PrOnto: Language Model Evaluations for 859 Languages

Luke Gessler

Introduction

- Pre-trained **language models (LMs)** foundational for most NLP systems
  - Learns things from unlabeled text that it wouldn't have learned otherwise
- LMs unobtainable for many low-resource languages, but things are changing
  - Gessler and Zeldes (2022): as little as 500K words
- **Evaluations**: scarce outside of English, critical for assessing LM capabilities, supporting LM development

Example: MasakhaNER (Adelani et al., 2022)

Named entity recognition (NER): **identify words** which refer to **named entities** such as people, organizations, or places

Mbiram ak Baraag Obamaa moo ko ubbil ay bunt .

Example: MasakhaNER (Adelani et al., 2022)

Named entity recognition (NER): **identify words** which refer to **named entities** such as people, organizations, or places

Mbiram ak **Baraag Obamaa** moo ko ubbil ay bunt .

PERSON

Example: AmericasNLI (Ebrahimi et al., 2021)

Natural language inference (NLI): given a **premise**, judge whether a **hypothesis** is **entailed**, **contradicted**, or **neutral**

Example: AmericasNLI (Ebrahimi et al., 2021)

Natural language inference (NLI): given a **premise**, judge whether a **hypothesis** is **entailed**, **contradicted**, or **neutral**

Language	Premise	Hypothesis
en	And he said, Mama, I'm home.	He told his mom he had gotten home.
es	Y él dijo: Mamá, estoy en casa.	Le dijo a su madre que había llegado a casa.
aym	Jupax sanwa: Mamita, utankastwa.	Utar purinxtwa sasaw mamaparux sanxa
bzd	Ena ie' iche: ãmı̀, ye' tso' ù a.	I ãmı̀ a iché irir tö ye' démine ù a.
cni	Iriori ikantiro: Ina, nosaiki pankotsiki.	Ikantiro iriniro yaretaja pankotsiki.
gn	Ha ha'e he'i: Mama, aime ógape.	He'ícuri isýpe oĝuahêhague hógape.
hch	metá mik+ petay+: ne mama kitá nepa yéka.	yu mama m+pa+ p+ra h+awe kai kename yu kitá he nuakai.
nah	huan yehhua quiihtoh: Nonantzin, niyetoc nochan	quiilih inantzin niehcoquia
oto	xi nydi biênâ: maMe dimi an ngû	bimâbi o ini maMe guê o ngû
quy	Hinaptinmi pay nirqa: Mamay wasipim kachkani.	Wasinman chayasqanmanta mamanta willarqa.
shp	Jara neskata iki: tita, xobonkoriki ea.	Jawen tita yoiaia iki moa xobon nokota.
tar	A'lí je aníli échiko: ku bitichí ne atíki Nana	Iyéla ku ruyéli, mapu bitichí ku nawáli.

Table 2: A parallel example in AmericasNLI with the *entailment* label.

Motivation

- Making evaluation datasets is expensive!
  - AmericasNLI covers 10 American languages
  - MasakhaNER covers 20 African languages

MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition

David Ifeoluwa Adelani[1,2,*], **Graham Neubig**[3], **Sebastian Ruder**[4], **Shruti Rijhwani**[3], **Michael Beukman**[5*], **Chester Palen-Michel**[6*], **Constantine Lignos**[6*], **Jesujoba O. Alabi**[1*], **Shamsuddeen H. Muhammad**[7*], **Peter Nabende**[8*], **Cheikh M. Bamba Dione**[9*], **Andiswa Bukula**[10], **Rooweither Mabuya**[10], **Bonaventure F. P. Dossou**[11*], **Blessing Sibanda***, **Happy Buzaaba**[12*], **Jonathan Mukiibi**[8*], **Godson Kalipe***, **Derguene Mbaye**[13*], **Amelia Taylor**[14*], **Fatoumata Kabore**[15*], **Chris Chinenye Emezue**[16*], **Anuoluwapo Aremu***, **Perez Ogayo**[3*], **Catherine Gitau***, **Edwin Munkoh-Buabeng**[17*], **Victoire M. Koagne***, **Allahsera Auguste Tapo**[18*], **Tebogo Macucwa**[19*], **Vukosi Marivate**[19*], **Elvis Mboning***, **Tajuddeen Gwadabe***, **Tosin Adewumi**[20*], **Orevaoghene Ahia**[21*], **Joyce Nakatumba-Nabende**[8*], **Neo L. Mokono**[19*], **Ignatius Ezeani**[22*], **Chiamaka Chukwuneke**[22*], **Mofetoluwa Adeyemi**[23*], **Gilles Q. Hacheme**[24*], **Idris Abdulmumin**[25*], **Odunayo Ogundepo**[23*], **Oreen Yousuf**[15*], **Tatiana Moteu Ngoli***, **Dietrich Klakow**[1]

