Dealing with other languages

- Other languages present some problems not seen in English at all!
- Some of our algorithms have been specified to English
  - Some structures like constituency parsing don’t make sense for other languages
  - Neural methods are typically tuned to English-scale resources, may not be the best for other languages where less data is available
- Question:
  1) What other phenomena / challenges do we need to solve?
  2) How can we leverage existing resources to do better in other languages without just annotating massive data?

This Lecture

- Morphological richness: effects and challenges
- Morphology tasks: analysis, inflection, word segmentation
- Cross-lingual tagging and parsing

Morphology
**What is morphology?**

- Study of how words form
- Derivational morphology: create a new *lexeme* from a base
  - *estrange (v) => estrangement (n)*
  - *become (v) => unbecoming (adj)*
- Inflectional morphology: word is inflected based on its context
  - I become / she becomes
  - Mostly applies to verbs and nouns

**Morphological Inflection**

**In English:**
- I arrive
- you arrive
- he/she/it arrives
- we arrive
- you arrive
- they arrive
- [X] arrived

**In French:**

<table>
<thead>
<tr>
<th>Tense</th>
<th>Indicative</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>present</td>
<td>arrive</td>
<td>arrive</td>
</tr>
<tr>
<td></td>
<td>/a.ri/</td>
<td>/a.ri/</td>
</tr>
<tr>
<td>imperfect</td>
<td>arrivers</td>
<td>arrivez</td>
</tr>
<tr>
<td></td>
<td>/a.ri.ve/</td>
<td>/a.ri.ve/</td>
</tr>
<tr>
<td>past</td>
<td>arrived</td>
<td>arrive</td>
</tr>
<tr>
<td></td>
<td>/a.ri.ve/</td>
<td>/a.ri.ve/</td>
</tr>
<tr>
<td></td>
<td>/a.ri.ve/</td>
<td>/a.ri.ve/</td>
</tr>
<tr>
<td>future</td>
<td>arriver</td>
<td>arriver</td>
</tr>
<tr>
<td></td>
<td>/a.ri.ve/</td>
<td>/a.ri.ve/</td>
</tr>
<tr>
<td></td>
<td>/a.ri.ve/</td>
<td>/a.ri.ve/</td>
</tr>
<tr>
<td>conditional</td>
<td>arrivers</td>
<td>arriver</td>
</tr>
<tr>
<td></td>
<td>/a.ri.ve/</td>
<td>/a.ri.ve/</td>
</tr>
<tr>
<td></td>
<td>/a.ri.ve/</td>
<td>/a.ri.ve/</td>
</tr>
</tbody>
</table>

**In Spanish:**

<table>
<thead>
<tr>
<th>Tense</th>
<th>Indicative</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>present</td>
<td>llego</td>
<td>llegamos</td>
</tr>
<tr>
<td>imperfect</td>
<td>llegaste</td>
<td>llegamos</td>
</tr>
<tr>
<td>preterite</td>
<td>llegué</td>
<td>llegaste</td>
</tr>
<tr>
<td>future</td>
<td>llegaste</td>
<td>llegaste</td>
</tr>
<tr>
<td>conditional</td>
<td>llegaron</td>
<td>llegaron</td>
</tr>
</tbody>
</table>

**Morphological Inflection**

**In Spanish:**

<table>
<thead>
<tr>
<th>Person</th>
<th>1st person</th>
<th>2nd person</th>
<th>3rd person</th>
<th>1st person</th>
<th>2nd person</th>
<th>3rd person</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yo</td>
<td>tú/usted</td>
<td>él/ella/ellos</td>
<td>nosotros/nosotras</td>
<td>vosotros/vosotras</td>
<td>ellos/ellas</td>
</tr>
<tr>
<td>indicative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>present</td>
<td>llego</td>
<td>llegaste</td>
<td>llega</td>
<td>llegamos</td>
<td>llegó</td>
<td>llegaron</td>
</tr>
<tr>
<td>imperfect</td>
<td>llegaba</td>
<td>llegabas</td>
<td>llegaba</td>
<td>llegábamos</td>
<td>llegaban</td>
<td>llegaron</td>
</tr>
<tr>
<td>preterite</td>
<td>llegué</td>
<td>llegaste</td>
<td>llegó</td>
<td>llegamos</td>
<td>llegaste</td>
<td>llegaron</td>
</tr>
<tr>
<td>future</td>
<td>llegaste</td>
<td>llegaste</td>
<td>llegó</td>
<td>llegamos</td>
<td>llegaste</td>
<td>llegaron</td>
</tr>
<tr>
<td>conditional</td>
<td>llegaron</td>
<td>llegaron</td>
<td>llegaron</td>
<td>llegamos</td>
<td>llegaron</td>
<td>llegaron</td>
</tr>
</tbody>
</table>

