Multilingual Language Models for Fine-tuning and Feature Extraction in Word-in-Context Disambiguation

Huiling You

Table of Contents

1 Introduction

Multilingual and Crosslingual Word-in-Context Disambiguation (MCL-WiC)

SemEval-2021 Task 2

- Extension from WiC - shared task at the IJCAI-19 SemDeep workshop (SemDeep-5)
- MCL-WiC Task
  - Given a sentence pair, each containing a polysemous target word
  - Determine the two target words are used in same meaning, or different meanings
  - Datasets -> multilingual and cross-lingual sentence pairs, 5 languages

Multilingual and Crosslingual Word-in-Context Disambiguation (MCL-WiC)

- In the multilingual setting, the two sentences are from the same language.
- In the cross-lingual setting, the two sentences are from different languages, English and one of the other four languages.
- Training data is only available for English--English, effectively leading to a zero-shot setting for the other languages.

Example	**Label**
The cat chases after the *mouse*. Click the right *mouse* button.	F
The cat chases after the *mouse*. La *souris* mange le fromage. (*The* mouse *is eating the cheese*)	T

Table 1: Examples for monolingual (top) and cross-lingual (bottom) word-in-context disambiguation.

Main Research Interest

Investigate the usefulness of pre-trained multilingual language models (LMs) in this MCL-WiC task, without resorting to sense inventories, dictionaries, or other resources

- **Fine-tune** the language models with a *span classification head*
- Using the multilingual language models as **feature extractors**, extracting contextual embeddings for the target word, and also adding syntactic information from a dependency parser.

We compare three different LMs: XLM-RoBERTa (XLMR), multilingual BERT (mBERT) and multilingual dis-tilled BERT (mDistilBERT).

2 Related Work

Related Work

WiC at SemDeep-5

- LMMS: BERT + WordNet 3.0
- ElMo + Classifier
- SuperGLUE benchmark: fine-tune

SensEmBERT - knowledge-based approach

3

Multilingual Language Models

Multilingual Language Models

XLMR

- XLMR (XLM-RoBERTa) is a scaled cross-lingual sentence encoder
- trained on 2.5T of data obtained from Common Crawl that covers more than 100 languages

mBERT

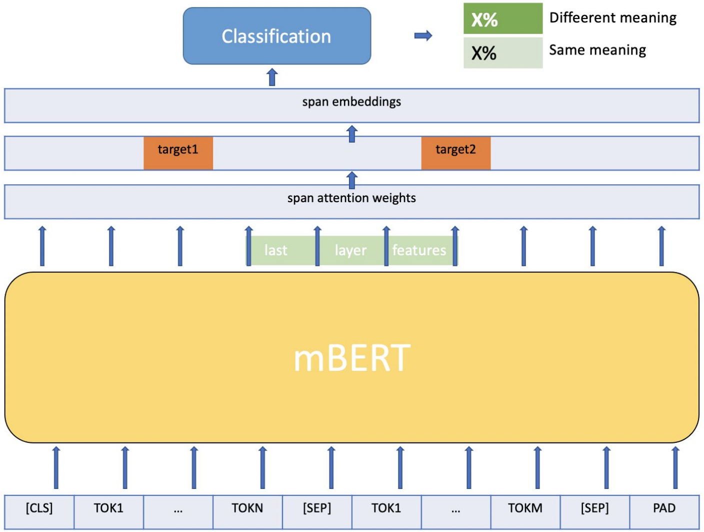
- pre-trained on the largest Wikipedias
- It is a multilingual extension of BERT that provides word and sentence representations for 104 languages

mDistilBERT

- a light Transformer trained by distilling mBERT
- reduces the number of parameters in mBERT by 40%

4 System Description

Fine-Tuning



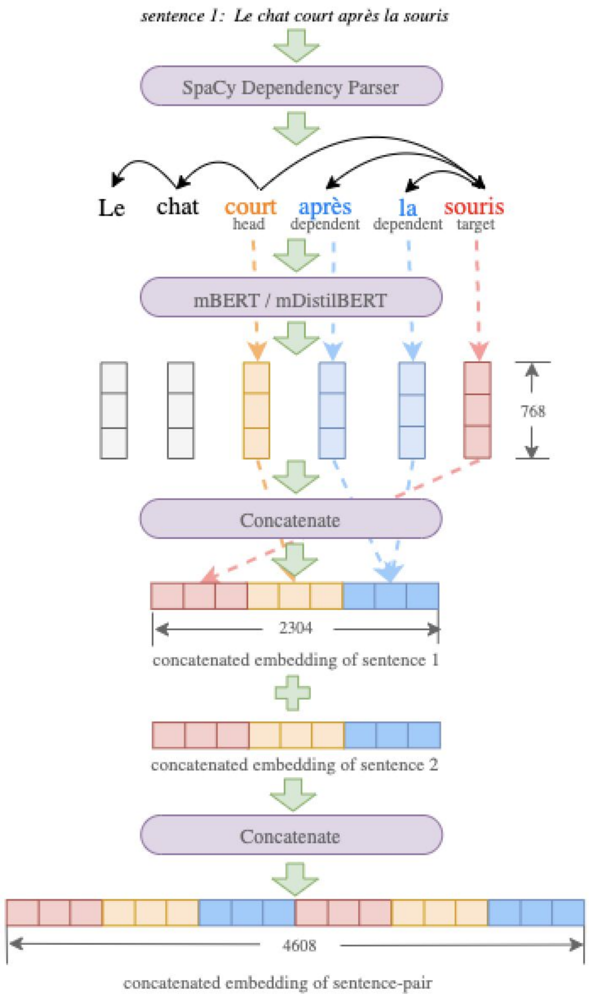
- A ***span classification head*** is stacked on top of pre-trained language models, and attends only to the target words.
- The ***span classification head*** consists of a ***span attention extractor*** and a classifier.
- Target spans get a weighted representation of the last-layer hidden states of either mBERT, mDistilBERT or XLMR.