*Masakhane NLP, [1]Saarland University, Germany, [2]University College London, UK, [3]Carnegie Mellon University, USA, Google Research, [5]University of the Witwatersrand, South Africa, [6] Brandeis University, USA, [7]LIAAD-INESC TEC, Portugal, [8]Makerere University, Uganda [9]University of Bergen, Norway, [10]SADiLaR, South Africa, [11]Mila Quebec AI Institute, Canada, [12]RIKEN Center for AI Project, Japan, [13]Baamtu, Senegal, [14]Malawi University of Business and Applied Science, Malawi, [15]Uppsala University, Sweden, [16]TU Munich, Germany, [17]TU Clausthal, Germany, [18]Rochester Institute of Technology, USA, [19]University of Pretoria, South Africa, [20]Luleå University of Technology, Sweden, [21]University of Washington, USA, [22]Lancaster University, UK, [23]University of Waterloo, Canada, [24]Ai4innov, France, [25]Ahmadu Bello University, Nigeria.

AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages

Abteen Ebrahimi[♢] **Manuel Mager**[♠] **Arturo Oncevay**[♡] **Vishrav Chaudhary**[§] **Luis Chiruzzo**[△] **Angela Fan**[▽] **John E. Ortega**[Ω] **Ricardo Ramos**[η] **Annette Rios**[ψ] **Ivan Meza-Ruiz**[♯] **Gustavo A. Giménez-Lugo**[♣] **Elisabeth Mager**[♯] **Graham Neubig**[⋈] **Alexis Palmer**[♢] **Rolando Coto-Solano**[♂] **Ngoc Thang Vu**[♠] **Katharina Kann**[♢]

[⋈]Carnegie Mellon University [♂]Dartmouth College [§]Microsoft Turing [▽]Facebook AI Research [Ω]New York University [△]Universidad de la República, Uruguay [η]Universidad Tecnológica de Tlaxcala [♯]Universidad Nacional Autónoma de México [♣]Universidade Tecnológica Federal do Paraná [♢]University of Colorado Boulder [♡]University of Edinburgh [♠]University of Stuttgart [ψ]University of Zurich

Motivation

- Making evaluation datasets is expensive!
  - AmericasNLI covers 10 American languages
  - MasakhaNER covers 20 African languages
- How many people would be needed to cover more languages?
- Could we **trade quality for quantity**?

Approach

- **Observation 1**: New Testament (NT) translations extremely common
- **Observation 2**: OntoNotes (Weischedel et al., 2013) has a richly annotated NT translation

Approach

- **Observation 1**: New Testament (NT) translations extremely common
- **Observation 2**: OntoNotes (Weischedel et al., 2013) has a richly annotated NT translation
- **Idea**: Could we automatically project (Yarowsky and Ngai, 2001) annotations from OntoNotes' NT onto others?