**Noun Inflection**

- Not just verbs either; gender, number, case complicate things
  - Nominative: I/he/she, accusative: me/him/her, genitive: mine/his/hers
  - Dative: merged with accusative in English, shows recipient of something
  - I taught the children => Ich unterrichte die Kinder
  - I give the children a book => Ich gebe den Kindern ein Buch

- In English:
  - I arrive
  - you arrive
  - he/she/it arrives
  - we arrive
  - you arrive
  - they arrive
  - [X] arrived
Irregular Inflection

- Common words are often irregular
  - I am / you are / she is
  - Je suis / tu es / elle est
  - Soy / está / es
- Less common words typically fall into some regular paradigm — these are somewhat predictable

Agglutinating Languages

- Finnish/Hungarian (Finno-Ugric), also Turkish: what a preposition would do in English is instead part of the verb
- Many possible forms — and in newswire data, only a few are observed

Morphologically-Rich Languages

- Many languages spoken all over the world have much richer morphology than English
- CoNLL 2006 / 2007: dependency parsing + morphological analyses for ~15 mostly Indo-European languages
- Great resources for challenging your assumptions about language and for understanding multilingual models!

Morphologically-Rich Languages

- Finnish/Hungarian (Finno-Ugric), also Turkish: what a preposition would do in English is instead part of the verb
- Many possible forms — and in newswire data, only a few are observed

halata: “hug”

illative: “into”
adessive: “on”
Morphological Analysis/Inflection

‣ In English, not that many word forms, lexical features on words and word vectors are pretty effective

‣ When we’re building systems, we probably want to know base form + morphological features explicitly

‣ How to do this kind of morphological analysis?

Morphological Analysis

‣ Given a word in context, need to predict what its morphological features are

‣ Basic approach: combines two modules:
  ‣ Lexicon: tells you what possibilities are for the word
  ‣ Analyzer: statistical model that disambiguates

‣ Models are largely CRF-like: score morphological features in context

‣ Lots of work on Arabic inflection (high amounts of ambiguity)
Predicting Inflection

- Inflection: given base form + features, inflect the word
- Hard for unknown words — need models that generalize

\[ \text{to wind (de)} \]

Durrett and DeNero (2013)

Predicting Inflection

- Inflection: given base form + features, inflect the word
- Hard for unknown words — need models that generalize
- Take a bunch of existing verbs from Wiktionary, extract these change rules using character alignments

\[ \text{Change describes how } i \text{ changes for 1st person sg, 2nd person sg, ...} \]

Durrett and DeNero (2013)

Morphological Reinfection

- Machine translation where phrase table is defined in terms of lemmas
- “Translate-and-inject”: translate into uninflected words and predict inflection based on source side

Chahuneau et al. (2013)

Word Segmentation
Morpheme Segmentation

- Can we do something unsupervised rather than these complicated analyses?
- unbecoming => un+becom+ing — we should be able to recognize these common pieces and split them off
- How do we do this?

