PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification

Weber Huang (黃彥鈞)

Graduate Student of Data Science Institute, TMU

About the Author

- Google Research
- Yinfei Yang , Yuan Zhang , Chris Tar , Jason Baldridge
- Google AI blog

Outline

Introduction

- Why PAWS-X ?
- What is PAWS-X?

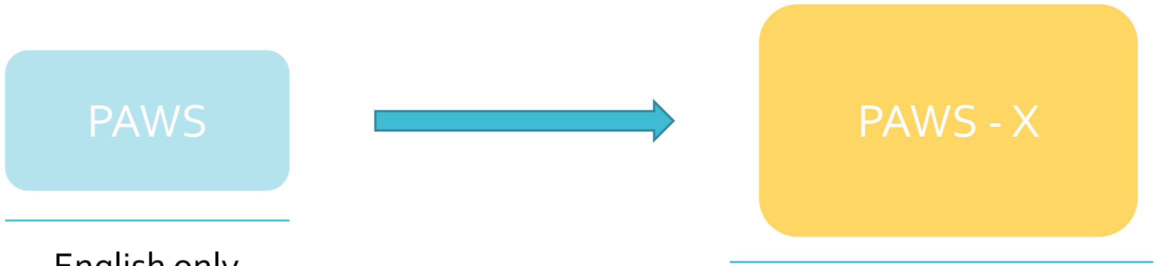
- Adversarial examples have effectively highlighted the deficiencies of state-of-the-art models for many natural language processing tasks
- PAWS, which has adversarial paraphrase identification pairs with high lexical overlap.
  - E.g. flights from New York to Florida vs flights from Florida to New York
- Research on adversarial examples has generally shown that augmenting training data with good adversarial examples can boost performance for some models

Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

- Most previous work focuses only on English despite the fact that the problems highlighted by adversarial examples are shared by other languages.
  - E.g. Multi30K , Opusparcus



