Recent Advances in Transferable Representation Learning

Muhao Chen, Kai-Wei Chang, Dan Roth

Cross-Lingual Representations and Transfer

AAAI 2020 Tutorial

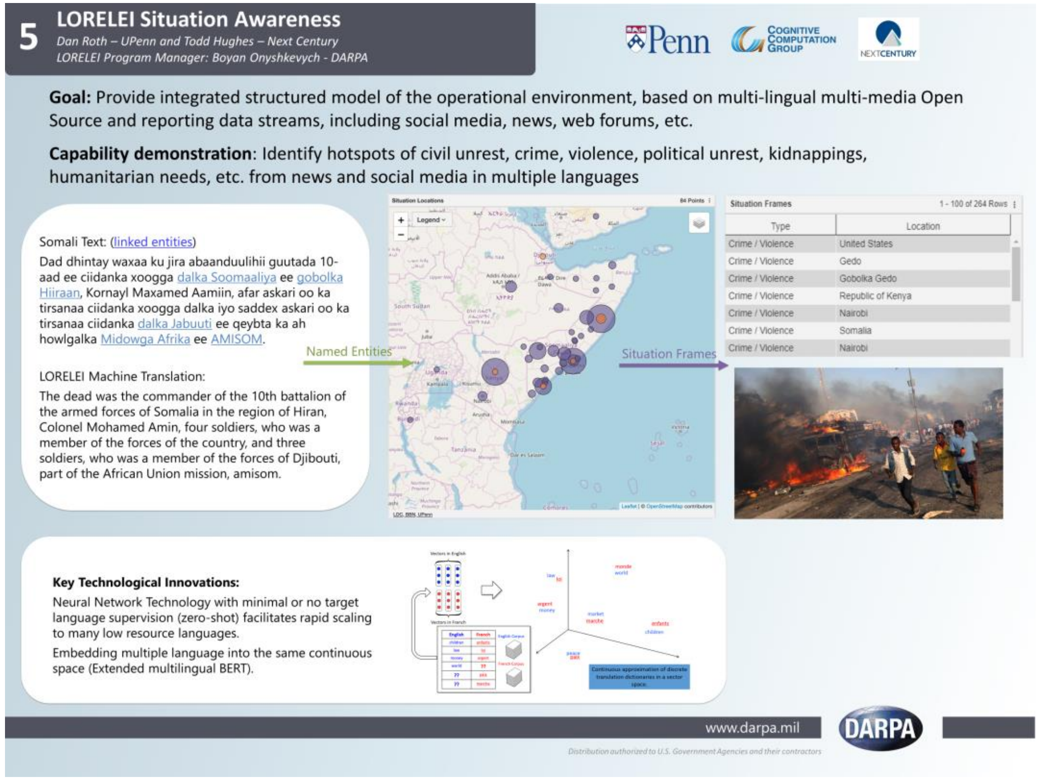
Cross-Lingual Natural Language Processing

- Goal: Given text data in a low resource language,
  - ☐ Can we "understand" it even if you only know English?
  - ☐ **No training data** in the low resource language!

Somali streaming data → Situation Awareness (described in English)

Goal

- "Understand" a situation described in Target Language
  - ☐ Identify Entities & Concepts (NER)
  - ☐ Ground in English Resources (EDL)
  - ☐ Type the situation



- We know how to develop models when it’s possible to give a lot of good annotated data
- The goal is not really to solve NER for low resource languages; not even for 100 LRLs
  - NER is “relatively easy” to annotate, and this can be done in a cheaper way by hiring enough annotators

- The goal is to advance tasks for which it’s not realistic to annotate exhaustively
  - Situation Frames is a much better representative: Involves event identification and textual entailment.

- Over the last few years, the LORELEI program of DARPA funded many teams to develop *methods* and *insights* and use these to develop capabilities for tasks for which there will never be enough directly annotated data

Success:

- **EDL, SF:** no training on target language data
  - Only incidental signals: Pre-training, data in other languages, Wikipedia, language specific knowledge
- **NER:** + Bootstrapping with non-speaker, weak supervision.

Final LORELEI Evaluation Results

- Setting:
  - Two surprise languages
    - 2019: Odia, Ilocano
  - You are given a week to develop solutions for several tasks.
  - **No annotated data**; some target language data; a (limited quality) dictionary.
  - Minimal (4 hours) remote exposure to native speakers (NIs)

	Named Entities	Entity Linking	Situation Frames
IL11 (Odia)	**79.4**	**56.6**	**37.8, 16.8,** 20.3 (Type, Type+Loc, nDCG)
IL12 (Ilocano)	**79.5**	54.7 (2nd)	34.1, 9.73, **25.65**

Results in **blue** are the **top scores** among all participants

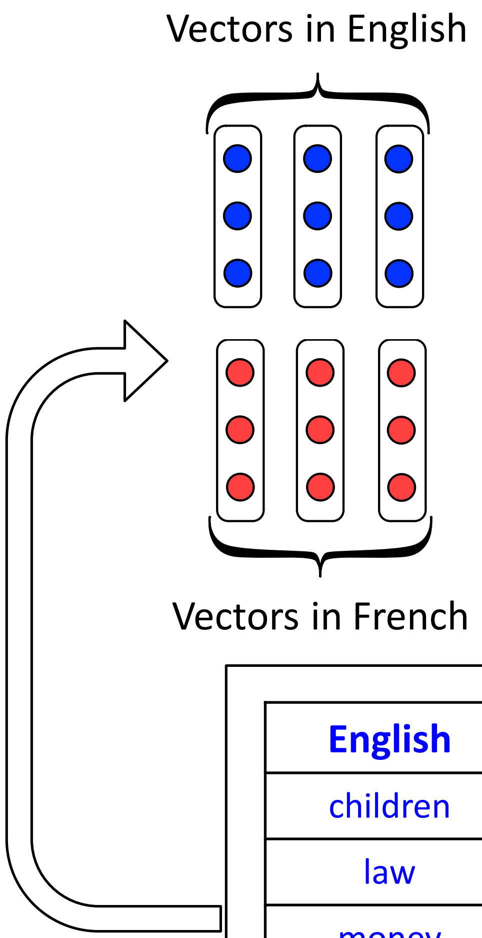
- Use of many incidental signals:
- Human knowledge:
  - non-speaker annotation (+IL) needed to bootstrap models; declarative knowledge
- Use of data that is out there, unrelated to the task: other languages & other tasks
  - Using "cheap" translation; existing textual entailment data; Wikipedia; Google query logs
- (Unsupervised) Pre-training of representations (extended M-BERT)

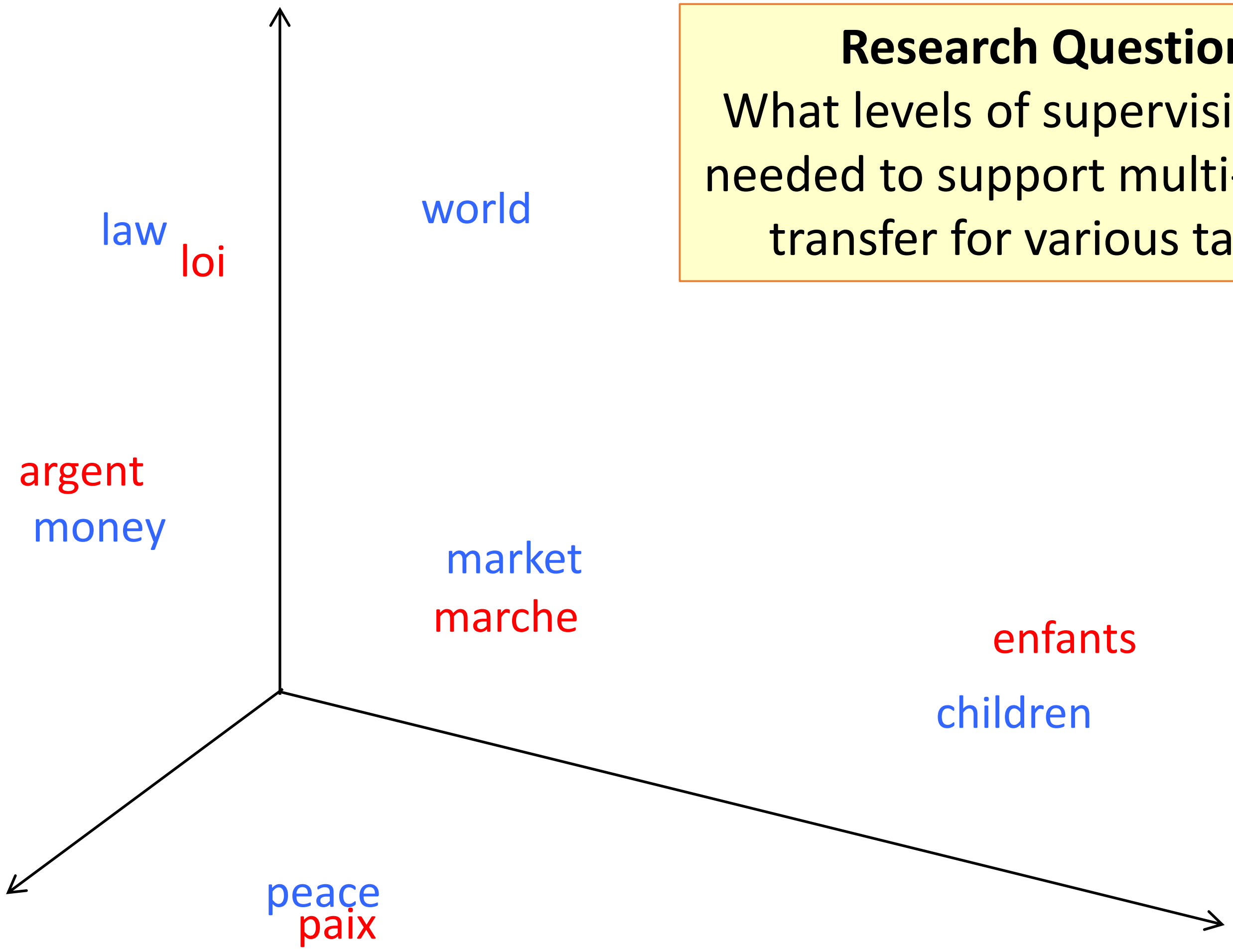
Outline

- A perspective
  - What happened and where we are
- Weak signals from humans
  - The role of non-speakers in low-resource languages
  - A sanity check
- Towards understanding M-BERT
  - Looking at what makes a difference and what doesn’t

What Happened?

