# Cross-Lingual Information Extraction

601.764

2/2/23

Note that many of the examples come Bikel and Zitouni 2012

# Information Extraction (IE)

- Identifying
- Extracting
- Useful *text information*

# Useful?

- User/Task/Application specific
- "*who* did *what* to *whom* at *when* and/or for what reason (*why*)"
- Arbitrarily broad
- Even up to world knowledge

# 2 Main Subtasks

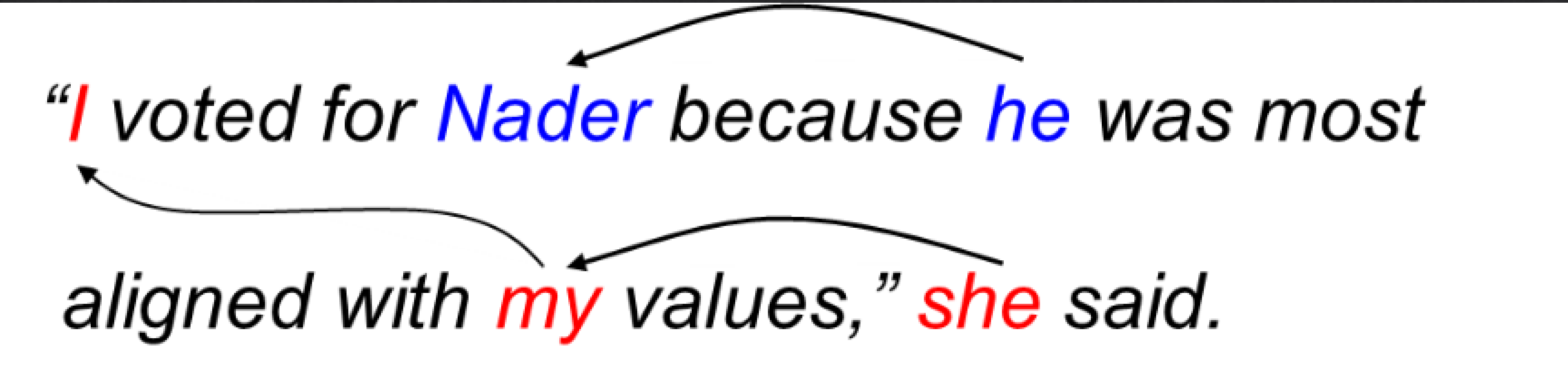
- Mention Detection
- Coreference Resolution

# Mention Detection

- Detecting the boundary of a mention
- Possibly detecting semantic type (PERSON, ORGANIZATION, PLACE, etc)
- Other attributes (Named, Nominal, Pronomial, etc.)

# Coreference Resolution

- Cluster mentions referring to the same entity into equivalence classes




Stanford NLP CoRef Project

# Anaphora Resolution

- Related to Coreference Resolution
- Much of the literature is at odds and I would argue incorrect
- For simplicity, we will just say coreference in this lecture

# President Ford said that he has no comments

# Mention Detection

**President Ford said that he has no comments**

Nominal

**President Ford said that he has no comments**

Nominal Named

**President Ford said that he has no comments**

Nominal Named Pronominal

**President Ford said that he has no comments**

When asked about what Nixon had to say:

President Ford said that he has no comments

# Named Entity Resolution

- NER
- Historically, NER has been central
- Message Understanding Conference (MUC-6) in 1995:
  - Person
  - Organization
  - Location
  - Time
  - Percent
  - Money

# spaCy English Entity List

```
PERSON:      People, including fictional.
NORP:        Nationalities or religious or political groups.
FAC:         Buildings, airports, highways, bridges, etc.
ORG:         Companies, agencies, institutions, etc.
GPE:         Countries, cities, states.
LOC:         Non-GPE locations, mountain ranges, bodies of water.
PRODUCT:     Objects, vehicles, foods, etc. (Not services.)
EVENT:       Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART: Titles of books, songs, etc.
LAW:         Named documents made into laws.
LANGUAGE:    Any named language.
DATE:        Absolute or relative dates or periods.
TIME:        Times smaller than a day.
PERCENT:     Percentage, including ”%“.
MONEY:       Monetary values, including unit.
QUANTITY:    Measurements, as of weight or distance.
ORDINAL:     “first”, “second”, etc.
CARDINAL:    Numerals that do not fall under another type.
```

# Feature Based Approaches

- Still Used …
- … but more historical*

# Feature Based Approaches

- Still Used …
- … but more historical*
  - Lexical Features (n-Grams, 3)
  - Syntactic Features (POS)
  - Gazetteer-based (list of names
  - Cross-Language Features (Word Alignments)

# Cross-Lingual

# Cross-Lingual

- "Instead, language-dependent phenomena are handled by either a preprocessing step …. Space-delimited words may not be a good unit for [entity detection and tracking], and a **morph** is often chosen to counter the data-sparseness problem."

# Cross-Lingual

- “Instead, language-dependent phenomena are handled by either a preprocessing step …. Space-delimited words may not be a good unit for [entity detection and tracking], and a **morph** is often chosen to counter the data-sparseness problem.”

**What’s wrong with this?**

# Automatic Content Extraction (ACE)

- DARPA program for IE
- Evaluations in 2005, 2007, 2008
- Training data from 2004 and 2005

# Automatic Content Extraction (ACE)

- DARPA program for IE
- Evaluations in 2005, 2007, 2008
- Training data from 2004 and 2005

**Do we still use this?**

# Automatic Content Extraction (ACE)

- DARPA program for IE
- Evaluations in 2005, 2007, 2008
- Training data from 2004 and 2005

**Do we still use this?**

Yes

# ACE 2004

| | English | | Chinese | | | Arabic | |
|---|---|---|---|---|---|---|---|
| Genre | Files | Words | Files | Words | Characters | Files | Words |
| Broadcast News | 220 | 60,291 | 314 | 67,702 | 135,405 | 304 | 63,238 |
| Newswire | 128 | 59,840 | 226 | 60,251 | 120,502 | 253 | 63,122 |
| Chinese Treebank | 37 | 12,337 | 106 | 25,749 | 51,499 | | |
| Arabic Treebank | 58 | 12,855 | | | | 132 | 25,010 |
| Fisher CTS | 8 | 12,630 | | | | | |
| Totals | 451 | 157,953 | 646 | 153,703 | 307,406 | 689 | 151,360 |

