Neural models as typologists

Can neural NLP models discover language generalizations?

Robert Östling

Institutionen för lingvistik
Stockholms universitet

2023-10-12

Language representations

Traditional Linguistic Typology

Painstaking collection of linguistic data and careful research on generalizations of language structure.

The Age of `whatever2vec`

Feed low-grade data to the machine and sift through its excrements.

Long overdue project

Step 1: Helsinki 2016

We (Östling & Tiedemann 2017) trained a character-level LSTM LM with language embeddings on 1k languages

Step 2: COVID + child no. 3

"I'll just finish this up" (yeah, right...)

Step 3: publication

Östling & Kurfalı (2023) — finally

The Question

If we give a massively multilingual neural model a small* per-language parameter space, how does its language encodings align with established typological generalizations?

*(*Small* = 100 dimensions, on par with the number of features in WALS or Grambank)

The Problem

Why is this question difficult to answer?

1. There are many types of neural models (obvious, but this turns out to be important)

The Problem

Why is this question difficult to answer?

1. There are many types of neural models (obvious, but this turns out to be important)
2. "Gold standards" are pretty noisy
   - Skirgård et al. (2023): Grambank inter-coder reliability study, out of 7876 pairwise double-annotated codings...
     - 48%: identical labels (very good)
     - 20%: both agree that there is insufficient data to make a decision (good)
     - 25%: disagreement on whether there is enough data (bad)
     - 7%: different labels (very bad)
   - Error analysis in typological feature prediction (e.g. Östling & Wälchli 2019) reveal database errors as a major contributor

The Problem

Why is this question difficult to answer?

1. There are many types of neural models (obvious, but this turns out to be important)
2. "Gold standards" are pretty noisy
   - Skirgård et al. (2023): Grambank inter-coder reliability study, out of 7876 pairwise double-annotated codings...
     - 48%: identical labels (very good)
     - 20%: both agree that there is insufficient data to make a decision (good)
     - 25%: disagreement on whether there is enough data (bad)
     - 7%: different labels (very bad)
   - Error analysis in typological feature prediction (e.g. Östling & Wälchli 2019) reveal database errors as a major contributor
3. Languages are **not** independent samples from a universal parameter space, and the relationships are notoriously difficult to model:
   - Genealogical relationships (imperfectly documented)
   - Language contact (even more imperfectly documented)

The Data

- Bible translations (surprise!)

The Data

- Bible translations (surprise!)
- Full corpus: 1,846 translations in 1,401 languages
- Good enough: 1,707 translations in 1,299 languages
- Projection targets: 1,664 translations in 1,295 languages

The Data

- Bible translations (surprise!)
- Full corpus: 1,846 translations in 1,401 languages
- Good enough: 1,707 translations in 1,299 languages
- Projection targets: 1,664 translations in 1,295 languages
- Did we consider that the Bible as a corpus has numerous problems? Yes! But look at the distribution of language families...

Why Bibles?

Macro-area	Bible	mT5
North America	17	0
South America	39	0
Eurasia	19	13
Africa	15	2
Papunesia	36	1
Australia	6	0
Total	132	16

Number of *language families* per linguistic macro-area (Glottolog)

The Models

A number of 1664-doculect (1295 ISO languages) models with language embeddings and joint multilingual training:

- Word-level language model (details follow)
- Character-level language model (character-level LSTM)
- Morphological reinflection model (OpenNMT, LSTM + attention)
- Word form encoder (LSTM)
- English-to-X NMT (OpenNMT, LSTM + attention)
- X-to-English NMT (OpenNMT, LSTM + attention)

Word-level LM

- Unusual beast, not meant to be a practical LM

Word-level LM

- Unusual beast, not meant to be a practical LM
- Using corpus tokenization, no subwords: 18M vocabulary

Word-level LM

- Unusual beast, not meant to be a practical LM
- Using corpus tokenization, no subwords: 18M vocabulary
- Fixed multilingual word embeddings obtained through multi-source projection from 32 aligned high-resource language embeddings

Word-level LM

- Unusual beast, not meant to be a practical LM
- Using corpus tokenization, no subwords: 18M vocabulary
- Fixed multilingual word embeddings obtained through multi-source projection from 32 aligned high-resource language embeddings
- Cosine loss function

Word-level LM

- Unusual beast, not meant to be a practical LM
- Using corpus tokenization, no subwords: 18M vocabulary
- Fixed multilingual word embeddings obtained through multi-source projection from 32 aligned high-resource language embeddings
- Cosine loss function
- 512-dimensional LSTM model (but note the fixed 18M $\times$ 300 = 5.4B fixed embedding parameters!)

Projections

We perform co-occurrence based word alignment and project several different annotations through the 1664 translations:

- Concept labels (source: Intercontinental Dictionary Series)
- Universal POS tags (source: Turku NLP pipeline)
- Universal Dependencies relations (source: Turku NLP pipeline)
- Multilingual word embeddings (source: MUSE)

Projections

We perform co-occurrence based word alignment and project several different annotations through the 1664 translations:

- Concept labels (source: Intercontinental Dictionary Series)
- Universal POS tags (source: Turku NLP pipeline)
- Universal Dependencies relations (source: Turku NLP pipeline)
- Multilingual word embeddings (source: MUSE)

...which are used to compute:

- Noun and verb inflectional paradigms (POS tags + concept labels)
- Affix lists, including dominant affix position (paradigms)
- Word order statistics (dependencies + concept labels)
- Word lists (concept labels)

Baseline representations

These should...

- Mirror the complex correlations between natural languages, due to to genealogical and contact relationships
- *Not* be influenced by grammatical or morphological structure

Baseline representations

These should...

- Mirror the complex correlations between natural languages, due to to genealogical and contact relationships
- *Not* be influenced by grammatical or morphological structure

The best approximation we could come up with are *lexical* language representations:

1. Take a word list (ASJP, or our own projected ones)
2. Compute pairwise normalized Levenshtein distance between corresponding words, for all pairs of languages
3. Reduce the distance matrix to 100 columns (SVD)

Evaluation

- Logistic regression classifier language vector $\rightarrow$ feature value
- But what exactly are we interested in measuring?

