Multilingual Universal Sentence Encoder for Semantic Retrieval

Yinfei Yang[a†], Daniel Cer[a†], Amin Ahmad[a], Mandy Guo[a], Jax Law[a], Noah Constant[a], Gustavo Hernandez Abrego[a], Steve Yuan[b], Chris Tar[a], Yun-Hsuan Sung[a], Brian Strope[a], Ray Kurzweil[a]

[a]Google AI
Mountain View, CA

[c]Google
Cambridge, MA

Abstract

We introduce two pre-trained retrieval focused multilingual sentence encoding models, respectively based on the Transformer and CNN model architectures. The models embed text from *16 languages* into a single semantic space using a multi-task trained dual-encoder that learns tied representations using translation based bridge tasks (Chidambaram et al., 2018). The models provide performance that is competitive with the state-of-the-art on: semantic retrieval (SR), translation pair bitext retrieval (BR) and retrieval question answering (ReQA). On English transfer learning tasks, our sentence-level embeddings approach, and in some cases exceed, the performance of monolingual, English only, sentence embedding models. Our models are made available for download on TensorFlow Hub.

1 Introduction

We introduce three new members in the *universal sentence encoder* (USE) (Cer et al., 2018) family of sentence embedding models. Two multilingual models, one based on CNN (Kim, 2014) and the other based on the Transformer architecture (Vaswani et al., 2017), target performance on tasks requiring models to capture multilingual semantic similarity. The third member introduced is an alternative interface to our multilingual Transformer model for use in retrieval question answering (ReQA). The *16 languages* supported by our multilingual models are given in Table 1.[1]

Languages	Family
Arabic (ar)	**Semitic**
Chinese (PRC) (zh) Chinese (Taiwan) (zh-tw)	**Sino-Tibetan**
Dutch(nl) English(en) German (de)	**Germanic**
French (fr) Italian (it) Portuguese (pt) Spanish (es)	**Latin**
Japanese (ja)	**Japonic**
Korean (ko)	**Koreanic**
Russian (ru) Polish (pl)	**Slavic**
Thai (th)	**Kra–Dai**
Turkish (tr)	**Turkic**

Table 1: Supported languages (ISO 639-1).

2 Model Toolkit

Models are implemented in TensorFlow (Abadi et al., 2016) and made publicly available on TensorFlow Hub.[2] Listing 1 illustrates the generation of sentence embeddings using one of our multilingual models. Listing 2 demonstrates using the question answering interface. Responses are encoded with additional context information such that the resulting embeddings have a high dot product similarity score with the questions they answer. This allows for retrieval of indexed candidates using efficient nearest neighbor search.[3]

```
import tensorflow_hub as hub

module = hub.Module("https://tfhub.dev/google/"
    "universal-sentence-encoder-multilingual/1")

multilingual_embeddings = module([
 "Hola Mundo!", "Bonjour le monde!", "Ciao mondo!"
 "Hello World!", "Hallo Welt!", "Hallo Wereld!",
 "你好世界!", "Привет, мир!", "!مرحبا بالعالم"])
```

Listing 1: Encoding for STS/Bitext retrieval.

† Corresponding authors:
{yinfeiy, cer}@google.com

[1]Due to character set differences, we treat Simplified Chinese, zh, and Traditional Chinese, zh-tw, prominently used in Taiwan, as two languages within our model.

[2] https://www.tensorflow.org/hub/, Apache 2.0 license, with models available as saved TF graphs.

[3]Popular efficient search tools include FAISS https://github.com/facebookresearch/faiss, Annoy https://github.com/spotify/annoy, or FLANN https://www.cs.ubc.ca/research/flann.

```
module = hub.Module("https://tfhub.dev/google/"
    "universal-sentence-encoder-multilingual-qa/1")

query_embeddings = module(
    dict(text=["What is your age?"]),
    signature="question_encoder", as_dict=True)

candidate_embeddings = module(
    dict(text=["I am 20 years old."],
        context=["I will be 21 next year."]),
    signature="response_encoder", as_dict=True)
```

Listing 2: Encoding for QA retrieval.

3 Encoder Architecture

3.1 Multi-task Dual Encoder Training

Similar to Cer et al. (2018) and Chidambaram et al. (2018), we target broad coverage using a multi-task dual-encoder training framework, with a single shared encoder supporting multiple downstream tasks. The training tasks include: a multi-feature question-answer prediction task,[4] a translation ranking task, and a natural language inference (NLI) task. Additional task specific hidden layers for the question-answering and NLI tasks are added after the shared encoder to provide representational specialization for each type of task.