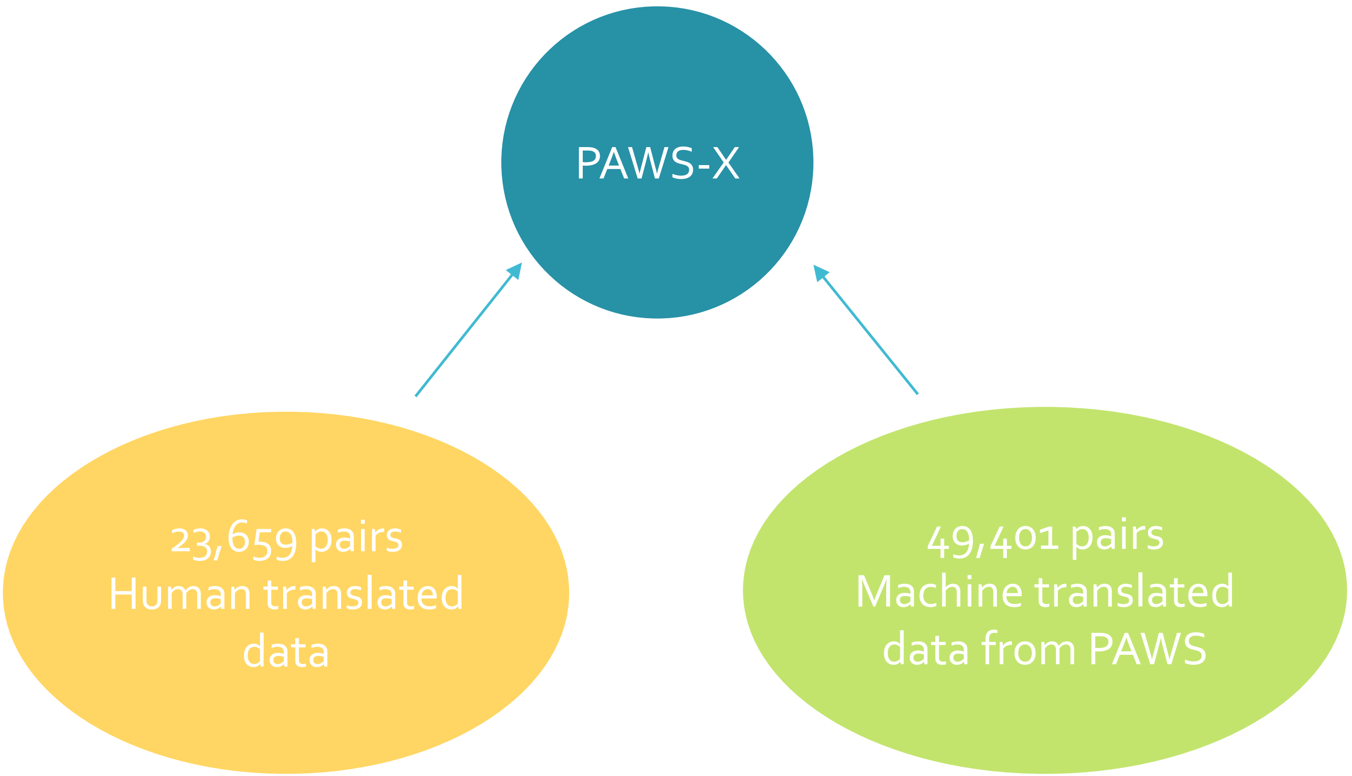
English only

Spanish, French, German, Chinese, Japanese, and Korean

Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

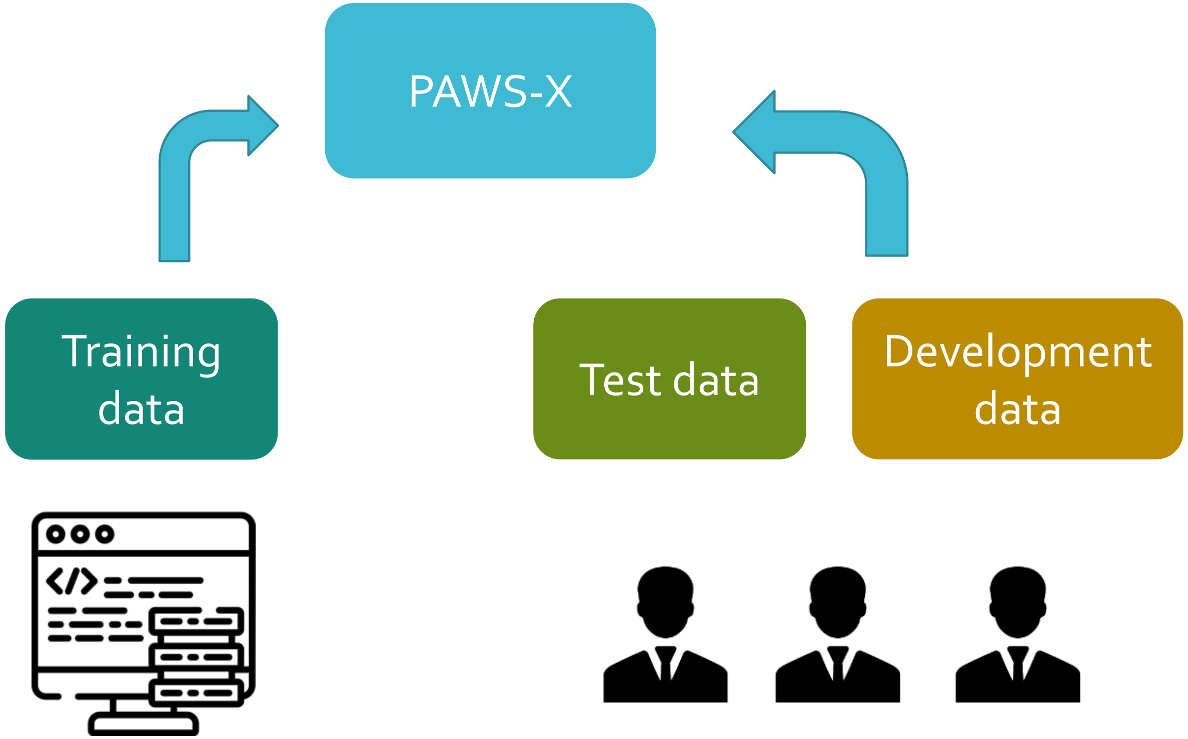
Conclusion



Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion



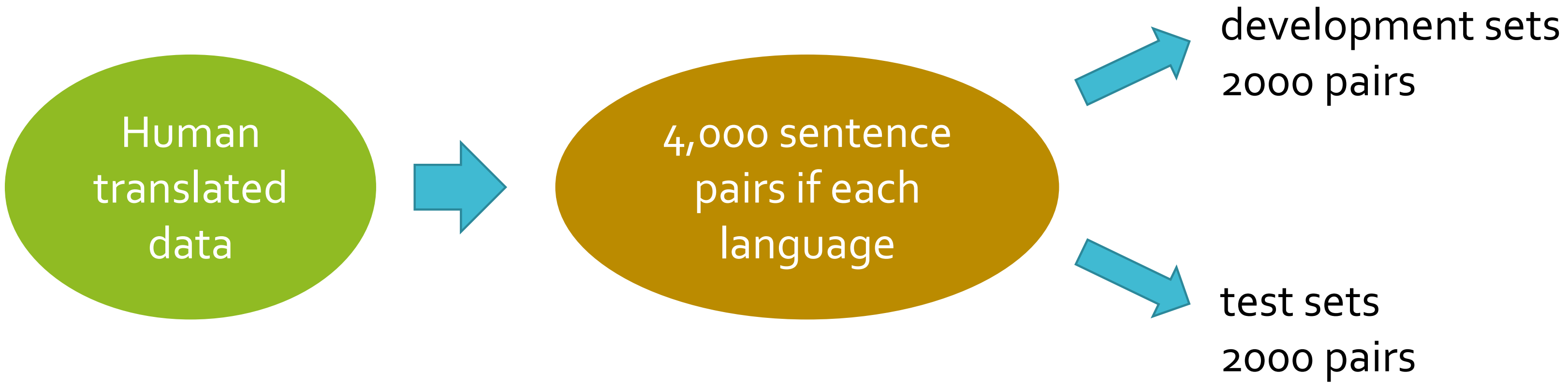
Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