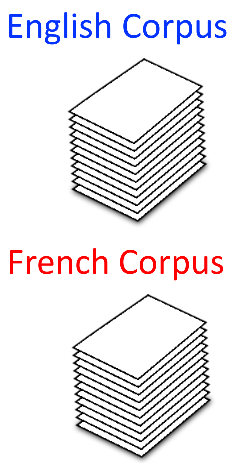
Cross-lingual Representations [Upadhyay et al. (ACL'16)]





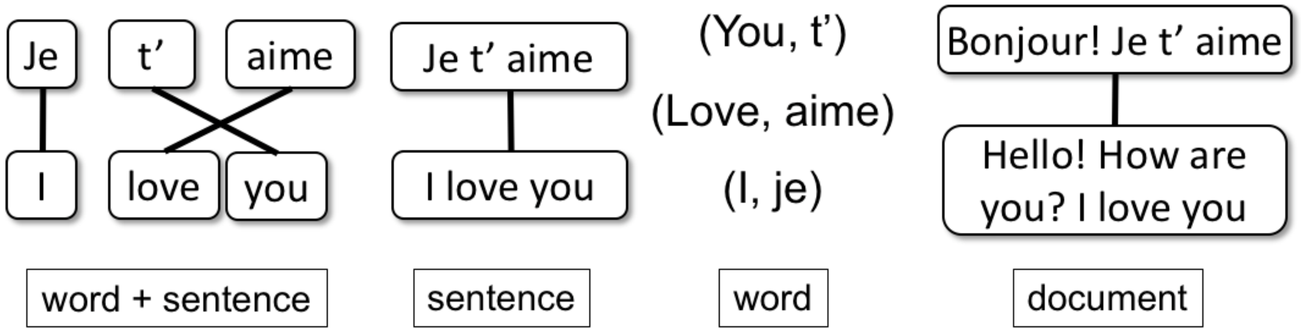
Research Question
What levels of supervision are needed to support multi-lingual transfer for various tasks?

English	French
children	enfants
law	loi
money	argent
world	??
??	paix
??	marche



Continuous approximation of discrete translation dictionaries in a vector space.

Recap: Learning Cross-lingual Representations



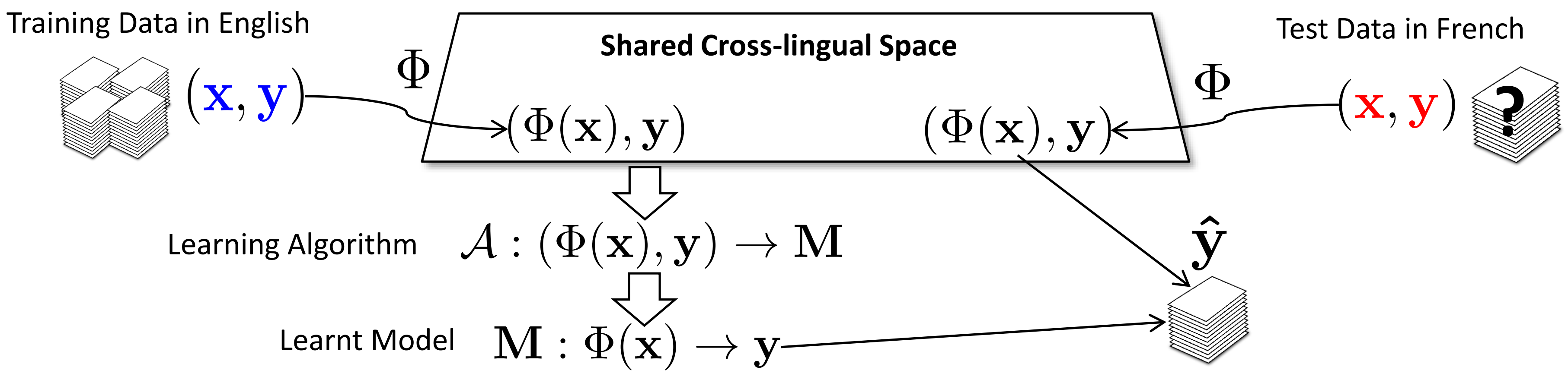
Upadhyay et al. (ACL 2016)

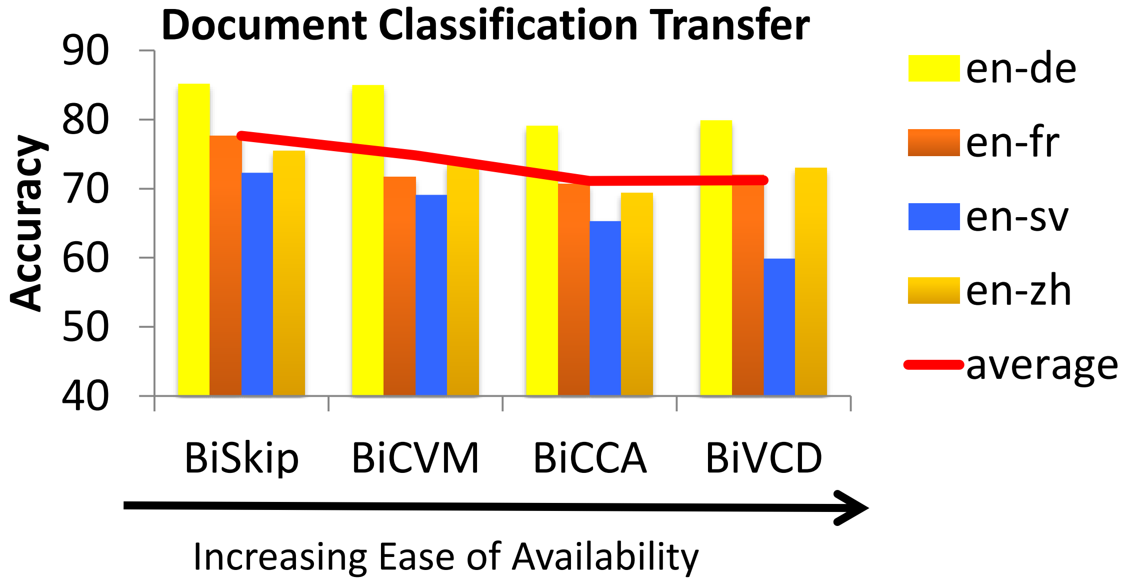
High-Resource Languages (e.g., English)

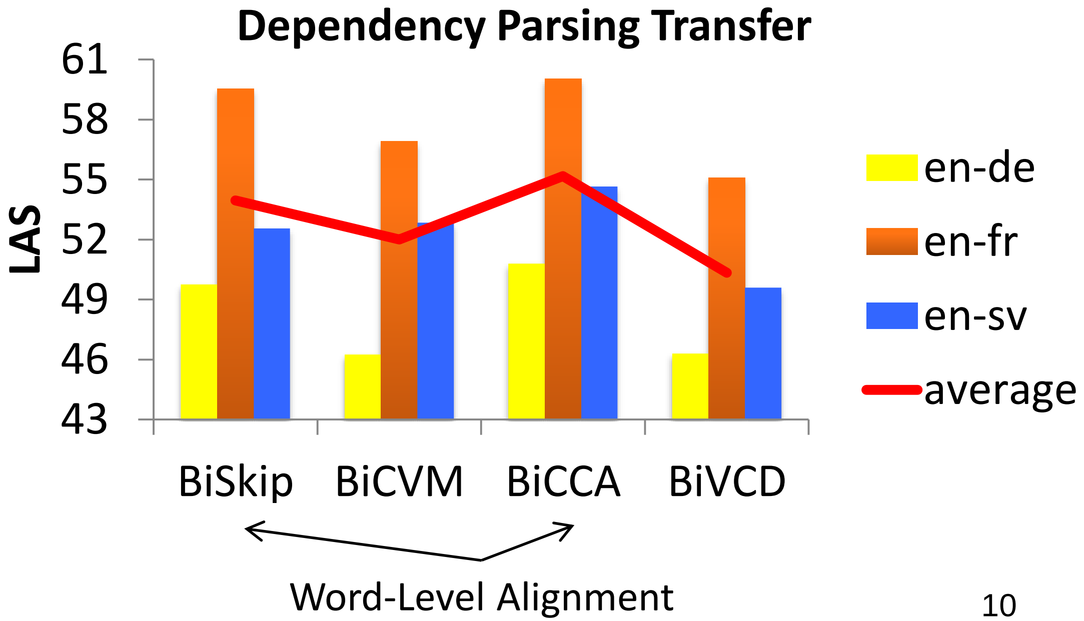
Other Languages

$$\underbrace{\mathbf{W}^*}_{\text{Vectors for English}}, \overbrace{\mathbf{V}^*}^{\text{Vectors for French}} \leftarrow \operatorname{argmin} \alpha \underbrace{A(\mathbf{W})}_{\text{Mono. Obj. for English}} + \beta \overbrace{B(\mathbf{V})}^{\text{Mono. Obj. for French}} + \underbrace{C(\mathbf{W}, \mathbf{V})}_{\text{Cross-lingual Obj.}}$$

Facilitating Model Transfer







1st Cross-lingual Entity Linking (XEL) [Tsai & Roth NAACL' 16]

Given a non-English document, extract named entities and disambiguate into the English Wikipedia (KB)

… its lead singer Nunn left Berlin to audition for Star Wars …

... sein Leadsänger Nunn verließ Berlin, um für Star Wars vorzuspielen ...

… 其主唱 纳恩 离开 柏林 去参加 星球大战 的试镜 …

Terri Nunn

From Wikipedia, the free encyclopedia

Terri Kathleen Nunn (born June 26, 1961[1]), is an American singer and actress. She is best known as the lead vocalist of the new wave/synthpop band Berlin.

Terri Nunn

Terri Nunn performing at Oracle OpenWorld 2010.

Contents [hide]

Berlin (band)

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(August 2013) (Learn how and when to remove this template message)*

Berlin is an American new wave band. The group was formed in Orange County in 1978 by John Crawford (bass guitar). Band members included Crawford, Terri Nunn (vocals), David Diamond (keyboards), Ric Olsen (guitar), Matt Reid (keyboards) and Rod Learned (drums). The band gained mainstream-commercial success in the early 1980s with singles including "The Metro", "Sex (I'm A...)", "No More Words" and then in

Berlin

Star Wars

From Wikipedia, the free encyclopedia

This article is about the film series and media franchise. For the original 1977 film, see Star Wars (film). For other uses, see Star Wars (disambiguation).

