XNLI: Evaluating Cross-lingual Sentence Representations

Roadmap

- What is this about?
- What does this look like?
- Literature review
- How is the data developed?

What is this about?

- Data for cross lingual language understanding (XLU) and low-resource cross-language transfer.
- Data annotation for all languages is unrealistic.
- Evaluation set for XLU:
  - Extending the development and test sets of the Multi-Genre Natural Language Inference Corpus (MultiNLI) to 15 languages,
  - Including low-resource languages such as Swahili and Urdu.
- Train with one language and evaluate with multiple languages.

What does this look like?

- XNLI consists of 7500 human-annotated development and test examples in NLI
    - Three-way classification format in English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu
    - 112,500 annotated pairs.

Language	Premise / Hypothesis	Genre	Label
English	You don’t have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d’offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction
Arabic	تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح. لا يمكننأللوكالات أ اتعرف ما إذا كانت ناجحة أم لا	Nine-Eleven	Contradiction

Literature Review

- Multi-lingual word embeddings
- Sentence Representation Learning
  - Continuous bag-of-words (CBOW)
  - Unsupervised SkipThought model
- Multilingual Sentence Representations
- Cross-lingual Evaluation Benchmarks
  - Reuters crosslingual document classification corpus
  - document level,
  - the comparison between different sentence embeddings is difficult.
  - distribution of classes is highly unbalanced
  - dataset does not provide a development set in the target language

How is the XNLI data developed?

- crowdsourcing-based procedure to collect and validate 750 new examples from each of the ten text sources used in NLI corpus for a total of 7500 examples.
- translate data into ten target languages
  - ensures that the data distributions are maximally similar across languages.
  - same trusted pool of workers as was used prior NLI crowdsourcing efforts
  - for any premise, this process allows for a corresponding hypothesis in any language.

Data Collection

- English:
  - We sample 250 sentences from each of the ten sources that were used in that corpus, ensuring that none of those selected sentences overlap with the distributed corpus.
  - MultiNLI worker pool from a crowdsourcing platform produce three hypotheses for each premise, one for each possible label
  - each pair of sentences is relabeled by four other workers.
  - for each validated sentence pair, assign a gold label representing a majority vote between the initial label assigned to the pair by the original annotator, and the four additional labels assigned by validation annotators
  - three-vote consensus for 93% of the data.

Translation

- translate the premises and hypotheses separately, to ensure that no context is added to the hypothesis that was not there originally, and simply copy the labels from the English source text.
- Two annotators reannotate 100 samples of English and French without seeing the source English text for any language they annotate
- consensus label 85% of the time on the original English data and 83% of the time on the translated French,

Resulting corpus

- The gold label for some of the sentence pairs changes as a result of information added or removed in the translation process.
  - It recovers the English consensus label 85% of the time on the original
  - 83% of the time on the translated French
- It does not tackle domain-adaptation that occurs when handling this the change in style from one language to another.
- the resulting corpus has similar properties to the MultiNLI corpus.
  - For all languages, on average, the premises are twice as long as the hypotheses.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Premise	21.7	24.1	22.1	21.1	21.0	20.9	19.6	16.8	20.7	27.6	22.1	21.8	23.2	18.7	24.1
Hypothesis	10.7	12.4	10.9	10.8	10.6	10.4	9.7	8.4	10.2	13.5	10.4	10.8	11.9	9.0	12.3

Table 2: Average number of tokens per sentence in the XNLI corpus for each language.

Baseline Approaches

- Translate train;
- Translate test
- Multilingual Word Embeddings
  - pretrained universal multilingual sentence embeddings based on the average of word embeddings (X-CBOW) (Evaluate transfer learning),
  - BiLSTM sentence encoders (Evaluate NLI encoders)
  - fixed-size embeddings for source; fine tuning embedding for target so that they are close in embedding space
  - back-propagate through the target encoder when optimizing L_{align} such that all 14 encoders live in the same English embedding space.

$$\mathcal{L}_{\text{align}}(x, y) = dist(x, y) - \lambda(dist(x_c, y) + dist(x, y_c))$$

where (x, y) corresponds to the source and target sentence embeddings, (x_c, y_c) is a contrastive term (i.e. negative sampling), λ controls

