.26	37.75	29.95
MS2	(case)	151,410	14,343	4.35	53.89	37.21	28.51
MS6	(case,poss)	156,390	12,009	3.49	54.10	37.29	28.19
MS7	(case,poss,cop)	157,927	11,519	3.18	55.05	37.73	27.67
MS11	(case,poss,cop,v.pers)	168,135	10,450	**2.54**	**56.23**	36.59	26.37

- segmentation minimizes differences in word granularity between TR and EN
- reduces dictionary size and data sparseness
- important OOV decrease and consequent BLEU improvement

FONDAZIONE BRUNO KESSLER

Experiments

Preprocessing		Training set Tokens	Training set Dict.	SMT (on devset2) %OOV	SMT (on devset2) %BLEU	SMT (on devset2) %WER	SMT (on devset2) %PER
baseline		139,514	17,619	6.16	52.26	37.75	29.95
MS2	(case)	151,410	14,343	4.35	53.89	37.21	28.51
MS6	(case,poss)	156,390	12,009	3.49	54.10	37.29	28.19
MS7	(case,poss,cop)	157,927	11,519	3.18	55.05	37.73	27.67
MS11	(case,poss,cop,v.pers)	168,135	10,450	**2.54**	**56.23**	**36.59**	**26.37**

- segmentation minimizes differences in word granularity between TR and EN
- reduces dictionary size and data sparseness
- important OOV decrease and consequent BLEU improvement
- WER figures not very significant, but PER constantly lowers
  $\rightarrow$ positive effect on lexical choice rather than on reordering

FONDAZIONE BRUNO KESSLER

Experiments

Varying the distortion limit (DL):

Preprocess.	DL	%BLEU	Δ	%WER	%PER
baseline	6	52.26	1.3%	37.75	29.95
	∞	52.96		37.18	29.71
MS6	6	54.10	1.4%	37.29	28.19
	∞	54.87		36.69	28.35
MS11	6	56.23	3.0%	36.59	26.37
	∞	57.91		33.70	25.69

- because task is simple, unlimited distortion has reasonable decoding time
- the more segmented the text, the more improvement possible

FONDAZIONE BRUNO KESSLER

Experiments

Lexical approximation:

Preprocess.	DL	%BLEU
MS11	∞	57.91
MS11 & lex.approx.	∞	58.12

- work in progress
- promising results in particular setting $\rightarrow$ room for improvement
- in final submission dropping OOV words gave better BLEU scores

Outline

- Turkish & SMT
- Morphological Segmentation
  - Preprocessing chain
  - Segmentation rules
- Lexical Approximation
- Experiments
- **Future Work & Conclusions**

FONDAZIONE BRUNO KESSLER

Future Work & Conclusions

- refine segmentation schemes by better addressing verbal suffixation
- improve lexical approximation technique:
  - test different similarity functions
  - feed the decoder with multiple options of replacement
- repeat experiments on a more complex task

FONDAZIONE BRUNO KESSLER

Future Work & Conclusions

- refine segmentation schemes by better addressing verbal suffixation
- improve lexical approximation technique:
  - test different similarity functions
  - feed the decoder with multiple options of replacement
- repeat experiments on a more complex task

- linguistic preprocessing crucial for morphologically rich language like Turkish
- split/removing suffixes from morph.analyzed text yields large improvements
- linguistic knowledge guides hypothesis formulation before empirical validation

Thanks for your attention!

Preprocessing scripts available at : http://hlt.fbk.eu/people/bisazza

FONDAZIONE BRUNO KESSLER

IWSLT09 TR-EN Outputs Compared

	Japon Büyükelçiliği ile irtibata geçmek istiyorum .
Ref:	*I'd like to contact the Japanese Embassy .*
baseline:	I'd like to contact with Japanese büyükelçiliği .
MS11:	I'd like to contact with Japanese embassy .
	Bu film rulolarını banyo ettirip basabilir miydiniz ?
Ref:	*Could you develop and print these rolls of film ?*
baseline:	Could you reissue ettirip rulolarını this film developed ?
MS11:	Could you reissue roll of film developed ?
	Onu bulmaktan ümidi hemen hemen kestim .
Ref:	*I've just about given up finding it .*
baseline:	bulmaktan ümidi cut it right away .
MS11:	I cut almost hope from find it .
	Şimdi kirazların çiçek açma mevsimi .
Ref:	*It's cherry blossom season .*
baseline:	kirazların buds mail seasons now .
MS11:	cherry blossoms bloom season now .

References

[1] N. Habash and F. Sadat, "Arabic Preprocessing Schemes for Statistical Machine Translation," in *Proc. of NAACL HLT*. New York City, USA: Association for Computational Linguistics, June 2006, pp. 49–52.

[2] K. Oflazer and I. D. El-Kahlout, "Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation," in *Proc. of Workshop on SMT*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 25–32.

[3] K. Oflazer, "Two-level Description of Turkish Morphology," *Literary and Linguistic Computing*, vol. 9, no. 2, pp. 137–148, 1994.

[4] T. G. H. Sak and M. Saraclar, "Morphological Disambiguation of Turkish Text with Perceptron Algorithm," in *Proc.of CICLing*, 2007, pp. 107–118.

[5] H. K. C. Mermer and M. U. Dogan, "The TUBITAK-UEKAE Statistical Machine Translation System for IWSLT 2007," in *Proc. of IWSLT*, Trento, Italy, 2007, pp. 176–179.

[6] T. A. W. Shen, B. Delaney and R. Slyh, "The MIT-LL/AFRL IWSLT-2008 MT System," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 69–76.