Translating Evaluation Sets

- Obtaining human translations on a random sample of 4,000 sentence pairs from the PAWS development set for each of the six languages
- A randomly sampled subset is presented and validated by a second worker. The final delivery is guaranteed to have less than 5% word level error rate.
- The sampled 4,000 pairs are split into new development and test sets, 2,000 pairs for each.



Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

	fr	es	de	zh	ja	ko
dev	1,992	1,962	1,932	1,984	1,980	1,965
test	1,985	1,999	1,967	1,975	1,946	1,972

Table 2: Examples translated per language.

- Some sentences could not be translated. Table 2 shows the final counts translated to each language.
  - Incompleteness, ambiguities
  - likely from the Adversarial generation process when creating PAWS
  - < 2%
- Original PAWS labels (paraphrase or not paraphrase) are mapped to the translations. Positive pairs account for 44.0% of development sets and 45.4% of test respectively–close to the PAWS label distribution.
- Entity mention problem

Method

- Evaluated Methods
- Experiments and Results

Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

PAWS-X : Probe models' ability to capture structure and context in a multilingual setting.

BOW encoder with COS similarity

- Unigram to bigram token
- Cosine value > 0.5 as a paraphrase

ESIM(Enhanced Sequential Inference Model)

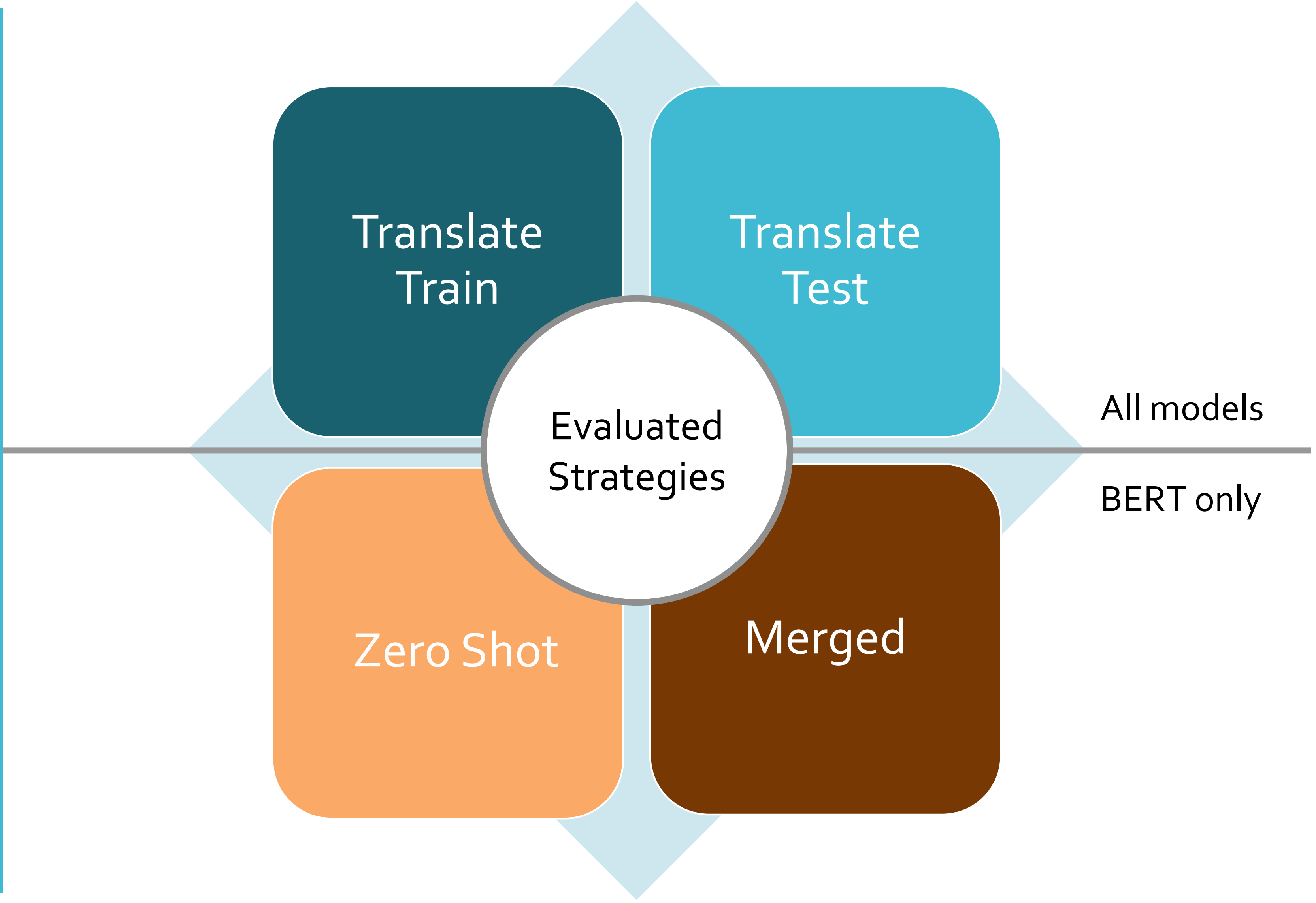
- BiLSTM
- feed-forward layer

BERT(Bidirectional Encoder Representations from Transformers)

Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion



Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

- **Translate Train:**
  - English training data is machine-translated into each target language to provide data to train each model.
- **Translate Test:**
  - Train a model using the English training data, and machine-translate all test examples to English for evaluation.
- **Zero Shot:**
  - The model is trained on the PAWS English training data, and then directly evaluated on all others. Machine translation is not involved in this strategy.
- **Merged:**
  - Train a multilingual model on all languages, including the original English pairs and machine-translated data in all other languages.

Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

- **BERT :**