Creutz and Lagus (2002)

\[
\text{Cost(Source text)} = \sum_{\text{morph tokens}} - \log p(m_i)
\]

\[
p(m_i) = \text{count(token)}/\text{count(all tokens)}
\]

- Train with EM: E-step involves estimating best segmentation with Viterbi, M-step: collect token counts
- allowed expected need needed all+owed expected ed need ed ed ed
- M0: ed has count 3
- Some heuristics: reject rare morphemes, one-letter morphemes
- Doesn’t handle stem changes: becoming => becom + ing

Creutz and Lagus (2002)

Chinese Word Segmentation

- Some languages including Chinese are totally untokenized
- LSTMs over character embeddings / character bigram embeddings to predict word boundaries
- Having the right segmentation can help machine translation

- 冬天 (winter), 能 (can) 穿 (wear) 多少 (amount) 穿 (wear) 多少 (amount); 夏天 (summer), 能 (can) 穿 (wear) 多 (more) 少 (little) 穿 (wear) 多 (more) 少 (little).
  Without the word “夏天 (summer)” or “冬天 (winter)”, it is difficult to segment the phrase “能穿多少穿多少”.

- separating nouns and pre-modifying adjectives:
  高血压 (high blood pressure)
  → 高(high) 血压 (blood pressure)

- separating compound nouns:
  内政部 (Department of Internal Affairs)
  → 内政 (Internal Affairs) 部 (Department).

Chen et al. (2015)

Cross-Lingual Tagging and Parsing
Cross-Lingual Tagging

- Labeling POS datasets is expensive
- Can we transfer annotation from high-resource languages (English, etc.) to low-resource languages?

![Diagram](image)

- English POS data + Spanish: POS data + Raw text → Spanish tagger
- Raw text + Malagasy bitext → Malagasy tagger

Cross-Lingual Tagging

- Multilingual POS induction
- Generative model of two languages simultaneously, joint alignment + tag learning
- Complex generative model, requires Gibbs sampling for inference

![Diagram](image)

Cross-Lingual Tagging

- Rather than doing unsupervised learning, can we use supervised learning in combination with alignments?
- Tag with English tagger, project across bitext, train French tagger?
- Can do something smarter

![Diagram](image)

Cross-Lingual Tagging

- Das and Petrov (2011)
- Das and Petrov (2011)

- I like it
- I love it
- he loves it
- she loves it
- I’ aime beaucoup
- I’ adore un peu
- I’ adore beaucoup

- Add links between words in similar contexts on each side. Can help resolve words that otherwise would be tricky
Cross-Lingual Tagging

- EM-HMM/feature HMM: unsupervised methods with a greedy mapping from learned tags to gold tags
- Projection: project tags across bitext to make pseudogold corpus, train on that
- LP: add monolingual connections and run "label propagation"

Das and Petrov (2011)

<table>
<thead>
<tr>
<th>Model</th>
<th>Danish</th>
<th>Dutch</th>
<th>German</th>
<th>Greek</th>
<th>Italian</th>
<th>Portuguese</th>
<th>Spanish</th>
<th>Swedish</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM-HMM</td>
<td>68.7</td>
<td>57.0</td>
<td>75.9</td>
<td>65.8</td>
<td>63.7</td>
<td>62.9</td>
<td>71.5</td>
<td>68.4</td>
<td>66.7</td>
</tr>
<tr>
<td>Feature-HMM</td>
<td>69.1</td>
<td>65.1</td>
<td>81.3</td>
<td>71.8</td>
<td>68.1</td>
<td>78.4</td>
<td>80.2</td>
<td>70.1</td>
<td>73.0</td>
</tr>
<tr>
<td>Projection</td>
<td>73.6</td>
<td>77.0</td>
<td>83.2</td>
<td>79.3</td>
<td>79.7</td>
<td>82.6</td>
<td>80.1</td>
<td>74.7</td>
<td>78.8</td>
</tr>
<tr>
<td>our approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No LP</td>
<td>79.0</td>
<td>78.8</td>
<td>82.4</td>
<td>76.3</td>
<td>84.8</td>
<td>87.0</td>
<td>82.8</td>
<td>79.4</td>
<td>81.3</td>
</tr>
<tr>
<td>With LP</td>
<td>83.2</td>
<td>79.5</td>
<td>82.8</td>
<td>82.5</td>
<td>86.8</td>
<td>87.9</td>
<td>84.2</td>
<td>80.5</td>
<td>83.4</td>
</tr>
<tr>
<td>oracles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TB Dictionary</td>
<td>93.1</td>
<td>94.7</td>
<td>93.5</td>
<td>96.6</td>
<td>96.4</td>
<td>94.0</td>
<td>95.8</td>
<td>85.5</td>
<td>93.7</td>
</tr>
<tr>
<td>Supervised</td>
<td>96.9</td>
<td>94.9</td>
<td>98.2</td>
<td>97.8</td>
<td>95.8</td>
<td>97.2</td>
<td>96.8</td>
<td>94.8</td>
<td>96.6</td>
</tr>
</tbody>
</table>