# Some of the Sources

## Copyright

Portions (c) 1994-1998, 2000 Xinhua News Agency (c) 1997 Department of Information Services, Hong Kong Special Administrative Region (c) 1996-1998, 2000-2001 Sinorama Magazine (c) 2000 Agence France-Presse, (c) 2000 New York Times, (c) 2000 Associated Press, (c) 2000 SPH AsiaOne, Ltd. (Zaobao), (c) 2000 An-Nahar, (c) 2000 Al-Hayat, (c) 2000 Nile TV, (c) 2000 Cable News Network, All Rights Reserved, (c) 2000 American Broadcasting Corporation, (c) 2000 National Broadcasting Company, Inc., (c) 2000 China National Radio, (c) 2000 China Television System, (c) 2000 China Central TV, (c) 2000 China Broadcasting System, (c) 2000 Public Radio International., (c) 2005 Trustees of the University of Pennsylvania

# Some of the Sources

### Copyright

Portions (c) 1994-1998, 2000 Xinhua News Agency (c) 1997 Department of Information Services, Hong Kong Special Administrative Region (c) 1996-1998, 2000-2001 Sinorama Magazine (c) 2000 Agence France-Presse, (c) 2000 New York Times, (c) 2000 Associated Press, (c) 2000 SPH AsiaOne, Ltd. (Zaobao), (c) 2000 An-Nahar, (c) 2000 Al-Hayat, (c) 2000 Nile TV, (c) 2000 Cable News Network, All Rights Reserved, (c) 2000 American Broadcasting Corporation, (c) 2000 National Broadcasting Company, Inc., (c) 2000 China National Radio, (c) 2000 China Television System, (c) 2000 China Central TV, (c) 2000 China Broadcasting System, (c) 2000 Public Radio International., (c) 2005 Trustees of the University of Pennsylvania

**Domains? Biases?**

# Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction

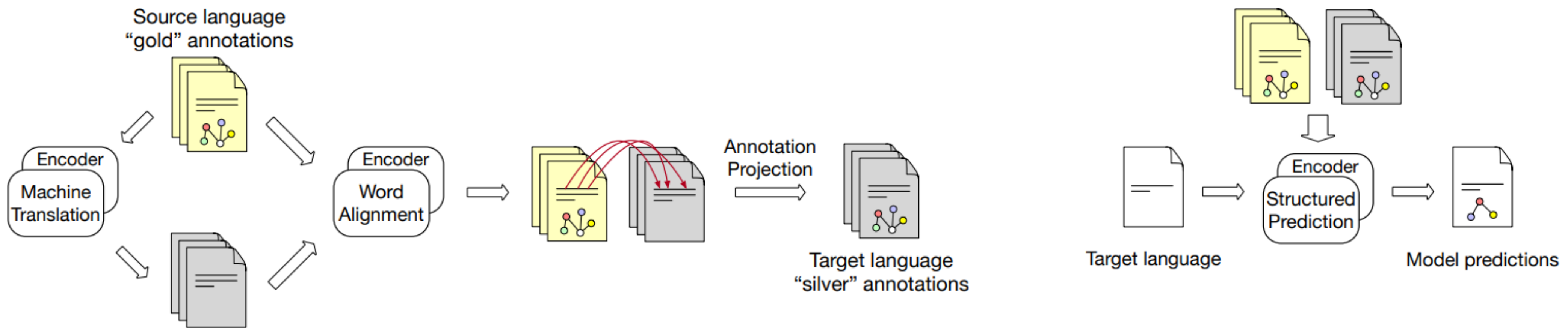



Figure 1: Process for creating projected "silver" data from source "gold" data (left). Downstream models are trained on a combination of gold and silver data (right). Components in boxes have learned parameters.

Yarmohammadi et al., 2021

# Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction

| | **Base** | **Large** |
|---|---|---|
| **Multilingual** | mBERT (Devlin et al.) | XLM-R (Conneau et al.) |
| **Bilingual** | GBv4 (Lan et al.) | L64K & L128K **(Ours)** |

Table 1: Encoders supporting English and Arabic.

Base models are 12-layer Transformers (d_model = 768), and large models are 24-layer Transformers (d_model = 1024)

Yarmohammadi et al., 2021

# GigaBERT

| Models | Training Data | | Vocabulary | | | Configuration | |
|---|---|---|---|---|---|---|---|
| | source | #tokens (all/en/ar) | tokenization | size (all/en/ar) | cased | size | #parameters |
| AraBERT | newswire | 2.5B/ – /2.5B | SentencePiece | 64k/ – / 58k | no | base | 136M |
| mBERT | Wiki | 21.9B/2.5B/153M | WordPiece | 110k/53k/5k | yes | base | 172M |
| XLM-$R_{base}$ | CommonCrawl | 295B/55.6B/2.9B | SentencePiece | 250k/80k/14k | yes | base | 270M |
| XLM-$R_{large}$ | CommonCrawl | 295B/55.6B/2.9B | SentencePiece | 250k/80k/14k | yes | large | 550M |
| GiagBERT-v0 | Gigaword | 4.7B/3.6B/1.1B | SentencePiece | 50k/28k/19k | yes | base | 125M |
| GigaBERT-v1 | Gigaword, Wiki | 7.4B/6.1B/1.3B | WordPiece | 50k/25k/23k | yes | base | 125M |
| GigaBERT-v2/3 | Gigaword, Wiki, Oscar | 10.4B/6.1B/4.3B | WordPiece | 50k/21k/26k | no | base | 125M |
| GigaBERT-v4 | Gigaword, Wiki, Oscar (+ code-switch) | 10.4B/6.1B/4.3B | WordPiece | 50k/21k/26k | no | base | 125M |

Table 1: Configuration comparisons for AraBERT (Antoun et al., 2020), mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020a), and GigaBERT (this work).

Wan et al., 2020

# Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction

| Encoder | BLEU |
|---|---|
| Public | 12.7 |
| None | 14.9 |
| mBERT | 15.7 |
| GBv4 | 15.7 |
| XLM-R | 16.0 |
| L64K | **16.2** |
| L128K | 15.8 |

Table 2: BLEU scores of MT systems with different pre-trained encoders on English–Arabic IWSLT'17.

Yarmohammadi et al., 2021

# Everything Is All It Takes: A Multipronged Strategy for Zero-Shot Cross-Lingual Information Extraction

| Model | Layer† | AER | P | R | F |
|---|---|---|---|---|---|
| fast-align* | n/a | 47.4 | 53.9 | 51.4 | 52.6 |
| *Awesome-align w/o FT* | | | | | |
| mBERT | 8 | 35.6 | 78.5 | 54.5 | 64.4 |
| GBv4 | 8 | **32.7** | **85.6** | 55.4 | **67.3** |
| XLM-R | 16 | 40.1 | 78.6 | 48.4 | 59.9 |
| L64K | 17 | 34.0 | 81.5 | **55.5** | 66.0 |
| L128K | 17 | 35.1 | 80.0 | 54.5 | 64.9 |
| *Awesome-align w/ FT* | | | | | |
| mBERT$_{ft}$ | 8 | 30.0 | 81.9 | **61.2** | 70.0 |
| GBv4$_{ft}$ | 8 | 29.3 | 86.9 | 59.7 | 70.7 |
| XLM-R$_{ft}$ | 18 | **27.8** | **90.3** | 60.2 | **72.2** |
| L64K$_{ft}$ | 17 | 29.1 | 84.9 | 60.9 | 70.9 |
| L128K$_{ft}$ | 16 | 32.2 | 80.3 | 58.7 | 67.8 |
| *Awesome-align w/ FT & supervision* | | | | | |
| XLM-R$_{ft.s}$ | 16 | **23.3** | 92.5 | **65.6** | **76.7** |
| L128K$_{ft.s}$ | 17 | 23.5 | **93.7** | 64.6 | 76.5 |

Table 3: Alignment performance on GALE EN–AR. *Trained on MT bitext. †We report the best layer of each encoder based on dev alignment error rate (AER).

