Recap: BERT

Announcements

- FP check-in due today, will be returned soon
- A4 back today, A5 back soon
- eCIS evaluations: please fill these out

Multilinguality

Dealing with other languages

- Other languages present some challenges 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
- Cross-lingual word representations

What is morphology?

- Study of how words form
- Derivational morphology: create a new *lexeme* from a base
  - strange (v) => estrangement (n)
  - become (v) => unbecoming (adj)
  - May not be totally regular: enflame => inflammable
- Inflectional morphology: word is inflected based on its context
  - I become / she becomes
  - Mostly applies to verbs and nouns

Morphology

- What is morphology?
- Derivational morphology: create a new *lexeme* from a base
  - strange (v) => estrangement (n)
  - become (v) => unbecoming (adj)
  - May not be totally regular: enflame => inflammable
- 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
  - we arrive
  - you arrive
  - he/she/it arrives
  - [X] arrived
  - we arrive
  - you arrive
  - they arrive

<table>
<thead>
<tr>
<th>Tense</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>first</td>
<td>second</td>
</tr>
<tr>
<td>indicative</td>
<td>je (I')</td>
<td>tu</td>
</tr>
<tr>
<td>present</td>
<td>arrive</td>
<td>arrives</td>
</tr>
<tr>
<td>imperfect</td>
<td>arrivais</td>
<td>arrivais</td>
</tr>
<tr>
<td>past historic1</td>
<td>arrive</td>
<td>arrivaie</td>
</tr>
<tr>
<td>future</td>
<td>arriverai</td>
<td>arriveras</td>
</tr>
<tr>
<td>conditional</td>
<td>arriverais</td>
<td>arriverais</td>
</tr>
</tbody>
</table>
Morphological Inflection

- In Spanish:
  - **Indicative**
    - singular 1st person: yo, 2nd person: tú, 3rd person: él/ella/el, 4th/5th person: usted
    - plural 1st person: nosotros, 2nd person: vosotros, 3rd person: ellos/ellas, ustedes
    - present tense: llevo, lleves, lleva, llevamos, lleváis, llevan
    - imperfect tense: llevo, llevas, lleva, llevamos, lleváis, llevan
    - preterite tense: llevé, llevaste, llevó, llevamos, llevasteis, llevaron
    - future tense: llevará, llevarás, llevará, llevaremos, llevaréis, llevarán
    - conditional tense: llevaría, llevarías, llevaría, llevaríamos, llevaríais, llevarían

Noun Inflection

- Not just verbs either; gender, number, case complicate things

<table>
<thead>
<tr>
<th>Declension of Kind</th>
<th>singular</th>
<th>plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>indefinite</td>
<td>nominative</td>
<td>die</td>
</tr>
<tr>
<td></td>
<td>genitive</td>
<td>Kindes, Kind</td>
</tr>
<tr>
<td></td>
<td>dative</td>
<td>Kind, Kind</td>
</tr>
<tr>
<td></td>
<td>accusative</td>
<td>Kind</td>
</tr>
</tbody>
</table>

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

- halata: “hug”
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
- Word piece / byte-pair encoding models for MT are pretty good at handling these if there’s enough data

Morphological Analysis

- In English, lexical features on words and word vectors are pretty effective
- In other languages, lots more unseen words due to rich morphology! Affects parsing, translation, ...
- 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: Hungarian

But the government does not recommend reducing taxes.
Ám a kormány egyetlen adó csökkentését sem javasolja.

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)

Morphological Inflection

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

Durrett and DeNero (2013)

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

Chahuneau et al. (2013)
Chinese Word Segmentation

- 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

Cross-Lingual Tagging and Parsing

- Having the right segmentation can help machine translation

Cross-Lingual Tagging

- Labeling POS datasets is expensive
- Can we transfer annotation from high-resource languages (English, etc.) to low-resource languages? (Chen et al. 2015)

Cross-Lingual Tagging

- Can we leverage word alignment here?

I like it a lot → I like it a lot

Je l’aime beaucoup → Je l’aime beaucoup

Tag with English tagger, project across bitext, train French tagger?

Works pretty well

Das and Petrov (2011)
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

<table>
<thead>
<tr>
<th></th>
<th>best-source</th>
<th>gold-POS</th>
<th>avg-source</th>
<th>gold-POS</th>
<th>pred-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>it</td>
<td>48.6</td>
<td>46.3</td>
<td>48.9</td>
<td>49.5</td>
</tr>
<tr>
<td>de</td>
<td>nl</td>
<td>55.8</td>
<td>48.9</td>
<td>56.7</td>
<td>56.6</td>
</tr>
<tr>
<td>el</td>
<td>en</td>
<td>63.9</td>
<td>51.7</td>
<td>60.1</td>
<td>65.1</td>
</tr>
<tr>
<td>es</td>
<td>it</td>
<td>68.4</td>
<td>53.2</td>
<td>64.2</td>
<td>64.5</td>
</tr>
<tr>
<td>fr</td>
<td>pt</td>
<td>69.1</td>
<td>58.5</td>
<td>64.1</td>
<td>65.0</td>
</tr>
<tr>
<td>nl</td>
<td>el</td>
<td>62.1</td>
<td>49.9</td>
<td>55.8</td>
<td>65.7</td>
</tr>
<tr>
<td>pt</td>
<td>it</td>
<td>74.8</td>
<td>61.6</td>
<td>74.0</td>
<td>75.6</td>
</tr>
<tr>
<td>sv</td>
<td>pt</td>
<td>66.8</td>
<td>54.8</td>
<td>65.3</td>
<td>68.0</td>
</tr>
<tr>
<td>avg</td>
<td></td>
<td>63.7</td>
<td>51.6</td>
<td>61.1</td>
<td>63.8</td>
</tr>
</tbody>
</table>