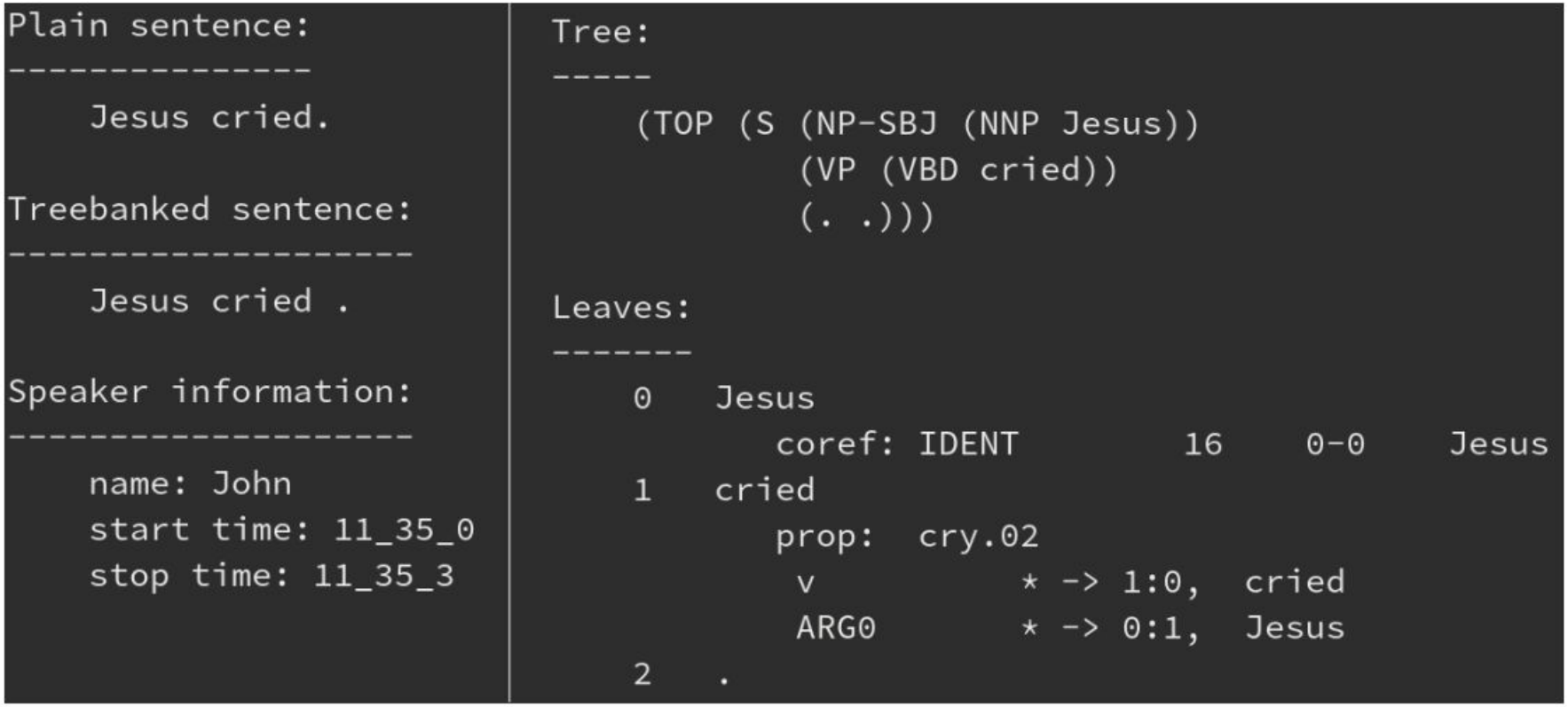


Figure 6.1: A sample verse, John 11:35, taken from OntoNotes. Note the annotations for tokenization, part-of-speech, constituency syntax, coreference, and argument structure. This file is in "OntoNotes Normal Form" (ONF), a human-readable format which OntoNotes provides its annotations in.

Annotation Projection: Referential NP Counting

Then the high priest asked Stephen, “Are these charges true?”

Annotation Projection: Referential NP Counting

3 Then **[the high priest]** asked **[Stephen]**, “Are **[these charges]** true?”

Annotation Projection: Referential NP Counting

3 Then **[the high priest]** asked **[Stephen]**, “Are **[these charges]** true?”

? महायाजक ने उनसे पूछा, “क्या यह आरोप सच है?”

Annotation Projection: Referential NP Counting

3 Then **[the high priest]** asked **[Stephen]**, “Are **[these charges]** true?”

3 **[महायाजक]** ने **[उनसे]** पूछा, “क्या यह **[आरोप]** सच है?”

Methods

- We frame **5 sequence classification tasks** and produce data for them via **annotation projection**
- Use all CC Bibles at ebible.org: **1051 translations** in **859 languages**

Projection is Not Perfect

- How likely is it that one of these annotations would also be appropriate for a translation of John 11:35?
  - Would sentential mood always be declarative?
  - Would an NP and a VP always be top-level sister constituents?

