A Robust Abstractive System for Cross-Lingual Summarization

Jessica Ouyang and Boya Song and Kathleen McKeown
Deptartment of Computer Science
Columbia University
New York, NY 10027
{ouyangj, kathy}@cs.columbia.edu, bs3065@columbia.edu
Highlights

Construct corpora using low-resource languages and MT with NYT;

Test on an unseen language;

Higher fluency than copy-attention summarizer on translated inputs.

Cross-lingual summarization: summarize in one language a doc available only in another language: summ then translate, or translate then summarization.
Dataset

NYT: translate them into low-resource articles;

Then translate BACK into English, noisy articles.

Pair noisy articles with clean references.

---- learns to take ‘bad’ input to generate ‘good’ output
Discussion

Still relies on an existing MT system: is it possible to combine MT and summarization together? Both are seq2seq models, can we let them share some features, like multi-task learning?

Mixed model: depends on how similar the languages are!

A short page, no significant model change.
MASS: Masked Sequence to Sequence Pre-training for Language Generation

Kaitao Song *\(^1\)  Xu Tan *\(^2\)  Tao Qin \(^2\)  Jianfeng Lu \(^1\)  Tie-Yan Liu \(^2\)
Highlights

MAsked Sequence to Sequence learning (MASS)

Inspired by BERT: pre-training and fine-tuning

Rich-resource to low-resource

Tested on multiple tasks with text generation: MT, summarization and conversational response generation.

Text generation: data hungry, MASS can ‘transfer’ knowledge from other domain

Pre-train on unpaired data, fine-tune on low-resource paired data.
Method

Transformer pre-train on WMT monolingual corpus, fine-tune on 3 tasks.

Pre-train in an unsupervised way, similar to BERT, to generate a sequence of segments, instead of a single token. K is an important parameter!

Figure 1. The encoder-decoder framework for our proposed MASS. The token “-” represents the mask symbol [M].
Discussion

BERT: designed for language understanding, not for generation.

MASS: jointly pre-train encoder and decoder for generation tasks.

- Only predicting the masked tokens through a sequence to understand unmasked tokens, encourage the decoder to extract useful info from encoder.
- Predicting consecutive tokens on the decoder side, decoder will be a better LM.
- Predict 3456, only 345 are masked.
Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting

Yen-Chun Chen and Mohit Bansal
UNC Chapel Hill
{yenchun, mbansal}@cs.unc.edu
Highlight

Compresses and paraphrasing method

Abstractiveness Scores, and other metrics on the CNN/DM dataset

Sentence-level and word-level: parallel decoding

Hybrid extractive-abstractive architecture with policy-based RL to bridge together the two networks.
Model: Extractor

Hierarchical Sentence Representation: CNN, LSTM-RNN to get doc embeddings

Another LSTM to get pointer network (for sent selection)

Abstractor Network: paraphrasing (encoder-aligner-decoder + copy)
Model: Abstractor

Rouge score-guided

Able to provide abstractive and extractive summaries.

Interesting Abstractiveness scores wrt pointer-generator paper.

Figure 2: Reinforced training of the extractor (for one extraction step) and its interaction with the abstractor. For simplicity, the critic network is not shown. Note that all $d$’s and $s_t$ are raw sentences, not vector representations.
Cross-lingual Language Model Pretraining

Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau*
Facebook AI Research
Université Le Mans
aconneau@fb.com
Cross-lingual language models (XLMs): a supervised model and an unsupervised

Tested on two tasks: cross-lingual classification and MT (WMT’16 dataset)

The cross-lingual pretrained model works well on low-resource languages.

Inspired by BERT: masked model.

Align sentences in an unsupervised way!

Google BERT: https://github.com/google-research/bert
Related Work

Cross-lingual: monolingual embeddings --- orthogonal transformations are sufficient to align these word distributions!

Aligning sentence representations from multiple languages: [Zero-Shot Cross-lingual Classification Using Multilingual Neural Machine Translation](#)
MLM: language embedding
TLM: mask on both sides
Cross-lingual classification

Dataset: cross-lingual natural language inference (XLNI), scripts to download 15 languages.

Fine-tune on pre-trained English Transformer, take the first hidden state then add a linear layer.

Compared with MT baselines also.

