SYNTAX-TO-MORPHOLOGY Mapping IN FACTORED PHRASE-BASED STATISTICAL MACHINE TRANSLATION

KemalOflazer
(Joint work with
ReyyanYeniterzi@CMU-LTI)

Click to edit Master subtitle style

Turkish

- Turkish is an Altaic language with over 60 Million speakers (> 150 M for Turkic Languages: Azeri, Turkoman, Uzbek, Kirgiz, Tatar, etc.)

- Agglutinative Morphology
  - Morphemes glued together like "beads-on-a-string"
  - Morphophonemic processes (e.g., vowel

Turkish Morphology

- Productive inflectional and derivational suffixation.
  - Many derivational suffixes
  - Possibly multiple derivations in a word form
  - Derivations applicable to almost all roots in a POS-class

- No prefixation, and no productive compounding

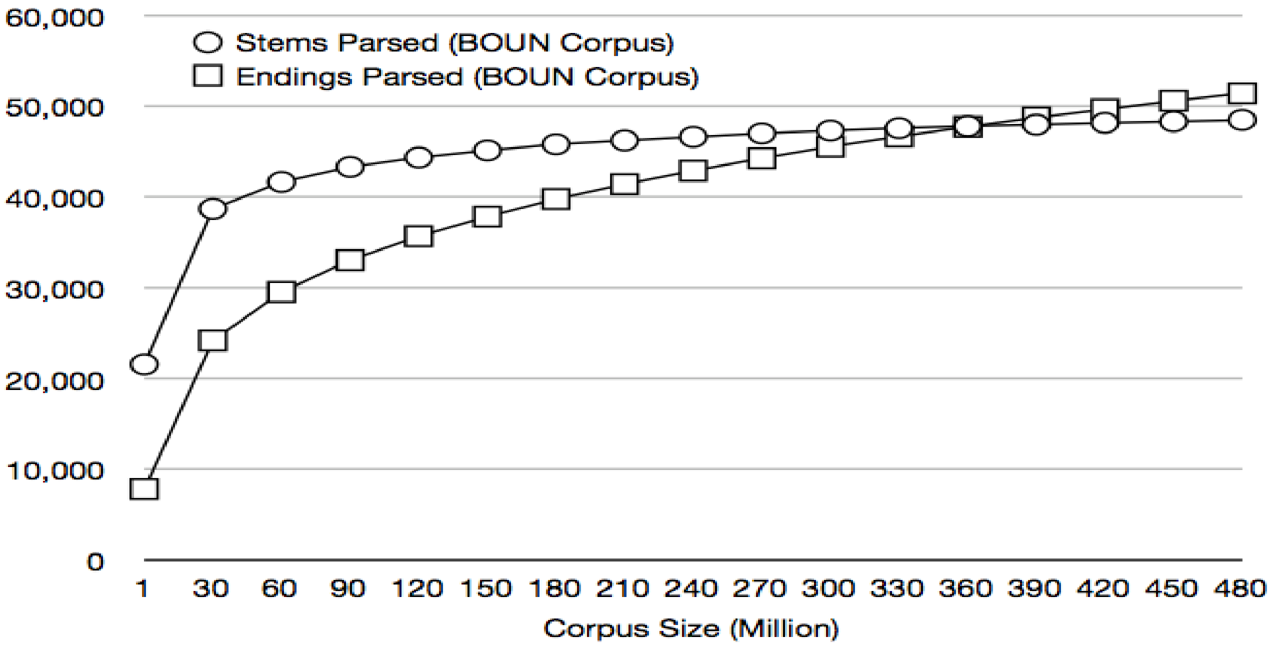
Turkish Morphology

- Basic root lexicon has about 30,000 entries
  - ~100,000 roots with proper nouns
- But each noun/verb root word can generate a very large number of forms
  - Nouns have about 100 different forms w/o any derivations
  - Verbs have about 500 again w/o any derivations

Some Statistics

- HasimSak and Murat Saraclar of Bogazici University have recently compiled a 491Mword corpus
  - About 4.1M types
  - Going from 490M to 491M adds about 5,000 new types
  - Most frequent 50K types cover 89%
  - Most frequent 300K types cover 97%
  - 3.4M Types occur less than 10 times