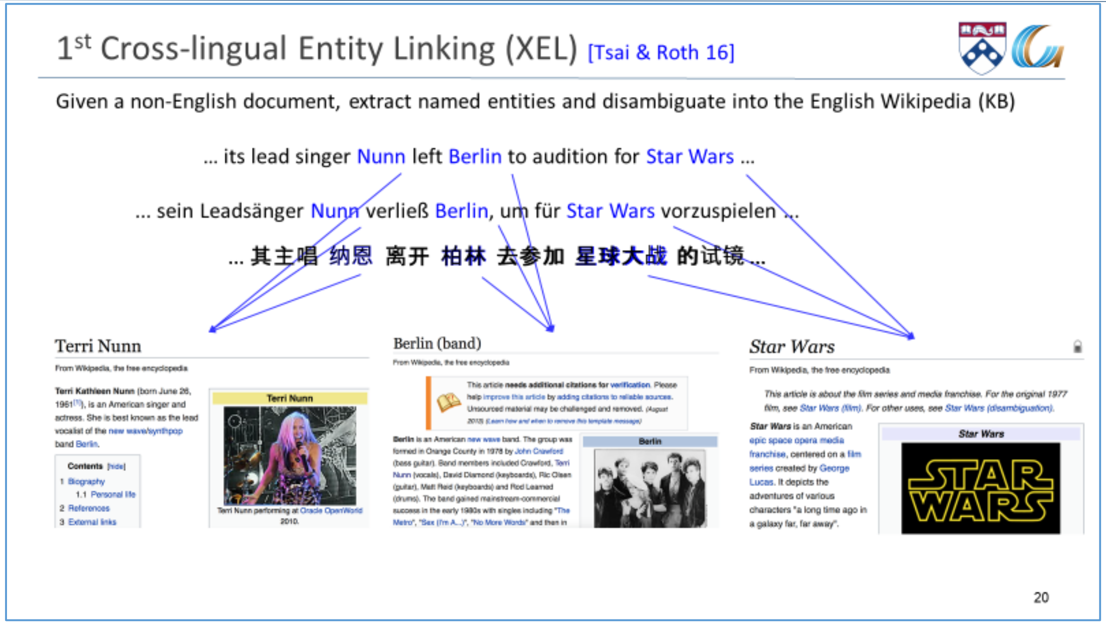
Star Wars is an American epic space opera media franchise, centered on a film series created by George Lucas. It depicts the adventures of various characters "a long time ago in a galaxy far, far away".

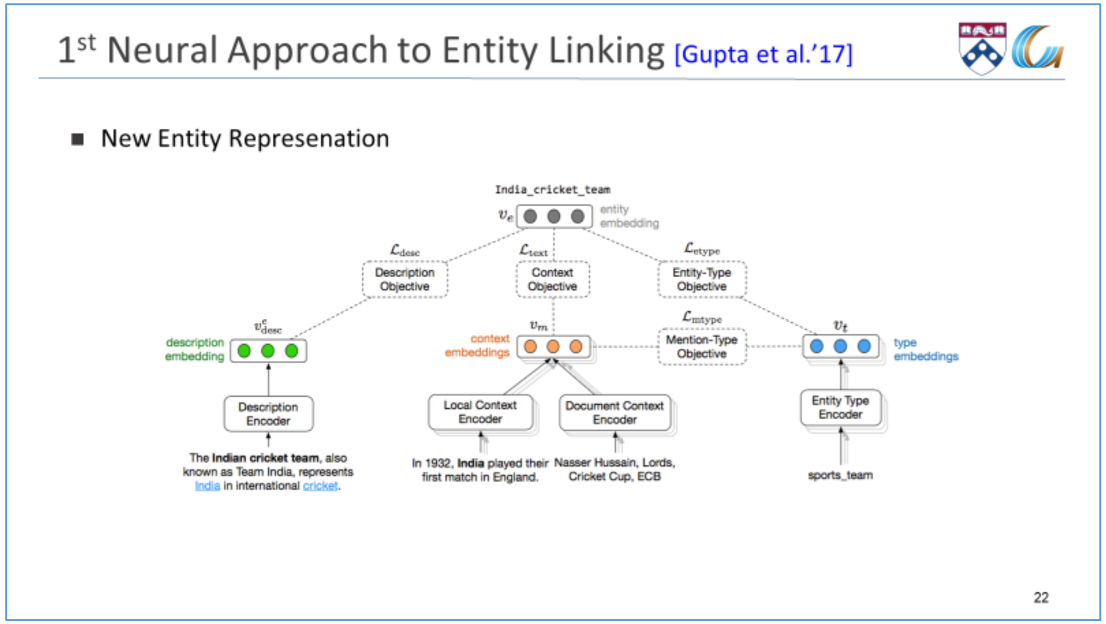
Star Wars

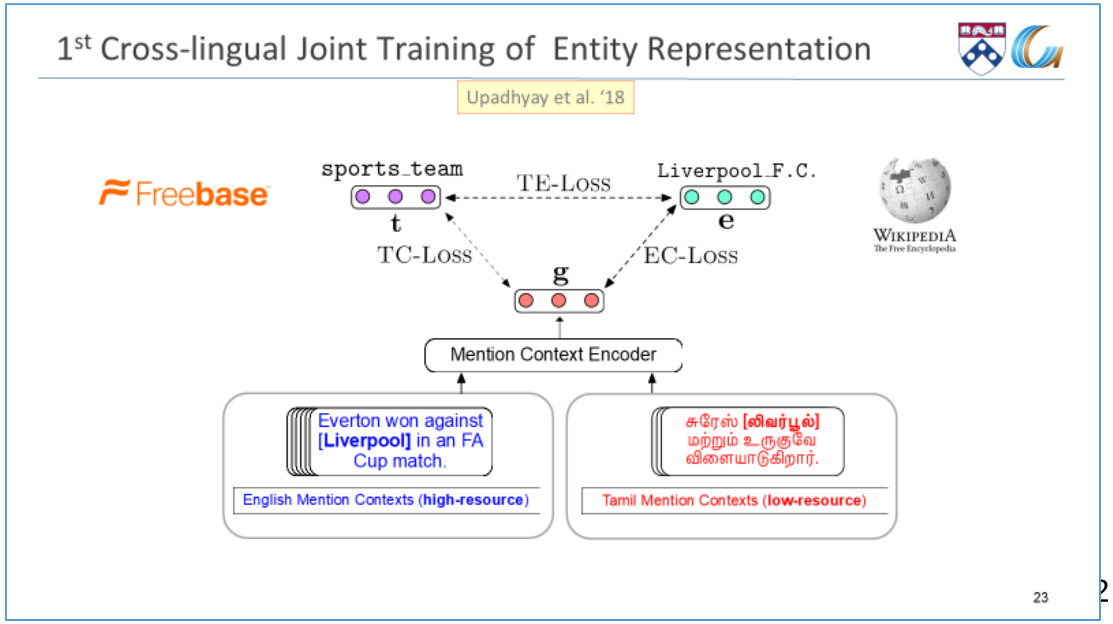
Joint Multilingual Supervision for Cross-lingual EDL

[Upadhyay et al. (EMNLP'18)]

Trains a single entity linking model for multiple languages by combining supervision from multiple languages thus facilitating EDL on low resource languages.

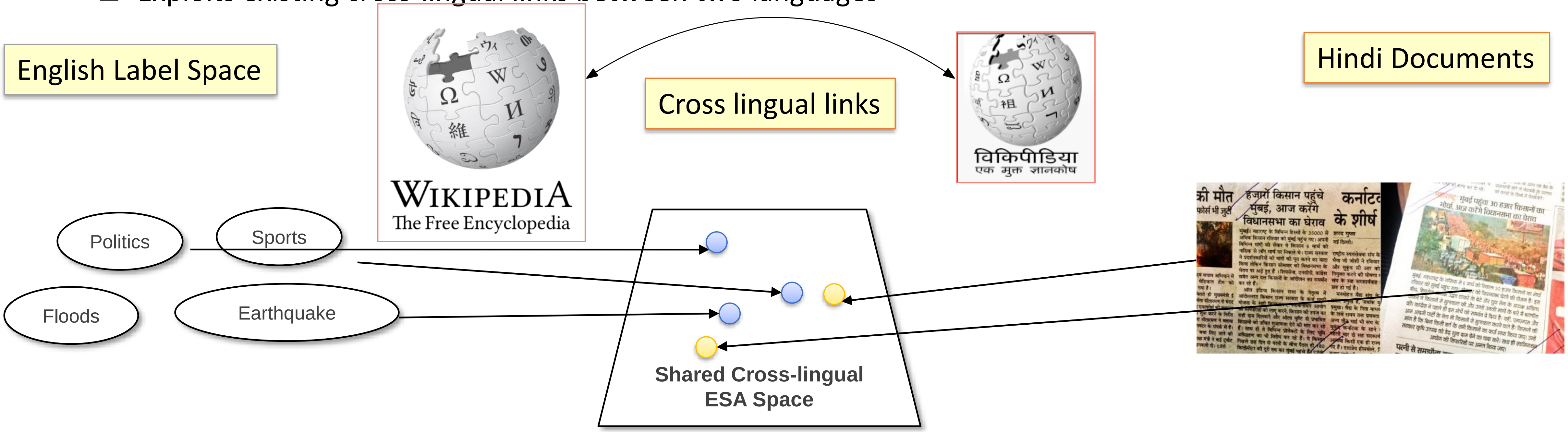






Cross-lingual Textual Classification [Song et al. IJCAI'16]

- Text Classification with No Annotated data
- A Wikipedia Based Representation: Cross-lingual Explicit Semantic Analysis (CL-ESA)
  - Exploits existing cross-lingual links between two languages



  - Represent low-resource language documents and English Labels in the corresponding CL-ESA Space
  - Map representation via cross-lingual links.
  - Quality depends on the (size of the) intersection of the title spaces

Single Document Classification (88 Language shown)