- MulE-dir: transfer a parser trained on several source treebanks to the target language
- MulE-proj: more complex annotation projection approach

McDonald et al. (2011)

Cross-Lingual Word Representations

- Input: corpora in many languages. Output: embeddings where similar words in different languages have similar embeddings

I have an apple
47 24 18 427
ID: 24
ai have

J’ai des oranges
47 24 89 1981
ID: 47
I Je J’

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

Works okay but not all that well

Ammar et al. (2016)
MulElingual Sentence Embeddings

- Form BPE vocabulary over all corpora (50k merges); will include characters from every script
- Take a bunch of bitexts and train an MT model between a bunch of language pairs with shared parameters, use W as sentence embeddings

Artetxe et al. (2019)

Multilingual Sentence Embeddings

<table>
<thead>
<tr>
<th>EN</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>el</th>
<th>bg</th>
<th>ru</th>
<th>tr</th>
<th>ar</th>
<th>vi</th>
<th>th</th>
<th>zh</th>
<th>hi</th>
<th>sw</th>
<th>ur</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-XX</td>
<td>90.70</td>
<td>69.74</td>
<td>77.36</td>
<td>73.59</td>
<td>96.82</td>
<td>89.40</td>
<td>85.91</td>
<td>91.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Zero-Shot Transfer, one NLI system for all languages:
- Conneau et al. X-BiLSTM 73.9 67.7 68.7 67.7 68.9 67.9 65.4 64.2 64.8 66.4 64.1 65.8 64.1 55.7 58.4
- (2018b) X-CBOW 64.5 60.3 60.7 61.0 60.5 60.4 57.8 58.7 57.5 58.8 56.9 59.8 56.3 50.4 52.2
- BERT uncased 81.4 74.3 70.5 74.2 72.1 69.7 71.4 72.0 69.2 71.4 65.5 62.2 61.0
- Proposed method BiLSTM 73.9 71.9 72.9 72.6 74.2 72.1 69.7 71.4 72.0 69.2 71.4 65.5 62.2 61.0

Train a system for NLI (entailment/neutral/contradiction of a sentence pair) on English and evaluate on other languages

Artetxe et al. (2019)

MulElingual BERT

- Take top 104 Wikipedias, train BERT on all of them simultaneously
- What does this look like?
  Beethoven may have proposed unsuccessfully to Therese Malfatti, the supposed dedicatee of "Für Elise"; his status as a commoner may again have interfered with those plans.

当人们在马尔法蒂身后发现这部小曲的手稿时，便误认为上面写的是“Für Elise”（即《给爱丽丝》）[51]。

Китай (официально — Китайская Народная Республика, сокращённо — КНР；кит. трад. 中華人民共和國, упр. 中华人民共和国, пиньинь: Zhōnghuá Rénmín

Devlin et al. (2019)

Multilingual BERT: Results

<table>
<thead>
<tr>
<th>Fine-tuning</th>
<th>Eval</th>
<th>EN</th>
<th>DE</th>
<th>NL</th>
<th>ES</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>90.70</td>
<td>69.74</td>
<td>77.36</td>
<td>73.59</td>
<td>96.82</td>
<td>89.40</td>
</tr>
<tr>
<td>DE</td>
<td>73.83</td>
<td>82.00</td>
<td>76.25</td>
<td>70.03</td>
<td>83.99</td>
<td>93.99</td>
</tr>
<tr>
<td>NL</td>
<td>65.46</td>
<td>65.68</td>
<td>89.36</td>
<td>72.10</td>
<td>81.64</td>
<td>88.77</td>
</tr>
<tr>
<td>ES</td>
<td>65.38</td>
<td>59.40</td>
<td>64.39</td>
<td>87.18</td>
<td>86.79</td>
<td>87.82</td>
</tr>
</tbody>
</table>

Table 1: NER F1 results on the CoNLL data.

<table>
<thead>
<tr>
<th>Fine-tuning</th>
<th>Eval</th>
<th>EN</th>
<th>DE</th>
<th>NL</th>
<th>ES</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>96.82</td>
<td>89.40</td>
<td>85.91</td>
<td>91.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>83.99</td>
<td>93.99</td>
<td>86.32</td>
<td>88.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NL</td>
<td>81.64</td>
<td>88.77</td>
<td>96.71</td>
<td>93.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td>86.79</td>
<td>87.82</td>
<td>91.28</td>
<td>98.11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: POS accuracy on a subset of UD languages.

Can transfer BERT directly across languages with some success

...but this evaluation is on languages that all share an alphabet

Pires et al. (2019)
Multilingual BERT: Results

<table>
<thead>
<tr>
<th></th>
<th>HI</th>
<th>UR</th>
<th>EN</th>
<th>BG</th>
<th>JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HI</td>
<td>97.1</td>
<td>85.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UR</td>
<td>91.1</td>
<td>93.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

- Urdu (Arabic script) => Hindi (Devanagari). Transfers well despite different alphabets!
- Japanese => English: different script and very different syntax

Pires et al. (2019)

Scaling Up: XLM-R

- Larger “Common Crawl” dataset, better performance than mBERT
- Low-resource languages benefit from training on other languages
- High-resource languages see a small performance hit, but not much

Conneau et al. (2019)

Where are we now?

- Universal dependencies: treebanks (+ tags) for 70+ languages
- Many languages are still small, so projection techniques may still help
- More corpora 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
- Multilingual models seem to be working better and better — but still many challenges for low-resource settings

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