Introduction

- **Problem:** How can we improve parsing when there are several, potentially heterogeneous treebanks for a language?
- **Treebank diversity**
  - Annotation scheme
    - Language variant
    - Spoken/written language
    - Genres and domains
  - Treebank size
  - Annotation quality and consistency
- **This work:**
  - Investigate previously proposed strategies
  - Introduce treebank embeddings

Strategies

**Single**
- One model per treebank
  - + Simple
  - – Does not take advantage of all data
  - – Separate models for each treebank

**Concatenation**
- One model per language, on concatenated data
  - + Simple
  - – Does not take treebank differences into account
  - + A single model per language

**Concatenation + fine-tuning**
- Fine-tune a different model for each treebank, based on the concatenation (Che et al., 2017, Shi et al., 2017)
  - Needs more training than previous models
  - – Separate models for each treebank
  - + Takes treebank differences into account

System Architecture

Treebank embeddings
- Train a single model per language, but use a treebank embedding to represent the treebank each word comes from.
- Similar to language embeddings (Ammar et al., 2016)
  - + Simple
  - + Takes treebank differences into account
  - + A single model per language

Other approaches (not in this paper)
- 1-hot treebank representation: similar to our approach, but with 1-hot representation rather than embedding (Lim & Poibeau, 2017).
- Adversarial learning: combine treebank specific models with a joint model where treebank identification is an adversarial task (Sato et al., 2017). Effective, especially on small treebanks, but more complicated than our model.

Parsing Unseen Data

When parsing unseen data, we need to choose an existing treebank: proxy treebank
- **Single:** the treebank used to train a model
- **Concatenation:** N/A
- **Concatenation + fine-tuning:** the treebank used for fine-tuning
- **Treebank embeddings:** the treebank embedding to use in the model

Experiments

- **Universal dependencies version 2.1**
- **Standardized annotation scheme, but still many differences**
- **9 languages:**
  - At least 2 training treebanks
  - – test set without training data (PUD)

Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
<th>Size</th>
<th>Single</th>
<th>Concat</th>
<th>Single</th>
<th>Concat</th>
<th>PUD (unseen)</th>
<th>Single</th>
<th>Concat</th>
<th>PUD (unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDT</td>
<td>68495</td>
<td></td>
<td>83.7</td>
<td>87.5</td>
<td>88.3</td>
<td>87.2</td>
<td>81.7</td>
<td>84.6</td>
<td>81.2</td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>23478</td>
<td></td>
<td>80.0</td>
<td>87.8</td>
<td>88.5</td>
<td>86.5</td>
<td>75.0</td>
<td>81.3</td>
<td>81.1</td>
<td></td>
</tr>
<tr>
<td>Tv-Tree</td>
<td>10160</td>
<td></td>
<td>84.3</td>
<td>89.3</td>
<td>89.5</td>
<td>80.2</td>
<td>66.1</td>
<td>79.8</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>CLTT</td>
<td>860</td>
<td></td>
<td>72.5</td>
<td>86.2</td>
<td>86.0</td>
<td>86.0</td>
<td>42.1</td>
<td>80.8</td>
<td>80.9</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>12444</td>
<td></td>
<td>92.4</td>
<td>92.2</td>
<td>92.5</td>
<td>92.5</td>
<td>90.5</td>
<td>88.0</td>
<td>87.1</td>
<td></td>
</tr>
<tr>
<td>PaTUT</td>
<td>1781</td>
<td></td>
<td>80.5</td>
<td>83.5</td>
<td>85.4</td>
<td>85.7</td>
<td>68.0</td>
<td>78.1</td>
<td>76.9</td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>14381</td>
<td></td>
<td>76.4</td>
<td>74.4</td>
<td>74.7</td>
<td>74.7</td>
<td>65.3</td>
<td>73.0</td>
<td>54.6</td>
<td></td>
</tr>
<tr>
<td>TDT</td>
<td>12217</td>
<td></td>
<td>78.1</td>
<td>79.6</td>
<td>80.4</td>
<td>80.3</td>
<td>76.6</td>
<td>80.9</td>
<td>50.3</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>14554</td>
<td></td>
<td>84.6</td>
<td>84.6</td>
<td>85.6</td>
<td>85.6</td>
<td>79.1</td>
<td>80.2</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td>Sequoia</td>
<td>2231</td>
<td></td>
<td>84.0</td>
<td>86.7</td>
<td>89.8</td>
<td>89.8</td>
<td>69.5</td>
<td>79.1</td>
<td>77.8</td>
<td></td>
</tr>
<tr>
<td>PaTUT</td>
<td>803</td>
<td></td>
<td>79.8</td>
<td>80.5</td>
<td>80.9</td>
<td>90.3</td>
<td>63.4</td>
<td>78.8</td>
<td>77.5</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>12058</td>
<td></td>
<td>87.7</td>
<td>87.9</td>
<td>87.9</td>
<td>87.9</td>
<td>85.4</td>
<td>85.5</td>
<td>86.0</td>
<td></td>
</tr>
<tr>
<td>MoTUTWITA</td>
<td>2808</td>
<td></td>
<td>71.4</td>
<td>76.7</td>
<td>76.8</td>
<td>77.0</td>
<td>68.5</td>
<td>85.7</td>
<td>85.3</td>
<td></td>
</tr>
<tr>
<td>PaTUT</td>
<td>1781</td>
<td></td>
<td>84.4</td>
<td>82.9</td>
<td>83.3</td>
<td>88.8</td>
<td>77.4</td>
<td>85.8</td>
<td>86.1</td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>35664</td>
<td></td>
<td>35.0</td>
<td>58.2</td>
<td>58.3</td>
<td>58.3</td>
<td>74.0</td>
<td>72.9</td>
<td>71.2</td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>8331</td>
<td></td>
<td>84.7</td>
<td>84.2</td>
<td>86.2</td>
<td>86.3</td>
<td>75.2</td>
<td>77.5</td>
<td>77.6</td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>48814</td>
<td></td>
<td>91.2</td>
<td>91.2</td>
<td>91.2</td>
<td>91.2</td>
<td>90.2</td>
<td>66.0</td>
<td>66.1</td>
<td></td>
</tr>
<tr>
<td>GSD</td>
<td>3650</td>
<td></td>
<td>74.7</td>
<td>73.4</td>
<td>79.8</td>
<td>80.3</td>
<td>70.1</td>
<td>77.6</td>
<td>78.0</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>14305</td>
<td></td>
<td>87.2</td>
<td>87.2</td>
<td>87.5</td>
<td>87.6</td>
<td>75.2</td>
<td>77.7</td>
<td>77.6</td>
<td></td>
</tr>
<tr>
<td>AinCor</td>
<td>14187</td>
<td></td>
<td>84.7</td>
<td>83.0</td>
<td>85.8</td>
<td>86.2</td>
<td>78.2</td>
<td>80.8</td>
<td>80.9</td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>41303</td>
<td></td>
<td>76.8</td>
<td>79.6</td>
<td>80.5</td>
<td>80.5</td>
<td>70.3</td>
<td>73.2</td>
<td>73.0</td>
<td></td>
</tr>
<tr>
<td>LinES</td>
<td>2738</td>
<td></td>
<td>74.3</td>
<td>76.8</td>
<td>77.3</td>
<td>77.1</td>
<td>64.0</td>
<td>70.0</td>
<td>69.0</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>81.4</td>
<td></td>
<td>82.7</td>
<td>84.9</td>
<td>84.9</td>
<td>84.9</td>
<td>77.9</td>
<td>85.0</td>
<td>80.1</td>
<td></td>
</tr>
</tbody>
</table>

| &nbsp; | **Significantly better than single** | **Significantly better than concat** | **Significantly better than single+concat** |

Conclusion

- Combining treebanks is beneficial, especially for small treebanks
- Treebank embeddings successful
  - At least on par with other methods
  - Simple model
  - Works for many different scenarios
  - Choice of proxy treebank very important

References

Wolfgang Ammer, George Maloueis, Miguel Ballste`n, Chris Dyres, and Noah Smith. 2018. Many languages, one parser. TACL, 4:444-444.