CS388: Natural Language Processing

Lecture 26: Multilingual, Multimodal Models

Greg Durrett

TEXAS
The University of Texas at Austin

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?

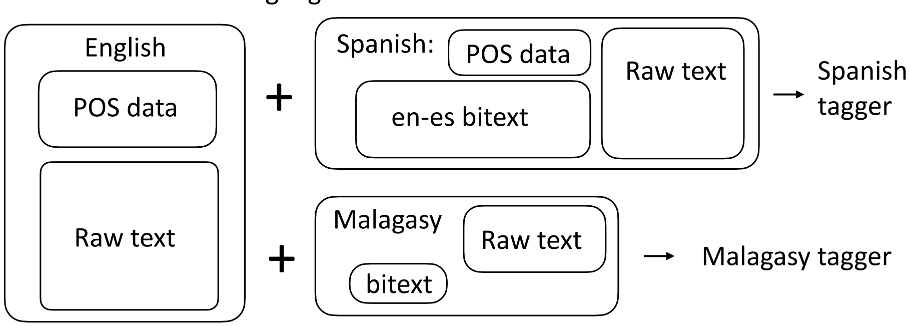
This Lecture

- Cross-lingual tagging and parsing
- Multilingual pre-training

Cross-Lingual Tagging and Parsing

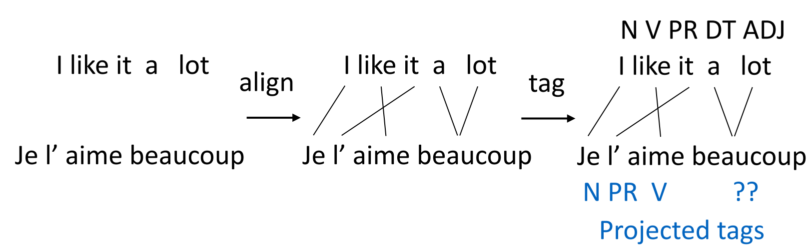
Cross-Lingual Tagging

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



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

PRON VERB NOUN
I like tomatoes

PRON VERB PRON
I like them

train

VERB is the head of PRON and NOUN

Parser trained to accept tag input

parse new data

PRON PRON VERB
Je les aime

McDonald et al. (2011)

Cross-Lingual Parsing

	best-source		avg-source	gold-POS	
	source	gold-POS	gold-POS	multi-dir.	multi-proj.
da	it	48.6	46.3	48.9	49.5
de	nl	55.8	48.9	56.7	56.6
el	en	63.9	51.7	60.1	65.1
es	it	68.4	53.2	64.2	64.5
it	pt	69.1	58.5	64.1	65.0
nl	el	62.1	49.9	55.8	65.7
pt	it	74.8	61.6	74.0	75.6
sv	pt	66.8	54.8	65.3	68.0
avg		63.7	51.6	61.1	63.8

- 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

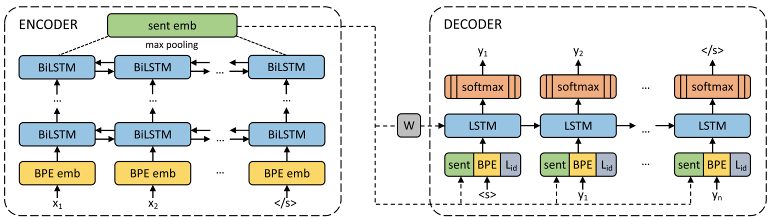
ID: 24
ai have

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)

Multilingual 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

		EN	fr	es	de	el	bg	ru
Zero-Shot Transfer, one NLI system for all languages:								
Conneau et al.	X-BiLSTM	73.7	67.7	68.7	67.7	68.9	67.9	65.4
(2018b)	X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8
BERT uncased*	Transformer	81.4	–	74.3	70.5	–	–	–
Proposed method	BiLSTM	73.9	**71.9**	72.9	72.6	**72.8**	**74.2**	**72.1**

- 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. 中華人民共和國, упр. 中华人民共和国, пиньинь: Zhōnghuá Rénmín Gònghéguó, палл.: Чжунхуа Жэньминь Гунхэго) — государство в Восточной Аз