Evaluation

- Logistic regression classifier language vector $\rightarrow$ feature value
- But what exactly are we interested in measuring?
- Good language vectors should be able to predict properties of a language isolate discovered tomorrow

Evaluation

- Logistic regression classifier language vector $\rightarrow$ feature value
- But what exactly are we interested in measuring?
- Good language vectors should be able to predict properties of a language isolate discovered tomorrow
- Leave-one-out cross validation with a world of simulated isolates. Training set constraints:
  1. Avoid test language family (Glottolog)
  2. Avoid test language macro-area (Glottolog)
  3. Avoid test language long-distance contact (SegBo)
  4. One representative per family
- Family-weighted F_1 is the most relevant metric

Labels

- We have two sets of typological feature labels:
  1. URIEL (sourced from WALS + Ethnologue)
  2. The ones we projected

Labels

- We have two sets of typological feature labels:
  1. URIEL (sourced from WALS + Ethnologue)
  2. The ones we projected
- Projected labels have a larger coverage than URIEL, so we can use them for classifier *training*

Labels

- We have two sets of typological feature labels:
  1. URIEL (sourced from WALS + Ethnologue)
  2. The ones we projected
- Projected labels have a larger coverage than URIEL, so we can use them for classifier *training*
- Projected labels are derived from the training data, so only URIEL labels are used for evaluation
- Not unproblematic: URIEL contains coding errors and the underlying reference grammars might describe a different language variety than the Bible translation

Points of comparison

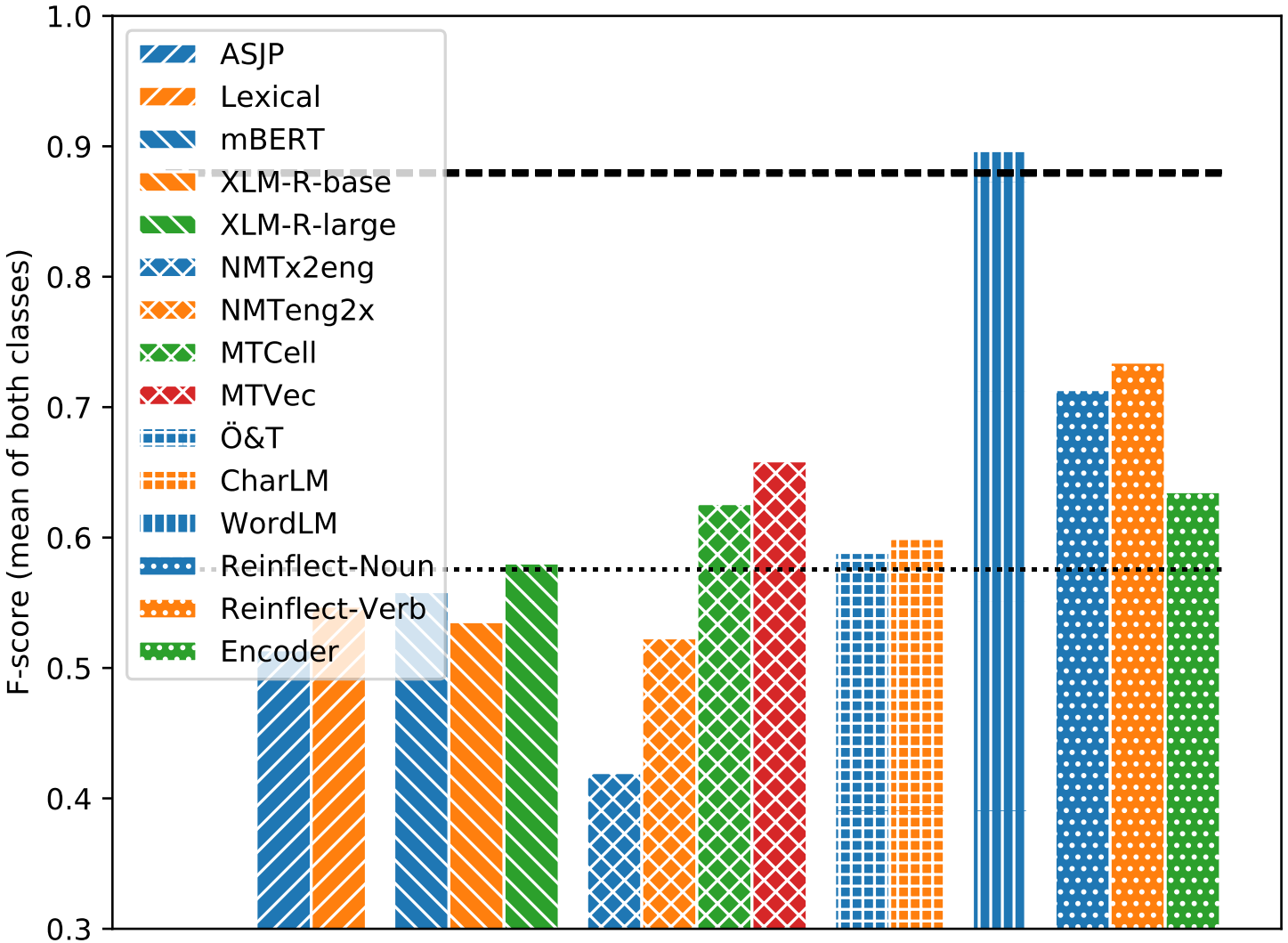
"Upper bound"

Performance of the projection-based method with respect to URIEL. Not a true upper bound since the projection may work poorly for some features, but usually close to the level of human agreement.