3.2 SentencePiece

SentencePiece tokenization (Kudo and Richardson, 2018) is used for all of the 16 languages supported by our models. A single 128k SentencePiece vocabulary is trained from 8 million sentences sampled from our training corpus and balanced across the 16 languages. For validation, the trained vocab is used to process a separate development set, also sampled from the sentence encoding model training corpus. We find the character coverage is higher than 99% for all languages, which means less than 1% output tokens are out of vocabulary. Each token in the vocab is mapped to a fixed length embedding vector.[5]

3.3 Shared Encoder

Two distinct architectures for the sentence encoding models are provided: (i) transformer (Vaswani et al., 2017), targeted at higher accuracy at the cost of resource consumption; (ii) convolutional neural network (CNN) (Kim, 2014), designed for efficient inference but obtaining reduced accuracy.

Transformer The transformer encoding model embeds sentences using the *encoder* component of the transformer architecture (Vaswani et al., 2017). Bi-directional self-attention is used to compute context-aware representations of tokens in a sentence, taking into account both the ordering and the identity of the tokens. The context-aware token representations are then averaged together to obtain a sentence-level embedding.

CNN The CNN sentence encoding model feeds the input token sequence embeddings into a convolutional neural network (Kim, 2014). Similar to the transformer encoder, average pooling is used to turn the token-level embeddings into a fixed-length representation. Sentence embeddings are then obtain by passing the averaged representation through additional feedforward layers.

4 Training and Configuration

4.1 Training Corpus

Training data consists of mined question-answer pairs,[6] mined translation pairs,[7] and the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015).[8] SNLI only contains English data. The number of mined questions-answer pairs also varies across languages with a bias toward a handful of top tier languages. To balance training across languages, we use Google's translation system to translate SNLI to the other 15 languages. We also translate a portion of question-answer pairs to ensure each language has a minimum of 60M training pairs. For each of our datasets, we use 90% of the data for training, and the remaining 10% for development/validation.

4.2 Model Configuration

Input sentences are truncated to 256 tokens for the CNN model and 100 tokens for the transformer. The CNN encoder uses 2 CNN layers with filter width of [1, 2, 3, 5] and a filter size of 256.

[4]Question-answer prediction is similar to conversational-response prediction (Yang et al., 2018). We treat the question as the conversational input and the answer as the response. For improved answer selection, we provide a bag-of-words (BoW) context feature as an additional input to the answer encoder. The context could be the surrounding text or longer version of answer that provides more information. The context feature is encoded using a separate DAN encoder.

[5]Out-of-vocabulary characters map to an <UNK> token.

[6]QA pairs are mined from online forums and QA websites, including Reddit, StackOverflow, and YahooAnswers.

[7]The translation pairs are mined using a system similar to the approach described in Uszkoreit et al. (2010).

[8]MultiNLI (Williams et al., 2018), a more extensive corpus, contains examples from multiple sources but with different licences. Employing SNLI avoids navigating the licensing complexity of using MultiNLI to training public models.

Model	Quora	AskUbuntu	Average
Gillick et al. (2018)	87.5	37.3	62.4
USE_{CNN}	89.2	39.9	64.6
$\text{USE}_{\text{Trans}}$	89.1	42.3	65.7

Table 2: MAP@100 on SR (English). Models are compared with the best models from Gillick et al. (2018) that do not benefit from in-domain training data.

The Transformer encoder employs 6 transformer layers, with 8 attentions heads, hidden size 512, and filter size 2048. Model hyperparameters are tuned on development data sampled from the same sources as the training data. We export sentence encoding modules for our two encoder architectures: $\textbf{USE}_{\textbf{Trans}}$ and $\textbf{USE}_{\textbf{CNN}}$. We also export a larger graph for QA tasks from our Transformer based model that includes QA specific layers and support providing context information from the larger document as $\textbf{USE}_{\textbf{QA Trans+Cxt}}$.[9]

5 Experiments on Retrieval Tasks

In this section we evaluate our multilingual encoding models on semantic retrieval, bitext and retrieval question answer tasks.

5.1 Semantic Retrieval (SR)

Following Gillick et al. (2018), we construct semantic retrieval (SR) tasks from the Quora question-pairs (Hoogeveen et al., 2015) and AskUbuntu (Lei et al., 2016) datasets. The SR task is to identify all sentences in the retrieval corpus that are semantically similar to a query sentence.[10]

For each dataset, we first build a graph connecting each of the positive pairs, and then compute its transitive closure. Each sentence then serves as a test query that should retrieve all of the other sentences it is connected to within the transitive closure. Mean average precision (MAP) is employed to evaluate the models. More details on the constructed datasets can be found in Gillick et al. (2018). Both datasets are English only.

