Part D: multilingual dependency parsing
Motivation

- A difficult syntactic ambiguity in one language may be easy to resolve in another language (bilingual constraints)
- A more accurate parser on one language may help a less accurate one on another language (semi-supervised)
- Rich labeled resources in one language can be transferred to build parsers of another language (unsupervised)
Obstacles

- Syntactic non-isomorphism across languages
- Different annotation choices (guideline)
- Partial (incomplete) parse trees resulted from projection
- Parsing errors on the source side
- Word alignment errors
Research map (two perspectives)

- **Methods**
  - Delexicalized (not rely on bitext)
  - Projection (rely on bitext)
    - Project hard edges
    - Project edge weights (distributions)
    - As training objective (bilingual constraints)

- **Supervision** (existence of target treebank)
  - Semi-supervised
  - Unsupervised
Bilingual word reordering features (Huang+ 09)

- Supervised with bilingual constraints
- Use the word reordering information in source language as extra features when parsing target text (using automatic word alignment as bridge)

Two characteristics
- Bilingual text with target-side labels
- No parse tree in the source side

I [ saw Bill ] [ with a telescope ].

wo [ yong wangyuanjin] [kandao le Bi’er].
Build local classifiers via projection (Jiang & Liu 10)

- Semi-supervised; project edges
  - Step 1: projection to obtain dependency/non-dependency classification instances
  - Step 2: build a target-language local dependency/non-dependency classifier
  - Step 3: feed the outputs of the classifier into a supervised parser as extra weights during test phase.
Prune projected structures with target-language marginal (Li+ 14)

- Semi-supervised; project edges from EN to CH
- The goal is to acquire extra large-scale high-quality labeled training instances from bitext

- Step 1: parse English sentences of bitext
- Step 2: project English trees into Chinese, and filter unlikely dependencies with target-language marginal probabilities (handling cross-language non-isomorphism, or parsing and alignment errors)
- Step 3: train with extra training instances with partial annotations.
Prune projected structures with target-language marginal (Li+ 14)

(a) Source tree and word alignments

(b) Projected incomplete tree

(c) Forest (ambiguous labelings)
Delexicalized transfer (no bitext) (McDonald+ 11)

- Unsupervised; delexicalized
- Direct delexicalized transfer (Zeman & Resnik 08, Sogaard 11, Cohen+ 11)
  - Train a delexicalized parser on source treebank, ignoring words, only using universal POS tags (Petrov+ 11)
Delexicalized transfer (no bitext) (McDonald+ 11)

- **Direct** delexicalized transfer
  - Step 1: replace the language-specific POS tags in source treebank with the universal POS tags (Petrov+ 11)
    - NOUN (nouns); VERB (verbs); ADJ (adjectives); ADV (adverbs)
    - PRON (pronouns); DET (determiners); ADP (pre/postpositions);
      NUM (numerals); CONJ (conjunctions); PRT (particles);
      PUNC (punctuation marks); X (a catch-all tag)
  - Step 2: train a delexicalized source-language parser
    - Only use POS tag features (no lexical features)
    - 89.3 (lex) => 82.5 (delex) on English test set
  - Step 3: parse target-language sentences with universal POS tags
Delexicalized transfer (no bitext) (McDonald+ 11)

- **Multi-source** delexicalized transfer
  - Use multiple source treebanks (besides English), and concatenate them as a large training data.

- Can improve parsing accuracy.
Delexicalized + projected transfer (with bitext) (McDonald+ 11)

• Unsupervised; delexicalized + project (as training objective)

• Workflow (Danish as target language)
  • Step 1: parse a set of Danish sentences with the delexicalized transfer parser (trained on En treebank).
  • Step 2: train a lexicalized Danish parser using the predicted parses as gold-standard.
  • Step 3: improve the lexicalized Danish parser on bitext
    • Extrinsic objective: the parser’s outputs are more consistent with the outputs of the lexicalized English parser (training with multiple objectives, Hall+ 11).
Delexicalized transfer with cross-lingual word clusters (Tackstrom+ 12)

- Unsupervised; delexicalized + cross-lingual word clusters

- In delexicalized dependency parsers
  - Only POS tag features are used

- In other previous work on dependency parsing
  - Features based on word clusters work very well
Delexicalized transfer with cross-lingual word clusters (Tackstrom+ 12)

- Word clusters derived from large-scale unlabeled data can alleviate data sparseness.
- POS tags < word clusters < words
Delexicalized transfer with cross-lingual word clusters (Tackstrom+, 2012)

- Re-lexicalize the delexicalized transfer parser by using cross-lingual word clusters.
- Cross-lingual word clusters need to be consistent across languages.
- A cross-lingual clustering algorithm that leverages
  - large amounts of monolingual unlabeled data, and
  - parallel data through which the cross-lingual word cluster constrains are enforced.
Delexicalized transfer with cross-lingual word clusters (Tackstrom+ 12)

Cluster | Lang. | Sample words
---|---|---
60 | EN | was, wasn’t, wasn’t,
60 | ES | estaba, estarán, est
101 | EN | very, mildly, wholly,
101 | ES | muchomás, fuerte, f
153 | EN | chicken, bird, ostrich
153 | ES | pollo, achote, manza
195 | EN | The, ...
195 | ES | El, La, Los, Las, Lo
dry, wet, moist, life
236 | ES | seco, secos, semisecco
Selectively multi-source delexicalized transfer (Tackstrom+ 13)