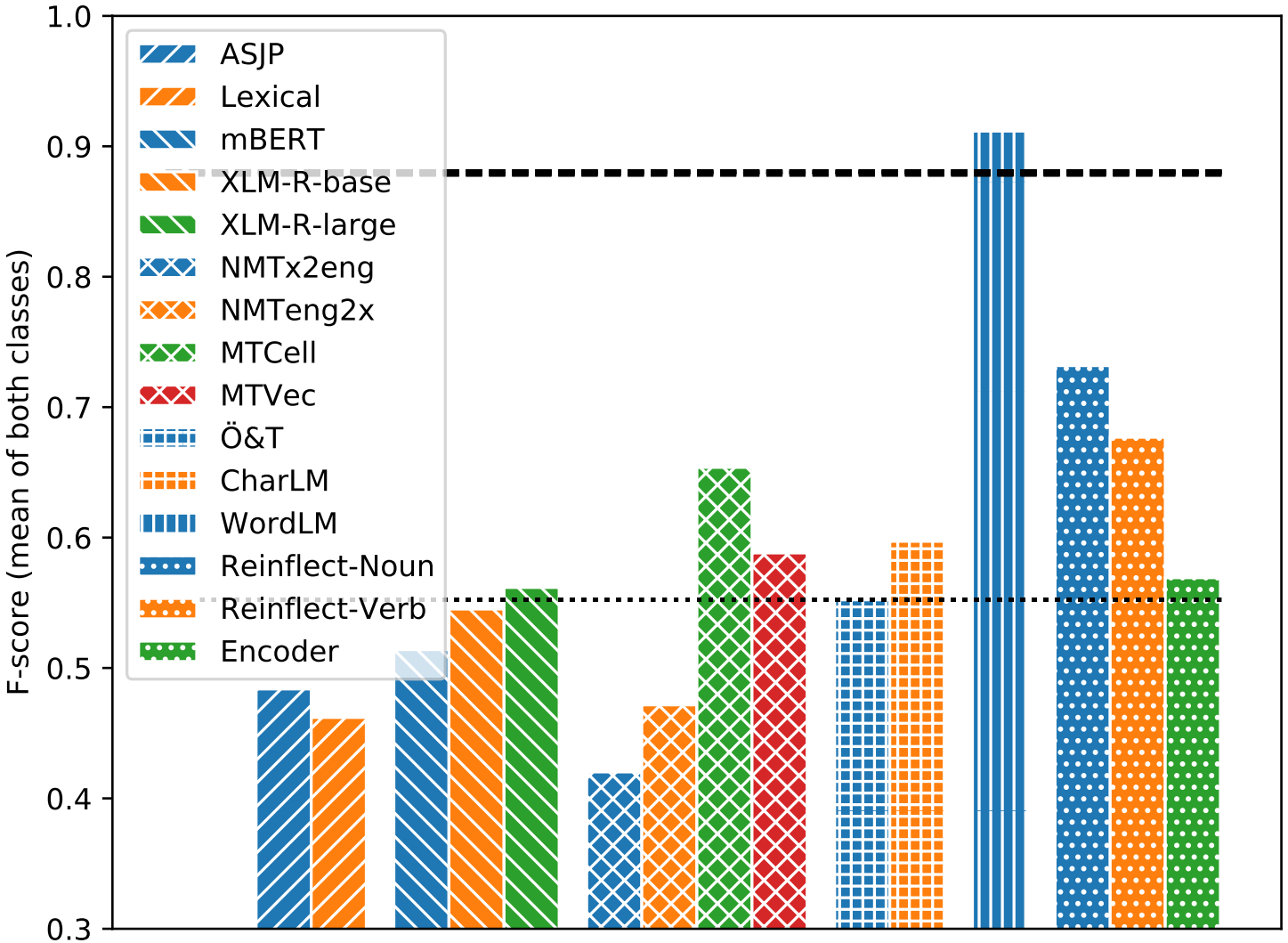
"Lower bound"

99th percentile in classifications with real data but shuffled labels. This model seems to somewhat underestimate the variance. (Using noise to guess at majority class?)

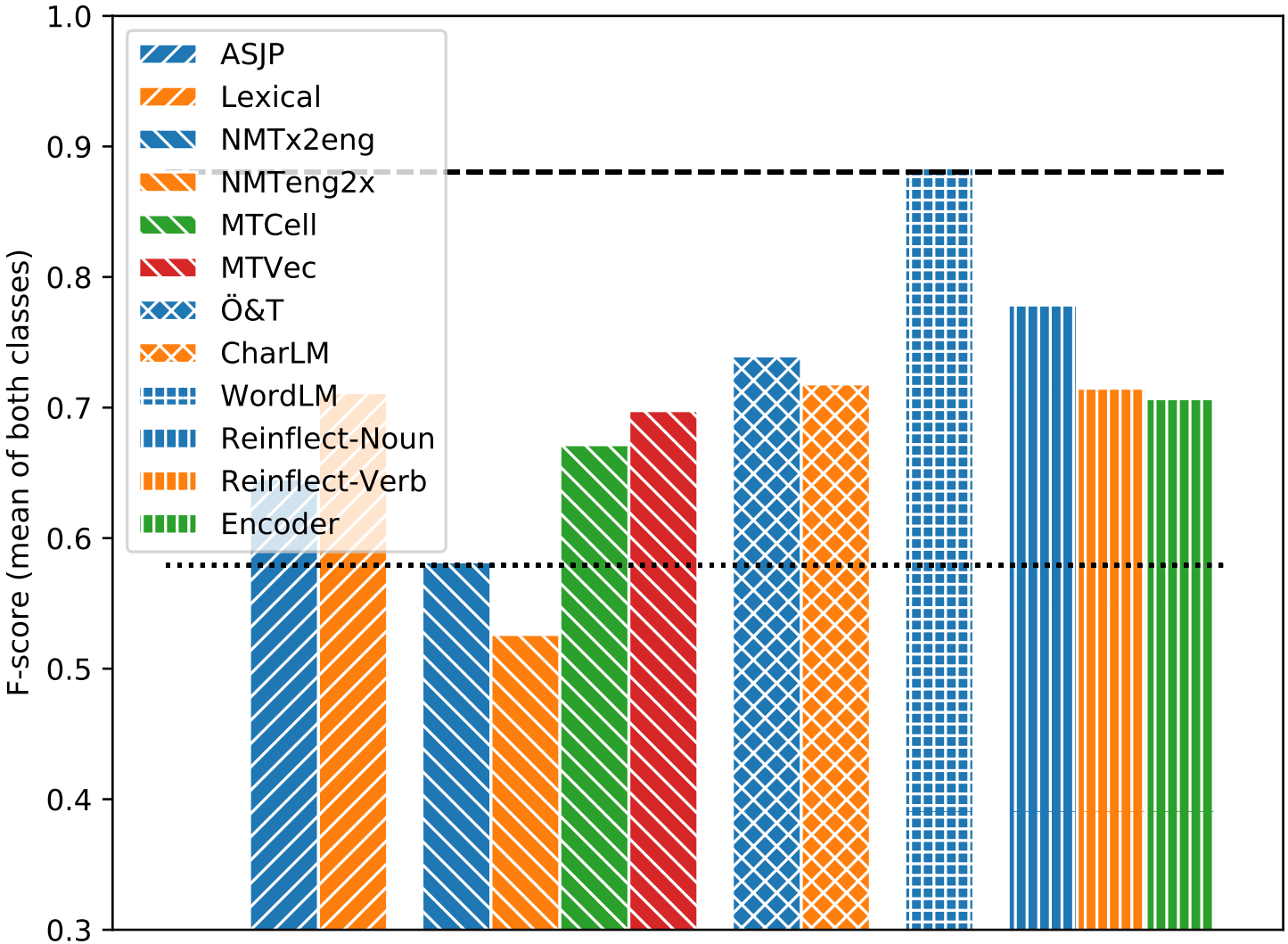
Object/verb order (trained on: URIEL)