Yarmohammadi et al., 2021

| | MT |
|---|---|
| | |
| (Z) | - |
| (A) | public |
| (B) | public |
| (B) | public |
| (C) | public |
| (C) | public |
| (C) | public |
| (D) | public |
| (D) | public |
| (D) | public |
| (E) | GBv4 |
| (E) | GBv4 |
| (E) | L128K |
| (E) | L128K |
| (S) | public |
| | |
| (Z) | - |
| (C) | public |
| (C) | public |
| (C) | public |
| (E) | GBv4 |
| (E) | GBv4 |
| (E) | L128K |
| (E) | L128K |
| (F) | GBv4 |
| (F) | GBv4 |
| (F) | L128K |
| (F) | L128K |
| (F) | L128K |
| (S) | public |

[redacted]

| | MT | Align | ACE | NER | POS | Parsing | BET. | ACE | NER | POS | Parsing | BET. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | [redacted] | | | | | | | | | | |
| (Z) | - | | | | | | | | | | | |
| (A) | public | | | | | | | | | | | |
| (B) | public | | | | | | | | | | | |
| (B) | public | | | | | | | | | | | |
| (C) | public | | | | | | | | | | | |
| (C) | public | | | | | | | | | | | |
| (C) | public | | | | | | | | | | | |
| (D) | public | | | | | | | | | | | |
| (D) | public | | | | | | | | | | | |
| (D) | public | | | | | | | | | | | |
| (E) | GBv4 | | | | | | | | | | | |
| (E) | GBv4 | | | | | | | | | | | |
| (E) | L128K | | | | | | | | | | | |
| (E) | L128K | | | | | | | | | | | |
| (S) | public | | | | | | | | | | | |
| | | | | | | | | | | | | |
| (Z) | - | | | | | | | | | | | |
| (C) | public | | | | | | | | | | | |
| (C) | public | | | | | | | | | | | |
| (C) | public | | | | | | | | | | | |
| (E) | GBv4 | | | | | | | | | | | |
| (E) | GBv4 | | | | | | | | | | | |
| (E) | L128K | | | | | | | | | | | |
| (E) | L128K | | | | | | | | | | | |
| (F) | GBv4 | | | | | | | | | | | |
| (F) | GBv4 | | | | | | | | | | | |
| (F) | L128K | | | | | | | | | | | |
| (F) | L128K | | | | | | | | | | | |
| (F) | L128K | | | | | | | | | | | |
| (S) | public | | | | | | | | | | | |

| | MT | Align | ACE | NER | POS | Parsing | BET. | ACE | NER | POS | Parsing | BET. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | [redacted] | | | | | | | | | |
| (Z) | - | - | [redacted] | | | | | | | | | |
| (A) | public | FA | [redacted] | | | | | | | | | |
| (B) | public | mBERT | [redacted] | | | | | | | | | |
| (B) | public | XLM-R | [redacted] | | | | | | | | | |
| (C) | public | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (C) | public | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (C) | public | XLM-R$_{ft.s}$ | [redacted] | | | | | | | | | |
| (D) | public | GBv4$_{ft}$ | [redacted] | | | | | | | | | |
| (D) | public | L128K$_{ft}$ | [redacted] | | | | | | | | | |
| (D) | public | L128K$_{ft.s}$ | [redacted] | | | | | | | | | |
| (E) | GBv4 | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | GBv4 | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | L128K | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | L128K | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (S) | public | ST | [redacted] | | | | | | | | | |
| | | | [redacted] | | | | | | | | | |
| (Z) | - | - | [redacted] | | | | | | | | | |
| (C) | public | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (C) | public | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (C) | public | XLM-R$_{ft.s}$ | [redacted] | | | | | | | | | |
| (E) | GBv4 | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | GBv4 | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | L128K | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | L128K | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | GBv4 | GBv4$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | GBv4 | L128K$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | L128K | GBv4$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | L128K | L128K$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | L128K | L128K$_{ft.s}$ | [redacted] | | | | | | | | | |
| (S) | public | ST | [redacted] | | | | | | | | | |

| | MT | Align | ACE | NER | POS | Parsing | BET. | ACE | NER | POS | Parsing | BET. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *mBERT (base, multilingual)* | | | | | [redacted] | | | | |
| (Z) | - | - | 27.0 | 41.6 | 59.7 | 29.2 | 39.9 | [redacted] | | | | |
| (A) | public | FA | +2.5 | -3.8 | +8.5 | +7.3 | +2.6 | [redacted] | | | | |
| (B) | public | mBERT | +6.5 | +0.2 | +8.5 | +7.6 | +2.3 | [redacted] | | | | |
| (B) | public | XLM-R | +0.9 | -2.9 | +9.5 | +9.0 | -1.2 | [redacted] | | | | |
| (C) | public | mBERT$_{ft}$ | +7.8 | **+5.6** | +7.7 | +10.0 | +4.1 | [redacted] | | | | |
| (C) | public | XLM-R$_{ft}$ | +7.7 | +4.9 | +6.2 | +9.3 | +4.5 | [redacted] | | | | |
| (C) | public | XLM-R$_{ft.s}$ | +7.3 | +1.5 | +10.1 | **+12.4** | +4.8 | [redacted] | | | | |
| (D) | public | GBv4$_{ft}$ | +8.5 | +4.3 | +5.9 | +8.9 | +5.0 | [redacted] | | | | |
| (D) | public | L128K$_{ft}$ | +6.4 | +3.1 | +6.5 | +8.2 | +1.6 | [redacted] | | | | |
| (D) | public | L128K$_{ft.s}$ | +7.0 | +3.7 | **+10.3** | +11.8 | **+5.4** | [redacted] | | | | |
| (E) | GBv4 | mBERT$_{ft}$ | +8.4 | +3.2 | +7.7 | +9.9 | +4.7 | [redacted] | | | | |
| (E) | GBv4 | XLM-R$_{ft}$ | +9.6 | +1.8 | +7.0 | +9.5 | +5.2 | [redacted] | | | | |
| (E) | L128K | mBERT$_{ft}$ | **+12.1** | +3.3 | +7.9 | +9.9 | +4.7 | [redacted] | | | | |
| (E) | L128K | XLM-R$_{ft}$ | +10.2 | -1.9 | +6.1 | +9.4 | +4.8 | [redacted] | | | | |
| (S) | public | ST | - | +5.5 | +0.1 | -20.3 | +0.3 | [redacted] | | | | |
| | | | [redacted] | | | | | | | | | |
| (Z) | - | - | [redacted] | | | | | | | | | |
| (C) | public | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (C) | public | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (C) | public | XLM-R$_{ft.s}$ | [redacted] | | | | | | | | | |
| (E) | GBv4 | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | GBv4 | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | L128K | mBERT$_{ft}$ | [redacted] | | | | | | | | | |
| (E) | L128K | XLM-R$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | GBv4 | GBv4$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | GBv4 | L128K$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | L128K | GBv4$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | L128K | L128K$_{ft}$ | [redacted] | | | | | | | | | |
| (F) | L128K | L128K$_{ft.s}$ | [redacted] | | | | | | | | | |
| (S) | public | ST | [redacted] | | | | | | | | | |