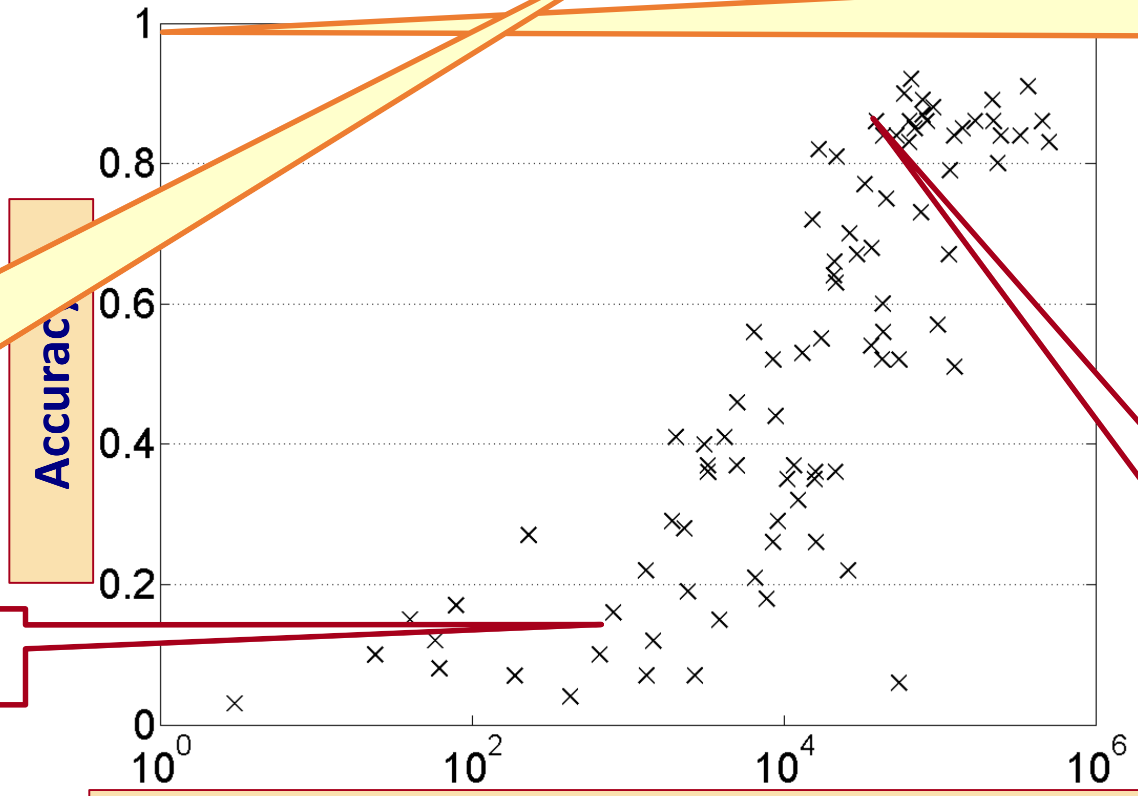
Extended to all 292 languages
represented in Wikipedia

Results in English are equivalent to having **1000-2000 labels per category**. In cross-lingual classification, equivalent to having **> 200 labels per category.**

Challenge I
Language Transfer

Challenge II
Beating supervised approaches
Better composition of ESA representations.

Dataless classification for English



Hausa

Hindi

Size of shared English-Language L title space

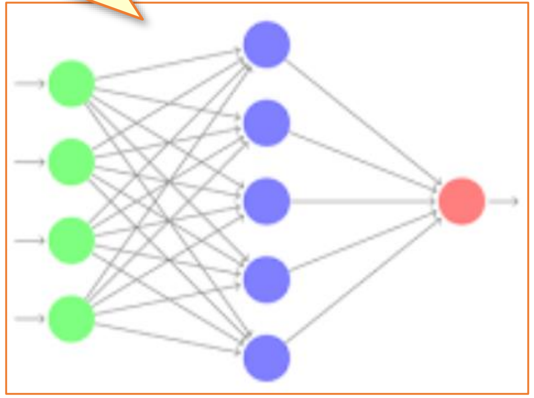
What Happened?

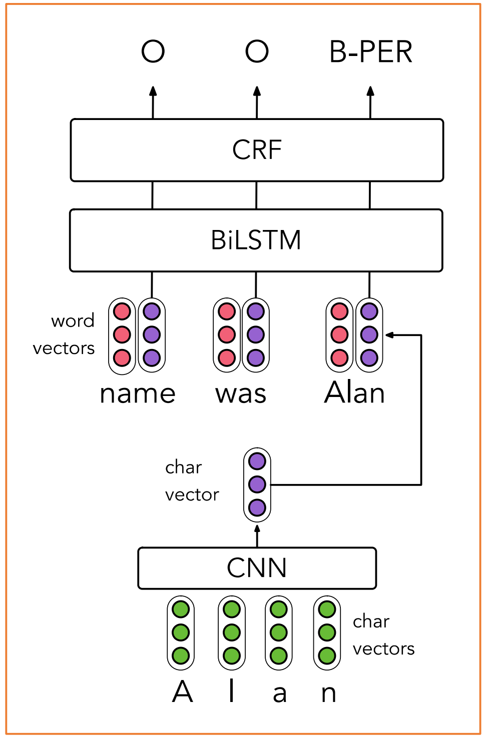
- It's been established that *multilingual embeddings* are essential to Low Resource work.
  - NER, EDL, SF all rely on these representations.
- BERT
  - A powerful **contextual** language model
  - M-BERT: a multilingual version – multilingual embeddings
  - A single multilingual embedding for many languages.
  - **No direct supervision** – only needs sufficient data in each languages.

- Neural Everywhere
  - Earlier approaches were replaced by neural models that necessitate the use of embeddings
  - Some simpler, robust, methods disappear

Bi-LSTM-CRF

Averaged Perceptron





- Many questions remains
  - Some are addressed next in the context of NER

Massively Multilingual Analysis of NER

- Low-resource NER:
  - different methods, parameters, languages
- Evaluation in 22 languages (LORELEI)
  - 10 different scripts
  - 10 language families (Niger-Congo most popular)
- Methods:
  1. Monolingual
  2. Transfer with cross-lingual embeddings
  3. Transfer with Cheap Translation [Mayhew et al. EMNLP ‘17]
  4. Transfer with M-BERT

Language	3 letter code
Akan (Twi)	aka
Amharic	amh
Arabic	ara
Bengali	ben
Farsi	fas
Hindi	hin
Hungarian	hun
Indonesian	ind
Chinese	cmn
Russian	rus
Somali	som
Spanish	spa
Swahili	swa
Tagalog	tgl
Tamil	tam
Thai	tha
Turkish	tur
Uzbek	uzb
Vietnamese	vie
Wolof	wol
Yoruba	yor
Zulu	zul

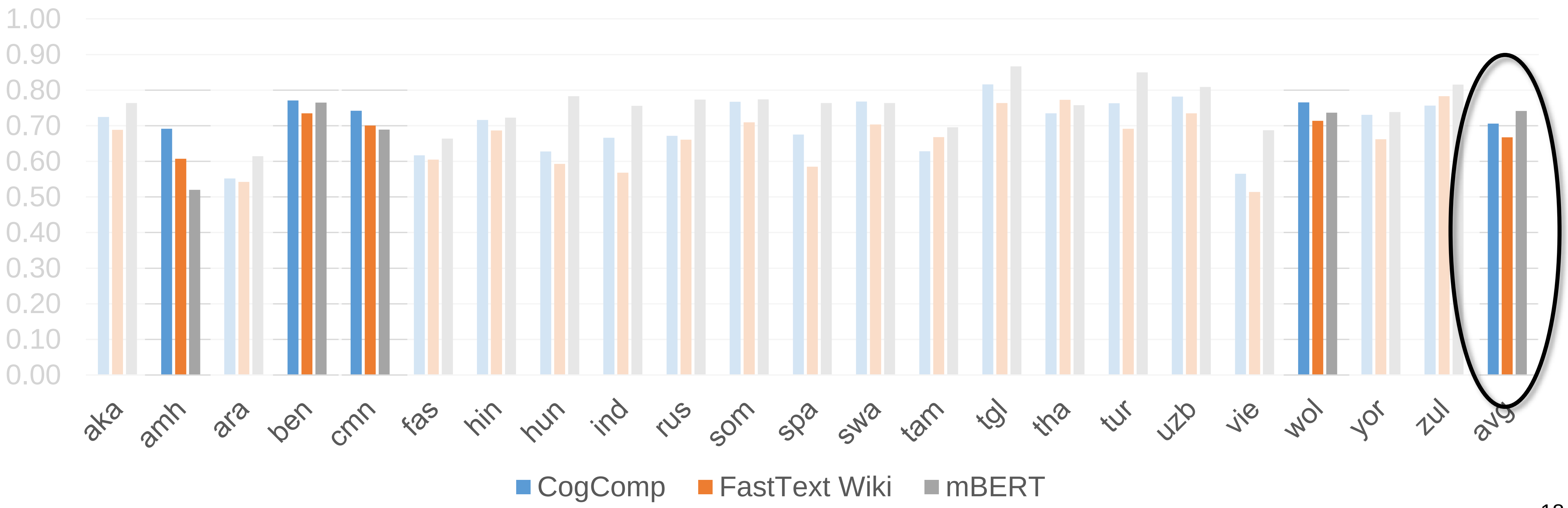
(i) Monolingual experiments

- Train on target language text
- Useful as an upper bound
- CogComp & BiLSTM-CRF (2 embs)

Averaged Perceptron based [Ratinov & Roth 09]

Takeaways:

- Don't discount non-neural systems
- Average about 70 F1

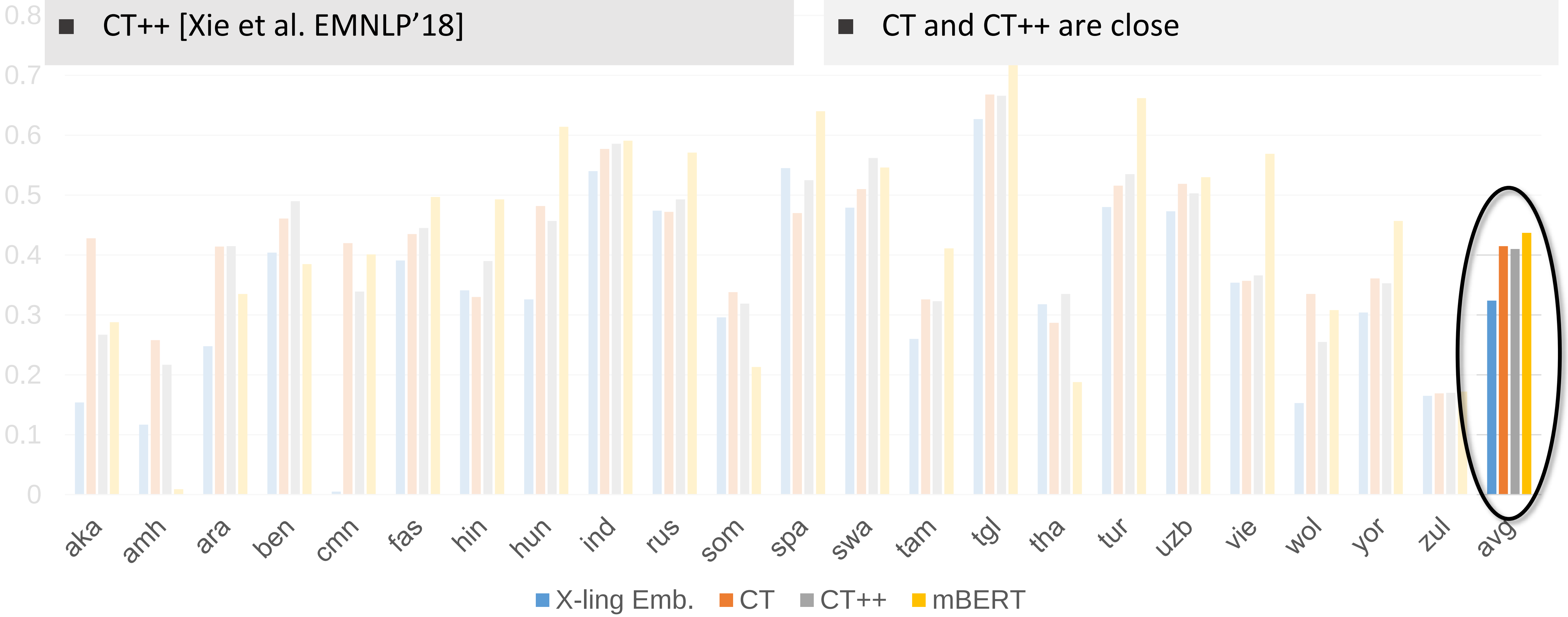


(ii) Cross-Lingual Results

- English Annotation as input
- Cheap Translation [Mayhew et al. EMNLP '17]
- CT++ [Xie et al. EMNLP'18]