Some Statistics



Word Structure

- A word can be seen as a sequence of inflectional groups (IGs) separated by derivational boundaries (^DB)

  Root+Infl1^DB+Infl2^DB+… ^DB+Infln

- sağlamlaştırdığımızdaki ((existing) at the time we caused (something) to become strong.)
- sağlam+laş+tır+dığ+ımız+da+ki
- sağlam+Adj^DB+Verb+Become(+laş)

How does English become Turkish?

if we are going to be able to make [something] become pretty

if we are going to be able to make become pretty

güzel +leş +tir +ebil +ecek +se +k

güzelleştirebileceksek

English phrases vs. Turkish words

- Verb complexes/Adverbial clauses
  - Iwould not be able todo(something)
  - yap+ama+yacak+tı+m
  - if wewillbe able to do (something)
  - yap+abil+ecek+se+k

discontinuity

  - when/at the time wehad (someone) have (someone else) do (something)
  - yap+tır+t+tığ+ımız+da

English phrases vs. Turkish words

- Possessive constructions/prepositional phrases
  - my .... magazines
  - dergi+ler+im
  - with your .... magazines
  - dergi+ler+iniz+le
  - due-to theirclumsi+ness
  - sakar+lık+ları+ndan

How bad can it potentially get?

- Finlandiyalılaştıramadıklarımızdanmışsınızcasına
  - (behaving) as if you have beenone of thosewhomwecouldnotconvertintoaFinn(ish citizen)/someone from Finland
  - Finlandiya+lı+laş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına

- Finlandiya+Noun+Prop+A3sg+Pnon+Nom
  - ^DB+Adj+With/From
  - ^DB+Verb+Become
  - ^DB+Verb+Caus
  - ^DB+Verb+Able+Neg

But it gets better!-Finnish Numerals

- Finnish numerals are written as one word and all components inflect and agree morphologically with the head noun they modify.

  - second tenth eighth
    Kahdensienkymmenensienkahdeksansien
    kaksi+Ord+Pl+Genkymmenen+Ord+Pl+Genkahdeksan+Ord+Pl+Gen
    - Twenty-eighth
      kahdens i en kymmenens i en kahdeksans i en

Example Courtesy of Lauri Karttunen

But it gets better!

- Aymara
  - ch'uñüwinkaskiriÿätwa
  - ch'uñu +: +wi +na -ka +si -ka -iri +: +ya:t(a) +wa
- I was (one who was) always at the place for making ch'uñu'

ch'uñu	N	'freeze-dried potatoes'
+:	N>V	be/make …
+wi	V>N	place-of
+na		in (location)
-ka	N>V	be-in (location)
+si		continuative
-ka		imperfect
-iri	V>N	one who

Example Courtesy of Ken Beesley

Polysynthetic Languages

- Inuktikut uses morphology to combine syntactically related components (e.g. verbs and their arguments) of a sentence together
  - Parismunngaujumaniralauqsimanngittunga
  - Paris+mut+nngau+juma+niraq+lauq+si+ma+nngit+jun

Example Courtesy of Ken Beesley

Back to English – Turkish SMT

- Previous work in English-to-Turkish SMT relied segmenting Turkish into morphemes and translated at the levels of morphemes. (Durgar-El Kahlout and Oflazer (2010))
  - Selective morpheme segmentation
  - Morpheme and word-based LMs
  - Post-processing to occasionally correct malformed words

English – Turkish SMT: Problems

- Sentences get longer for alignment
  - Many sentences getting close to 100 tokens after morpheme segmentation
- Morphemes attach to incompatible roots; incorrect morphotactics
  - Decoder handles both syntactic reordering and morphotactics using the same statistics
    - Intuitively this did not look right

English – Turkish SMT: Highlights

- Two phrase translations coming together to form a new word
  - Source: promote protection of children's rights in line with eu and international standards .
  - Translation:çocukhak+larh+nhn koru+hn+ma+sh+nhn**abveuluslar+arasistandart+lar+ya uygunşekil+da** geliş+dhr+hl+ma+sh .
    - Lit. develop protection of children's rights in