| | MT | Align | ACE | NER | POS | Parsing | BET. | ACE | NER | POS | Parsing | BET. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *mBERT (base, multilingual)* | | | | | *XLM-R (large, multilingual)* | | | | |
| (Z) | - | - | 27.0 | 41.6 | 59.7 | 29.2 | 39.9 | **45.1** | 46.4 | 73.3 | **48.0** | 50.8 |
| (A) | public | FA | +2.5 | -3.8 | +8.5 | +7.3 | +2.6 | -7.5 | -0.1 | -7.7 | -9.5 | -1.6 |
| (B) | public | mBERT | +6.5 | +0.2 | +8.5 | +7.6 | +2.3 | -4.4 | +6.9 | -6.1 | -8.4 | -2.6 |
| (B) | public | XLM-R | +0.9 | -2.9 | +9.5 | +9.0 | -1.2 | -10.0 | +0.0 | -5.9 | -8.8 | -6.3 |
| (C) | public | mBERT$_{ft}$ | +7.8 | **+5.6** | +7.7 | +10.0 | +4.1 | -0.6 | +7.4 | -8.0 | -6.8 | +0.3 |
| (C) | public | XLM-R$_{ft}$ | +7.7 | +4.9 | +6.2 | +9.3 | +4.5 | -2.6 | +7.0 | -9.0 | -7.6 | +1.0 |
| (C) | public | XLM-R$_{ft.s}$ | +7.3 | +1.5 | +10.1 | **+12.4** | +4.8 | -3.0 | +9.1 | -3.8 | -3.7 | **+2.3** |
| (D) | public | GBv4$_{ft}$ | +8.5 | +4.3 | +5.9 | +8.9 | +5.0 | -1.5 | +7.7 | -9.4 | -9.1 | -0.1 |
| (D) | public | L128K$_{ft}$ | +6.4 | +3.1 | +6.5 | +8.2 | +1.6 | -1.6 | +6.1 | -9.0 | -9.4 | -3.6 |
| (D) | public | L128K$_{ft.s}$ | +7.0 | +3.7 | **+10.3** | +11.8 | **+5.4** | -0.3 | +5.2 | -4.4 | -4.6 | +2.1 |
| (E) | GBv4 | mBERT$_{ft}$ | +8.4 | +3.2 | +7.7 | +9.9 | +4.7 | -1.5 | +3.2 | -7.1 | -6.7 | +0.7 |
| (E) | GBv4 | XLM-R$_{ft}$ | +9.6 | +1.8 | +7.0 | +9.5 | +5.2 | -0.4 | +1.4 | -8.3 | -7.7 | +1.4 |
| (E) | L128K | mBERT$_{ft}$ | **+12.1** | +3.3 | +7.9 | +9.9 | +4.7 | -1.4 | +7.2 | -8.1 | -6.7 | +1.3 |
| (E) | L128K | XLM-R$_{ft}$ | +10.2 | -1.9 | +6.1 | +9.4 | +4.8 | -0.5 | +4.6 | -9.8 | -7.5 | +2.0 |
| (S) | public | ST | - | +5.5 | +0.1 | -20.3 | +0.3 | - | **+10.0** | **+1.8** | -29.6 | +1.2 |

| | | | |
|---|---|---|---|
| | | | [redacted] |
| (Z) | - | - | [redacted] |
| (C) | public | mBERT$_{ft}$ | [redacted] |
| (C) | public | XLM-R$_{ft}$ | [redacted] |
| (C) | public | XLM-R$_{ft.s}$ | [redacted] |
| (E) | GBv4 | mBERT$_{ft}$ | [redacted] |
| (E) | GBv4 | XLM-R$_{ft}$ | [redacted] |
| (E) | L128K | mBERT$_{ft}$ | [redacted] |
| (E) | L128K | XLM-R$_{ft}$ | [redacted] |
| (F) | GBv4 | GBv4$_{ft}$ | [redacted] |
| (F) | GBv4 | L128K$_{ft}$ | [redacted] |
| (F) | L128K | GBv4$_{ft}$ | [redacted] |
| (F) | L128K | L128K$_{ft}$ | [redacted] |
| (F) | L128K | L128K$_{ft.s}$ | [redacted] |
| (S) | public | ST | [redacted] |