Table 2 shows the MAP@100 on the Quora/AskUbuntu retrieval tasks. We use Gillick et al. (2018) as the baseline model, which is trained using a similar dual encoder architecture. The numbers listed here are from the models without in-domain training data [11].

[9]While $\text{USE}_{\text{QA Trans+Cxt}}$ uses the same underlying shared encoder as $\text{USE}_{\text{Trans}}$ but with additional task specific layers, we anticipate that the models could diverge in the future.

[10]The task is related to paraphrase identification (Dolan et al., 2004) and Semantic Textual Similarity (STS) (Cer et al., 2017), but with the identification of meaning similarity being assessed in the context of a retrieval task.

Model	en-es	en-fr	en-ru	en-zh
Yang et al. (2019)	89.0	86.1	89.2	87.9
USE_{CNN}	85.8	82.7	87.4	79.5
$\text{USE}_{\text{Trans}}$	86.1	83.3	88.9	78.8

Table 3: P@1 on UN Bitext retrieval task.

Model	SQuAD Dev	SQuAD Train
Paragraph Retrieval		
$\text{USE}_{\text{QA Trans+Cxt}}$	63.5	53.3
BM25 (baseline)	61.6	52.4
Sentence Retrieval		
$\text{USE}_{\text{Trans}}$	47.1	37.2
$\text{USE}_{\text{QA Trans+Cxt}}$	53.2	43.3

Table 4: P@1 for SQuAD ReQA. Models are not trained on SQuAD. Dev and Train only refer to the respective sections of the SQuAD dataset.

5.2 Bitext Retrieval (BR)

Bitext retrieval performance is evaluated on the United Nation (UN) Parallel Corpus (Ziemski et al., 2016), containing 86,000 bilingual document pairs matching English (en) documents with with their translations in five other languages: French (fr), Spanish (es), Russian (ru), Arabic (ar) and Chinese (zh). Document pairs are aligned at the sentence-level, which results in 11.3 million aligned sentence pairs for each language pair.

Table 3 shows precision@1 (P@1) for the proposed models as well as the current state-of-the-art results from Yang et al. (2019), which uses a dual-encoder architecture trained on mined bilingual data. $\text{USE}_{\text{Trans}}$ is generally better than USE_{CNN}, performing lower than the SOTA but not by too much with the exception of en-zh.[12]

5.3 Retrieval Question Answering (ReQA)

Similar to the data set construction used for the SR tasks, the SQuAD v1.0 dataset (Rajpurkar et al., 2016) is transformed into a retrieval question answering (ReQA) task.[13] We first break all docu-

[11]The model for Quora is trained on Paralex (http://knowitall.cs.washington.edu/paralex) and AskUbuntu data. The model for AskUbuntu is trained on Paralex and Quora.

[12]Performance is degraded from Yang et al. (2019) due to using a single sentencepiece vocabulary to cover 16 languages. Languages like Chinese, Korean, Japanese have much more characters. To ensure the vocab coverage, sentencepiece tends to split the text of these languages into single characters, which increases the difficulty of the task.

[13]The retrieval question answering task was suggested by Chen et al. (2017) and then recently explored further by Cakaloglu et al. (2018). However, Cakaloglu et al. (2018)'s

Model	en	ar	de	es	fr	it	ja	ko	nl	pt	pl	ru	th	tr	zh / zh-t
Cross-lingual Semantic Retrieval (cl-SR)															
Quora															
USE_{CNN}	89.2	79.9	83.7	85.0	85.0	85.5	82.4	77.6	81.3	85.2	78.3	83.8	83.5	79.9	81.9
$\text{USE}_{\text{Trans}}$	89.1	83.1	85.5	86.3	86.7	86.8	85.1	82.5	83.8	86.5	82.1	85.7	85.8	82.5	84.8
AskUbuntu															
USE_{CNN}	39.9	33.0	35.0	35.6	35.2	36.1	35.5	35.1	34.5	35.6	32.9	35.2	35.2	32.8	34.6
$\text{USE}_{\text{Trans}}$	42.3	38.2	40.0	39.9	39.3	40.2	40.6	40.3	39.5	39.8	38.4	39.6	40.3	37.7	40.1
Average															
USE_{CNN}	64.6	56.5	59.4	60.3	60.1	60.8	59.0	56.4	57.9	60.4	55.6	59.5	59.4	56.4	58.3
$\text{USE}_{\text{Trans}}$	65.7	60.7	62.8	63.1	63.0	63.5	63.8	62.4	61.7	63.2	60.7	62.7	63.1	60.1	62.5
Cross-lingual Retrieval Question Answering (cl-ReQA)															
SQuAD train															
$\text{USE}_{\text{QA Trans+Cxt}}$	43.3	33.2	35.2	37.2	37.0	37.0	32.9	31.1	36.6	37.7	34.5	33.2	36.9	32.3	32.7