English – Turkish SMT: Highlights

- Two phrase translations coming together to form a new word
  - Source: promote protection of children's rights in line with eu and international standards .
  - Translation:çocukhak+larh+nhn koru+hn+ma+sh+nhn**abveuluslar+arasistandart+lar+ya uygunşekil+da**geliş+dhr+hl+ma+sh .
    - Lit. develop protection of children's rights in

English – Turkish SMT: Highlights

- Mining the phrase-table, one finds similar interesting phrase pairs like
  - afterexamine +vvg, +acc incele +dhk +abl sonra
- One can think of this as a structural transfer rulelike
  - afterexamine +vvgNP$_{eng}$ □

NPturk+acc incele +dhk +abl sonra

Now for a completely different approach

- Examples such as
  - Iwould not be able todo(something)
  - yap+ama+yacak+tı+m ➔ yapamayacaktım
  - if wewillbe able to do (something)
  - yap+abil+ecek+se+k ➔yapabileceksek
  - when/at the time wehad (someone) have (someone else) do (something)
  - yap+tır+t+tığ+ımız+da ➔yaptırttığımızda

Now for a completely different approach

- Instead of segmenting Turkish, can we map syntactic structures in English to complex words in Turkish directly ?
  - Recognize certain local and nonlocal syntactic structures on the English side
  - Package those structures and attach to heads to obtain parallel morphological structures
  - Use factored PB-SMT