| | MT | Align | ACE | NER | POS | Parsing | BET. | ACE | NER | POS | Parsing | BET. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *mBERT (base, multilingual)* | | | | | *XLM-R (large, multilingual)* | | | | |
| (Z) | - | - | 27.0 | 41.6 | 59.7 | 29.2 | 39.9 | **45.1** | 46.4 | 73.3 | **48.0** | 50.8 |
| (A) | public | FA | +2.5 | -3.8 | +8.5 | +7.3 | +2.6 | -7.5 | -0.1 | -7.7 | -9.5 | -1.6 |
| (B) | public | mBERT | +6.5 | +0.2 | +8.5 | +7.6 | +2.3 | -4.4 | +6.9 | -6.1 | -8.4 | -2.6 |
| (B) | public | XLM-R | +0.9 | -2.9 | +9.5 | +9.0 | -1.2 | -10.0 | +0.0 | -5.9 | -8.8 | -6.3 |
| (C) | public | mBERT$_{ft}$ | +7.8 | **+5.6** | +7.7 | +10.0 | +4.1 | -0.6 | +7.4 | -8.0 | -6.8 | +0.3 |
| (C) | public | XLM-R$_{ft}$ | +7.7 | +4.9 | +6.2 | +9.3 | +4.5 | -2.6 | +7.0 | -9.0 | -7.6 | +1.0 |
| (C) | public | XLM-R$_{ft.s}$ | +7.3 | +1.5 | +10.1 | **+12.4** | +4.8 | -3.0 | +9.1 | -3.8 | -3.7 | **+2.3** |
| (D) | public | GBv4$_{ft}$ | +8.5 | +4.3 | +5.9 | +8.9 | +5.0 | -1.5 | +7.7 | -9.4 | -9.1 | -0.1 |
| (D) | public | L128K$_{ft}$ | +6.4 | +3.1 | +6.5 | +8.2 | +1.6 | -1.6 | +6.1 | -9.0 | -9.4 | -3.6 |
| (D) | public | L128K$_{ft.s}$ | +7.0 | +3.7 | **+10.3** | +11.8 | **+5.4** | -0.3 | +5.2 | -4.4 | -4.6 | +2.1 |
| (E) | GBv4 | mBERT$_{ft}$ | +8.4 | +3.2 | +7.7 | +9.9 | +4.7 | -1.5 | +3.2 | -7.1 | -6.7 | +0.7 |
| (E) | GBv4 | XLM-R$_{ft}$ | +9.6 | +1.8 | +7.0 | +9.5 | +5.2 | -0.4 | +1.4 | -8.3 | -7.7 | +1.4 |
| (E) | L128K | mBERT$_{ft}$ | **+12.1** | +3.3 | +7.9 | +9.9 | +4.7 | -1.4 | +7.2 | -8.1 | -6.7 | +1.3 |
| (E) | L128K | XLM-R$_{ft}$ | +10.2 | -1.9 | +6.1 | +9.4 | +4.8 | -0.5 | +4.6 | -9.8 | -7.5 | +2.0 |
| (S) | public | ST | - | +5.5 | +0.1 | -20.3 | +0.3 | - | **+10.0** | **+1.8** | -29.6 | +1.2 |
| | | | *GBv4 (base, bilingual)* | | | | | | | | | |
| (Z) | - | - | 46.0 | 45.4 | 64.7 | 33.2 | 41.7 | | | | | |
| (C) | public | mBERT$_{ft}$ | +0.6 | +3.7 | +2.6 | +6.9 | +7.5 | | | | | |
| (C) | public | XLM-R$_{ft}$ | -1.4 | **+4.5** | +1.8 | +6.0 | +8.4 | | | | | |
| (C) | public | XLM-R$_{ft.s}$ | -0.1 | +3.4 | **+5.1** | **+9.2** | +8.0 | | | | | |
| (E) | GBv4 | mBERT$_{ft}$ | -0.1 | +0.1 | +3.3 | +7.2 | +8.1 | | | | | |
| (E) | GBv4 | XLM-R$_{ft}$ | +0.1 | +0.4 | +1.5 | +6.0 | **+9.7** | | | | | |
| (E) | L128K | mBERT$_{ft}$ | -0.6 | +1.0 | +2.6 | +6.1 | +7.4 | | | | | |
| (E) | L128K | XLM-R$_{ft}$ | +0.9 | -2.1 | +1.1 | +5.5 | +7.8 | | | | | |
| (F) | GBv4 | GBv4$_{ft}$ | +0.0 | -1.9 | +1.6 | +4.5 | +9.1 | | | | | |
| (F) | GBv4 | L128K$_{ft}$ | -0.9 | -1.4 | +1.5 | +4.1 | +5.7 | | | | | |
| (F) | L128K | GBv4$_{ft}$ | -4.3 | -1.0 | +0.4 | +4.1 | +7.4 | | | | | |
| (F) | L128K | L128K$_{ft}$ | -3.5 | -1.1 | +0.3 | +3.8 | + 4.5 | | | | | |
| (F) | L128K | L128K$_{ft.s}$ | **+1.9** | +0.2 | +3.3 | +7.4 | +7.2 | | | | | |
| (S) | public | ST | - | -2.5 | -1.3 | -18.6 | +1.9 | | | | | |

