# 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

# 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)

  - May not be totally regular: enflame => inflammable
- Inflectional morphology: word is inflected based on its context

  I become / she become**s**

  - 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:

| | | singular | | | plural | | |
|---|---|---|---|---|---|---|---|
| | | first | second | third | first | second | third |
| indicative | | je (j') | tu | il, elle | nous | vous | ils, elles |
| (simple tenses) | present | arrive /a.ʁiv/ | arrives /a.ʁiv/ | arrive /a.ʁiv/ | arrivons /a.ʁi.vɔ̃/ | arrivez /a.ʁi.ve/ | arrivent /a.ʁiv/ |
| | imperfect | arrivais /a.ʁi.vɛ/ | arrivais /a.ʁi.vɛ/ | arrivait /a.ʁi.vɛ/ | arrivions /a.ʁi.vjɔ̃/ | arriviez /a.ʁi.vje/ | arrivaient /a.ʁi.vɛ/ |
| | past historic[2] | arrivai /a.ʁi.vɛ/ | arrivas /a.ʁi.va/ | arriva /a.ʁi.va/ | arrivâmes /a.ʁi.vam/ | arrivâtes /a.ʁi.vat/ | arrivèrent /a.ʁi.vɛʁ/ |
| | future | arriverai /a.ʁi.vʁɛ/ | arriveras /a.ʁi.vʁa/ | arrivera /a.ʁi.vʁa/ | arriverons /a.ʁi.vʁɔ̃/ | arriverez /a.ʁi.vʁe/ | arriveront /a.ʁi.vʁɔ̃/ |
| | conditional | arriverais /a.ʁi.vʁɛ/ | arriverais /a.ʁi.vʁɛ/ | arriverait /a.ʁi.vʁɛ/ | arriverions /a.ʁi.və.ʁjɔ̃/ | arriveriez /a.ʁi.və.ʁje/ | arriveraient /a.ʁi.vʁɛ/ |

# Morphological Inflection

- In Spanish:

| | | singular | | | plural | | |
|---|---|---|---|---|---|---|---|
| | | 1st person | 2nd person | 3rd person | 1st person | 2nd person | 3rd person |
| indicative | | yo | tú<br>vos | él/ella/ello<br>usted | nosotros<br>nosotras | vosotros<br>vosotras | ellos/ellas<br>ustedes |
| | present | llego | llegas[tú]<br>llegás[vos] | llega | llegamos | llegáis | llegan |
| | imperfect | llegaba | llegabas | llegaba | llegábamos | llegabais | llegaban |
| | preterite | llegué | llegaste | llegó | llegamos | llegasteis | llegaron |
| | future | llegaré | llegarás | llegará | llegaremos | llegaréis | llegarán |
| | conditional | llegaría | llegarías | llegaría | llegaríamos | llegaríais | llegarían |

# Noun Inflection

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

Declension of Kind [hide ▲]

| | singular | | | plural | |
|---|---|---|---|---|---|
| | indef. | def. | noun | def. | noun |
| nominative | ein | das | Kind | die | Kinder |
| genitive | eines | des | Kindes,<br>Kinds | der | Kinder |
| dative | einem | dem | Kind,<br>Kinde[1] | den | Kindern |
| accusative | ein | das | Kind | die | Kinder |

- 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

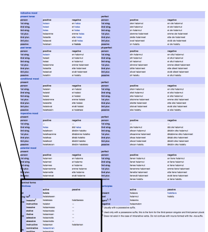
# Irregular Inflection

- 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 Langauges

- Finnish/Hungarian (Finno-Ugric), also Turkish: what a preposition would do in English is instead part of the verb

| | | active | passive |
|---|---|---|---|
| 1st | | halata | |
| long 1st[2] | | halatakseen | |
| 2nd | inessive[1] | halatessa | halattaessa |
| | instructive | halaten | — |
| 3rd | inessive | halaamassa | — |
| | elative | halaamasta | — |
| | illative | halaamaan | — |
| | adessive | halaamalla | — |
| | abessive | halaamatta | — |
| | instructive | halaaman | halattaman |
| 4th | nominative | halaaminen | |
| | partitive | halaamista | |
| 5th[2] | | halaamaisillaan | |



halata: "hug"

illative: "into"