Table 5: Cross-lingual performance on Quora/AskUbuntu cl-SR (MAP) and SQuAD cl-ReQA (P@1). Queries/questions are machine translated to the other languages, while retrieval candidates remain in English.

ments in the dataset into sentences using an off-the-shelf sentence splitter. Each question of the (question, answer spans) tuples in the dataset is treated as a query. The task is to retrieve the sentence designated by the tuple answer span. Search is performed on a retrieval corpus consisting of all of the sentences within the corpus. We contrast sentence and paragraph-level retrieval using our models, with the later allowing for comparison against a BM25 baseline (Jones et al., 2000).[14]

We evaluated ReQA using the SQuAD dev and train sets and without training on the SQuAD data.[15] The sentence and paragraph retrieval P@1 are shown in table 4. For sentence retrieval, we compare encodings produced using context from the text surrounding the retrieval candidate, $\text{USE}_{\text{QA Trans+Cxt}}$, to sentence encodings produced without contextual cues, $\text{USE}_{\text{Trans}}$. Paragraph retrieval contrasts $\text{USE}_{\text{QA Trans+Cxt}}$ with BM25.

5.4 Cross-lingual Retrieval

Our earlier experiments are extended to explore cross-lingual semantic retrieval (cl-SR) and cross-lingual retrieval question answering (cl-ReQA). SR queries and ReQA questions are machine translated into other languages, while keeping the retrieval candidates in English.[16] Table 5 provides our cross-lingual retrieval results. On all the languages, $\text{USE}_{\text{Trans}}$ outperforms USE_{CNN}. While cross-lingual performance lags the English only tasks, the performance is surprisingly close given the added difficulty of the cross-lingual setting.

6 Experiments on Transfer Tasks

For comparison with prior USE models, English task transfer performance is evaluated on SentEval (Conneau and Kiela, 2018). For sentence classification transfer tasks, the output of the sentence encoders are provided to a task specific DNN. For the pairwise semantic similarity task, the similarity of sentence embeddings u and v is assessed using $-\arccos\left(\frac{uv}{||u||\ ||v||}\right)$ following Yang et al. (2018). As shown in table 6, our multilingual models show competitive transfer performance comparing with state-of-the-art sentence embedding models. $\text{USE}_{\text{Trans}}$ performs better than USE_{CNN} in all tasks. Our new multilingual $\text{USE}_{\text{Trans}}$ even outperforms our best previously released English only model, $\text{USE}_{\text{Trans}}$ for English (Cer et al., 2018), on some tasks.

7 Resource Usage

Figure (1) provides compute and memory usage benchmarks for our models.[17] Inference times on

use of sampling makes it difficult to directly compare with their results and we provide our own baseline base on BM25.

[14] BM25 is a strong baseline for text retrieval tasks. Paragraph-level experiments use the BM25 implementation: https://github.com/nhirakawa/BM25, with default parameters. We exclude sentence-level BM25, as BM25 generally performs poorly at this granularity.

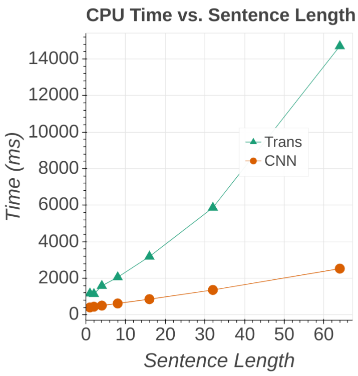
[15] For sentences, the resulting retrieval task for development set consists of 11,425 questions and 10,248 candidates, and the retrieval task for train set is consists of 87,599 questions and 91,703 candidates. For paragraph retrieval, there are 2,067 retrieval candidates in the development set and 18,896 in the training set. To retrieve paragraphs with our model, we first run sentence retrieval and use the retrieved nearest sentence to select the enclosing paragraph.

[16] Poor translations are detected and rejected when the original English text and English back translation have a cosine similarity < 0.5 according our previously released English $\text{USE}_{\text{Trans}}$ model (Cer et al., 2018).