- Unsupervised; delexicalized (also in Naseem+ 12)
- Some delexicalized features do not transfer well across typologically different languages
- Selectively parameter sharing based on typological and language-family features (Dryer and Haspelmath 11)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>81A</td>
<td>Order of Subject, Object and Verb</td>
</tr>
<tr>
<td>85A</td>
<td>Order of Adposition and Noun</td>
</tr>
<tr>
<td>86A</td>
<td>Order of Genitive and Noun</td>
</tr>
<tr>
<td>87A</td>
<td>Order of Adjective and Noun</td>
</tr>
</tbody>
</table>
Selectively multi-source delexicalized transfer (Tackstrom+ 13)

- Selectively use features for different source and target language pairs
  - Indo-European (Bulgarian, Catalan, Czech, Greek, English, Spanish, Italian, Dutch, Portuguese, Swedish)
  - Altaic languages (Japanese, Turkish)
Ambiguity-aware transfer (Tackstrom+ 13)

- Unsupervised; delexicalized

- Relexicalize delexicalized parsers using target-language unlabeled text.
  - First, parse the unlabeled data with the base delexicalized parser, and use the arc-marginals to construct ambiguous labels (forest) for each sentence.
  - Then, train a lexicalized parser on the unlabeled data via ambiguity-aware training.
**Ambiguity-aware transfer (Tackstrom+ 13)**

- Ambiguity-aware training
  - For each training instance $x$, a set of ambiguous labels $\tilde{y}(x)$ are provided, instead of a gold-standard label.

\[
\mathcal{L}(\theta; \tilde{D}) = \sum_{i=1}^{n} \log \left\{ \sum_{y \in \tilde{y}(x_i)} p_{\theta}(y | x_i) \right\} - \lambda \| \theta \|_2^2
\]
Ambiguity-aware transfer (Tackstrom+ 13)

- Ambiguity-aware self-training
- Ambiguity-aware ensemble training
  - Add the outputs of another parser (Naseem+, 2012) into the ambiguous label set.
  - Can improve accuracy.
Syntax projection as posterior regularization (Ganchev+ 09)

- Unsupervised; project high-probability edges (used in training objective)
- Key points
  - Filter unlikely projected dependencies according to source-side dependency probabilities and alignment probabilities (only project high-confidence dependencies)
  - Give the model more flexibility via posterior regularization

```
El debate iniciado ha sido apasionante.
the debate has been fascinating.
```
Syntax projection as posterior regularization (Ganchev+ 09)

- Posterior regularization
  - After training, the model should satisfy that the expectation of the number of conserved edges (among the projected edges) is at least 90%
Syntax projection as posterior regularization (Ganchev+ 09)

\[ \hat{E}[- \log p_\theta(x)] + R(\theta) + \hat{E}[\text{KL}(Q_x || p_\theta(z | x))] \]

- log-likelihood of the target-side bitext
- Model prior
- The set of distributions satisfying the desired posterior constrains
- Model posterior
Syntax projection as posterior regularization (Ganchev+ 09)

- Create several simple rules to handle different annotation decisions (annotation guideline).
  - E.g., Main verbs should dominate auxiliary verbs.

- Can largely improve parsing accuracy.
Combining unsupervised and projection objectives (Liu+ 13)

- Unsupervised; project edges
- Derive local dependency/non-dependency instances from projection (Jiang+, 10)
Combining unsupervised and projection objectives (Liu+ 13)

- Workflow
Combining unsupervised and projection objectives (Liu+ 13)

- Joint objective: linear interpolation of

Unsupervised objective

$$\theta(\lambda) = \prod_{d_e \in D_E} Pr(+|d_e) \prod_{d_e \in \tilde{D}_E} Pr(-|d_e)$$

Projection objective

$$\phi(\lambda) = \sum_{d_e \in D_P} \log Pr(+|d_e) + \sum_{d_e \in D_N} \log Pr(-|d_e)$$
Transfer distribution as bilingual guidance (Ma & Xia 14)

- Unsupervised; hybrid of projection and delexicalized; transfer edge weights (distribution)

\[
\tilde{w}(e^t, x^t_i) = \begin{cases} 
  w_E(e^s, x^s_i), & \text{if } e^t \xrightarrow{align} e^s \\
  w_E(e^t_{delex}, x^s_i), & \text{otherwise}
\end{cases}
\]

(a) Source tree and word alignments
(b) Projected incomplete tree
Transfer distribution as bilingual guidance (Ma & Xia 14)

- Use the target-language sentences to train a parser, by minimizing the KL divergence from the transferred distribution to the distribution of the learned model.

\[
\sum_{y_i} \tilde{p}(y_i | x_i) \log p_x(y_i | x_i)
\]
Transfer distribution as bilingual guidance (Ma & Xia 14)

- Use unlabeled data by minimizing entropy.
  - Not very helpful
Summary of Part D

- Help target-language parsing with source-language treebanks and bitext
  - Semi-supervised vs. unsupervised
  - Projection vs. delexicalized
  - Project hard edges, transfer edge weights, or use bilingual constraints
Summary of Part D

- Future thoughts
  - Word alignment errors (probabilities)
  - Source-language parsing errors (probabilities)
  - How to capture cross-language non-isomorphism
  - Joint word alignment and bilingual parsing?
References

- Xuezhe Ma and Fei Xia. 2014. Unsupervised Dependency Parsing with Transferring Distribution via Parallel Guidance and Entropy Regularization. In ACL.
References