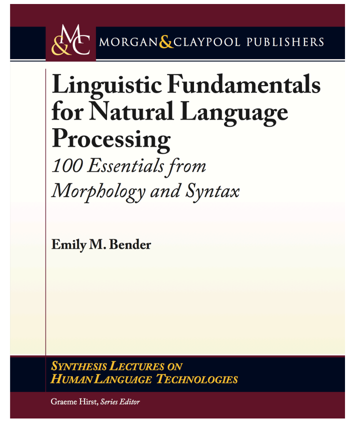
adessive: "on"

- 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
  - SPMRL shared tasks (2013-2014): Syntactic Parsing of Morphologically-Rich Languages
- Word piece / byte-pair encoding models for MT are pretty good at handling these if there's enough data

## Morphologically-Rich Languages




- Great resources for challenging your assumptions about language and for understanding multilingual models!

## Morphological Analysis/Inflection

## 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

w i n d e n →

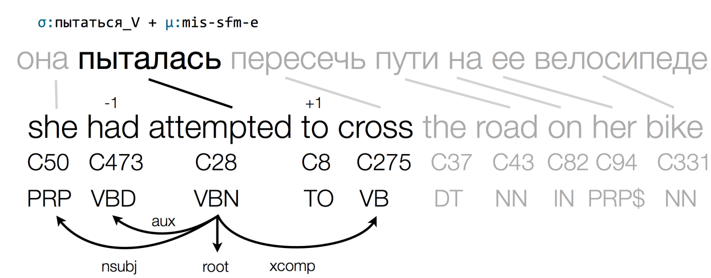
conjugation of winden [hide ▲]

| | | | | | |
|---|---|---|---|---|---|
| infinitive | winden | | | | |
| present participle | windend | | | | |
| past participle | gewunden | | | | |
| auxiliary | haben | | | | |
| | indicative | | | subjunctive | |
| present | ich winde | wir **winden** | i | ich winde | wir **winden** |
| | du windest | ihr windet | | du windest | ihr windet |
| | er windet | sie **winden** | | er winde | sie **winden** |
| preterite | ich wand | wir wanden | ii | ich wände | wir wänden |
| | du wandest | ihr wandet | | du wändest | ihr wändet |
| | er wand | sie wanden | | er wände | sie wänden |
| imperative | winde (du) | windet (ihr) | | | |

composed forms of winden [show ▼]

Durrett and DeNero (2013)

# Morphological Inflection




- 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

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".

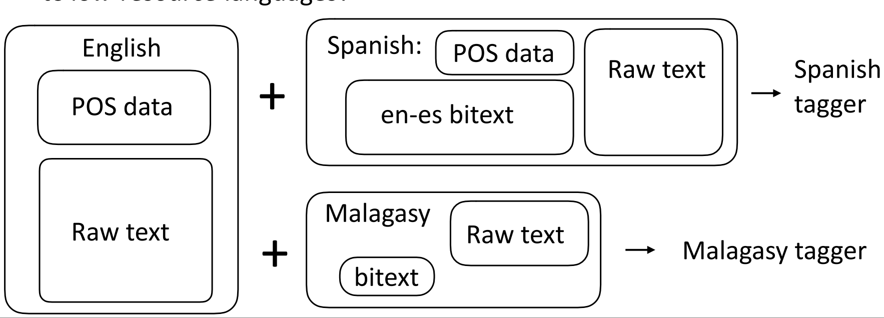
- separating nouns and pre-modifying adjectives:
  高血压 (*high blood pressure*)
  → 高(*high*) 血压(*blood pressure*)
- separating compound nouns:
  内政部 (*Department of Internal Affairs*)
  → 内政(*Internal Affairs*) 部(*Department*).

Chen et al. (2015)

## Cross-Lingual Tagging and Parsing

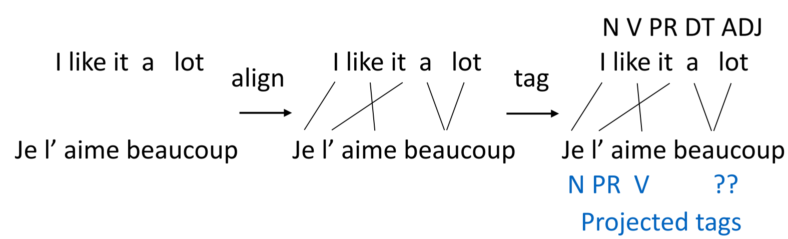
## Cross-Lingual Tagging

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




## Cross-Lingual Tagging

- Can we leverage word alignment here?

