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

KemalOflazer
(Joint work with
ReyyanYeniterzi@CMU-LTI)
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

Hankamer (1989) e.g., estimates few million forms per verbal root (counting derivations and inflections).
Hasim Sak 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

[Graph showing the relationship between corpus size and the number of stems and endings parsed.]
Word Structure

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

\[
\text{Root}+\text{Infl}_1^{\text{^DB}}+\text{Infl}_2^{\text{^DB}}+\ldots^{\text{^DB}}+\text{Infl}_n
\]

- sağlamlaştırdığımızdaki (existing) at the time we caused (something) to become strong.)
- sağlam+laş+tır+dığ+ımız+da+ki
- sağlam+Adj^{\text{^DB}}+Verb+Become(laş)
How does English become Turkish?

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

güz +leş +tir +ebil +ece +s +k
el

Güzelleştirebilecekse k
English phrases vs. Turkish words

- **Verb complexes/Adverbial clauses**
  - I would not be able to do (something)
  - yap + ama + yacak + tı + m
  - if we will be able to do (something)
  - yap + abil + ecek + se + k
  - when/at the time we had (someone) have (someone else) do (something)
  - yap + tır + t + tığ + imız + da
Possessive constructions/prepositional phrases

- my .... magazines
- dergi+ler+im
- with your .... magazines
- dergi+ler+iniz+le
- due-to theirclumsi+ness
- sakar+lık+lari+ndan
How bad can it potentially get?

- **Finlandiya**-la**ş**tir**a**madıklarımızdan**m**i**ş**s**i**n**iz**c**ası
  - (behaving) as if you have been one of those whom we could not convert into a Finn(ish citizen)/someone from Finland
  - Finlandiya**+I+Iaş+tır+ama+dık+lar+ımız+dan+miş+sini+z+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.

Kahdensienkymmenensienienkahdeksansie

Example Courtesy of Lauri Karttunen
But it gets better!

- Aymara
  - ch’uñüwinkaskirïyätwa
  - 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’

Example Courtesy of Ken Beesley
Inuktikut uses morphology to combine syntactically related components (e.g. verbs and their arguments) of a sentence together

Parismunngaujumaniralaauqsimanngittunga
Paris+mut+nngau+juma+niraq+lauq+si+ma+nngit+jun

Example Courtesy of Ken Beesley
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
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
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: çocuk hakları koruma standartlarıarasıda gelişmiştir. 

Lit. develop protection of children's rights in
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:** çocuk hak+lar+ınh koru+hn+ma+sh+ınh na+bveulus lar+ar as ısta ndart+lar+ya uygun şekil+da geliş+dhr+hl+ma+sh.

- **Lit:** develop protection of children's rights in
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 rule like

- afterexamine +vvgNP_{eng} NP_{turk}+acc incele +dhk +abl sonra
Now for a completely different approach

Examples such as

- I would not be able to do (something)
- yap + ama + yacak + tı + m → yapamayacaktım

- if we will be able to do (something)
- yap + abil + ecek + se + k → yapabileceksek

- when/at the time we had (someone) have (someone else) do (something)
- yap + tır + t + tığı + imiz + da → vaptırreactstrapımızda
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

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

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

Transformation

Tagger

Dependency Parser PMO POS

economic+JJ relation+NN_NNS their+PRP$_on+
Syntax-to-Morphology Mapping

economic+JJ relation+NN _NNS _their+PRP$ _on+IN

Syntax-to-morphology mapping

ekonominik+Adj ilişiği+Noun +A3pl +P3

Morphological Analyzer/Disambiguator

ekonominikilişkilerinde
A Constituency View

in their economic relations

economic relations in

Align Map

Case-Marked Poss-

ekonomik ilişki in

NP NP

PP PP PP

NP NP NP

PP PP PP
On both sides of the parallel data, each token now comprises of three factors:

- Surface (= Root+Tag)
- Root
- The complex tag

- Full morphology on the Turkish side (+any morphology)
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 of certain (syntactic) source phrases can be conflated during the SMT phrase extraction process.
Rest of Talk