Syntax-to-Morphology Mapping

ontheireconomicrelations

Tagger

on+IN their+PRP$ economic+JJ relation+NN_NNS

Dependency Parser

PMOD

POS

on+IN their+PRP$ economic+JJ relation+NN_NNS

Transformation

economic+JJ relation+NN_NNS_their+PRP$_on+

Syntax-to-Morphology Mapping

economic+JJrelation+NN_NNS_their+PRP$ _on+IN

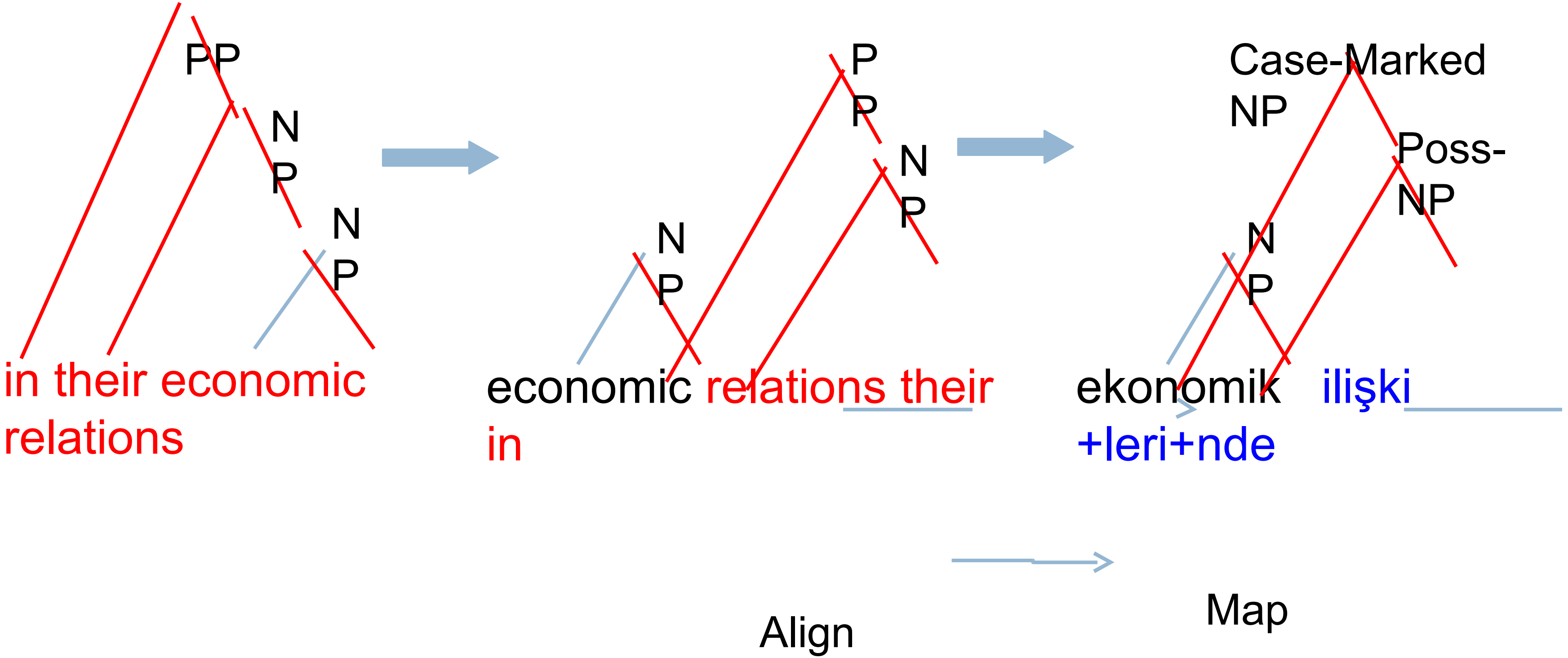
Syntax-to-morphology mapping

ekonomik+Adjilişki+Noun+A3pl+P3

Morphological Analyzer/Disambiguator

ekonomikilişkilerinde

A Constituency View



Syntax-to-Morphology Mapping

- On both sides of the parallel data, each token now comprises of three factors:
  - Surface (= Root+Tag)
  - Root
  - The complex tag
    - Local syntax on the English side(+any morphology)
    - Full morphology on the Turkish side

economic|economic|+JJ relations|relation|+NN_NNS their|their+PRP$ on+IN

ekonomik|ekonomik|+Adj ilişkilerinde|ilişki|+Noun+A3pl+P3sg+Loc

Observations

- We can identify and reorganize phrases on the English side, to “align” English syntax to Turkish morphology.
- The length of the English sentences can be dramatically reduced.
    - most function words encoding syntax are now abstracted into complex tags
- Continuous and discontinuous variants

Rest of Talk

- Another example
- Experimental Setup
- Experiments
- Additional Improvements
- Constituent Reordering
- Applications to Turkish-to-English SMT
- Conclusions

Syntax-to-Morphology Mapping

ifarequestismadeorallytheauthoritymustmakearecordofit

Tagger

if+INa+DTrequest+NNbe+VB_VBZmake+VB_VBNorally+RB

the+DTauthority+NNmust+MDmake+VBa+DTrecord+NNof+INit+PRP

Dependency Parser

VMOD NMOD VC

if+IN a+DT request+NN be+VB_VBZ make+VB_VBN orally+RB

NMOD VC NMOD PMOD

the+DT authority+NN must+MD make+VB a+DT record+NN of+IN it+PRP

Transformation

request+NN_a+DTmake+VB_VBN_be+VB_VBZ_if+INorally+RB

authority+NN_the+DTmake+VB_must+MDrecord+NN_a+DTit+PRP

Capturing Discontinuous Syntax

ifarequestismadeorallytheauthoritymustmakearecordofit

Tagger

if+INa+DTrequest+NNbe+VBVBZmake+VB_VBNorally+RB the+DTauthority+NNmust+MDmake+VBa+DTrecord+NNof+INit+PRP

Dependency Parser

VMOD

NMOD VC

if+IN a+DT request+NN be+VB_VBZ make+VB_VBN orally+RB the+DT authority+NN must+MD make+VB a+DT record+NN of+IN it+PRP

NMOD VC NMOD PMOD

Transformation

request+NN_a+DTmake+VB_VBN_be_VB+VBZ_if+INorally+RB authority+NN_the+DTmake+VB_must+MDrecord+NN_a+DTit+PRP

Syntax-to-Morphology Mapping

request+NN_a_DTmake+VB_VBN_be_VB_VBZ_if_INorally+RB
authority+NN_the_DTmake+VB_must_MDrecord+NN_a_DTit+PRP_of_IN