| | MT | Align | ACE | NER | POS | Parsing | BET. | ACE | NER | POS | Parsing | BET. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *mBERT (base, multilingual)* | | | | | *XLM-R (large, multilingual)* | | | | |
| (Z) | - | - | 27.0 | 41.6 | 59.7 | 29.2 | 39.9 | **45.1** | 46.4 | 73.3 | **48.0** | 50.8 |
| (A) | public | FA | +2.5 | -3.8 | +8.5 | +7.3 | +2.6 | -7.5 | -0.1 | -7.7 | -9.5 | -1.6 |
| (B) | public | mBERT | +6.5 | +0.2 | +8.5 | +7.6 | +2.3 | -4.4 | +6.9 | -6.1 | -8.4 | -2.6 |
| (B) | public | XLM-R | +0.9 | -2.9 | +9.5 | +9.0 | -1.2 | -10.0 | +0.0 | -5.9 | -8.8 | -6.3 |
| (C) | public | $\text{mBERT}_{ft}$ | +7.8 | **+5.6** | +7.7 | +10.0 | +4.1 | -0.6 | +7.4 | -8.0 | -6.8 | +0.3 |
| (C) | public | $\text{XLM-R}_{ft}$ | +7.7 | +4.9 | +6.2 | +9.3 | +4.5 | -2.6 | +7.0 | -9.0 | -7.6 | +1.0 |
| (C) | public | $\text{XLM-R}_{ft.s}$ | +7.3 | +1.5 | +10.1 | **+12.4** | +4.8 | -3.0 | +9.1 | -3.8 | -3.7 | **+2.3** |
| (D) | public | $\text{GBv4}_{ft}$ | +8.5 | +4.3 | +5.9 | +8.9 | +5.0 | -1.5 | +7.7 | -9.4 | -9.1 | -0.1 |
| (D) | public | $\text{L128K}_{ft}$ | +6.4 | +3.1 | +6.5 | +8.2 | +1.6 | -1.6 | +6.1 | -9.0 | -9.4 | -3.6 |
| (D) | public | $\text{L128K}_{ft.s}$ | +7.0 | +3.7 | **+10.3** | +11.8 | **+5.4** | -0.3 | +5.2 | -4.4 | -4.6 | +2.1 |
| (E) | GBv4 | $\text{mBERT}_{ft}$ | +8.4 | +3.2 | +7.7 | +9.9 | +4.7 | -1.5 | +3.2 | -7.1 | -6.7 | +0.7 |
| (E) | GBv4 | $\text{XLM-R}_{ft}$ | +9.6 | +1.8 | +7.0 | +9.5 | +5.2 | -0.4 | +1.4 | -8.3 | -7.7 | +1.4 |
| (E) | L128K | $\text{mBERT}_{ft}$ | **+12.1** | +3.3 | +7.9 | +9.9 | +4.7 | -1.4 | +7.2 | -8.1 | -6.7 | +1.3 |
| (E) | L128K | $\text{XLM-R}_{ft}$ | +10.2 | -1.9 | +6.1 | +9.4 | +4.8 | -0.5 | +4.6 | -9.8 | -7.5 | +2.0 |
| (S) | public | ST | - | +5.5 | +0.1 | -20.3 | +0.3 | - | **+10.0** | **+1.8** | -29.6 | +1.2 |
| | | | *GBv4 (base, bilingual)* | | | | | *L128K (large, bilingual)* | | | | |
| (Z) | - | - | 46.0 | 45.4 | 64.7 | 33.2 | 41.7 | 42.7 | 46.3 | 67.9 | 36.7 | 40.9 |
| (C) | public | $\text{mBERT}_{ft}$ | +0.6 | +3.7 | +2.6 | +6.9 | +7.5 | +2.7 | +8.2 | -0.9 | +4.9 | +11.7 |
| (C) | public | $\text{XLM-R}_{ft}$ | -1.4 | **+4.5** | +1.8 | +6.0 | +8.4 | +1.2 | **+9.0** | -2.5 | +3.9 | +10.5 |
| (C) | public | $\text{XLM-R}_{ft.s}$ | -0.1 | +3.4 | **+5.1** | **+9.2** | +8.0 | +2.7 | +7.0 | +1.2 | **+7.2** | **+12.1** |
| (E) | GBv4 | $\text{mBERT}_{ft}$ | -0.1 | +0.1 | +3.3 | +7.2 | +8.1 | +4.2 | -0.5 | -0.1 | +5.1 | +11.2 |
| (E) | GBv4 | $\text{XLM-R}_{ft}$ | +0.1 | +0.4 | +1.5 | +6.0 | **+9.7** | +2.4 | +0.0 | -1.3 | +4.2 | +10.8 |
| (E) | L128K | $\text{mBERT}_{ft}$ | -0.6 | +1.0 | +2.6 | +6.1 | +7.4 | **+5.5** | +0.8 | -0.7 | +4.7 | +10.6 |
| (E) | L128K | $\text{XLM-R}_{ft}$ | +0.9 | -2.1 | +1.1 | +5.5 | +7.8 | +4.4 | -3.6 | -2.2 | +4.1 | +11.3 |
| (F) | GBv4 | $\text{GBv4}_{ft}$ | +0.0 | -1.9 | +1.6 | +4.5 | +9.1 | +2.0 | -0.3 | -1.7 | +3.2 | +10.9 |
| (F) | GBv4 | $\text{L128K}_{ft}$ | -0.9 | -1.4 | +1.5 | +4.1 | +5.7 | +2.3 | -1.7 | -2.4 | +2.6 | +8.3 |
| (F) | L128K | $\text{GBv4}_{ft}$ | -4.3 | -1.0 | +0.4 | +4.1 | +7.4 | +4.1 | -3.6 | -2.1 | +2.3 | +11.4 |
| (F) | L128K | $\text{L128K}_{ft}$ | -3.5 | -1.1 | +0.3 | +3.8 | + 4.5 | +2.9 | +0.1 | -2.9 | +2.0 | +6.7 |
| (F) | L128K | $\text{L128K}_{ft.s}$ | **+1.9** | +0.2 | +3.3 | +7.4 | +7.2 | +2.8 | -1.8 | +0.8 | +6.0 | +11.8 |
| (S) | public | ST | - | -2.5 | -1.3 | -18.6 | +1.9 | - | +7.1 | **+1.5** | -21.7 | +8.1 |

# Frustratingly Easy Label Projection for Cross-lingual Transfer (EasyProject)


Figure 1: An example of translating and projecting English ACE event triggers and named entities to Chinese. (a) Label projection pipeline starts with machine translation of the English sentence to Chinese, followed by word-to-word alignment. Then, label spans are projected based on word alignments. (b) Markers are inserted around entity and event trigger spans in the text. The modified sentence with markers inserted is then fed as input to a machine translation system, projecting the label span markers to the target language as a byproduct of translation.

Chen et al., 2022

# Frustratingly Easy Label Projection for Cross-lingual Transfer (EasyProject)

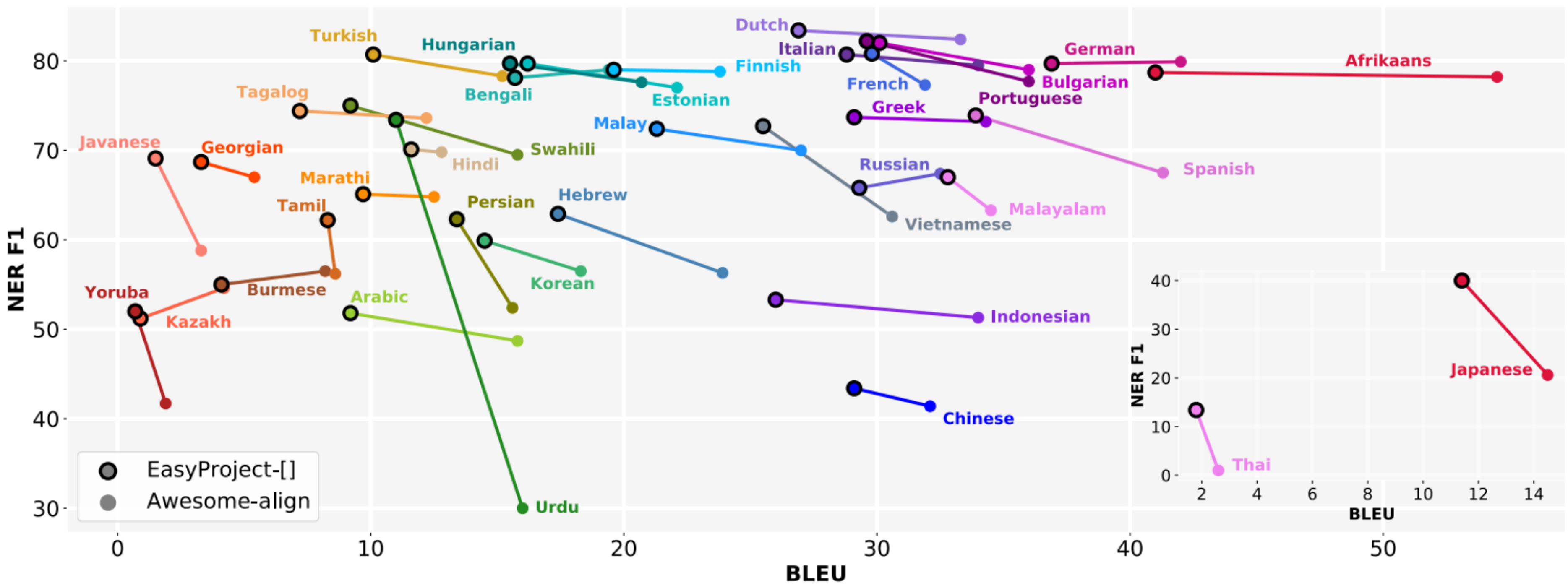



Figure 2: Comparison of translation quality and end-task performance for different label projection methods on the WikiANN dataset. EasyProject (§3.3) outperforms the alignment-based approach on $F_1$ scores for most languages, although inserting span markers degrade translation quality. The detailed experimental setting is in §4.1.

