Deep-synta...
Outline

- TectoMT architecture
- Development of a new language pair (English - Spanish)
  - Analysis
  - Transfer
  - Synthesis
- Evaluation
- Conclusions
Tecto layers

- TectoMT
  - transfer-based system which works at the deep tectogrammatical level
  - combines linguistic knowledge and statistical techniques, particularly during transfer
  - originally developed for the English-Czech language direction

- Stratification approach
  - Morphological layer
  - Analytical layer (shallow-syntax dependency tree)
  - Tectogrammatical layer (deep-syntax dependency tree)
Tectogrammatical layer

- Only autosemantic nodes are kepted
- Functional words represented by attributes
- Each t-node consists of:
  - Tectogrammatical lemma
  - Functor: semantic values of syntactic dependency relations (causal adjunt, actor, effect, etc.)
  - Grammatemes: semantically oriented morphological categories (tense, number, modality, etc.)
  - Formemes: values of the morphosyntactic form in the surface sentence (subject, direct object, etc.)
TectoMT architecture
Tecto blocks and scenarios

• Blocks: reusable components of NLP subtasks that can be listed in a specific sequence, that is, rules to define, set, change and move node-information in/across the layers

• Scenarios: specific sequences of blocks to be applied to relevant data

• TectoMT includes over a thousand blocks:
  - 224 blocks specific for English
  - 237 for Czech
  - 57 for English-to-Czech transfer
  - 129 for other languages
  - 467 language-independent
Developing a new pair

- We set to port the TectoMT system to work for the English-Spanish language pair in both directions.
  - English analysis and synthesis ready to use
  - Our focus: Spanish analysis and synthesis, and transfer stages
- TectoMT is integrated within Treex
  - Modules divided into language-specific and language independent blocks
Analysis

- From raw text to tecto-level
- English analysis solved
- Spanish analysis
  - Tokenization and sentence splitting: adapted modules in Treex
  - Lemmatisation and POS: integration of ixa-pipes tools (pos) in Treex
  - Dependency parsing: integration of ixa-pipes tools (srl) in Treex
    - Tagset compatibility: from AnCora to Interset
  - Spanish blocks:

<table>
<thead>
<tr>
<th>Block type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language-independent blocks</td>
<td>11</td>
</tr>
<tr>
<td>Adapted blocks</td>
<td>4</td>
</tr>
<tr>
<td>New language-specific blocks</td>
<td>1</td>
</tr>
</tbody>
</table>
Transfer

• Statistical transfer dictionary
  - trained on parallel corpora analyzed up to the t-level in both languages
    - lemmas, formemes and grammatemes
  - for each t-lemma and formeme in a source t-tree, the translation model assigns a score to all possible translations observed in the training data
  - probability estimate calculated as a linear combination of
    - Discriminative TM
    - Dictionary TM

• Static manual dictionary (priority resource)
  - Microsoft Terminology Collection - 22,475 entries
Transfer

• Blocks for grammateme equivalences
  – linguistically abstract, usually paralleled in the target language
• rules are inherently language-specific
• 5 blocks for English-to-Spanish direction:
  – lack of gender in English nouns (necessary in Spanish);
  – differences in definiteness and articles;
  – differences in structures such as “There is...” and relative clauses.
Synthesis

- From tecto-level to raw text
- English synthesis solved
- Spanish synthesis
  - Transform the t-tree into an a-tree
  - Transform the a-tree into word forms
  - Polish the output

<table>
<thead>
<tr>
<th>Block type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language-independent blocks</td>
<td>9</td>
</tr>
<tr>
<td>Adapted blocks</td>
<td>12</td>
</tr>
<tr>
<td>New language-specific blocks</td>
<td>3</td>
</tr>
</tbody>
</table>
Synthesis

- Transform the t-tree into an a-tree:
  - fill in morphological attributes that will be needed in the second step
  - add function words where necessary
  - remove superfluous nodes
  - add punctuation nodes

- Transform the a-tree into word forms
  - new Spanish models in Flect (statistical morphological generator)
  - corpus: subset of morphologically annotated (530K tokens)

- Polish the output: detokenization, contractions, ...
Evaluation

- **Compared systems:**
  - PBSMT (Moses)
    - Features: mGiza, SRILM
    - Corpora:
      - Bilingual: europarl (~2M sentences)
      - Monolingual: europarl (~2M sentences)
    - Tuning: 1,000 IT-domain Q&A set - 1
  - TectoMT
    - Language-independent blocks only
    - + Spanish blocks (new + adapted)
    - + domain-specific dictionary
Evaluation

- **Test-sets:**
  - 1,000 IT-domain Q&A set - 2
  - WMT11 newswire test-set

- **Results**
  - Moses outperforms the TectoMT systems
  - BLEU increases as TectoMT customisation increases
  - en->es scores higher than es->en in accordance with the development effort
  - Systems score better for the IT set

<table>
<thead>
<tr>
<th></th>
<th>English-Spanish</th>
<th>Spanish-English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IT</td>
<td>WMT11</td>
</tr>
<tr>
<td>Moses</td>
<td>28.12</td>
<td>26.91</td>
</tr>
<tr>
<td>TectoMT – language independent blocks</td>
<td>12.40</td>
<td>8.38</td>
</tr>
<tr>
<td>TectoMT – + domain dictionary</td>
<td>26.40</td>
<td>13.25</td>
</tr>
</tbody>
</table>
Conclusions

- Development of an entry-level deep-syntax system for the English-Spanish pair
  - Reuse of English analysis and synthesis modules
  - Integration of ixa-pipes for Spanish
  - Crafting of blocks for Spanish
  - Training of statistical models for transfer
  - Training of morphological models for Spanish synthesis

- Available at: https://github.com/ufal/treex

- BLEU scores still behind Moses (but close for En-Es on the IT domain!)
  - Flexible customization options
  - Further customization and tuning has potential for improvement
Thank you