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

if a request is made orally, the authority must make a record of it

Dependency Parser

Tagger

Transformation
Capturing Discontinuous Syntax

if a request is made orally, the authority must make a record of it.
Syntax-to-Morphology Mapping

request\textunderscore NN\_\textunderscore a\_\textunderscore DT make\_\textunderscore VB\_\textunderscore VBN\_\textunderscore be\_\textunderscore VB\_\textunderscore VBZ\_\textunderscore if\_\textunderscore IN oraly\_\textunderscore RB

authority\_\textunderscore NN\_\textunderscore the\_\textunderscore DT make\_\textunderscore VB\_\textunderscore must\_\textunderscore MDrecord\_\textunderscore NN\_\textunderscore a\_\textunderscore DT

PRP\_of\_IN

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

istek\_\textunderscore Noun sözlü\_\textunderscore Adjol\_\textunderscore Verb\_\textunderscore ByDoingSoyap\_\textunderscore Verb\_\textunderscore Pass\_\textunderscore N

yetkili\_\textunderscore Adjmakam\_\textunderscore Noun bu\_\textunderscore Pron\_\textunderscore Acc kaydet\_\textunderscore Verb\_\textunderscore Neces+

Cop

Morphological Analyzer/Disambiguator

istek sózlü olarak

vapılmişşayetkilmakambunu\textunderscore kaydetmelidir
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 continuous
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 development
  - Remaining 50712 sentences for training
Data Preparation

- **English**
  - POS tagging with Stanford Log-Linear Tagger
  - Dependency parsing with MaltParser
  - Additional stemming with TreeTagger

- **Turkish**
  - Perform full morphological analysis and morphological disambiguation
  - Remove any morphological features that are not explicitly marked by an overt morpheme
Experiments

- Moses toolkit
  - to encourage long distance reordering
    - distortion limit of $\infty$
    - 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
  - 3-gram LM initially for all factors
  - Modified Kneser-Ney discounting with interpolation

Evaluation

- Each experiment was repeated over the 10 data sets
- BLEU metric
  - Average, standard deviation, maximum and minimum values
Baseline Systems

- **Baseline System**
  - Surface form of the word relation+NN_NNS
  - 3-gram LM for surface words

- **Baseline-Factored System**
  - Surface | Lemma | ComplexTag
  - Aligned based on Lemma factor
  - Different 3-gram LMs are used for each factor

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Ave.</th>
<th>STD.</th>
<th>Max.</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>17.08</td>
<td>0.60</td>
<td>17.99</td>
<td>15.97</td>
</tr>
<tr>
<td>Baseline-Factored Model</td>
<td>18.61</td>
<td>0.76</td>
<td>19.41</td>
<td>16.80</td>
</tr>
</tbody>
</table>
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)
    - Auxiliary verbs, negation markers, modals, passive constructions etc.
  - Adverbs (Adv)
    - Various adverbial clauses formed with if, while, when etc.
  - Verbs and adverbs (Verb+Adv)

- Transformations on the Turkish side
  - Some lexical postpositions in Turkish correspond to English prepositions
  - These postpositions are treated as if they were case-markers and attached to the immediately preceding noun (PostP)
### Experiments with Transformations

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

28.57% points over baseline
18.00% points over factored baseline
# Experiments with Transformations

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-Factored Model</td>
<td>18.61</td>
</tr>
<tr>
<td>Noun+Adj</td>
<td>21.33</td>
</tr>
<tr>
<td>Verb</td>
<td>19.41</td>
</tr>
<tr>
<td>Adv</td>
<td>18.62</td>
</tr>
<tr>
<td>Verb+Adv</td>
<td>19.42</td>
</tr>
<tr>
<td>Noun+Adj+Verb+Adv</td>
<td>21.67</td>
</tr>
<tr>
<td>Noun+Adj+Verb+Adv+PostP</td>
<td>21.96</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>1-gr.</th>
<th>2-gr.</th>
<th>3-gr.</th>
<th>4-gr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>21.96</td>
<td>55.73</td>
<td>27.86</td>
<td>16.61</td>
</tr>
<tr>
<td>BLEU-R</td>
<td>27.63</td>
<td>68.60</td>
<td>35.49</td>
<td>21.08</td>
</tr>
<tr>
<td>BLEU-M</td>
<td>27.93</td>
<td>67.41</td>
<td>37.27</td>
<td>21.40</td>
</tr>
</tbody>
</table>

- 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 morphological tag factor.