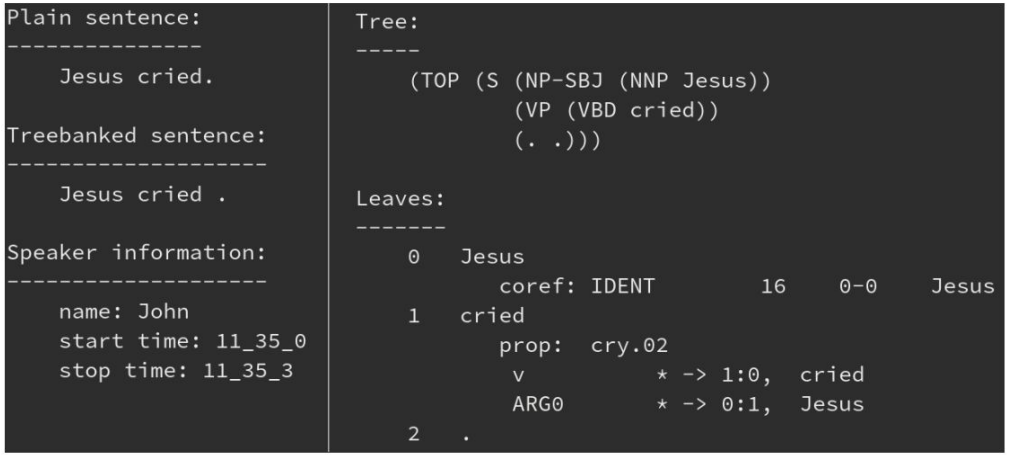
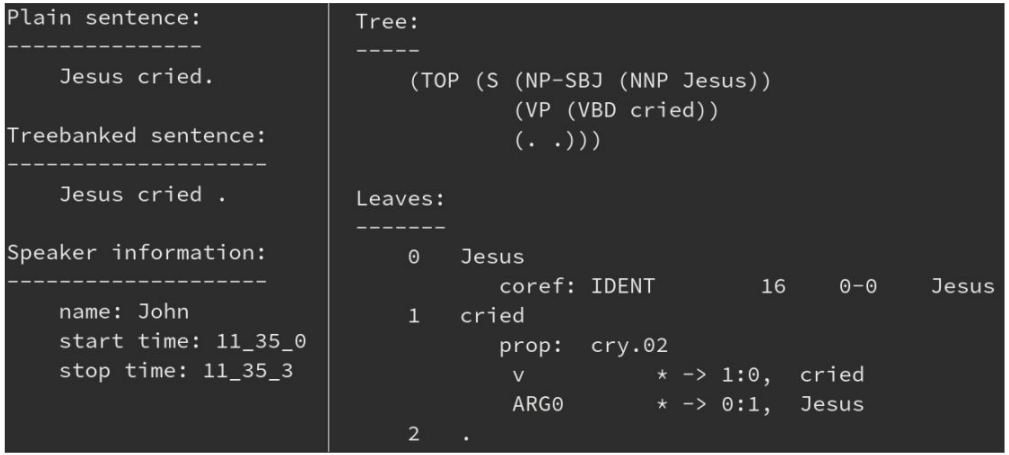


Figure 6.1: A sample verse, John 11:35, taken from OntoNotes. Note the annotations for tokenization, part-of-speech, constituency syntax, coreference, and argument structure. This file is in "OntoNotes Normal Form" (ONF), a human-readable format which OntoNotes provides its annotations in.

Projection is Not Perfect

- How likely is it that one of these annotations would also be appropriate for a translation of John 11:35?
- Projection **is imperfect** because of translation and grammatical differences
- But if some of the annotations are correct, the dataset might still be **useful**



```
Plain sentence:
---------------
    Jesus cried.

Treebanked sentence:
--------------------
    Jesus cried .

Speaker information:
--------------------
    name: John
    start time: 11_35_0
    stop time: 11_35_3
```

```
Tree:
-----
    (TOP (S (NP-SBJ (NNP Jesus))
            (VP (VBD cried))
            (. .)))

Leaves:
-------
    0   Jesus
           coref: IDENT        16    0-0    Jesus
    1   cried
           prop:  cry.02
            v          * -> 1:0,  cried
            ARG0       * -> 0:1,  Jesus
    2   .
```

Figure 6.1: A sample verse, John 11:35, taken from OntoNotes. Note the annotations for tokenization, part-of-speech, constituency syntax, coreference, and argument structure. This file is in "OntoNotes Normal Form" (ONF), a human-readable format which OntoNotes provides its annotations in.

Tasks (3 out of 5)

- **Non-pronominal Mention Counting (NMC)**: how many non-pronominal referential NPs in a verse?
    - *Non-pronominal*: because pronoun usage famously varies much between languages

Then **[the high priest]** asked **[Stephen]**, “Are **[these charges]** true?”