Devlin et al. (2019)

Multilingual BERT: Results

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

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

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

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

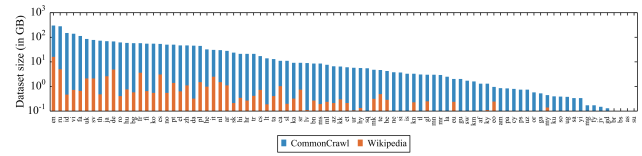
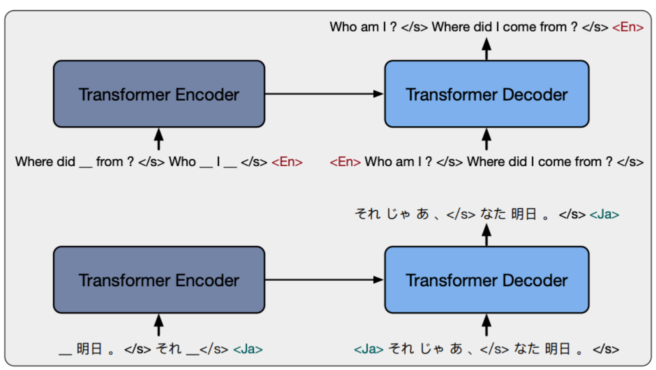


Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

- 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

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction
	MLQA			4,517–11,590	translations	7	Span extraction
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval

- 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
how far Uranus-Sg.Nom from
Земл-и?
Earth-Sg.Gen?
How far is Uranus from Earth?

A: Расстояние между Уран-ом
distance between Uranus-Sg.Instr
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

Sequence: 菊池は アメリカ大リーグ への参戦も視野に進路が注目されていたが、10月25日に日本のプロ野球に挑戦することを表明していた。...

Translation: Kikuchi was considering Major League Baseball as his next career, but he announced that he would play professional baseball in Japan ...

Predictions: *baseball, established, establishments, in the united states, organizations, sports*

Gold Types: *baseball, baseball leagues in the united states, bodies, established, establishments, events, in canada, in the united states, major league baseball, multi-national professional sports leagues, organizations, professional, sporting, sports...*

Precision: 100% **Recall**: 31.6%

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

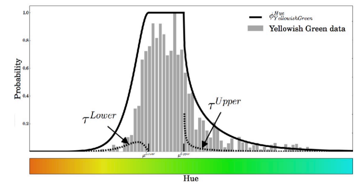
some slides from Eunsol Choi

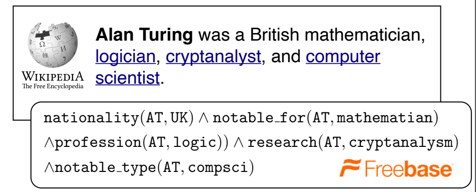
Language Grounding

- 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

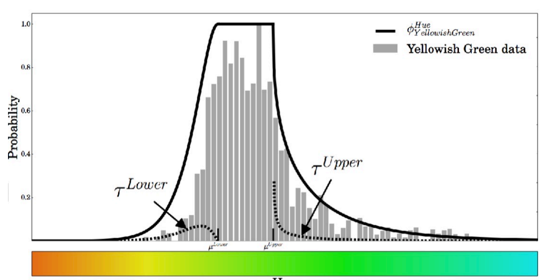
- How to associate words with sensory-motor experiences
- How to associate words with meaning representation





Language Grounding

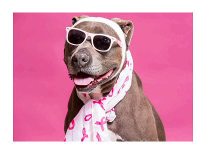
- What does "yellowish green" mean?
- Formal semantics: yellowish green is a predicate. Things are either yellowish green or not. No connection to real color
- Grounding in perceptual space:



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

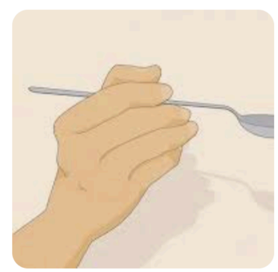
- How would you describe this image?
- What does the word "*spoon*" evoke?

the girl is licking the spoon of batter

Grounding Spoon

Winco 0005-03 7 3/8" Dinner Spoon...