  - Latest public multilingual BERT base model with 12 layers2 and apply the default fine-tuning strategy with batch size 32 and learning rate 1e-5.
- **BOW and ESIM :**
  - using their own implementations and 300 dimensional multilingual word embeddings from fastText.
- **Two metrics :**
  - Classification accuracy
  - Area-under-curve scores of precision-recall curves (AUC-PR)

Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

Method	Accuracy en	fr	es	de	zh	ja	ko	AUC-PR en	fr	es	de	zh	ja	ko
BOW														
Translate Train	55.8	51.7	47.9	50.2	54.5	55.1	56.7	41.1	48.9	46.8	46.4	50.0	48.7	49.3
Translate Test	–	54.9	54.7	55.2	55.3	55.9	55.2	–	46.3	45.5	45.8	50.9	46.8	48.5
ESIM														
Translate Train	67.2	66.2	66.0	63.7	60.3	59.6	54.2	69.6	67.0	64.2	59.2	58.2	56.3	50.5
Translate Test	–	66.2	66.3	66.0	62.0	62.3	60.6	–	68.4	69.5	68.2	62.3	61.8	60.3
BERT														
Translate Train	93.5	89.3	89.0	85.3	82.3	79.2	79.9	**97.1**	93.6	92.4	92.0	87.4	81.4	82.4
Translate Test	–	88.7	89.3	88.4	79.3	75.3	72.6	–	93.8	93.1	92.9	85.1	80.9	80.1
Zero shot	–	85.2	86.0	82.2	75.8	70.5	71.7	–	91.0	90.5	89.4	79.6	72.7	75.5
Merged	**93.8**	**90.8**	**90.7**	**89.2**	**85.4**	**83.1**	**83.9**	96.5	**94.0**	**92.9**	**92.9**	**88.9**	**86.0**	**86.3**

Table 4: Accuracy (%) and AUC-PR (%) of each approach. Best numbers in each column are marked in bold.

Method		Averaged Accuracy	Averaged AUC-PR
BOW	Translate Train	52.7	48.4
	Translate Test	55.2	47.3
ESIM	Translate Train	61.7	59.2
	Translate Test	63.9	65.1
BERT	Translate Train	84.2	88.2
	Translate Test	82.3	87.6
	Zero Shot	78.6	83.1
	Merged	**87.2**	**90.2**

Table 5: Average Accuracy (%) and AUC-PR (%) over the six languages.

- Model Comparisons
- Training/Evaluation Strategies
- Language Difference

Conclusion

Introduction
- Why PAWS-X ?
- What is PAWS-X?

Method
- Evaluated Methods
- Experiments and Results

Conclusion

Our experimental results showed that PAWS-X effectively measures sensitivity of models to word order and the efficacy of cross-lingual learning approaches.

It also leaves considerable headroom as a new challenging benchmark to drive multilingual research on the problem of paraphrase identification.

The PAWS-X dataset, including both the new human translated pairs and the machine translated examples, is available for download at https://github.com/google-research-datasets/paws.

Q & A