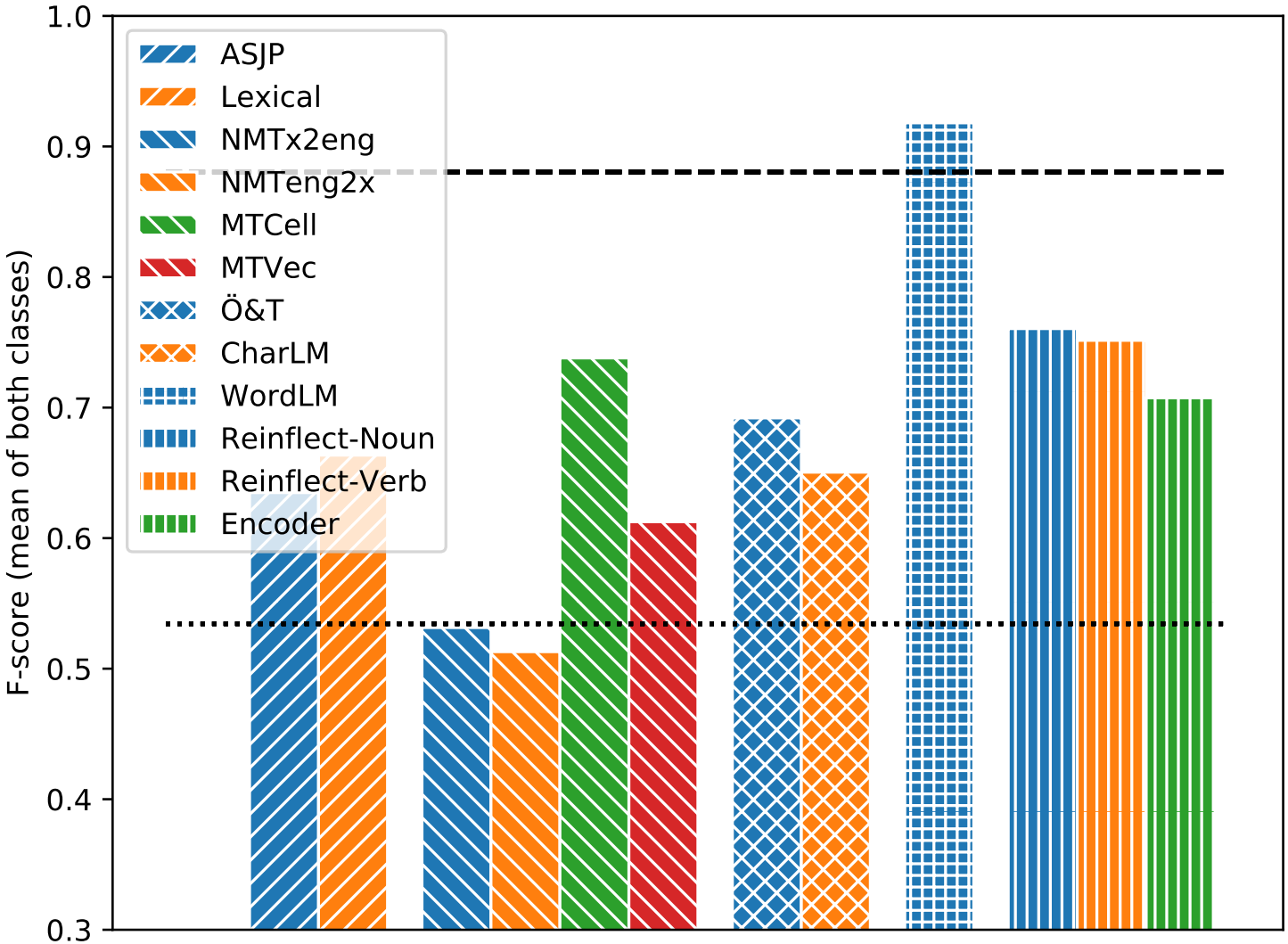
Object/verb order (trained on: projected)