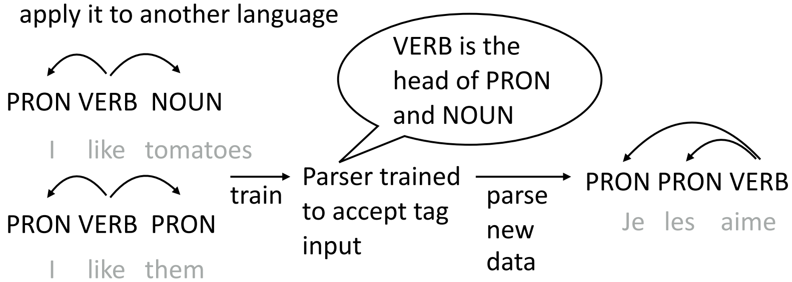



- Tag with English tagger, project across bitext, train French tagger? Works pretty well

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




McDonald et al. (2011)

# Cross-Lingual Parsing

| | best-source | | avg-source | gold-POS | | pred-POS | |
|---|---|---|---|---|---|---|---|
| | source | gold-POS | gold-POS | multi-dir. | multi-proj. | multi-dir. | multi-proj. |
| da | it | 48.6 | 46.3 | 48.9 | 49.5 | 46.2 | 47.5 |
| de | nl | 55.8 | 48.9 | 56.7 | 56.6 | 51.7 | 52.0 |
| el | en | 63.9 | 51.7 | 60.1 | 65.1 | 58.5 | 63.0 |
| es | it | 68.4 | 53.2 | 64.2 | 64.5 | 55.6 | 56.5 |
| it | pt | 69.1 | 58.5 | 64.1 | 65.0 | 56.8 | 58.9 |
| nl | el | 62.1 | 49.9 | 55.8 | 65.7 | 54.3 | 64.4 |
| pt | it | 74.8 | 61.6 | 74.0 | 75.6 | 67.7 | 70.3 |
| sv | pt | 66.8 | 54.8 | 65.3 | 68.0 | 58.3 | 62.1 |
| avg | | 63.7 | 51.6 | 61.1 | 63.8 | 56.1 | 59.3 |

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

McDonald et al. (2011)

# Cross-Lingual 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

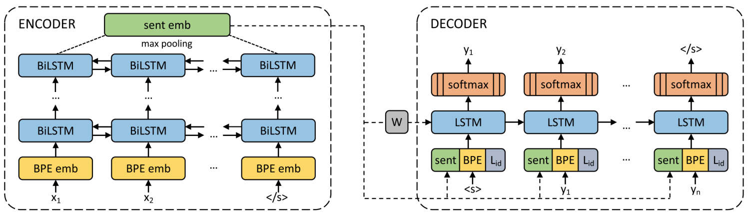
ID: 24
ai have

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 | EN → XX | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur |
| **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 | 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 | 58.8 | 56.3 | 50.4 | 52.2 |
| BERT uncased* | Transformer | 81.4 | – | 74.3 | 70.5 | – | – | – | – | 62.1 | – | – | 63.8 | – | – | 58.3 |
| Proposed method | BiLSTM | 73.9 | **71.9** | 72.9 | 72.6 | **72.8** | **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)

# 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

Devlin et al. (2019)

# Multilingual BERT: Results

| Fine-tuning \ Eval | EN | DE | NL | ES |
|---|---|---|---|---|
| EN | **90.70** | 69.74 | 77.36 | 73.59 |
| DE | 73.83 | **82.00** | 76.25 | 70.03 |
| NL | 65.46 | 65.68 | **89.86** | 72.10 |
| ES | 65.38 | 59.40 | 64.39 | **87.18** |

Table 1: NER F1 results on the CoNLL data.

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

Pires et al. (2019)

# Scaling Up: XLM-R

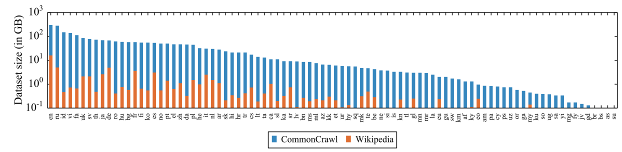



Figure 1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages.

- 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)

# 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