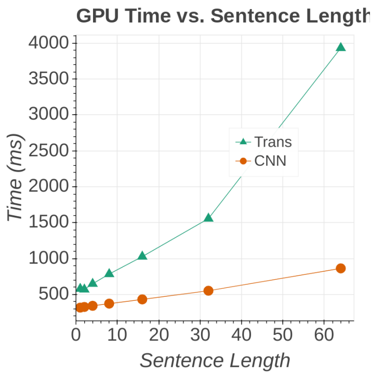
[17] CPU benchmarks are run on Intel(R) Xeon(R) Platinum 8173M CPU @ 2.00GHz. GPU benchmarks were run on an NVidia v100. Memory footprint was measured on CPU.

Model	MR	CR	SUBJ	MPQA	TREC	SST	STS Bench (dev / test)
USE mutlilingual models							
USE_{CNN}	73.8	83.2	90.1	87.7	96.4	78.1	0.829 / 0.809
$\text{USE}_{\text{Transformer}}$	78.1	**87.0**	92.1	89.9	**96.6**	80.9	**0.837 / 0.825**
The state-of-the-art English embedding models							
InferSent (Conneau et al., 2017)	81.1	86.3	92.4	**90.2**	88.2	84.6	0.801 / 0.758
Skip-Thought LN (Ba et al., 2016)	79.4	83.1	93.7	89.3	–	–	–
Quick-Thought (Logeswaran and Lee, 2018)	**82.4**	86.0	94.8	**90.2**	92.4	**87.6**	–
USE_{DAN} for English (Cer et al., 2018)	72.2	78.5	92.1	86.9	88.1	77.5	0.760 / 0.717
$\text{USE}_{\text{Transformer}}$ for English (Cer et al., 2018)	82.2	84.2	**95.5**	88.1	93.2	83.7	0.802 / 0.766

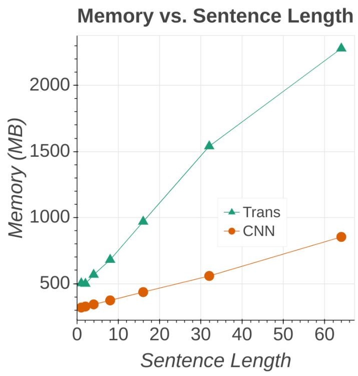
Table 6: Performance on English transfer tasks from SentEval (Conneau and Kiela, 2018).



(a) CPU Inference Time



(b) GPU Inference Time



(c) Memory Footprint

Figure 1: Resource usage for the multilingual Transformer and CNN encoding models.

GPU are 2 to 3 times faster than CPU. Our CNN models have the smallest memory footprint and are the fastest on both CPU and GPU. The memory requirements increase with sentence length, with the Transformer model increasing more than twice as fast as the CNN model.[18] While this makes CNNs an attractive choice for efficiently encoding longer texts, this comes with a corresponding drop in accuracy on many retrieval and transfer tasks.

8 Conclusion

We present two multilingual models for embedding sentence-length text. Our models *embed text from 16 languages into a shared semantic embedding space* and achieve performance on transfer tasks that approaches monolingual sentence embedding models. The models achieve good performance on semantic retrieval (SR), bitext retrieval (BR) and retrieval question answering (ReQA). They achieve performance on cross-lingual semantic retrieval (cl-SR) and cross-lingual retrieval question answering (cl-ReQA) that approaches monolingual SR and ReQA performance for many language pars. Our models are made freely available with additional documentation and tutorial colaboratory notebooks at: https://tfhub.dev/s?q=universal-sentence- encoder-multilingual.

[18] Transformer models are ultimately governed by a time and space complexity of $O(n^2)$. The benchmarks show for shorter sequence lengths the time and space requirements are dominated by computations that scale linearly with length and have a larger constant factor than the quadratic terms.

Acknowledgments

We thank our teammates from Descartes and other Google groups for their feedback and suggestions. Special thanks goes to Muthu Chidambaram for his early exploration of multilingual training, Taku Kudo for the SentencePiece model support, Chen Chen for the templates used to perform the transfer learning experiments and Mario Guajardo for an early version of the ReQA tutorial Colab.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of USENIX OSDI'16*, OSDI'16, pages 265–283.

Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Tolgahan Cakaloglu, Christian Szegedy, and Xiaowei Xu. 2018. Text embeddings for retrieval from a large knowledge base. *CoRR*, abs/1810.10176.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *CoRR*, abs/1810.12836.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *CoRR*, abs/1811.08008.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, pages 3:1–3:8.

K. Sparck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. *Inf. Process. Manage.*, 36(6):779–808.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1279–1289.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations (ICLR)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 6000–6010.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *CoRR*, abs/1902.08564.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534.