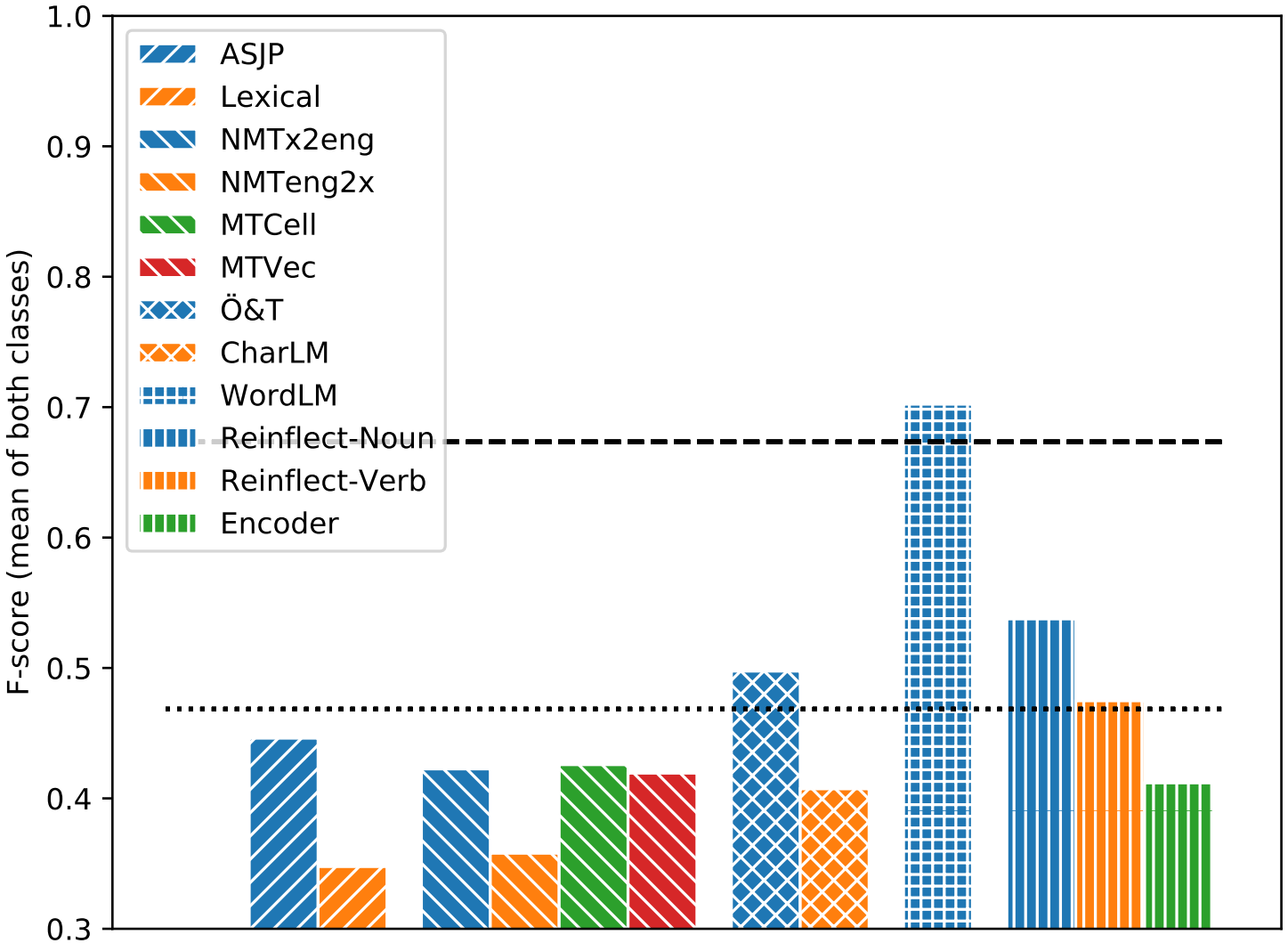
Object/verb order (naive cross-validation, URIEL)



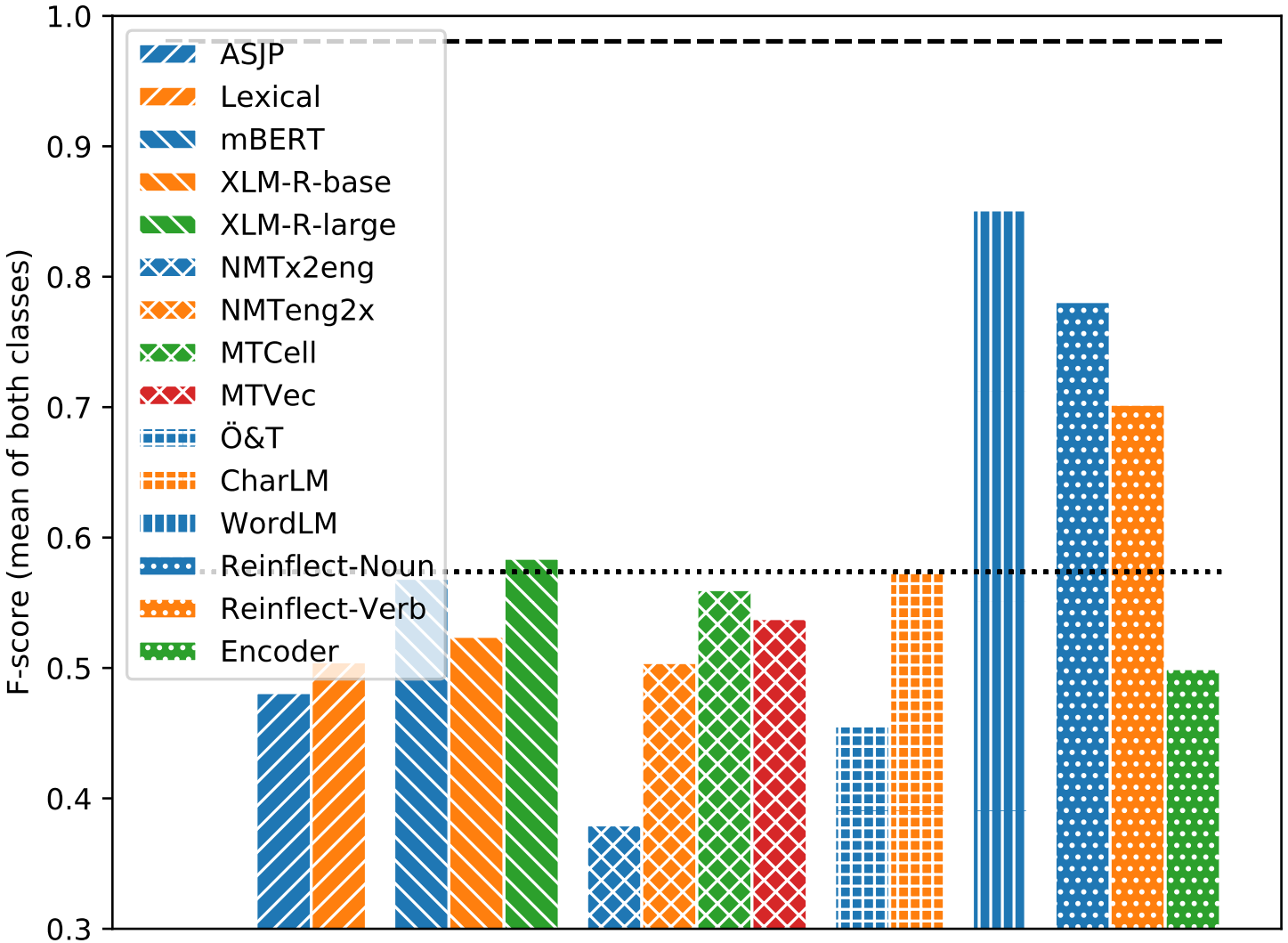
Object/verb order (naive cross-validation, projected)