Takeaways:

- M-BERT is best
- CT and CT++ are close

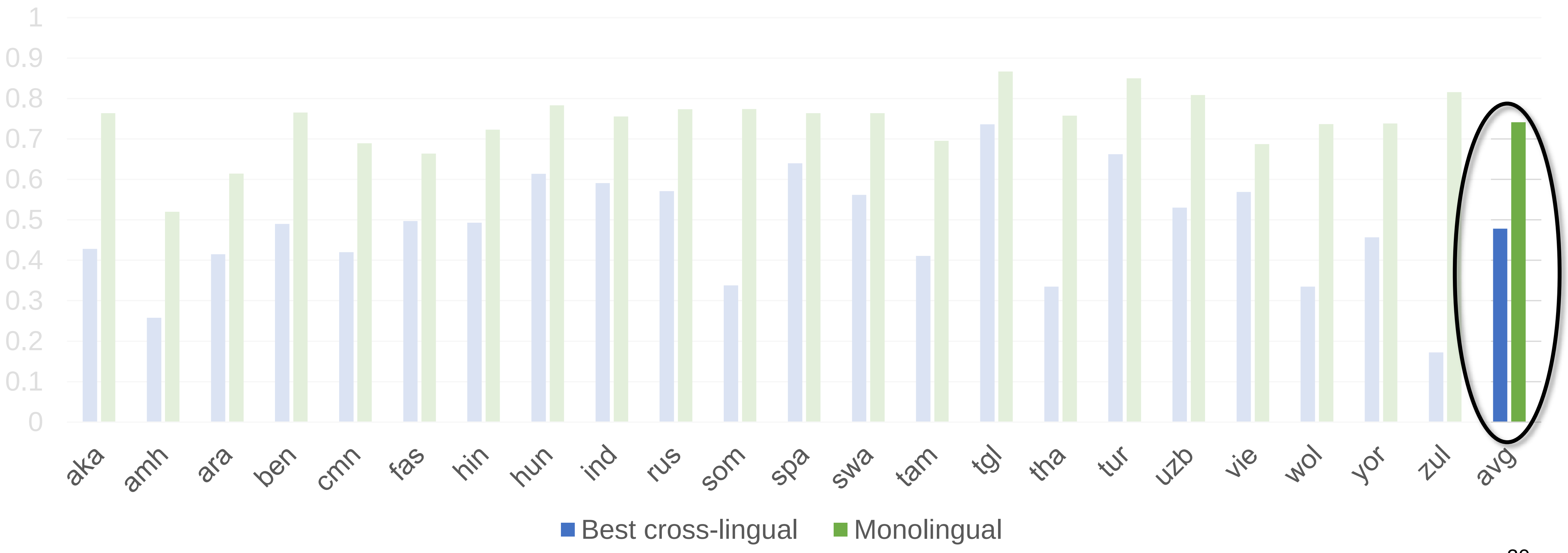


Overall: Still Ways to Go

- Average of best cross-lingual (47 F1) is still less than monolingual (74 F1)

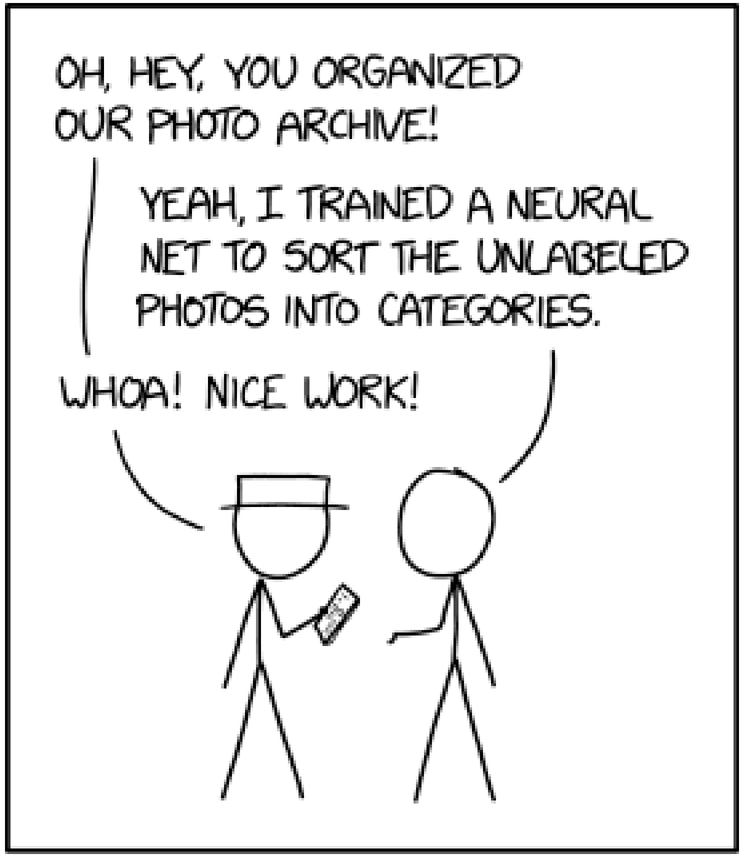
Takeaways:

- Cross-lingual transfer by itself isn't sufficient



Outline

- A perspective
  - What happened and where we are
- Weak signals from humans
  - The role on non-speakers in low-resource languages
  - A sanity check
- Towards understanding M-BERT
  - Looking at what makes a difference and what doesn't



ENGINEERING TIP:
WHEN YOU DO A TASK BY HAND, YOU CAN TECHNICALLY SAY YOU TRAINED A NEURAL NET TO DO IT.

NER with Partial Annotation

Romanized Bengali

ebisi’ra giliyyaana phinnddale aaja pyaalestaaina adhiinastha gaajaa theke aaja raate ekhabara jaaniyyechhena .

ABC’s Gillian Findale has reported from Gaza in Palestine today.

- Weak signals:
  - High precision, low recall signal
  - [Mayhew et al.’CoNLL 19] describes an algorithmic approach that allows training high quality NER from such partial annotation given by non-native speaker.

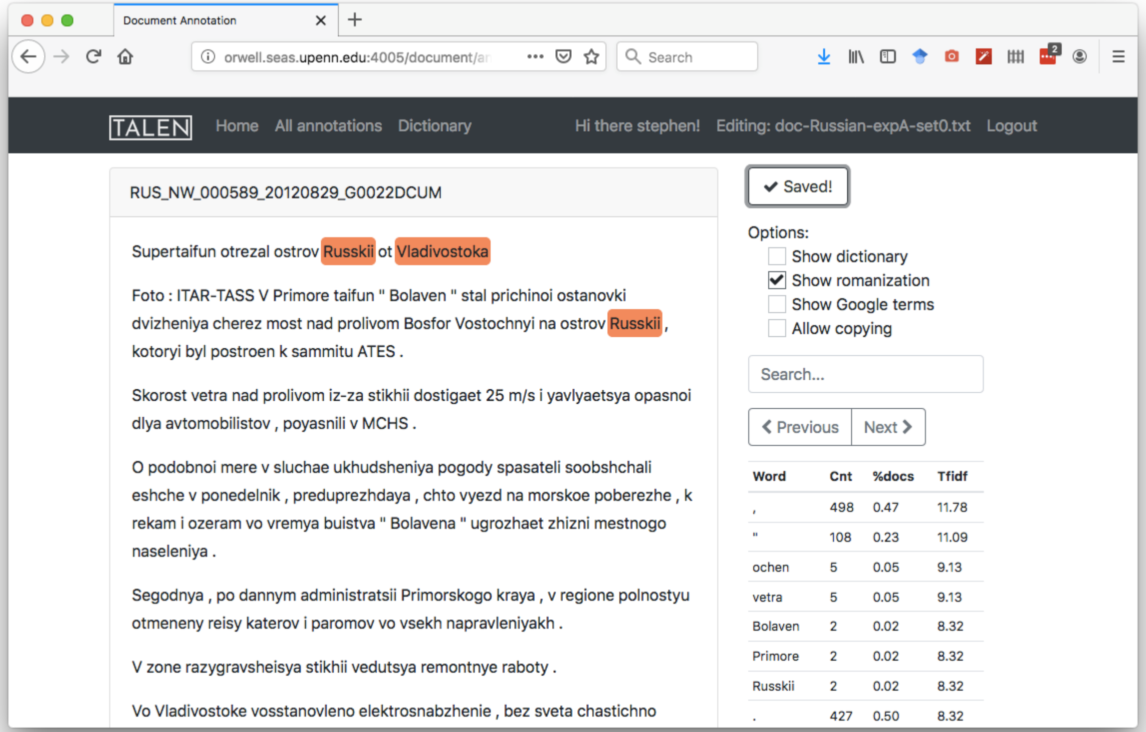
What Can We Get from Non-Speakers?

Experimental questions:

1. Can non-speaker (NS) annotators produce meaningful annotations?
2. How to compare NS annotations against a native informant (NI)
3. How best to combine NS/NI annotations?

Human annotation experiment:

- Annotating Russian data using TALEN [Mayhew & Roth, ACL-Demo’18]
  - Many bells and whistles
- All annotations are done on romanized gold annotated data
  - Gold annotations removed
- Completed over 4 sessions (45 min each)
- NI, 3 committed NS annotators, 4+ non-committed NS annotators



Non-Speaker (NS) Annotation: How Good Is It?