Chen et al., 2022

# Frustratingly Easy Label Projection for Cross-lingual Transfer (EasyProject)

| | *en* → Lang. | Fine-tune | | M2M+Word Aligner | | | M2M+Markers | | GMT+Word Aligner | | | GMT+Markers | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ref | $\text{XLM}_R$ | QAalign | awesome | $\text{awes}_{ft}$ | XML | EProj. ($\Delta_{\text{XLM}_R}$) | QAalign | awesome | $\text{awes}_{ft}$ | XML | EProj. ($\Delta_{\text{XLM}_R}$) |
| NER | *yo* | 41.3 | 37.1 | - | 46.8 | 48.3 | 56.0 | 58.3 (+21.2) | - | 72.1 | 66.1 | 71.8 | 73.8 (+36.7) |
| | *ja* | 18.3 | 18.0 | 19.3 | 21.1 | 24.9 | 40.2 | 40.8 (+22.8) | 19.3 | 23.0 | 22.6 | 42.0 | 43.5 (+25.5) |
| | *zh* | 25.8 | 27.1 | 43.1 | 40.5 | 39.5 | 42.4 | 44.4 (+17.3) | 45.2 | 43.3 | 39.6 | 43.8 | 45.9 (+18.8) |
| | *th* | 1.5 | 0.7 | - | 2.4 | 1.0 | 12.3 | 13.0 (+12.3) | - | 1.2 | 1.3 | 14.7 | 15.1 (+14.4) |
| | *ur* | 54.2 | 63.6 | - | 37.0 | 58.7 | 77.4 | 76.1 (+12.5) | - | 70.2 | 72.3 | 76.3 | 74.7 (+11.1) |
| | *he* | 54.1 | 56.0 | - | 56.6 | 56.3 | 60.1 | 62.1 (+6.1) | - | 59.6 | 60.2 | 63.7 | 67.1 (+11.1) |
| | *ms* | 69.8 | 64.1 | - | 70.7 | 71.8 | 74.2 | 73.9 (+9.8) | - | 73.0 | 73.8 | 73.2 | 74.1 (+10.0) |
| | *my* | 51.3 | 53.5 | - | 61.0 | 60.4 | 53.8 | 57.0 (+3.5) | - | 60.2 | 60.1 | 57.0 | 62.0 (+8.5) |
| | *ar* | 43.7 | 48.5 | 49.7 | 46.6 | 49.3 | 43.0 | 48.9 (+0.4) | 50.7 | 50.9 | 51.2 | 51.3 | 56.3 (+7.8) |
| | *jv* | 58.4 | 62.3 | - | 60.4 | 56.6 | 64.4 | 65.9 (+3.6) | - | 64.6 | 68.8 | 69.2 | 69.8 (+7.5) |
| | *tl* | 72.2 | 73.0 | - | 73.3 | 73.9 | 78.1 | 76.5 (+3.5) | - | 80.4 | 80.4 | 79.9 | 80.0 (+7.0) |
| | *hi* | 71.0 | 69.5 | - | 72.1 | 71.3 | 73.2 | 72.3 (+2.8) | - | 75.6 | 76.0 | 75.9 | 75.7 (+6.2) |
| | *ka* | 68.9 | 68.8 | - | 66.6 | 67.1 | 68.4 | 71.1 (+2.3) | - | 73.5 | 73.2 | 72.7 | 74.7 (+5.9) |
| | *bn* | 76.3 | 75.1 | - | 79.7 | 79.5 | 80.7 | 79.1 (+4.0) | - | 82.0 | 81.7 | 80.6 | 80.9 (+5.8) |
| | *ta* | 56.9 | 58.8 | - | 56.1 | 56.3 | 59.0 | 62.1 (+3.3) | - | 62.4 | 63.2 | 63.9 | 64.3 (+5.5) |
| | *eu* | 62.1 | 63.6 | - | - | - | - | - | - | 69.8 | 66.5 | 67.5 | 69.0 (+5.4) |
| | *ko* | 58.0 | 57.9 | - | 57.5 | 58.1 | 57.4 | 60.9 (+3.0) | - | 62.9 | 62.4 | 61.7 | 61.9 (+4.0) |
| | *mr* | 64.1 | 63.9 | - | 64.9 | 62.9 | 66.9 | 64.3 (+0.4) | - | 62.6 | 61.2 | 64.0 | 67.1 (+3.2) |
| | *sw* | 70.0 | 68.5 | - | 70.2 | 70.1 | 74.1 | 71.8 (+3.3) | - | 70.2 | 71.5 | 72.2 | 70.7 (+2.2) |
| | *te* | 52.3 | 55.6 | - | - | - | - | - | - | 57.4 | 56.8 | 57.6 | 57.4 (+1.8) |
| | *vi* | 77.2 | 74.2 | - | 64.1 | 62.7 | 75.4 | 74.9 (+0.7) | - | 70.4 | 67.2 | 77.5 | 76.0 (+1.8) |
| | *id* | 52.3 | 52.4 | - | 53.0 | 53.2 | 53.4 | 53.9 (+1.5) | - | 52.7 | 55.0 | 57.3 | 53.9 (+1.5) |
| | *ml* | 65.8 | 63.5 | - | 66.7 | 66.1 | 65.6 | 68.9 (+5.4) | - | 61.9 | 63.0 | 68.1 | 64.3 (+0.8) |
| | *es* | 68.8 | 74.8 | - | 69.8 | 68.2 | 71.8 | 73.5 (-1.3) | - | 71.3 | 72.6 | 73.5 | 75.6 (+0.8) |
| | *de* | 77.9 | 79.4 | 79.5 | 79.9 | 80.0 | 80.2 | 80.7 (+1.3) | 79.5 | 80.0 | 79.4 | 79.8 | 80.2 (+0.8) |
| | *kk* | 49.8 | 53.5 | - | 54.2 | 51.6 | 50.5 | 51.1 (-2.4) | - | 53.2 | 55.1 | 51.3 | 54.2 (+0.7) |
| | *fr* | 79.0 | 80.1 | 79.1 | 78.5 | 79.8 | 80.7 | 81.7 (+1.6) | 79.6 | 80.7 | 79.4 | 81.5 | 80.8 (+0.7) |
| | *af* | 77.6 | 78.6 | - | 77.5 | 78.4 | 78.9 | 79.1 (+0.5) | - | 79.1 | 78.9 | 79.0 | 79.2 (+0.6) |
| | *et* | 78.0 | 79.6 | - | 78.6 | 78.4 | 78.2 | 80.9 (+1.3) | - | 80.2 | 79.6 | 78.6 | 80.1 (+0.5) |
| | *hu* | 79.3 | 81.0 | - | 77.1 | 77.1 | 79.0 | 80.8 (-0.2) | - | 79.9 | 79.7 | 80.6 | 80.7 (-0.3) |
| | *fi* | 78.6 | 80.6 | - | 79.2 | 78.8 | 78.3 | 80.1 (-0.5) | - | 80.7 | 79.7 | 78.8 | 80.3 (-0.3) |
| | *it* | 81.1 | 81.3 | - | 80.2 | 80.5 | 80.7 | 81.1 (-0.2) | - | 80.3 | 80.4 | 81.1 | 80.9 (-0.4) |
| | *tr* | 78.9 | 80.3 | - | 78.4 | 77.8 | 82.7 | 82.0 (+1.7) | - | 80.1 | 80.2 | 81.5 | 79.6 (-0.7) |
| | *nl* | 84.3 | 84.1 | - | 82.1 | 82.3 | 83.1 | 84.2 (+0.1) | - | 83.5 | 82.9 | 83.0 | 83.1 (-1.0) |
| | *bg* | 81.2 | 82.1 | - | 80.3 | 79.9 | 82.1 | 81.3 (-0.8) | - | 80.9 | 79.7 | 82.5 | 80.6 (-1.5) |
| | *pt* | 79.6 | 82.0 | - | 76.8 | 78.0 | 81.9 | 82.3 (+0.3) | - | 79.0 | 80.2 | 80.6 | 80.1 (-1.9) |
| | *ru* | 71.5 | 71.1 | - | 67.4 | 67.9 | 67.6 | 67.8 (-3.3) | - | 67.4 | 66.8 | 67.4 | 68.2 (-2.9) |
| | *el* | 77.2 | 79.3 | - | 73.3 | 74.5 | 75.6 | 75.6 (-3.7) | - | 73.1 | 75.2 | 76.2 | 75.0 (-4.3) |
| | *fa* | 61.1 | 64.3 | - | 59.0 | 51.7 | 56.0 | 62.4 (-1.9) | - | 52.9 | 52.4 | 45.5 | 52.0 (-12.3) |
| AVG(-*eu*/*te*) | | 63.6 | 64.6 | - | 63.8 | 64.1 | 67.1 | 68.1 (+3.6) | - | 66.9 | 66.8 | 68.6 | 69.3 (+4.7) |