Baseline approaches Cont'd

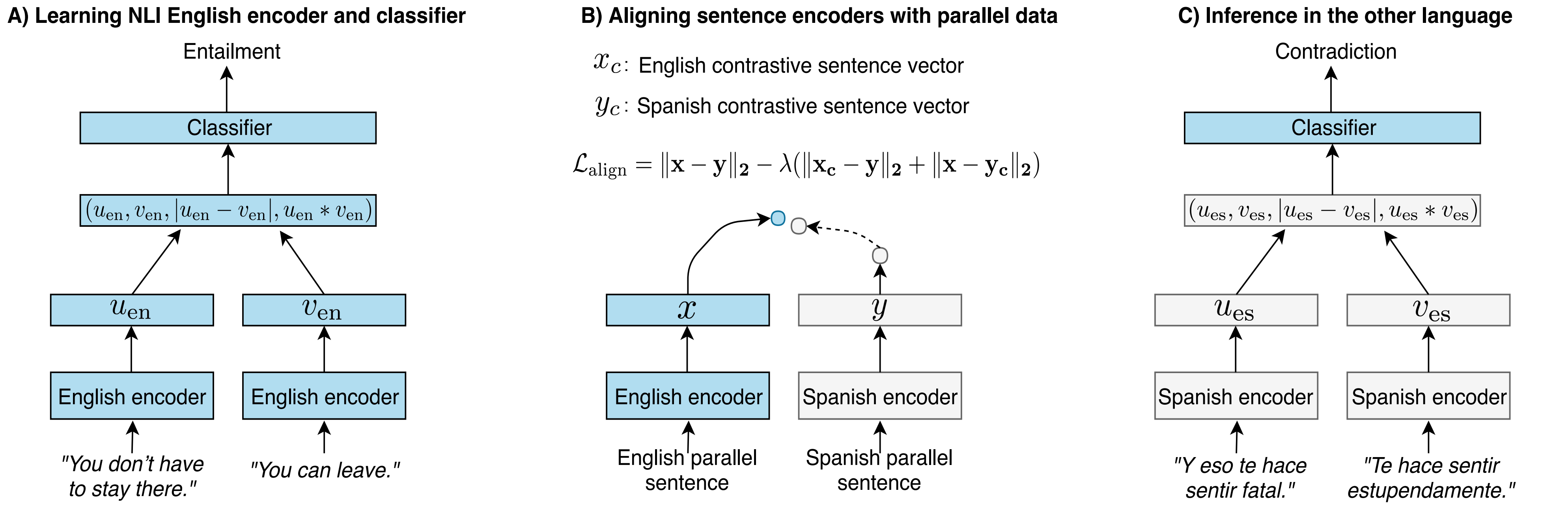

Figure 1: **Illustration of language adaptation by sentence embeddings alignment.** A) The English encoder and classifier in blue are learned on English (*in-domain*) NLI data. The encoder can also be pretrained (*transfer learning*). B) The Spanish encoder in gray is trained to mimic the English encoder using parallel data. C) After alignment of the encoders, the classifier can make predictions for Spanish.

Training details

- pretrained 300D aligned word embeddings for both X-CBOW and X-BILSTM
- 500,000 frequent words in the dictionary, which generally covers more than 98% of the words found in XNLI corpora.
- 512 hidden units of the BiLSTMs
- Adam optimizer with default parameters
- The classifier is a feed-forward neural network with one hidden layer of 128 hidden units, regularized with dropout rate 0.1

Parallel data

- United Nation corpora (French, Spanish, Russian, Arabic and Chinese publicly)
- Europarl corpora (German, Greek and Bulgarian)
- OpenSubtitles 2018 corpus (Turkish, Vietnamese and Thai)
- IIT Bombay corpus (Hindi)
- Bible and Quran transcriptions, the OpenSubtitles 2016 and 2018 corpora and LDC2010T21, LDC2010T23, corpora, 64k parallel sentences. (Urdu)
- 42k sentences using the Global Voices corpus and Tanzil Quran transcription corpus5 (Swahili)
- learn the alignment between English and target encoders.
- >=500,000 <= 2 million.