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>fr</th>
<th>es</th>
<th>de</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Machine translation baselines (TRANSLATE-TRAIN)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devlin et al. (2018)</td>
<td>81.9</td>
<td>-</td>
<td>77.8</td>
<td>75.9</td>
</tr>
<tr>
<td>XLM (MLM+TLM)</td>
<td>85.0</td>
<td>80.2</td>
<td>80.8</td>
<td>80.3</td>
</tr>
<tr>
<td><em>Machine translation baselines (TRANSLATE-TEST)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Devlin et al. (2018)</td>
<td>81.4</td>
<td>-</td>
<td>74.9</td>
<td>74.4</td>
</tr>
<tr>
<td>XLM (MLM+TLM)</td>
<td>85.0</td>
<td>79.0</td>
<td>79.5</td>
<td>78.1</td>
</tr>
<tr>
<td><em>Evaluation of cross-lingual sentence encoders</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conneau et al. (2018b)</td>
<td>73.7</td>
<td>67.7</td>
<td>68.7</td>
<td>67.7</td>
</tr>
<tr>
<td>Devlin et al. (2018)</td>
<td>81.4</td>
<td>-</td>
<td>74.3</td>
<td>70.5</td>
</tr>
<tr>
<td>Artetxe and Schwenk (2018)</td>
<td>73.9</td>
<td>71.9</td>
<td>72.9</td>
<td>72.6</td>
</tr>
<tr>
<td>XLM (MLM)</td>
<td>83.2</td>
<td>76.5</td>
<td>76.3</td>
<td>74.2</td>
</tr>
<tr>
<td>XLM (MLM+TLM)</td>
<td><strong>85.0</strong></td>
<td><strong>78.7</strong></td>
<td><strong>78.9</strong></td>
<td><strong>77.8</strong></td>
</tr>
</tbody>
</table>
**Multilingual-WE**: MUSE -- linear mapping

\[ W^* = \arg\min_{W \in M_d(\mathbb{R})} \|WX - Y\|_F \]

Adversarial way to learn \( W \), then fine-tune \( W \).

[Word translation without parallel data (2018)]
Unsupervised MWE: EMNLP 2018 code

Unsupervised Multilingual Word Embeddings

Xilun Chen
Department of Computer Science
Cornell University
Ithaca, NY, 14853, USA
xlchen@cs.cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
Ithaca, NY, 14853, USA
cardie@cs.cornell.edu
Motivation

BWE-Pivot and BWE-Direct

Goal: learn a single multi-lingual embedding space for N languages -- map N monolingual embeddings of N languages into a single space with efficiency. $O(N)$
Method

Learn linear encoders and decoders for each language:

Multilingual Adversarial Training (MAT)

Multilingual Pseudo-Supervised Refinement (MPSR)
MAT

A discriminator for each language.

Allows to train in an unsupervised way.

Train D and M jointly.

Discriminator D

Encoder M, decoder is transpose M.

(M is orthogonal)

Reasonable but can be improved.
Multilingual Pseudo-Supervised Refinement

Rare words are noisier, MPSR to induce a dictionary of highly confident word pairs: construct word pairs using mutual nearest neighbors from frequent words.

Mutual nearest neighbours + mean square loss

\[ J_r = \mathbb{E}_{(i,j) \sim \mathcal{L}^2} \mathbb{E}_{(x_i, x_j) \sim \text{Lex}(i,j)} L_r(M_i x_i, M_j x_j) \]

Cross-Lingual Similarity Scaling (CSLS) to construct the pseudo-supervised lexica.
Experiments

Two tasks:

- Multilingual Word Translation (6 languages, previous paper)
- SemEval2017 cross-lingual word similarity task

Pre-trained 300d fastText embeddings on Wikipedia corpus.
Context-Aware Cross-Lingual Mapping

Hanan Aldarmaki\textsuperscript{1} and Mona Diab\textsuperscript{1,2}
\textsuperscript{1}The George Washington University
\textsuperscript{2}AWS, Amazon AI
aldarmaki@gwu.edu, diabmona@amazon.com
Fitting an orthogonal matrix as a mapping.

Word-level mapping: reflects sentence-level cross-lingual similarity.

Two approaches:

- Based on ELMo (contextualized features): a parallel corpus with word-alignments.
- Learn a transformation between sentence embeddings rather than word embeddings.
Orthogonal Bilingual Mapping

Source language X to target Y:

\[ R = \arg \min_{\hat{R}} \| \hat{R} X - Y \| \quad \text{s. t.} \quad \hat{R}^T \hat{R} = I \]

Approximate solution:

\[ Y X^T = U \Sigma V^T \]

\[ R = U V^T \]
Methods

Contextualized Embeddings

IBM Model: a parallel corpus with word alignments, port into ELMo to get word embeddings (contextualized).