Fixed wall clock time: 3 hours

	NI	Combined NS
Annotation time	3 hrs	~15 person hrs

- **Not Good**, but comparable in performance with an NI model

How Best to Use the NI?

Don't get too excited:
- NS (and NI) annotation are not sufficient, but are essential to bootstrap a model.
- Our best models make use of it, but improve by 30+%, with cross-lingual training.

Giving the NI pre-annotated documents (by the NS) dramatically increases efficiency (2x)

While X-lingual embeddings aren't enough
- Other weak signals can help a lot.

	NI from scratch
Annotation time	1 hr
Dataset size (tokens)	**4k**
Annotation quality (F1)	75.7
Model performance (F1)	**30.6**

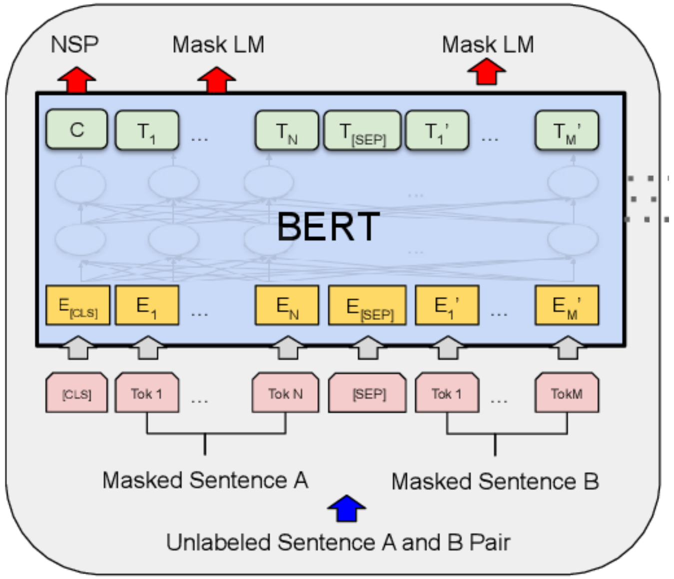
Talk Outline

- A perspective
  - What happened and where we are
- Weak signals from humans
  - The role on non-speakers in low-resource languages
- Towards understanding M-BERT
  - Looking at what makes a difference and what doesn't
  - Presentation based on:
    - **Cross-Lingual Ability of Multilingual BERT: An Empirical Study**, Karthikeyan K, Zihan Wang, Stephen Mayhew, Dan Roth, ICLR'20

Multilingual BERT

- BERT
  - A transformer-based pre-training language model.
  - Training objectives: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP)
  - Input: A pair of sentences S1 and S2, such that half of the time S2 comes after S1 in the original text and the other half of time S2 is a randomly sampled sentence.
  - Data: English Wikipedia and Books corpus

- Multilingual BERT (M-BERT) ?
  - Same training procedure as BERT except the data.
  - Data – Wikipedia text from top 104 languages
  - **No specific cross-lingual objectives or any cross-lingual data.**



Pre-training

Surprisingly Cross Lingual

- Cross-lingual: Train on one language and test on another language
- M-BERT is trained without any cross-lingual objectives but it is cross-lingual

System	English	Chinese	Spanish	German	Arabic	Urdu
XNLI Baseline - Translate	73.7	67.0	68.8	66.5	65.8	56.6
BERT - Translate Train	81.4	74.2	77.3	75.2	70.5	61.7
M-BERT - Transfer	**81.4**	**63.8**	**74.3**	**70.5**	**62.1**	**58.3**

- We can see that M-BERT transfers from English to other languages very well

[XNLI paper: Conneau et al.'18]

Why is M-BERT Cross-lingual?

- What components of M-BERT are important for its cross-lingual ability?

- We consider three dimensions:

  - **Linguistics**: What is the contribution of word-piece overlap and language similarity?

  - **Architecture**: How do depth, multi-head attention, and total number of parameters affect the cross-lingual ability of M-BERT?

  - **Input and Learning Objective**: Is Next Sentence Prediction or language identification really important? Is word or character vocabulary better than word-piece vocabulary ?

Experimental Setting

- Languages:
  - English is always the source language
  - 3 typologically different target languages: Spanish, Hindi, and Russian

- Bilingual BERT (B-BERT) – BERT trained on two languages.
  - B-BERT trained on language A and B is denoted as A-B
  - en-hi -- B-BERT trained on English (en) and Hindi (hi), similarly for Spanish (es) and Russian (ru)

- Tasks:
  - Two conceptually different tasks: Textual Entailment and named entity recognition (NER)
    - TE: the XNLI dataset
    - Cross-lingual NER: LORELEI dataset

Linguistics (1): Word-piece Overlap

- BERT's representation is based on word-pieces
- Hypothesis: M-BERT works due to overlapping word-pieces
  - M-BERT generalizes across languages because of **shared word-pieces across languages** that are mapped to a shared feature space.
- Indeed, texts in different languages share some common word-piece vocabulary
  - Numbers, named entities, even actual words (when the script is shared).
  - We refer to this as word-piece overlap.

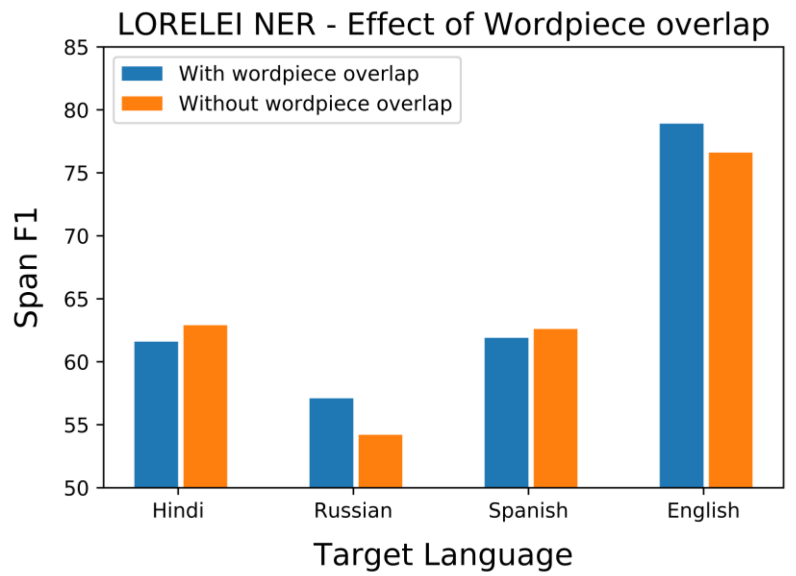
Removing word-piece overlap

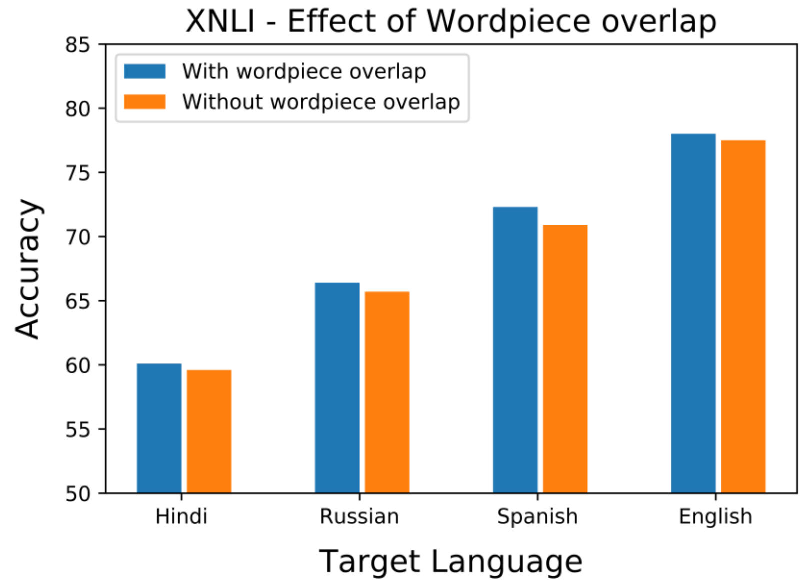
- To study the effect of word-pieces we should compare the models with and without it, but how to get a model without word-piece overlap?
  - Fake English (enfake)

- Fake-English
  - Shifting the Unicode of each character in English Wikipedia text by a large constant so that there is strictly no character overlap with any other Wikipedia text
  - English and Fake-English don't share any vocabulary/characters, but they have exactly the same structure.
  - We measure the contribution of word-piece overlap as the drop in performance when we use Fake-English instead of English.

Impact of word-piece overlap

- Setting:
  - Train a pair of B-BERTs
    - English-L vs. FakeEnglish-L
  - Tune on English (FakeEnglish, resp.) data
  - Test on the target Language
  - (Right most: Eng-FakeEnglish B-BERT; tune on FakeEnglish; test on both)





- B-BERT is cross-lingual even when there is absolutely no word-piece overlap

Linguistics (2): Structural Similarity

- Structure of a language
  - In today's context, we define "structure" to include every aspect of the language that is invariant to its script.
  - E.g., morphology, word-ordering, word frequency, word-pair frequency, etc.