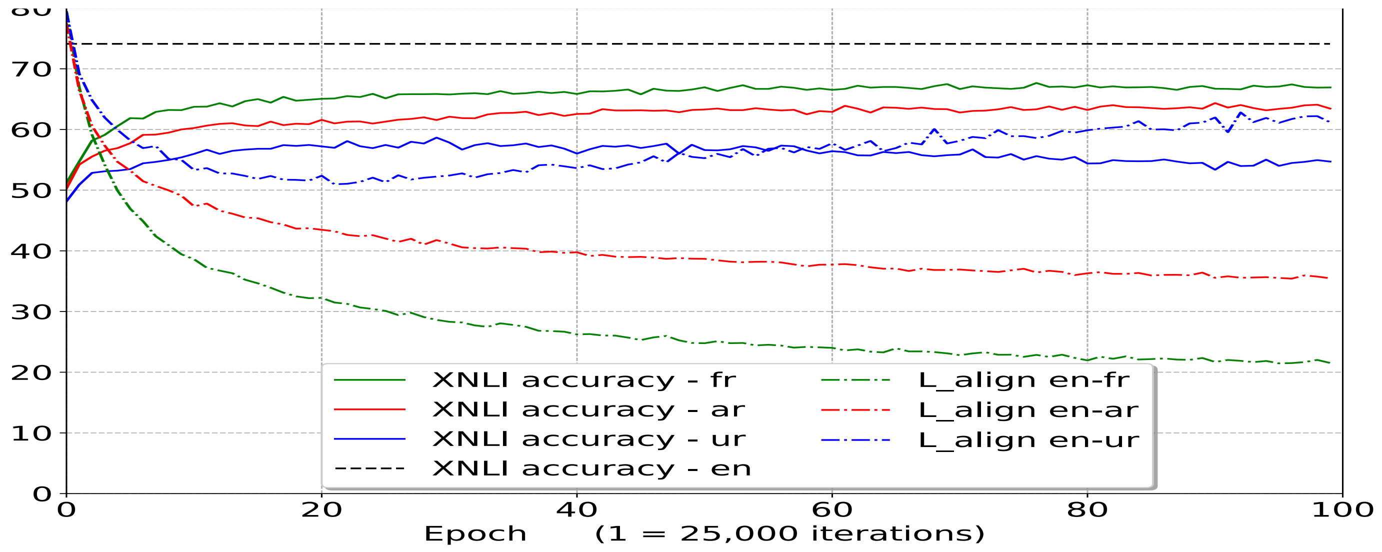

Figure 2: Evolution along training of alignment losses and X-BILSTM XNLI French (fr), Arabic (ar) and Urdu (ur) accuracies. Observe the correlation between $\mathcal{L}_{\text{align}}$ and accuracy.

Results

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Machine translation baselines (TRANSLATE TRAIN)															
BiLSTM-last	71.0	66.7	67.0	65.7	65.3	65.6	65.1	61.9	63.9	63.1	61.3	65.7	61.3	55.2	55.2
BiLSTM-max	**73.7**	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6
Machine translation baselines (TRANSLATE TEST)															
BiLSTM-last	71.0	68.3	68.7	66.9	67.3	68.1	66.2	64.9	65.8	64.3	63.2	66.5	61.8	60.1	58.1
BiLSTM-max	**73.7**	**70.4**	**70.7**	**68.7**	**69.1**	**70.4**	**67.8**	**66.3**	**66.8**	**66.5**	**64.4**	**68.3**	**64.2**	**61.8**	**59.3**
Evaluation of XNLI multilingual sentence encoders (in-domain)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**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
Evaluation of pretrained multilingual sentence encoders (transfer learning)															
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

Table 4: Cross-lingual natural language inference (XNLI) test accuracy for the 15 languages.

	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
XX-En BLEU	41.2	45.8	39.3	42.1	38.7	27.1	29.9	35.2	23.6	22.6	24.6	27.3	21.3	24.4
En-XX BLEU	49.3	48.5	38.8	42.4	34.2	24.9	21.9	15.8	39.9	21.4	23.2	37.5	24.6	24.1
Word translation P@1	73.7	73.9	65.9	61.1	61.9	60.6	55.0	51.9	35.8	25.4	48.6	48.2	-	-

Table 3: BLEU scores of our translation models (XX-En) P@1 for multilingual word embeddings.