English side now has less tokens (7 vs 14 originally)

istek+Nounsözlü+Adjol+Verb+ByDoingSoyap+Verb+Pass+Narr+Cond
yetkili+Adjmakam+Nounbu+Pron+Acckaydet+Verb+Neces+Cop

Morphological Analyzer/Disambiguator

isteksözlü olarak

yapılmışsayetkilimakambunukaydetmelidir

Syntax-to-Morphology Mapping

- We use about 20 linguistically motivated syntax-to-morphology transformations which handle the following cases:
  - Prepositions
  - Possessive pronouns
  - Possessive markers
  - Auxiliary verbs and modals
  - Forms of *be* used as predicates with adjectival or nominal dependents
  - Forms of *be* or *have* used to form passive voice, and forms of *be* used with -ing verbs to form present

Data Preparation

- Same data that has been used in Durgar-El-Kahlout and Oflazer, 2010
  - 52712 parallel sentences
  - Average of
    - 23 words in English sentences
    - 18 words in Turkish sentences
- Randomly generated 10 train, test and dev set combinations
  - 1000 sentences each for testing and

Data Preparation

- English
  - POS tagging with Stanford Log-Linear Tagger
  - Dependency parsing with MaltParser
  - Additional stemming with
- Turkish
  - Perform full morphological analysis and morphological disambiguation
  - Remove any morphological features that are not

Experiments

- Moses toolkit
  - to encourage long distance reordering
    - distortion limit of ∞
    - distortion weight of 0.1
    - Dual-path decoding
      - Translate surface if you can
      - Translate root and complex tag and conjoin to get the translated surface
      - Large generation table!
- SRILM Toolkit

Baseline Systems

- Baseline System
  - Surface form of the word
  - 3-gram LM for surface words

relation+NN_NNS
ilişki+Noun+A3pl

- Baseline-Factored System
  - Surface | Lemma | ComplexTag
  - Aligned based on Lemma factor
  - Different or each factor

Surface | Lemma | ComplexTag

Experiment	Ave.	STD.	Max.	Min
Baseline	17.08	0.60	17.99	15.97
Baseline-Factored Model	18.61	0.76	19.41	16.80

Experiments with Transformations

- Transformations on the English side
  - Nouns and adjectives (Noun+Adj)
    - Prepositions, possessive pronouns and markers, forms of be used as predicates with adjectives etc.
  - Verbs (Verb)
- Transformations on the Turkish side

Experiments with Transformations

Experiment	Ave.	STD.	Max.	Min
Baseline	17.08	0.60	17.99	15.97
Baseline-Factored Model	18.61	0.76	19.41	16.80
Noun+Adj	21.33	0.62	22.27	20.05
Verb	19.41	0.62	20.19	17.99
Adv	18.62	0.58	19.24	17.30
Verb+Adv	19.42	0.59	20.17	18.13
Noun+Adj+Verb+Adv	21.67	0.72	22.66	20.38
Noun+Adj+Verb+Adv+PostP	**21.96**	**0.72**	**22.91**	**20.67**

28.57% points over baseline

18.00% points over factored baseline

Experiments with Transformations

Experiment	Ave.
Baseline-Factored Model	18.61
Noun+Adj	21.33
Verb	19.41
Adv	18.62
Verb+Adv	19.42
Noun+Adj+Verb+Adv	21.67
Noun+Adj+Verb+Adv+PostP	21.96

2.72 BLEU points

0.8 BLEU points

BLEU Score vs. Number of Tokens

Correlation :
-0.99

n-gram Precision Components of BLEU Scores

- BLEU for words, roots (BLEU-R) and morphological tags (BLEU-M)

		1-gr.	2-gr.	3-gr.	4-gr.
BLEU	21.96	55.73	27.86	16.61	10.68
BLEU-R	27.63	68.60	35.49	21.08	13.47
BLEU-M	27.93	67.41	37.27	21.40	13.41

- We are getting most of the root words and the complex morphological tags correct, but not necessarily getting the combination equally as good

Experiments with Higher Order LMs

- Factored phrase-based SMT allows the use of multiple LMs for different factors during decoding
- Investigate the contribution of higher order n-gram language models (4-grams to 9-grams) for the