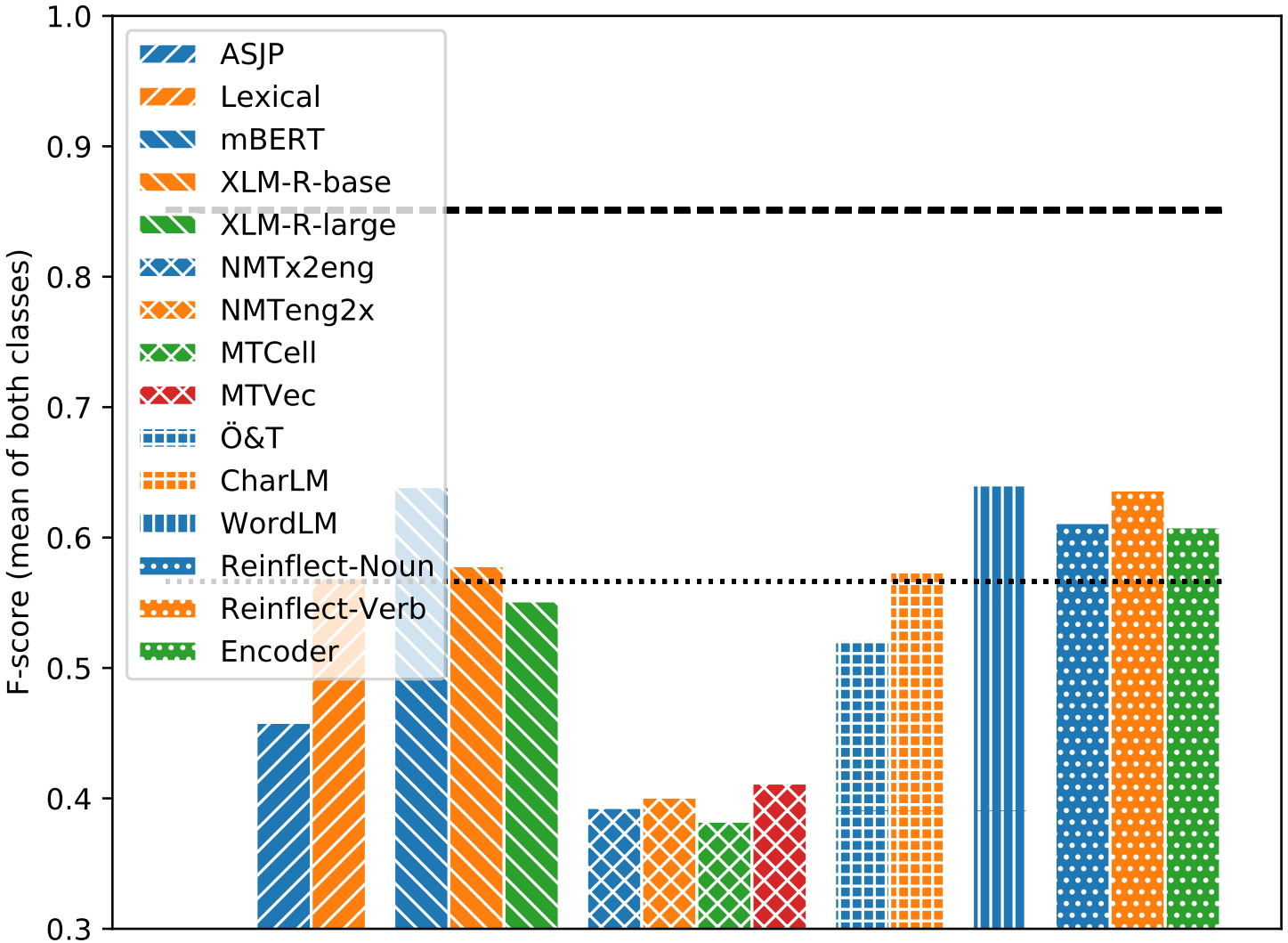
Subject/verb order (trained on: URIEL)



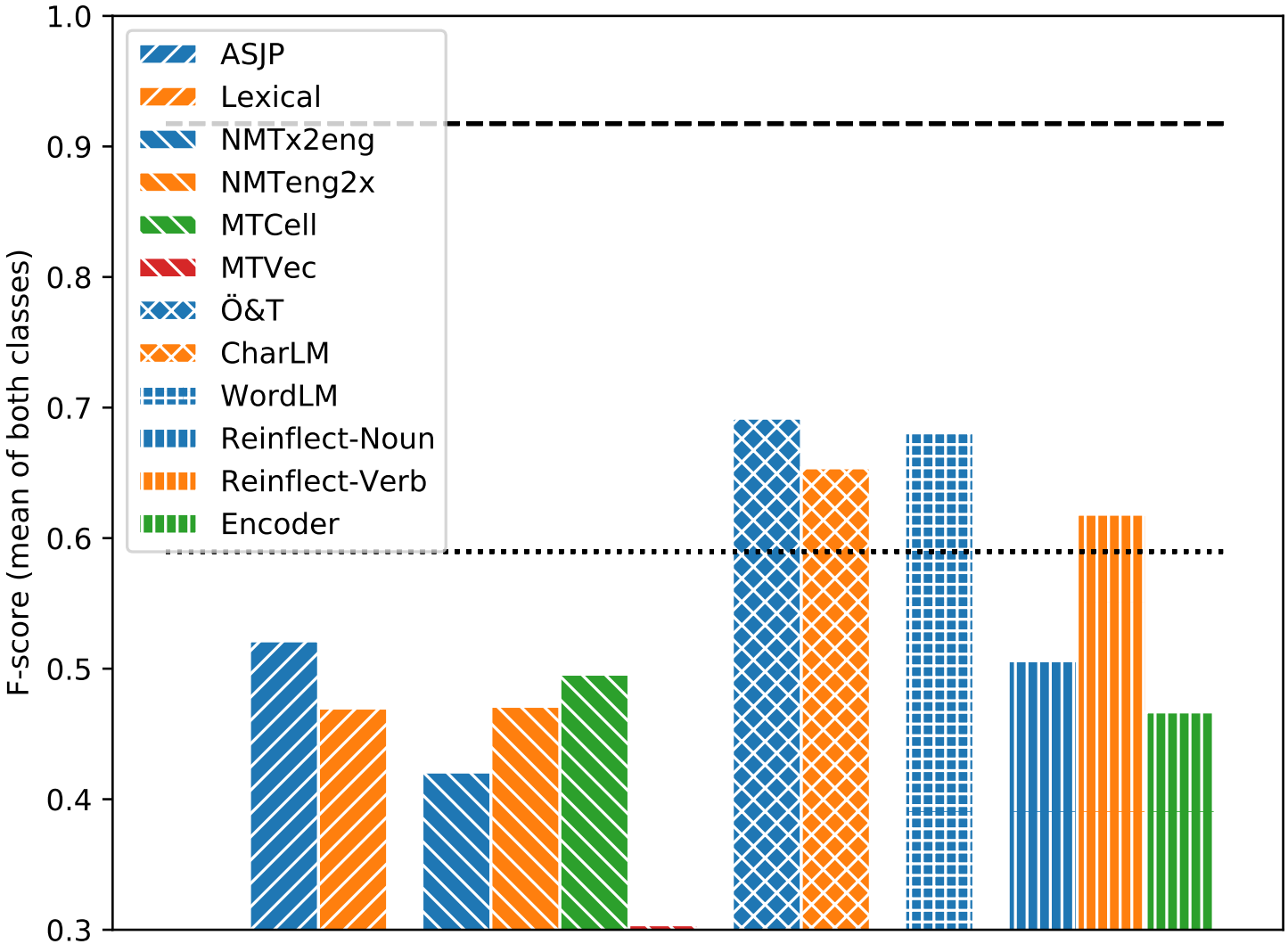
Prepositions/postpositions (trained on: URIEL)