Target Words Embeddings + Logistic Regression / MLP

- **The multilingual language models serve as pure feature extractors, to get target word embeddings from last-layer hidden states.**
- **We feed the two sentences separately to the models, and concatenate the embeddings for the two target words as input to the classifier.**
- **We experimented with two classifiers, logistic regression (LR) and a multi-layer perceptron (MLP).**

Dependency-based Syntax-Incorporated Embeddings

- **First, each sentence is parsed using the spaCy dependency parser**
- **Next, the sentence is passed to mBERT/mDistilBERT, and the corresponding target word embedding, head word embedding, and dependent word embedding(s) are retrieved**
- **Finally, the concatenated embeddings of two constituent sentences are further concatenated to form the sample feature vector of the sentence-pair, and fed to an MLP**



5 Experiment Setup

Experiment Setup

- **Dataset:** Only the datasets provided by SemEval-2021 Task 2 are used
- **Fine-tuning:** fine-tuned for three iterations, with batch size of 32, learning rate of 1e-5, and parameters optimized with AdamW
- **LR:** All LR models are trained for 150 iterations, with batch size of 32, learning rate of 0.0025 and parameters optimized with SGD
- **MLP**: 2-layer, trained for maximum 200 iterations, with learning rate of 0.001 and parameters optimized with Adam
- **Language model**: We use the base version of all multilingual language models, with 12 layers, 12 attention heads, and hidden dimension of 768.

	Train	Dev	Test
en-en	8000	500	1000
ar-ar	–	500	1000
fr-fr	–	500	1000
ru-ru	–	500	1000
zh-zh	–	500	1000
en-ar	–	–	1000
en-fr	–	–	1000
en-ru	–	–	1000
en-zh	–	–	1000

6 Results and Analysis

Result

	System	en-en	zh-zh	fr-fr	ru-ru	ar-ar	en-zh	en-fr	en-ru	en-ar
Fine-tune	XLMR	**84.5%**	**78.3%**	76.7%	73.1%	75.1%	**66.3%**	**70.9%**	**73.6%**	**65.2%**
	mBERT	82.9%	76.2%	**80.3%**	**73.6%**	**75.6%**	62.2%	66.3%	63.1%	59.4%
	mDistilBERT	75.5%	68.0%	66.8%	64.8%	68.9%	51.8%	53.4%	51.9%	50.9%
Feature Extractor	XLMR + LR	53.9%	55.4%	54.8%	57.2%	53.0%	58.2%	55.8%	55.4%	54.7%
	mBERT + LR	53.4%	53.5%	49.7%	51.7%	53.1%	52.0%	52.8%	52.8%	51.1%
	mDistilBERT + LR	55.7%	50.5%	52.6%	52.5%	51.9%	54.0%	52.5%	52.0%	51.6%
	mBERT + MLP	67.7%	51.4%	57.6%	54.2%	54.0%	47.4%	62.6%	55.6%	53.2%
	mDistilBERT + MLP	66.6%	59.1%	59.8%	61.8%	56.0%	48.2%	63.2%	57.4%	52.3%
	mBERT + Syntax + MLP	61.4%	52.7%	57.6%	57.0%	–	53.4%	57.8%	55.6%	–
	mDistilBERT + Syntax + MLP	67.0%	56.6%	58.2%	57.6%	–	54.0%	57.2%	56.2%	–

- We can see that the fine-tuning approach is preferable to the feature extraction approach. All feature extraction variants fall behind the fine-tuned systems by a large margin.