Tasks (3 out of 5)

- **Non-pronominal Mention Counting (NMC)**: how many non-pronominal referential NPs in a verse?
- **Proper Noun in Subject (PNS)**: is the subject of the first sentence a proper noun?

Then the high priest❌ asked Stephen, “Are these charges true?”

Tasks (3 out of 5)

- **Non-pronominal Mention Counting (NMC)**: how many non-pronominal referential NPs in a verse?
- **Proper Noun in Subject (PNS)**: is the subject of the first sentence a proper noun?
- **Sentence Mood (SM)**: mood of the first sentence declarative, interrogative, imperative, or other?

Then the high priest <u>asked</u> Stephen, “Are these charges true?”

DECL

Tasks (3 out of 5)

- **Non-pronominal Mention Counting (NMC)**: how many non-pronominal referential NPs in a verse?
- **Proper Noun in Subject (PNS)**: is the subject of the first sentence a proper noun?
- **Sentence Mood (SM)**: mood of the first sentence declarative, interrogative, imperative, or other?

- All available verses in each translation are used, and some data is held out
- Model to be evaluated is trained on each task, tested on the held-out data

Evaluation

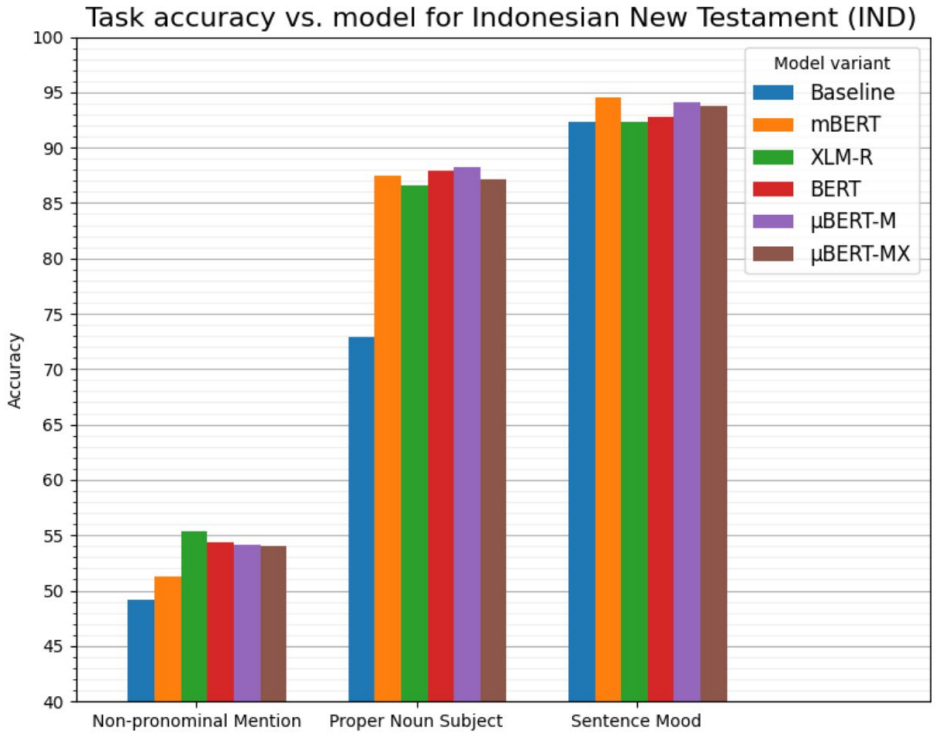
- Run tasks with a wide variety of pretrained models and Bibles
  - Some models are monolingual, others are multilingual
- Assess accuracy
- Compare to a “guess the most common answer” baseline

Hypotheses

- What are we hoping to see?
- If models **reproduce other findings** about their relative qualities, we have reason to believe the projected annotations are **partially correct**
  - Monolingual LMs tend to outperform multilingual ones (e.g. Angerri et al., 2020)
  - In “low-data” scenarios (between 0.5M and 10M words), much smaller LMs tend to match or exceed “normal-size” LMs (Gessler and Zeldes, 2022)

Example: Indonesian

- Monolingual LM (BERT) always outperforms one or both of the multilingual LMs (mBERT, XLM-R)
- Smaller monolingual LMs (μBERT-M, μBERT-MX) match or exceed larger LM (BERT)



Conclusion

- Evidence that projections are at least partially sound
- Method for creating LM evaluations that **only requires a NT translation**
- Pre-applied to 859 languages
- Aim: support LM work on “low-resource” languages

Is Your Language Included?