LM orders Surface|Lemma|Tag	Ave.	STD.	Max.	Min
3|3|3	21.96	0.72	22.91	20.67
3|3|8	22.61	0.72	23.66	21.37
3|4|8	22.80	0.85	24.07	21.57
3|4|8 + Lexical Reordering	23.76	0.93	25.16	22.49

Augmenting the Training Data

- Augment the training data with reliable phrase pairs obtained from a previous alignment
- Extract phrases from phrase table that satisfy
  - $0.9 \leq p(e|t)/p(t|e) \leq 1.1$ (phrases translate to each other)
  - $p(t|e) + p(e|t) \geq 1.5$ (and not much to

Experiment	Ave.	STD.	Max.	Min
3|4|8 + Lexical Reordering	23.76	0.93	25.16	22.49
Above+Augmentation	24.38	0.81	25.44	23.18

further bias the alignment process

Sentence Length vs Transformations

- Results after only the transformations (same LMs)
  - English Sentence length 1-10 in the original test set
    - Average BLEU 46.19
    - Average %Improvement over baseline 3% relative
  - English Sentence length 20-30 in the original test set

Constituent Reordering

- Syntax to morphology transformations do not perform any constituent level reordering
- We now reordered the source sentences, to bring English constituent order (SVO) more in line with the Turkish constituent order (SOV) at the top and embedded

Constituent Reordering

- Object reordering (ObjR)
  - from English SVO to Turkish SOV
- Adverbial phrase reordering (AdvR)
  - from English V AdvPto Turkish AdvP V
- Passive sentence agent reordering (PassAgR)
  - from English SBJ PassiveVC*by*VAgentto Turkish SBJ Vagent*by*PassiveVC
- Subordinate clause reordering (SubCR)
  - postnominal relative clauses and prepositional phrase modifiers

Experiments with Reordering

Experiment	Ave.	STD.	Max.	Min
Best Result from Previous Transformations (3-3-3/No-reordering/No Aug.)	21.96	0.72	22.91	20.67
ObjR	21.94	0.71	23.12	20.56
ObjR+AdvR	21.73	0.50	22.44	20.69
ObjR+PassAgR	21.88	0.73	23.03	20.51
ObjR+SubCR	21.88	0.61	22.77	20.92

- Although there were some improvements for certain cases, none of the reorderings gave consistent improvements for all the data sets
- Examination of the alignments produced after these reordering transformations indicated that the resulting

Turkish to English Translation

- Syntax-to-Morphology mapping can be applied in the reverse direction, but
  - The decoded English would have tags encoding syntax which would further have to be post-processed to put various function words in their right places.

economic+JJ relation+NN_NNS _their+PRP$ _on+IN

↓

on+IN their+PRP$ economic+JJ relation+NN_NNS

Turkish to English Translation

- Exactly the same set-up as English-to-Turkish system (except for decoding parms)
  - Post-processing with a Transformed English-to-English SMT
    - Train with transformed English train set as the source and the POS-tagged original English as the target language
  - Rule/Heuristics-based transformation undo

Turkish-to-English Translation

Experiment	Ave.	STD.	Max.	Min
Factored Baseline (3-3-3)	24.96	0.48	25.82	24.02
Syntax-to-Morphology Transformations (3-3-3)+Rule-based+SMT Undo (3-3-3)	27.59	0.62	28.47	26.72
Syntax-to-Morphology Transformations (3-3-3)+Only SMT Undo (3-3-3)	28.27	0.46	28.99	27.75
Syntax-to-Morphology Transformations (3-4-5)+Only SMT Undo (4-5-7)	29.67	0.61	30.60	28.75
Above + Lexical Reordering	30.31	0.72	31.35	29.34

Sentence Length vs Transformations

- Results after only the transformations (same LMs)
  - English Sentence length 1-10 in the original test set
    - Average BLEU 43.66
    - Average %Improvement over baseline 11% relative
  - English Sentence length 20-30 in the original test set

Conclusions: English-to-Turkish SMT

- A novel approach

Conclusions: Source-side Reordering

- We performed numerous additional syntactic reordering transformations on the source to further bring the constituent order in line with the target order
- These reorderings did not provide any tangible improvements when averaged over the 10 different data