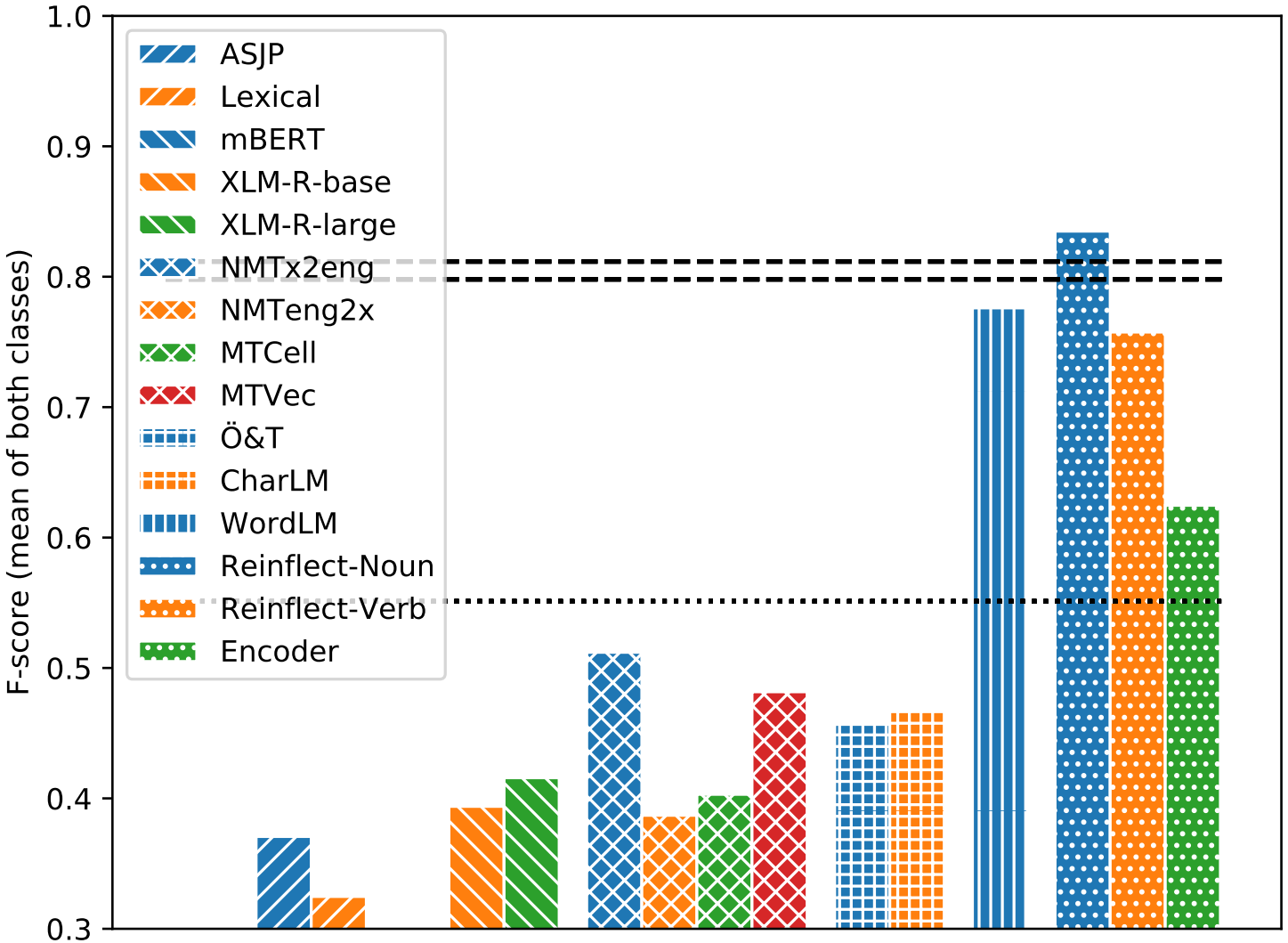
Adjective/noun order (trained on: URIEL)



Numeral/noun order (trained on: URIEL)



Prefix/suffix (trained on: URIEL)



Conclusions

Findings:

- Neural models (sometimes) discover typological generalizations
- Strong dependency on which task and model we use
- You need a diverse enough dataset for this kind of work

Results:

- Paper: https://doi.org/10.1162/coli_a_00491
- Data: https://zenodo.org/record/7506220
- Code: https://github.com/robertostling/parallel-text-typology

Features are not independent

- Typological features are correlated
- Still controversial to what extent this is due to (mainly) genealogical relationships

Features are not independent

- Typological features are correlated
- Still controversial to what extent this is due to (mainly) genealogical relationships
- Difficult to know *which* of correlated features is detected

Features are not independent

- Typological features are correlated
- Still controversial to what extent this is due to (mainly) genealogical relationships
- Difficult to know *which* of correlated features is detected
- A model trained to predict feature X is also tested on all other features Y, Z, W, ...
- If the F_1 for X is *not* the highest, this indicates that the language representations rather encode some other feature
- We can not exclude the possibility of even better, yet unknown, features

Effect of using projected features

- We normally train and evaluate using URIEL labels, in order to be independent of the LM training data (Bibles)

