CS388: Natural Language Processing

Lecture 26: Multilingual, Multimodal Models

Greg Durrett
Announcements

› FP due December 9

› Next lecture — ethics and the last written response

› eCIS evaluations: fill these out for extra credit!
Multilinguality
NLP in 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
- This lecture: How can we leverage existing resources to do better in other languages without just annotating massive data?
This Lecture

- Cross-lingual tagging and parsing
- Multilingual pre-training
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 showing the process of cross-lingual tagging with examples of English, Spanish, and Malagasy datasets.](image)
Cross-Lingual Tagging

- Can we leverage word alignment here?

- 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

Parser trained to accept tag input

VERB is the head of PRON and NOUN

I like tomatoes

PRON VERB NOUN

PRON VERB PRON

I like them

Parser trained to accept tag input

VERB is the head of PRON and NOUN

Je les aime

Parser trained to accept tag input

VERB is the head of PRON and NOUN

I like them

Parser trained to accept tag input

VERB is the head of PRON and NOUN

McDonald et al. (2011)
## Cross-Lingual Parsing

<table>
<thead>
<tr>
<th></th>
<th>best-source</th>
<th>gold-POS</th>
<th>avg-source</th>
<th>gold-POS</th>
<th>gold-POS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>gold-POS</td>
<td></td>
<td>multi-dir.</td>
<td>multi-proj.</td>
</tr>
<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>it</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>

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

McDonald et al. (2011)
Cross-Lingual, Multilingual
Word Representations
Multilingual Embeddings

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

  I have an apple
  47 24 18 427

  J’ai des oranges
  47 24 89 1981

- 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)
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>Zero-Shot Transfer, one NLI system for all languages:</th>
<th>EN</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>el</th>
<th>bg</th>
<th>ru</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conneau et al. (2018b)</td>
<td>X-BiLSTM</td>
<td>73.7</td>
<td>67.7</td>
<td>68.7</td>
<td>67.7</td>
<td>68.9</td>
<td>67.9</td>
</tr>
<tr>
<td>X-CBOW</td>
<td>64.5</td>
<td>60.3</td>
<td>60.7</td>
<td>61.0</td>
<td>60.5</td>
<td>60.4</td>
<td>57.8</td>
</tr>
<tr>
<td>BERT uncased*</td>
<td>Transformer</td>
<td>81.4</td>
<td>–</td>
<td>74.3</td>
<td>70.5</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Proposed method</td>
<td>BiLSTM</td>
<td>73.9</td>
<td><strong>71.9</strong></td>
<td>72.9</td>
<td><strong>72.6</strong></td>
<td><strong>72.8</strong></td>
<td><strong>74.2</strong></td>
</tr>
</tbody>
</table>

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

Artetxe et al. (2019)
Multilingual 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]。

Кита́й (официально — Кита́йская Наро́дная Респу́блика, сокращённо — КНР; кит. трад. 中华人民共和国, упр. 中华人民共和国, пиньинь: Zhōnghuá Rénmín Gònghéguó, палл.: Чжунхуа Жэнььминь Гунхэго) — государство в Восточной Аз

Devlin et al. (2019)
Multilingual BERT: Results

<table>
<thead>
<tr>
<th>Fine-tuning \ Eval</th>
<th>EN</th>
<th>DE</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>
</tr>
<tr>
<td>DE</td>
<td>83.99</td>
<td>93.99</td>
<td>86.32</td>
<td>88.39</td>
</tr>
<tr>
<td>ES</td>
<td>81.64</td>
<td>88.87</td>
<td>96.71</td>
<td>93.71</td>
</tr>
<tr>
<td>IT</td>
<td>86.79</td>
<td>87.82</td>
<td>91.28</td>
<td>98.11</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>96.8</td>
<td>87.1</td>
<td>49.4</td>
</tr>
<tr>
<td>UR</td>
<td>91.1</td>
<td>93.8</td>
<td>82.2</td>
<td>98.9</td>
<td>51.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>57.4</td>
<td>67.2</td>
<td>96.5</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/Nastaliq script) => Hindi (Devanagari). Transfers well despite different alphabets!