<table>
<thead>
<tr>
<th>LM orders</th>
<th>Ave.</th>
<th>STD.</th>
<th>Max.</th>
<th>Min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>Lemma</td>
<td>Tag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>21.96</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>8</td>
<td>22.61</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>8</td>
<td>22.80</td>
<td>0.85</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>8 + Lexical Reordering</td>
<td>23.76</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Augmenting the Training Data

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

These phrases are added to the training data to further bias the alignment process.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Ave.</th>
<th>STD.</th>
<th>Max.</th>
<th>Min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>8 + Lexical Reordering</td>
<td>23.76</td>
<td>0.93</td>
</tr>
<tr>
<td>Above+Augmentation</td>
<td>24.38</td>
<td>0.81</td>
<td>25.44</td>
<td>23.18</td>
</tr>
</tbody>
</table>
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 AdvP* to Turkish *AdvP V*

- Passive sentence agent reordering *(PassAgR)*
  - from English *SBJ PassiveVC by VAgent* to Turkish *SBJ Vagent by PassiveVC*

- Subordinate clause reordering *(SubCR)*
  - postnominal relative clauses and prepositional phrase modifiers
Experiments with Reordering

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 root alignments were not necessarily that close to being monotonic as we would have expected.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Ave.</th>
<th>STD.</th>
<th>Max.</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Result from Previous Transformations (3-3-3/No-reordering/No Aug.)</td>
<td>21.96</td>
<td>0.72</td>
<td>22.91</td>
<td>20.67</td>
</tr>
<tr>
<td>ObjR</td>
<td>21.94</td>
<td>0.71</td>
<td>23.12</td>
<td>20.56</td>
</tr>
<tr>
<td>ObjR+AdvR</td>
<td>21.73</td>
<td>0.50</td>
<td>22.44</td>
<td>20.69</td>
</tr>
<tr>
<td>ObjR+PassAgR</td>
<td>21.88</td>
<td>0.73</td>
<td>23.03</td>
<td>20.51</td>
</tr>
<tr>
<td>ObjR+SubCR</td>
<td>21.88</td>
<td>0.61</td>
<td>22.77</td>
<td>20.92</td>
</tr>
</tbody>
</table>
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.
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 with coupled with a second SMT system
## Turkish-to-English Translation

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Ave.</th>
<th>STD.</th>
<th>Max.</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factored Baseline (3-3-3)</td>
<td>24.96</td>
<td>0.48</td>
<td>25.82</td>
<td>24.02</td>
</tr>
<tr>
<td>Syntax-to-Morphology Transformations (3-3-3)+Rule-based+SMT Undo (3-3-3)</td>
<td>27.59</td>
<td>0.62</td>
<td>28.47</td>
<td>26.72</td>
</tr>
<tr>
<td>Syntax-to-Morphology Transformations (3-3-3)+Only SMT Undo (3-3-3)</td>
<td>28.27</td>
<td>0.46</td>
<td>28.99</td>
<td>27.75</td>
</tr>
<tr>
<td>Syntax-to-Morphology Transformations (3-4-5)+Only SMT Undo (4-5-7)</td>
<td>29.67</td>
<td>0.61</td>
<td>30.60</td>
<td>28.75</td>
</tr>
<tr>
<td>Above + Lexical Reordering</td>
<td>30.31</td>
<td>0.72</td>
<td>31.35</td>
<td>29.34</td>
</tr>
</tbody>
</table>
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 to map source syntactic structures to target morphological structures by encoding many local and nonlocal source syntactic structures as additional complex tag factors.

In our experiments, we performed syntax-to-morphology mapping transformations on the source side and a very small set of transformations on the target side.

Overall, with some additional techniques we got about 30% improvement of a factored baseline. A lot of the improvement is probably due to reduction in the number of English tokens during GIZA++ alignment.
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 sets.
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 numerals are written as one word and all components inflect and agree morphologically with the head noun they modify.

- ...of the twenty eighth olympics ....
- .... Kahdensienkymmenensienienkahdeksansa

Parse English and propagate any features (you can extract) to all components of the ordinal (e.g., other words, phrases).
Thanks
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.

\[
\text{if } (<X>+IN \text{ PMOD }<Z>+NN<TAG>) \text{ then } \{
\text{APPEND } <X>+IN \text{ TO } <Z>+NN<TAG>
\text{REMOVE } <X>+IN
\text{on}+IN
\text{relation}+NN\_NNS
\text{relation}+NN\_NNS\_o
\}
\]
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.
The intensifier adverbial *en* (most) modifies the intermediate derived adjective *akıl+lı* (with intelligence/intelligent)