- Note that English and Fake-English do not share any vocabulary or characters
  - but they have exactly the same structure

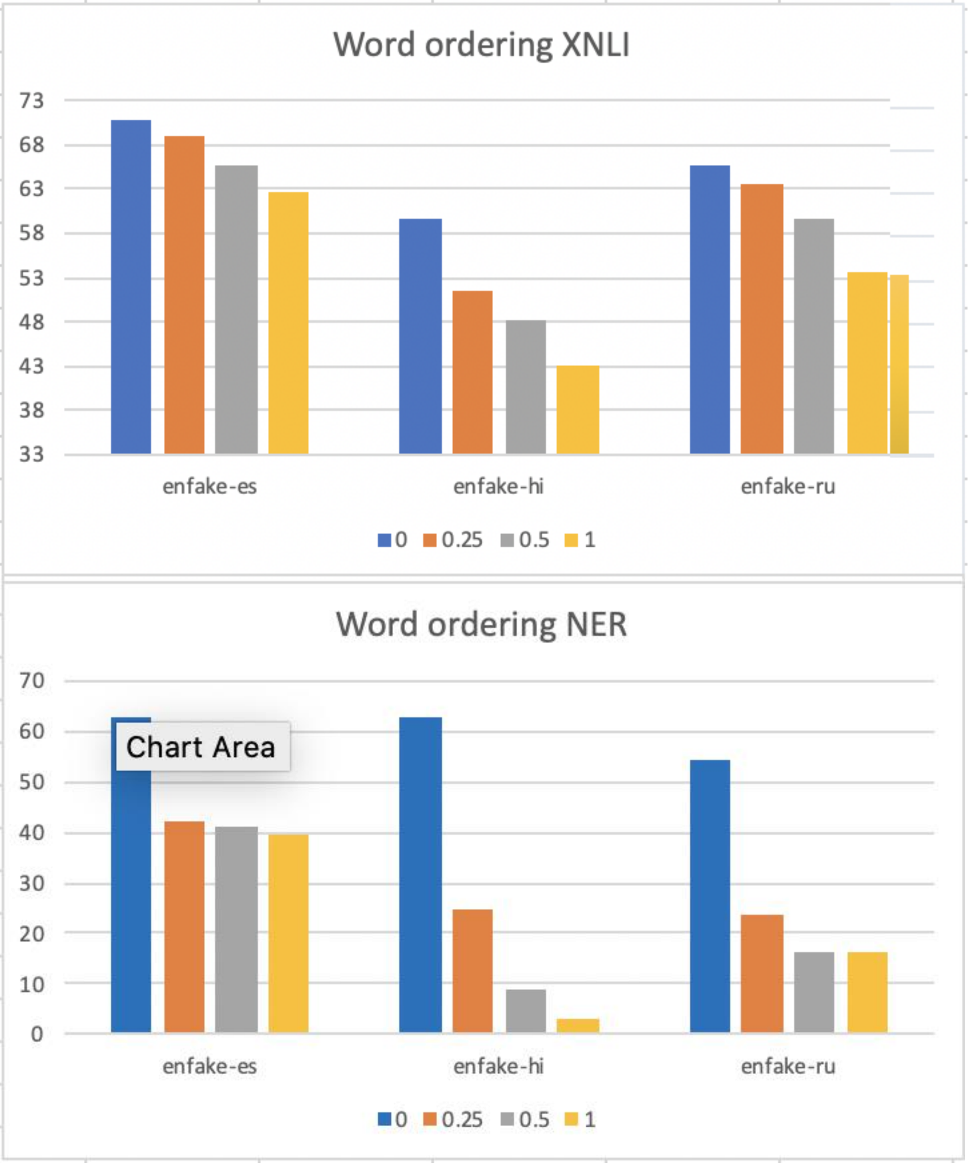
- Intuitively, English and Spanish are more "structurally similar" than (English, Hindi) and (English, Russian).

- Hypothesis: There is some similarity of **word ordering** in each language
- Hypothesis: Bert can rely on similarity of **frequency of words** to learn cross-lingually.

Eliminating word order

- We eliminate similarity of word ordering on a pair of languages (e.g. Fake English-Spanish) by randomly permuting each sentence.
  - Permuting: for a sentence of length L, we define permuting it with probability p as randomly choosing $p\binom{L}{2}$ pairs of indices from all $\binom{L}{2}$ pairs and swap them.
  - Finetune on un-permuted Fake English

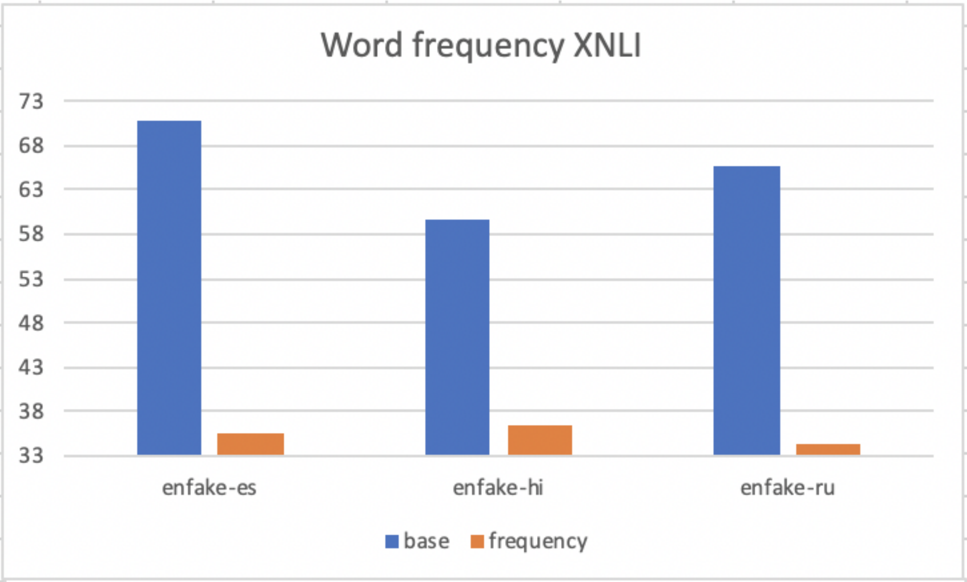
- Ordering is a main source of similarity, but there is still something beyond. BERT performs better than random when there is no order --- when the set of context words are the same.
- Drop in NER is a lot more significant then for XNLI
  - Says something about XNLI

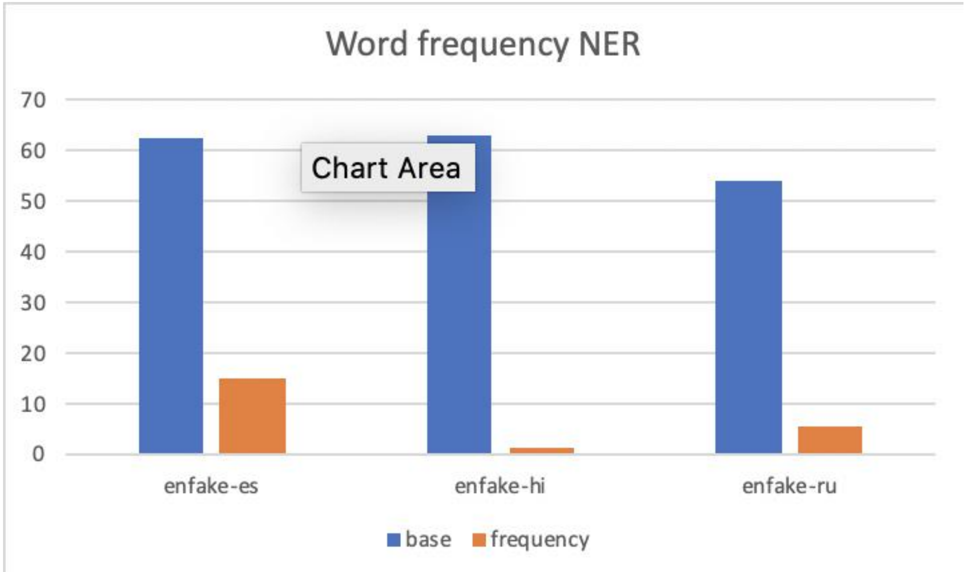


Preserving unigram word frequency

- By re-generating the pre-training text corpus based on distribution of vocabulary in the language, we can generate a frequency-based corpus such that BERT can only learn from similar frequencies between two languages --- when no context within sentence is preserved

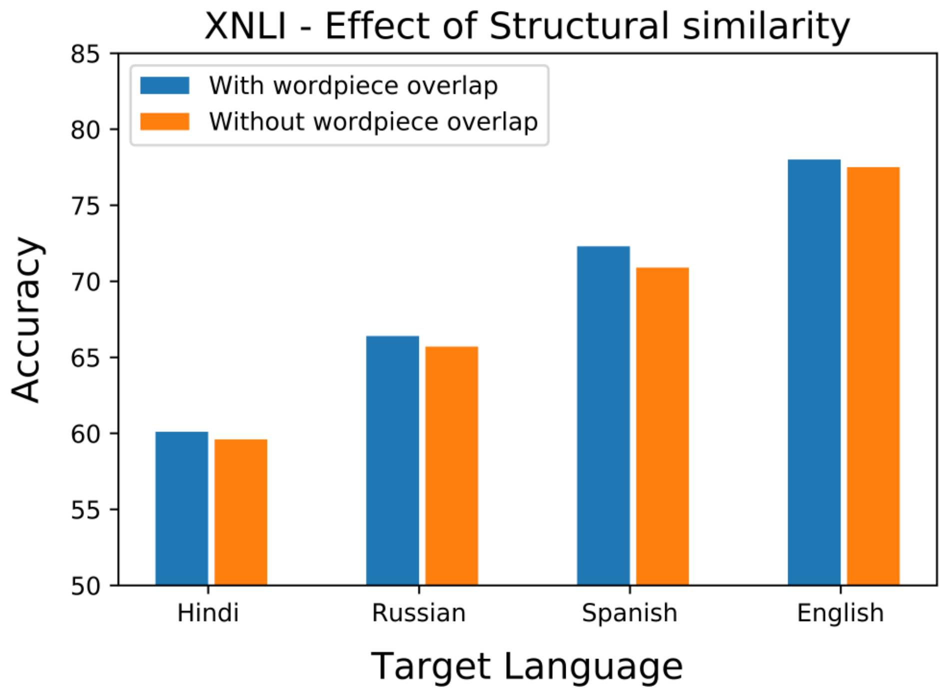
- BERT learns almost nothing with unigram frequency.
- Hypothesis: BERT can learn cross-lingual features with k-co-occurrences (k > 1) of words.





The Impact of Structural Similarity

- The same setting:
  - Train a pair of B-BERTs
    - **English-L** vs. **FakeEnglish-L**
  - Tune on the English (or FakeEnglish) data
  - Test on the target Language
- Fake-English transfers to English almost perfectly.
  - And transfers to other languages as English does.