- Japanese => English: different script and very different syntax

Pires et al. (2019)
Scaling Up: XLM-R

Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

- 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)
Scaling Up: mBART

Multilingual Denoising Pre-Training (mBART)

Liu et al. (2020)
### Scaling Up: Benchmarks

<table>
<thead>
<tr>
<th>Task</th>
<th>Corpus</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>XNLI</td>
<td>392,702</td>
<td>2,490</td>
<td>5,010</td>
<td>translations</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>PAWS-X</td>
<td>49,401</td>
<td>2,000</td>
<td>2,000</td>
<td>translations</td>
<td>7</td>
</tr>
<tr>
<td>Struct. pred.</td>
<td>POS</td>
<td>21,253</td>
<td>3,974</td>
<td>47-20,436</td>
<td>ind. annot.</td>
<td>33 (90)</td>
</tr>
<tr>
<td></td>
<td>NER</td>
<td>20,000</td>
<td>10,000</td>
<td>1,000-10,000</td>
<td>ind. annot.</td>
<td>40 (176)</td>
</tr>
<tr>
<td>QA</td>
<td>XQuAD</td>
<td>87,599</td>
<td>34,726</td>
<td>1,190</td>
<td>translations</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>MLQA</td>
<td>4,517–11,590</td>
<td>323–2,719</td>
<td>translations</td>
<td>7</td>
<td>Span extraction</td>
</tr>
<tr>
<td></td>
<td>TyDiQA-GoldP</td>
<td>3,696</td>
<td>634</td>
<td>-</td>
<td>ind. annot.</td>
<td>9</td>
</tr>
<tr>
<td>Retrieval</td>
<td>BUCC</td>
<td>-</td>
<td>-</td>
<td>1,896–14,330</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Tatoeba</td>
<td>-</td>
<td>-</td>
<td>1,000</td>
<td>-</td>
<td>33 (122)</td>
</tr>
</tbody>
</table>

- Many of these datasets are translations of base datasets, not originally annotated in those languages
- Exceptions: POS, NER, TyDiQA

Hu et al. (2021)
TyDiQA

- Typologically-diverse QA dataset
- Annotators write questions based on very short snippets of articles; answers may or may not exist, fetched from elsewhere in Wikipedia

Q: Как далеко Уран от Земл-и?
How far is Uranus from Earth?

A: Расстояние между Уран-ом и Земл-ёй меняется от 2,6 and Earth-SG.INSTR varies from 2,6 до 3,15 млрд км...
to 3,15 bln km...

The distance between Uranus and Earth fluctuates from 2.6 to 3.15 bln km...

Clark et al. (2021)
Cross-Lingual Typing

- Train an mBERT-based typing model on Wikipedia data in English, Spanish, German and Finnish
- Achieves solid performance even on totally new languages like Japanese that don’t share a character set with these

Selvaraj, Onoe, Durrett (2021)
Where are we now?

- Universal dependencies: treebanks (+ tags) for 70+ languages

- Datasets in other 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 — can even transfer to new languages “zero-shot”. But still many challenges for low-resource settings
Multimodality, Language Grounding
Language Grounding

- We’ve seen that we can learn representations that transfer across multiple languages

- What about different modalities of data?

- Can we view an (image, text) pair as two “languages” and train something like what we had for multilingual data?

- Ultimate goal: learn models that ground language in something other than symbols
Language Grounding

‣ How to associate words with sensory-motor experiences

‣ How to associate words with meaning representation

Alan Turing was a British mathematician, logician, cryptanalyst, and computer scientist.

\[ \text{nationality(AT,UK)} \land \text{notable for(AT,mathematician)} \land \text{profession(AT,logic)} \land \text{research(AT,cryptanalysm)} \land \text{notable type(AT,compsci)} \]
Language Grounding

- What does “yellowish green” mean?
- Formal semantics: yellowish green is a predicate. Things are either yellowish green or not. No connection to real color
- Grounding in perceptual space:
Perception

- Visual: $green = [0,1,0]$ in RGB

- Auditory: $loud = >120$ dB

- Taste: sweet = $>some\ threshold\ level\ of\ sensation\ on\ taste\ buds$

- High-level concepts:
  - cat
  - dog
  - running
  - eating
Learning from Interaction

1. Use feedback from control application to understand language

Walk across the bridge

Reward +1

Alleviate dependence on large scale annotation

2. Use language to improve performance in control applications

Score: 7

1. Ghosts chase and try to kill you
2. Collect all the pellets
3. ...

Score: 107
Other Grounding

- **Temporal concepts**
  - *late evening* = after 6pm
  - *fast, slow* = describing rates of change

- **Relations**
  - **Spatial:**
    - *left, on top of, in front of*

- **Functional:**
  - *Jacket:* keeps people warm
  - *Mug:* holds water

- **Size:**
  - Whales are *larger* than lions

- **Focus today:** grounding in images
Grounding in Images

‣ How would you describe this image?