Result

	System	en-en	zh-zh	fr-fr	ru-ru	ar-ar	en-zh	en-fr	en-ru	en-ar
Fine-tune	XLMR	**84.5%**	**78.3%**	76.7%	73.1%	75.1%	**66.3%**	**70.9%**	**73.6%**	**65.2%**
	mBERT	82.9%	76.2%	**80.3%**	**73.6%**	**75.6%**	62.2%	66.3%	63.1%	59.4%
	mDistilBERT	75.5%	68.0%	66.8%	64.8%	68.9%	51.8%	53.4%	51.9%	50.9%
Feature Extractor	XLMR + LR	53.9%	55.4%	54.8%	57.2%	53.0%	58.2%	55.8%	55.4%	54.7%
	mBERT + LR	53.4%	53.5%	49.7%	51.7%	53.1%	52.0%	52.8%	52.8%	51.1%
	mDistilBERT + LR	55.7%	50.5%	52.6%	52.5%	51.9%	54.0%	52.5%	52.0%	51.6%
	mBERT + MLP	67.7%	51.4%	57.6%	54.2%	54.0%	47.4%	62.6%	55.6%	53.2%
	mDistilBERT + MLP	66.6%	59.1%	59.8%	61.8%	56.0%	48.2%	63.2%	57.4%	52.3%
	mBERT + Syntax + MLP	61.4%	52.7%	57.6%	57.0%	–	53.4%	57.8%	55.6%	–
	mDistilBERT + Syntax + MLP	67.0%	56.6%	58.2%	57.6%	–	54.0%	57.2%	56.2%	–

- Among the fine-tuned systems, XLMR and mBERT give the best results, whereas mDistilBERT falls behind by quite a large margin in most cases, in several cases by more than 10 percentage points.

Result

	System	en-en	zh-zh	fr-fr	ru-ru	ar-ar	en-zh	en-fr	en-ru	en-ar
Fine-tune	XLMR	**84.5%**	**78.3%**	76.7%	73.1%	75.1%	**66.3%**	**70.9%**	**73.6%**	**65.2%**
	mBERT	82.9%	76.2%	**80.3%**	**73.6%**	**75.6%**	62.2%	66.3%	63.1%	59.4%
	mDistilBERT	75.5%	68.0%	66.8%	64.8%	68.9%	51.8%	53.4%	51.9%	50.9%
Feature Extractor	XLMR + LR	53.9%	55.4%	54.8%	57.2%	53.0%	58.2%	55.8%	55.4%	54.7%
	mBERT + LR	53.4%	53.5%	49.7%	51.7%	53.1%	52.0%	52.8%	52.8%	51.1%
	mDistilBERT + LR	55.7%	50.5%	52.6%	52.5%	51.9%	54.0%	52.5%	52.0%	51.6%
	mBERT + MLP	67.7%	51.4%	57.6%	54.2%	54.0%	47.4%	62.6%	55.6%	53.2%
	mDistilBERT + MLP	66.6%	59.1%	59.8%	61.8%	56.0%	48.2%	63.2%	57.4%	52.3%
	mBERT + Syntax + MLP	61.4%	52.7%	57.6%	57.0%	–	53.4%	57.8%	55.6%	–
	mDistilBERT + Syntax + MLP	67.0%	56.6%	58.2%	57.6%	–	54.0%	57.2%	56.2%	–

- Among the systems with feature extraction, the relative performance of the three sets of contextual embeddings differ from the fine-tuning. Here, mDistilBERT are competitive to the other two embeddings.
- Using an MLP is preferable to LR, leading to large improvements in most cases.
- The addition of syntax leads to mixed results

Result

	System	en-en	zh-zh	fr-fr	ru-ru	ar-ar	en-zh	en-fr	en-ru	en-ar
Fine-tune	XLMR	**84.5%**	**78.3%**	76.7%	73.1%	75.1%	**66.3%**	**70.9%**	**73.6%**	**65.2%**
	mBERT	82.9%	76.2%	**80.3%**	**73.6%**	**75.6%**	62.2%	66.3%	63.1%	59.4%
	mDistilBERT	75.5%	68.0%	66.8%	64.8%	68.9%	51.8%	53.4%	51.9%	50.9%
Feature Extractor	XLMR + LR	53.9%	55.4%	54.8%	57.2%	53.0%	58.2%	55.8%	55.4%	54.7%
	mBERT + LR	53.4%	53.5%	49.7%	51.7%	53.1%	52.0%	52.8%	52.8%	51.1%
	mDistilBERT + LR	55.7%	50.5%	52.6%	52.5%	51.9%	54.0%	52.5%	52.0%	51.6%
	mBERT + MLP	67.7%	51.4%	57.6%	54.2%	54.0%	47.4%	62.6%	55.6%	53.2%
	mDistilBERT + MLP	66.6%	59.1%	59.8%	61.8%	56.0%	48.2%	63.2%	57.4%	52.3%
	mBERT + Syntax + MLP	61.4%	52.7%	57.6%	57.0%	–	53.4%	57.8%	55.6%	–
	mDistilBERT + Syntax + MLP	67.0%	56.6%	58.2%	57.6%	–	54.0%	57.2%	56.2%	–

- We also note that the performance is stronger for English--English than for the other languages in most settings. This is expected, since we only have English--English training data.

7 Conclusion

Conclusion

- Fine-tuning the language models is preferable to using them as feature extractors
- Add dependency-based syntax information in the MLP gave mixed results.
- XLMR performed better than mBERT in the cross-lingual setting, both with fine-tuning and feature extraction,
- mDistilBERT did not perform well with fine-tuning, but was competitive to the other models in the feature extraction setting.

Future work

- Hypothesis: XLMR has a better representation of words across languages than mBERT and mDistilBERT.
- Explore sub-word models of XLMR and mBERT
- Using representations from different layers of the pre-trained multilingual language models.

ANY QUESTIONS?