Cross-Lingual Parsing

- Now that we can POS tag other languages, can we parse them too?
- Direct transfer: train a parser over POS sequences in one language, then apply it to another language

McDonald et al. (2011)

<table>
<thead>
<tr>
<th>Model</th>
<th>best-source</th>
<th>avg-source</th>
<th>gold-POS</th>
<th>pred-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>48.6</td>
<td>49.5</td>
<td>46.2</td>
<td>47.5</td>
</tr>
<tr>
<td>de</td>
<td>55.8</td>
<td>56.6</td>
<td>51.7</td>
<td>52.0</td>
</tr>
<tr>
<td>el</td>
<td>63.9</td>
<td>65.1</td>
<td>58.5</td>
<td>63.0</td>
</tr>
<tr>
<td>es</td>
<td>68.4</td>
<td>64.5</td>
<td>55.6</td>
<td>56.5</td>
</tr>
<tr>
<td>it</td>
<td>69.1</td>
<td>65.0</td>
<td>56.8</td>
<td>58.9</td>
</tr>
<tr>
<td>nl</td>
<td>62.1</td>
<td>65.7</td>
<td>54.3</td>
<td>64.4</td>
</tr>
<tr>
<td>pt</td>
<td>74.8</td>
<td>75.6</td>
<td>67.7</td>
<td>70.3</td>
</tr>
<tr>
<td>sv</td>
<td>66.8</td>
<td>68.0</td>
<td>58.3</td>
<td>62.1</td>
</tr>
<tr>
<td>avg</td>
<td>63.7</td>
<td>63.8</td>
<td>56.1</td>
<td>59.3</td>
</tr>
</tbody>
</table>

Cross-Lingual Embeddings

- Learn a shared multilingual embedding space so any neural system can transfer over
- multiCluster: use bilingual dictionaries to form clusters of words that are translations of one another, replace corpora with cluster IDs, train "monolingual" embeddings over all these corpora

Ammar et al. (2016)
Cross-Lingual Embeddings

- Word vectors work pretty well at “intrinsic” tasks, some improvement on things like document classification and dependency parsing as well.
- CCA = canonical correlation analysis

<table>
<thead>
<tr>
<th>Task</th>
<th>multiCluster</th>
<th>multiCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>dependency parsing</td>
<td>48.4 [72.3]</td>
<td>48.8 [69.3]</td>
</tr>
<tr>
<td>doc. classification</td>
<td>90.3 [52.3]</td>
<td>91.6 [55.6]</td>
</tr>
<tr>
<td>mono. wordsim</td>
<td>14.9 [71.8]</td>
<td>43.0 [71.8]</td>
</tr>
<tr>
<td>cross. wordsim</td>
<td>12.8 [78.2]</td>
<td>66.8 [78.2]</td>
</tr>
<tr>
<td>word translation</td>
<td>30.0 [38.9]</td>
<td>83.6 [31.8]</td>
</tr>
</tbody>
</table>

Ammar et al. (2016)

Where are we now?

- Universal dependencies: treebanks (+ tags) for 70+ languages
- Many languages are still small, so projection techniques may still help
- More corpora are getting annotated in other languages, less and less reliance on structured tools like parsers, and pretraining on unlabeled data means that performance on other languages is better than ever
- BERT has pretrained multilingual models that seem to work pretty well (trained on a whole bunch of languages)

Takeaways

- Many languages have richer morphology than English and pose distinct challenges
- Problems: how to analyze rich morphology, how to generate with it
- Can leverage resources for English using bitexts
- Next time: wrapup + discussion of ethics