Chen et al., 2022

# Frustratingly Easy Label Projection for Cross-lingual Transfer (EasyProject)

| Event Extraction | | Fine-tune | M2M+Word Aligner | | | M2M+Markers | | GMT+Word Aligner | | | GMT+Markers | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\text{XLM}_R$ | QAalign | awesome | $\text{awes}_{ft}$ | XML | EProj. ($\Delta_{\text{XLM}_R}$) | QAalign | awesome | $\text{awes}_{ft}$ | XML | EProj. ($\Delta_{\text{XLM}_R}$) |
| *Arabic* | Entity | 69.2 | 74.3 | 73.7 | 74.1 | 73.7 | 73.4 (+4.2) | 74.4 | 74.3 | 74.0 | 73.7 | 74.0 (+4.8) |
| | Relation | 28.1 | 35.0 | 32.9 | 33.9 | 30.3 | 31.3 (+3.2) | 34.8 | 33.1 | 34.2 | 31.8 | 33.7 (+5.6) |
| | Trig-I | 42.7 | 43.9 | 44.0 | 44.0 | 44.5 | 44.1 (+1.4) | 43.6 | 44.2 | 43.7 | 43.8 | 44.0 (+1.3) |
| | Trig-C | 40.0 | 42.2 | 42.0 | 42.3 | 42.7 | 42.3 (+2.3) | 41.8 | 42.6 | 42.0 | 41.5 | 42.0 (+2.0) |
| | Arg-I | 33.5 | 37.9 | 36.3 | 37.8 | 37.8 | 38.1 (+4.6) | 37.7 | 37.9 | 37.6 | 36.9 | 37.8 (+4.3) |
| | Arg-C | 30.8 | 34.9 | 33.6 | 34.7 | 35.1 | 35.1 (+4.3) | 34.6 | 35.2 | 34.5 | 34.1 | 35.2 (+4.4) |
| | AVG | 40.7 | 44.7 | 43.7 | 44.5 | 44.0 | 44.1 (+3.4) | 44.5 | 44.5 | 44.3 | 43.6 | 44.4 (+3.7) |
| *Chinese* | Entity | 59.1 | 67.5 | 70.2 | 69.0 | 69.0 | 70.5 (+11.4) | 67.1 | 68.8 | 70.6 | 70.2 | 71.0 (+11.9) |
| | Relation | 20.4 | 33.5 | 30.0 | 26.2 | 25.6 | 35.0 (+14.6) | 30.7 | 28.2 | 30.1 | 35.6 | 28.4 (+8.0) |
| | Trig-I | 25.0 | 45.0 | 52.2 | 50.9 | 41.9 | 46.8 (+21.8) | 43.7 | 53.5 | 50.0 | 50.7 | 52.6 (+27.6) |
| | Trig-C | 23.9 | 42.5 | 49.0 | 47.1 | 37.6 | 43.4 (+19.5) | 40.8 | 50.0 | 46.6 | 47.4 | 49.3 (+25.4) |
| | Arg-I | 28.6 | 38.1 | 40.2 | 39.1 | 37.6 | 41.7 (+13.1) | 38.7 | 39.6 | 39.4 | 39.8 | 40.1 (+11.5) |
| | Arg-C | 28.1 | 36.7 | 39.4 | 38.0 | 34.8 | 40.0 (+11.9) | 37.3 | 38.4 | 38.2 | 38.2 | 38.2 (+10.1) |
| | AVG | 30.8 | 43.9 | 46.8 | 45.0 | 41.1 | 46.2 (+15.4) | 43.0 | 46.4 | 45.8 | 47.0 | 46.6 (+15.8) |

Chen et al., 2022

# Other Major Programs

- DARPA LORELEI
- IARPA BETTER
- DARPA TIDES

# What am I missing from this lecture?

# What am I missing from this lecture?

- Modeling!

# What am I missing from this lecture?

- Modeling!
- CRFs
- Seq2Seq
- Encoders