Conclusions: Turkish-to-English SMT

- We obtained similar improvements in the reverse direction using a second straight-forward SMT system to undo transformations.
  - There is still more room there
    - Augmentation
    - LM’s using much larger English data
    - Experiments with reordering

Future Work

- Can we learn transformation rules from a pre-processed / parsed corpora with some minimal additional information about relative morphology?
- Other languages
  - English-to-Finnish would be interesting

Finnish: Some ideas

- Finnish numerals are written as one word and all components inflect and agree morphologically with the head noun they modify.
  - ...of the twenty eighth olympics ....
  - .... Kahdensienkymmenensienkahdeksansien...
- Parse English and propagate any features (you can extract) to all components of the ordinal

second tenth eighth

kaksi+Ord+Pl+Gen kymmenen+Ord+Pl+Gen kahdeksan+Ord+

Thanks

Syntax-to-Morphology Mapping

- These rules are based on the morphological structure of the target language words.
- These transformations are handled by scripts that process dependency parser's output

if (<X>+IN PMOD <Z>+NN<TAG>)

then {

APPEND <X>+IN TO <Z>+NN<TAG>

REMOVE <X>+IN

}

PMOD

on+IN relation+NN_NNS

Root

Complex Tag

relation+NN_NNS_on

Syntactic Annotation

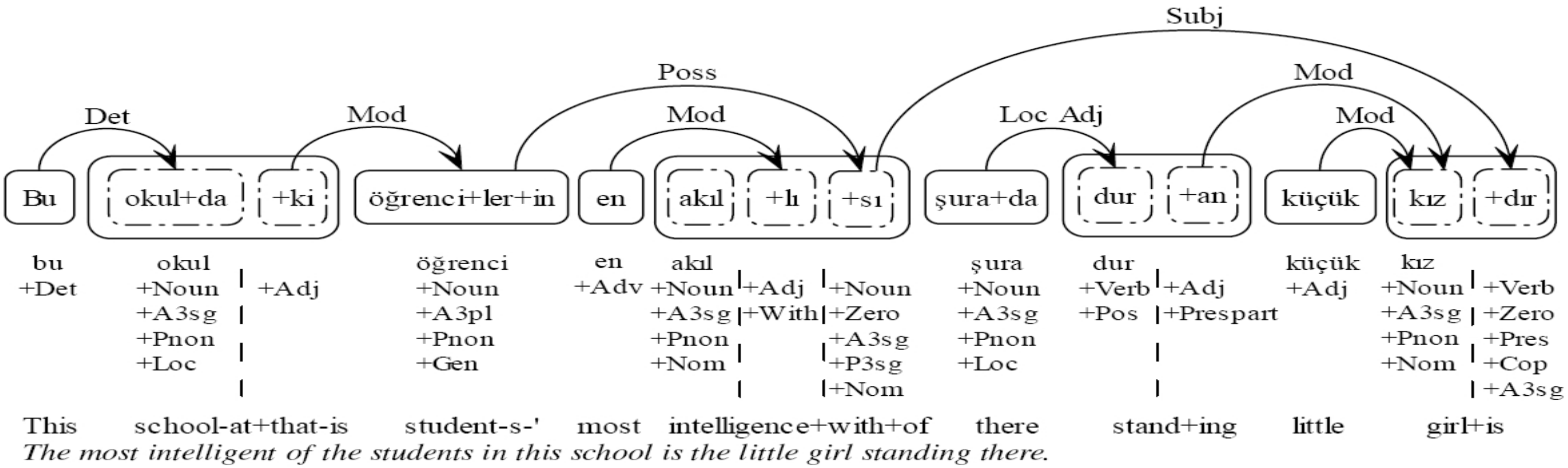


Figure 1
Dependency links in an example Turkish sentence.
+'s indicate morpheme boundaries. The rounded rectangles show words while IGs within words that have more than one IG are indicated by the dashed rounded rectangles. The inflectional features of each IG as produced by the morphological analyzer are listed below the IG.

SyntacticAnnotation

- The intensifier adverbial *en* (most) modifies the intermediate derived adjective *akıl+lı*(with intelligence/intelligent)