Sentence-level Embeddings

A sentence is less ambiguous than a single word.

Average word vectors to get a sentence-level embedding in a parallel corpus.
Experiments

Code and data

Monolingual training: 1 Billion Word benchmark for English.

WMT’13 common crawl data for cross-lingual mapping.

Eval: accuracy of retrieving the correct translation from the target side…

Word-level evaluation: the precision of correctly retrieving a translation from the vocab of another language.
Conclusion

Contextualized mappings work better.

Word-level mappings work better with smaller parallel corpora, sent-level may increase when more data is available.

More variations may be considered in the future.
Zero-Shot Cross-Lingual Opinion Target Extraction

Soufian Jebbara and Philipp Cimiano
Semalytix GmbH, Bielefeld, Germany
Semantic Computing Group, CITEC - Bielefeld University, Bielefeld, Germany
soufian.jebbara@semalytix.com
cimiano@cit-ec.uni-bielefeld.de
Highlight

Aspect-based sentiment analysis: opinion target expressions (OATEs)

Train on source language, make predictions on target language without using labeled samples.

Two methods for obtaining cross-lingual word embeddings on the task:

SVD: classic method [git](#)

Unsupervised version (MUSE) [git](#)

Method

Using IOE scheme to label a sentence.

Apply a multi-layer CNN.

**Monolingual Model:**

Prediction as a classification task.

**Cross-lingual Model:**

Train monolingual embeddings on monolingual datasets, the model is adapted to any target language.
Evaluation

Dataset: SemEval 16' task 5, restaurant domain on 5 languages: en,ru,es,tr,nl (Dutch).

<table>
<thead>
<tr>
<th>System</th>
<th>en</th>
<th>es</th>
<th>nl</th>
<th>ru</th>
<th>tr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toh and Su (2016)</td>
<td>0.723</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Álvarez-López et al. (2016)</td>
<td>0.666</td>
<td>0.685</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Kumar et al. (2016)</td>
<td>0.685</td>
<td>0.697</td>
<td>0.644</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Pontiki et al. (2016)*</td>
<td>0.441</td>
<td>0.520</td>
<td>0.506</td>
<td>0.493</td>
<td>0.419</td>
</tr>
<tr>
<td>Li and Lam (2017)</td>
<td>0.734</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>all→target (Ours)</td>
<td>0.660</td>
<td>0.687</td>
<td>0.624</td>
<td>0.567</td>
<td>0.490</td>
</tr>
</tbody>
</table>
Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification

Learn language-invariant features that generalize across languages using only monolingual data.

Adversarial part: minimize the distances between two languages (on feature spaces).

Pre-trained bi-lingual embeddings.

Multi-lingual version?
Dataset for X-lingual classification

Multilingual multi-domain Amazon review dataset [link]

English-Chinese Yelp Hotel Reviews [link]

SemEval 2016 workshop [link]: aspect-based sentiment datasets (8 languages, 7 domains)

Google translation
MWEs

BilBOWA [git](https://github.com/)

MUSE by facebook: [git](https://github.com/)


United Nation Parallel Corpus
Neural Cross-Lingual Named Entity Recognition with Minimal Resources

Unsupervised transfer, words and word order across languages.

Finds translations based on bilingual word embeddings; improve using self-attention (to word order).

Combines embedding method and dictionary method: limited resource, char-level

Method: 1) train separate embeddings; 2) map into one space; 3) translate each word into nearest neighbor using the common space; 4) train NER using translated words in En. (En is larger, how to utilize BERT?)

Use a word dictionary to learn better transformation from X to Y. (math!)---&gt; cross-domain similarity local scaling.
NER Model Architecture

Char-level network (**RNN, CNN**): capture subword information like morphological features

Word-level network (LSTM): context sensitive hidden representations

A linear-chain CRF: dependency between labels and performs inference

Translate Es->En, then apply english NER.

But missing Es order
Datasets, Embeddings, Tools

- Multilingual multi-domain Amazon review dataset [link]
- Annotated hotel reviews dataset on 4 languages (blse dataset) [link]
- English-Chinese Yelp Hotel Reviews [link]
- SemEval 2016 workshop [link]: aspect-based sentiment datasets (8 languages, 7 domains)
- BilBOWA [git]
- MUSE by facebook: [git]
- Joint Bilingual Sentiment Embeddings and Classifier (2018): [git paper]
- United Nation Parallel Corpus
- Google Translation