$7.16



wikiHow

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

Indiegogo

Spoon that Elevates Taste ...

Grounding Language in Images

- 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

Image encoder (CNN, Transformer)

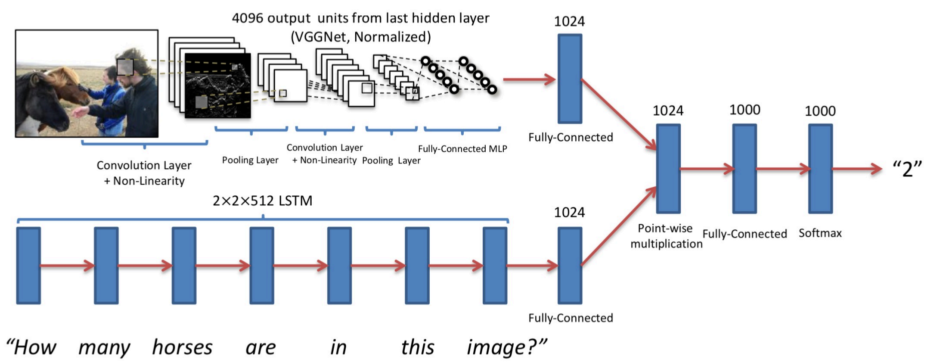
the girl is licking the spoon of batter

Language encoder (LSTM, Transformer)

Cross-attention/joint layer

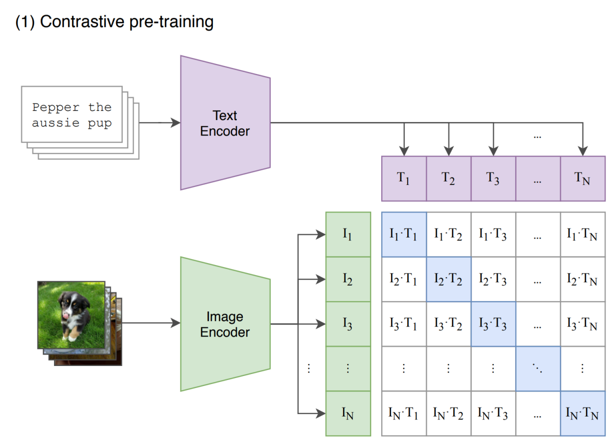
Prediction

Visual Question Answering



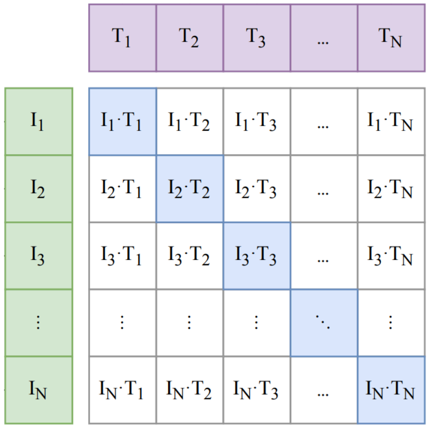
Agrawal et al., 2015

Language-vision Pre-training



Radford et al., 2021

Language-vision Pre-training

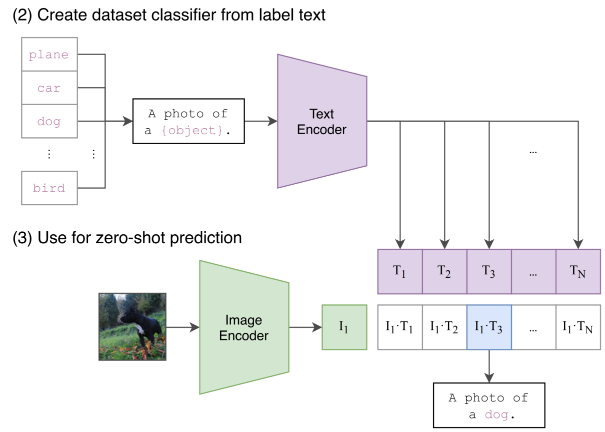


- Contrastive objective: each image should be more similar to its correspond caption than to other captions

maximize softmax($I_1^T T_i$)[1] + softmax($I_2^T T_i$)[2] + ...