Effect of using projected features

- We normally train and evaluate using URIEL labels, in order to be independent of the LM training data (Bibles)
- Still interesting to look at the different combinations of train/test labels
- Large disparity would indicate systematic differences between typological databases and Bible texts

Three-way confusion matrix

- Matrix interpretation:
  - URIEL value: upper/lower matrix
  - Projected value: row within matrix
  - Classifier prediction: column
- Models either trained on URIEL or projected labels

OV/VO URIEL

$$\begin{pmatrix} \begin{pmatrix} 54.6 & 0.4 \\ 9.9 & 0.2 \end{pmatrix} \\ \begin{pmatrix} 1.3 & 0 \\ 7.3 & 26.4 \end{pmatrix} \end{pmatrix}$$

OV/VO projected

$$\begin{pmatrix} \begin{pmatrix} 53.5 & 1.5 \\ 8.7 & 1.4 \end{pmatrix} \\ \begin{pmatrix} 1.3 & 0 \\ 3.9 & 29.8 \end{pmatrix} \end{pmatrix}$$

AdpN/NAdp URIEL

$$\begin{pmatrix} \begin{pmatrix} 35.8 & 5.4 \\ 0.1 & 0.0 \end{pmatrix} \\ \begin{pmatrix} 0.0 & 1.8 \\ 5.5 & 51.3 \end{pmatrix} \end{pmatrix}$$

AdpN/NAdp projected

$$\begin{pmatrix} \begin{pmatrix} 37.5 & 4 \\ 0.1 & 0.0 \end{pmatrix} \\ \begin{pmatrix} 0.0 & 1.8 \\ 5.8 & 51.1 \end{pmatrix} \end{pmatrix}$$

Three-way confusion matrix (2)

- Matrix interpretation:
  - URIEL value: upper/lower matrix
  - Projected value: row within matrix
  - Classifier prediction: column
- All models are trained on URIEL labels

RelN/NRel URIEL

$$\begin{pmatrix} \begin{pmatrix} 15.2 & 0.1 \\ 9.5 & 0.0 \end{pmatrix} \\ \begin{pmatrix} 0.0 & 0.0 \\ 29.6 & 45.5 \end{pmatrix} \end{pmatrix}$$

NumN/NNum URIEL

$$\begin{pmatrix} \begin{pmatrix} 44.4 & 10.8 \\ 0.7 & 2.2 \end{pmatrix} \\ \begin{pmatrix} 1.6 & 3.4 \\ 8.5 & 28.3 \end{pmatrix} \end{pmatrix}$$

AdjN/NAdj URIEL

$$\begin{pmatrix} \begin{pmatrix} 29.0 & 4.9 \\ 1.9 & 1.4 \end{pmatrix} \\ \begin{pmatrix} 8.7 & 2.5 \\ 20.8 & 30.7 \end{pmatrix} \end{pmatrix}$$

SV/VS URIEL

$$\begin{pmatrix} \begin{pmatrix} 75.1 & 14.7 \\ 0.0 & 1.1 \end{pmatrix} \\ \begin{pmatrix} 0.4 & 6.1 \\ 0.2 & 2.2 \end{pmatrix} \end{pmatrix}$$

High-confidence labels

Feature	Mean F_1 score All	Proj = Pred
Order of adjective and noun	0.639	0.880
Order of numeral and noun	0.762	0.947
Order of relative clause and noun	0.648	0.999
Order of adposition and noun	0.866	1.000
Order of object and verb	0.896	0.980
Order of subject and verb	0.702	0.865

- Errors seem to be complementary between projected/predicted labels
- We have a useful method to extend typological databases!

  ...for word order features, anyway

Remaining errors

- Adposition/Noun
  - Serbian: apparent mistake in URIEL

Remaining errors

- Adposition/Noun
  - Serbian: apparent mistake in URIEL
- Object/Verb
  - Mbyá Guaraní (Tupian): Ethnologue (VO) partly disagrees with Martins (2004): OV and VO (Projected OV ratio: 0.82)
  - Yine (Arawakan): Ethnologue (OV) disagrees with Hanson (2010): "The predicate-first order is somewhat more common than argument-first in verbal clauses." (Projected OV ratio: 0.31)
  - Purépecha (isolate): Dryer (VO), but Friedrich (1984): "Short objects and, often, pronominal ones are generally preverbal. [...] Objects with two or more words, especially long words, tend to be placed after the verb." (Projected OV ratio: 0.57)
  - Koreguaje (Tucanoan): clear disagreement with Dryer and Grambank
  - Luwo (Nilotic): clear disagreement, Storch (2010): "the basic word order in transitive sentences is always O-V-S"

Remaining errors

- Different definitions of features cause some systematic errors
- Example: noun/adjective order, where we actually measure noun/*core* adjective order (in order to increase chances of actually capturing adjectives across languages)
- For instance, Romance languages tend to have a different order for roughly this set of adjectives

Future work

- Language embeddings are well-explored by now, where do we head from here for massively multilingual models?
  - Hierarchical parameter sharing?
  - (Hiererchical) adapters?
  - Monolithic model with language embeddings?

Future work

- Language embeddings are well-explored by now, where do we head from here for massively multilingual models?
  - Hierarchical parameter sharing?
  - (Hiererchical) adapters?
  - Monolithic model with language embeddings?
- Additional consideration: which models give interesting by-products for linguists?
  - For instance, interpretable information on language relationships (sound changes, lexical replacement, grammatical changes, etc.)

Future work

- Language embeddings are well-explored by now, where do we head from here for massively multilingual models?
  - Hierarchical parameter sharing?
  - (Hiererchical) adapters?
  - Monolithic model with language embeddings?
- Additional consideration: which models give interesting by-products for linguists?
  - For instance, interpretable information on language relationships (sound changes, lexical replacement, grammatical changes, etc.)
- Do the models encode interesting language generalizations that we have not thought about?

Future work

- Language embeddings are well-explored by now, where do we head from here for massively multilingual models?
  - Hierarchical parameter sharing?
  - (Hiererchical) adapters?
  - Monolithic model with language embeddings?
- Additional consideration: which models give interesting by-products for linguists?
  - For instance, interpretable information on language relationships (sound changes, lexical replacement, grammatical changes, etc.)
- Do the models encode interesting language generalizations that we have not thought about?
- In computational typology, how do we move towards fine-grained automatic analyses? This may be the best application of high-accuracy parsers for low-resource languages

The End

Questions?