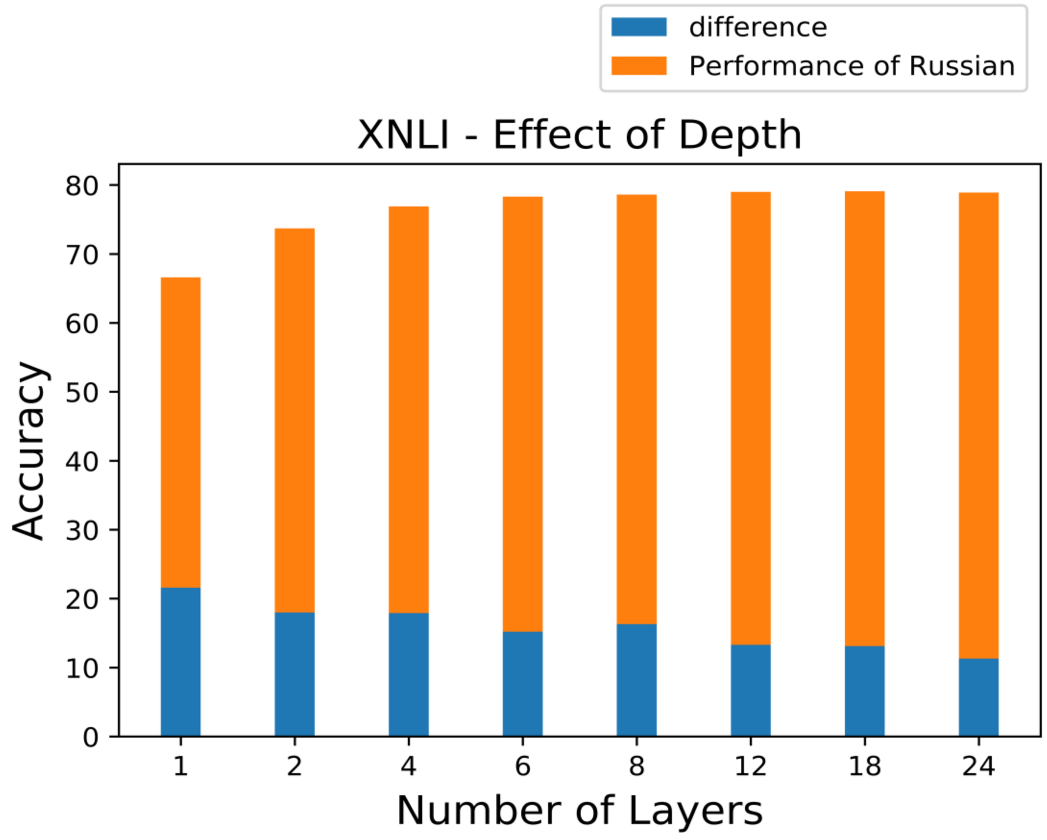
- Quality of transfer:
  - (English, Hindi) < (English, Russian) < (English, Spanish) < (English, FakeEnglish)
- In all pairs: no word-piece overlap
  - The transferability is due to the structural similarity between language L and (Fake)-English.
- Leaves more questions on the specific aspects of structural similarity that matters.

Architecture

- Here we study:
  - The depth of the Transformer structure
  - Number of attention heads
  - Total number of parameters

Architecture (1): Depth is critical

- We vary depth
  - Fix #(attention heads)
  - Fix #(parameters)
    - the size of the hidden and intermediate units is changed so that the total number of parameters remains almost the same
- Measure cross-lingual transfer as the **difference** between the performance on Fake-English and vs. **performance on Russian**.
  - Similar results in other languages.
- Deeper models (1) perform better and (2) transfer better.

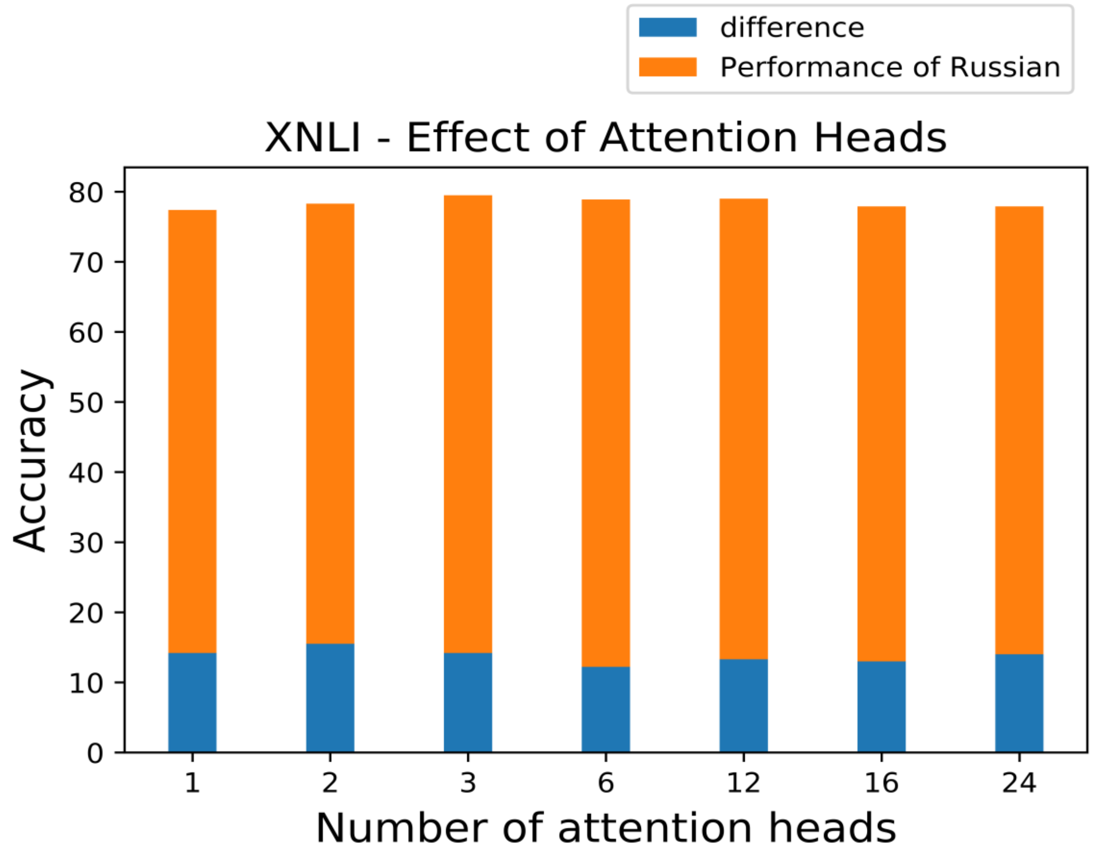


Architecture (2) #(Attention Heads)

- We vary #(attention heads)
  - Fix Depth
  - Fix #(parameters)

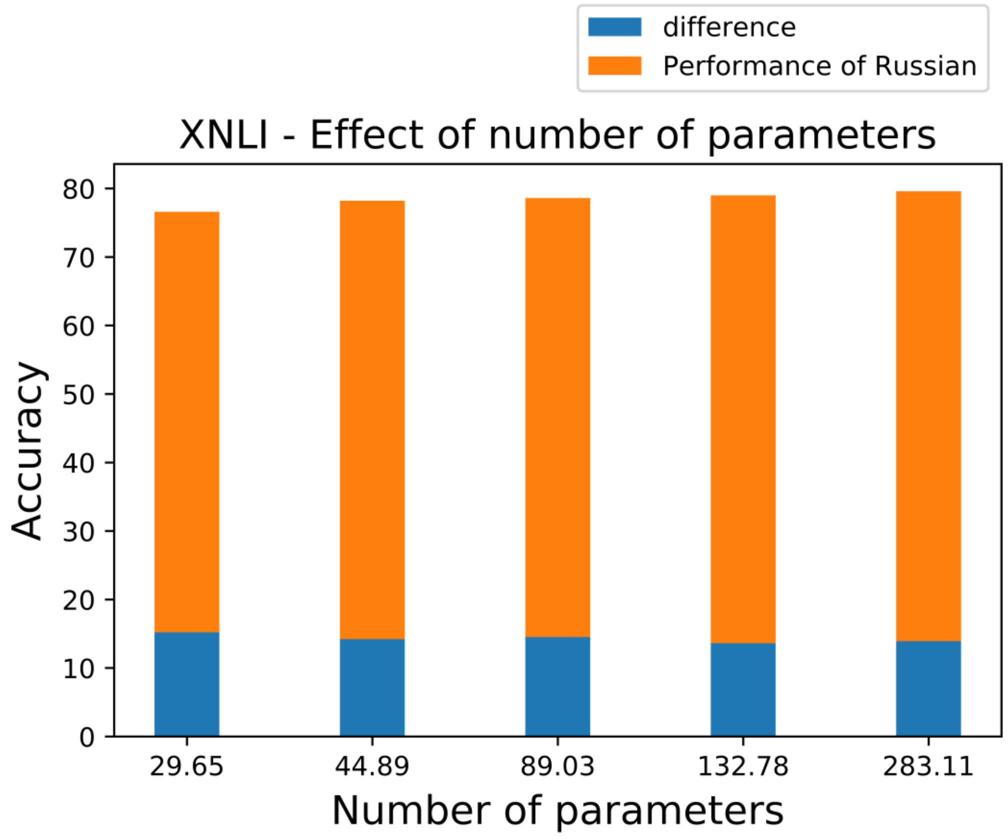
- Measure cross-lingual transfer as the **difference** between the performance on Fake-English and vs. **performance on Russian**.
  - Similar results in other languages.

- B-BERT ability to transfer exists even with a single attention head



Architecture (3) #(Parameters)

- We vary #(parameters)
  - by changing size of hidden and intermediate units
  - Fix #(attention heads)
  - Fix Depth
- Measure cross-lingual transfer as the **difference** between the performance on Fake-English and vs. **performance on Russian**.
  - Similar results in other languages.
- The #(parameters) isn't as significant as depth
  - But, below some threshold, #(parameters) seems significant (not shown)



Architecture (3) #(Parameters)

- We vary #(parameters)
  - by changing size of hidden and intermediate units
  - Fix #(attention heads)
  - Fix Depth
- The #(parameters) isn't as significant as depth
  - But, below some threshold, #(parameters) seems significant

Parameters (M)	Depth	Multi-head Attn	XNLI Fake-English	XNLI Russian	XNLI Gap
7.87	3	3	0.685	0.432	0.253
12.19	3	3	0.701	0.441	0.260
16.78	3	3	0.708	0.504	0.204
8.40	6	6	0.702	0.497	0.205
13.37	6	6	0.724	0.562	0.162
18.87	6	6	0.733	0.544	0.189
29.65	12	12	0.766	0.614	0.152
44.89	12	12	0.782	0.640	0.142
89.03	12	12	0.786	0.641	0.145
283.11	12	12	0.796	0.654	0.142
132.78	**12**	**12**	**0.790**	**0.657**	**0.133**

Input and Learning Objective

- Does Next Sentence Prediction affect cross-lingual ability ?
  - Yes, it hurts performance. Even more than in the monolingual case.

- Can we include Language Identity in the input to B-BERT ?
  - Adding language identify markers to the input does not make a difference.

- What is the impact of token representation (Character vs Word-piece vs Word)?
  - word-piece tokenized input is better than both
  - It seems to carry more information than characters, and address unseen words better than words.

Conclusion

Thank You!

- Huge progress in Low Resource NLP
  - Mostly in “easy” tasks
  - Some in Event-related and TE tasks
  - But higher-level tasks still a challenge
- Discussed some of the work in this context
  - Importance of Contextual Embeddings!
    - And some understanding of it
  - Embedding are not enough – weak signals are strong!
    - Cheap Translation
    - Bootstrapping from non-speakers
- We still have ways to go.

Somali streaming data → Situation Awareness (described in English)