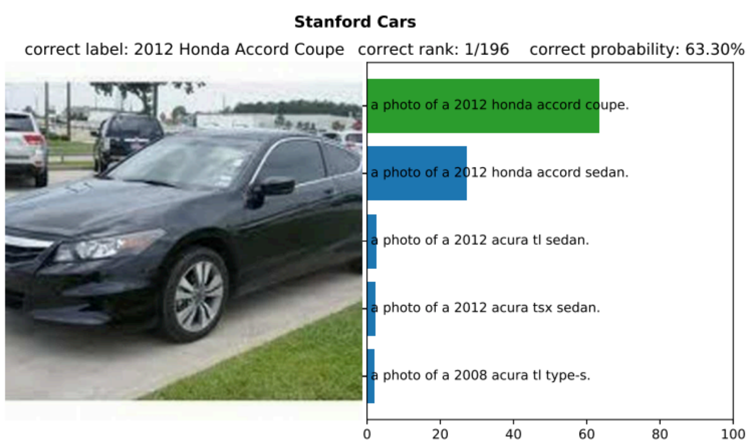
Radford et al., 2021

Language-vision Pre-training

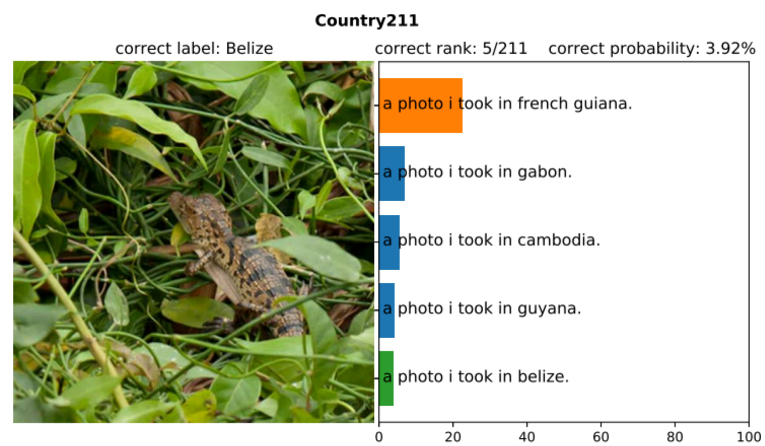


Radford et al., 2021

CLIP: Zero-shot Results



CLIP: Zero-shot Results



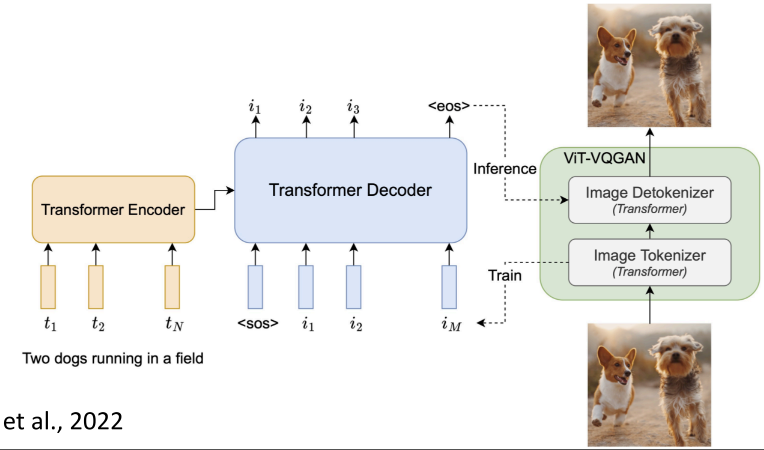
Parti

- Autoregressive text-to-image model (differs from the diffusion models you may have seen, like Stable Diffusion or DALL-E)

A. *A photo of a frog reading the newspaper named "Toaday" written on it. There is a frog printed on the newspaper too.*

Yu et al., 2022

Parti



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