‣ What does the word “spoon” evoke?

the girl is licking the spoon of batter
Winco 0005-03 73/8" Dinner Spoon...

$7.16

wikiHow
How to Hold a Spoon: 13 Steps ...

Indiegogo
Spoon that Elevates Taste ...
More broadly,

- Nouns: objects
- Verbs: actions
- Sentences: whole scenes or things happening

Tasks:

- Object recognition (pick out one most salient object or detect all of them)
- Image captioning: produce a whole sentence for an image
Language-vision Models

The girl is licking the spoon of batter

Image encoder
(CNN, Transformer)

Cross-attention/joint layer

Language encoder
(LSTM, Transformer)

Prediction
Visual Question Answering

“How many horses are in this image?”

Agrawal et al., 2015
(1) Contrastive pre-training

Pepper the aussie pup

Text Encoder

Image Encoder

T_1, T_2, T_3, ..., T_N

I_1, I_2, I_3, ..., I_N

Radford et al., 2021
Language-vision Pre-training

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>…</th>
<th>$T_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$</td>
<td>$I_1 \cdot T_1$</td>
<td>$I_1 \cdot T_2$</td>
<td>$I_1 \cdot T_3$</td>
<td>…</td>
<td>$I_1 \cdot T_N$</td>
</tr>
<tr>
<td>$I_2$</td>
<td>$I_2 \cdot T_1$</td>
<td>$I_2 \cdot T_2$</td>
<td>$I_2 \cdot T_3$</td>
<td>…</td>
<td>$I_2 \cdot T_N$</td>
</tr>
<tr>
<td>$I_3$</td>
<td>$I_3 \cdot T_1$</td>
<td>$I_3 \cdot T_2$</td>
<td>$I_3 \cdot T_3$</td>
<td>…</td>
<td>$I_3 \cdot T_N$</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>$I_N$</td>
<td>$I_N \cdot T_1$</td>
<td>$I_N \cdot T_2$</td>
<td>$I_N \cdot T_3$</td>
<td>…</td>
<td>$I_N \cdot T_N$</td>
</tr>
</tbody>
</table>

- Contrastive objective: each image should be more similar to its corresponding caption than to other captions

  maximize $\text{softmax}(I_1^T T_i)[1]$
  + $\text{softmax}(I_2^T T_i)[2]$
  + …

Radford et al., 2021
Language-vision Pre-training

(2) Create dataset classifier from label text

(3) Use for zero-shot prediction

Radford et al., 2021
CLIP: Zero-shot Results

Stanford Cars

correct label: 2012 Honda Accord Coupe  correct rank: 1/196  correct probability: 63.30%

- a photo of a 2012 honda accord coupe.
- a photo of a 2012 honda accord sedan.
- a photo of a 2012 acura tl sedan.
- a photo of a 2012 acura tsx sedan.
- a photo of a 2008 acura tl type-s.
CLIP: Zero-shot Results

**Country211**

- correct label: Belize
- correct rank: 5/211
- correct probability: 3.92%

- a photo I took in French Guiana.
- a photo I took in Gabon.
- a photo I took in Cambodia.
- a photo I took in Guyana.
- a photo I took in Belize.
• Autoregressive text-to-image model (differs from the diffusion models you may have seen, like Stable Diffusion or DALL-E)
Two dogs running in a field

Yu et al., 2022
Where are we today

- Explosion of multimodal pre-training for \{video, audio, images, text\}
- Many of these methods are Transformer-based
- Still haven’t seen large-scale pre-training of this form advance text-only tasks, but there’s potential!
Takeaways

‣ Cross-lingual methods allow us to transfer resources from English to other languages

‣ Multilingual models can be learned in a bitext-free way and can transfer between languages

‣ Multimodal methods can allow us to learn representations for images as well as text and provide a path towards language grounding

‣ Next time: wrapup + discussion of ethics