
DEEP LEARNING BASED SEMANTIC TEXTUAL SIMILARITY FOR APPLICATIONS IN MACHINE TRANSLATION DOMAIN

THARINDU RANASINGHE

A thesis submitted in partial fulfilment of the requirements of the
University of Wolverhampton for the degree of Doctor of Philosophy

2021

This work or any part thereof has not previously been presented in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

The right of Tharindu Ranasinghe to be identified as author of this work is asserted in accordance with ss.77 and 78 of the Copyright, Designs and Patents Act 1988. At this date copyright is owned by the author.

Signature:

Date:

ABSTRACT

An abstract is a synopsis of the thesis, and it goes in the file `abstract.tex`.

ACKNOWLEDGEMENTS

Your acknowledgements should go in `ack.tex`.

We would like to acknowledge Donald Craig at Memorial University, Newfoundland who published the meta-thesis on which this template is based. You can find Donald's work on his web site, here: <http://www.cs.mun.ca/~donald/metathesis/>.

CONTENTS

Abstract	ii
Acknowledgements	iv
List of Tables	x
List of Figures	xii
List of Code Listings	xiv
I Semantic Textual Similarity	1
1 Introduction	3
1.1 What is Semantic Textual Similarity?	3
1.2 Related Work	3
1.3 Datasets	3
1.3.1 English Datasets	3
1.3.2 Datasets on Other Languages	9
1.4 Applications	10
2 State of the Art Methods	11
2.1 Introduction	11
2.2 Related Work	11
2.3 Improving State of the Art STS Methods	11
2.3.1 Portability to Other Languages	11
2.3.2 Portability to Other Domains	11

2.4	Conclusions	11
3	Sentence Encoders	13
3.1	Introduction	13
3.2	Related Work	13
3.3	Exploring Sentence Encoders in English STS	13
3.4	Portability to Other Languages	13
3.5	Portability to Other Domains	13
3.6	Conclusions	13
4	Siamese Neural Networks	15
4.1	Introduction	15
4.2	Related Work	15
4.3	MAGRU: Improving Siamese Neural Networks	15
4.3.1	Portability to Other Languages	15
4.3.2	Portability to Other Domains	15
4.4	Conclusions	15
5	Transformers	17
5.1	Introduction	17
5.2	Related Work	17
5.3	Exploring Transformers in English STS	17
5.4	Exploring Transformers for STS in Other Languages	17
5.5	Exploring Transformers for STS in Other Domains	17
5.6	Conclusions	17

II	Applications - Translation Memories	19
1	Introduction	21
1.1	What is Translation Memory?	21
1.2	Datasets	21
1.3	Related Work	21
1.4	STS for Translation Memories	21
2	Sentence Encoders for Translation Memories	23
2.1	Introduction	23
2.2	Methodology	23
2.3	Results and Evaluation	23
III	Applications - Translation Quality Estimation	25
1	Introduction	27
1.1	What is Translation Quality Estimation?	27
1.2	Datasets	27
1.3	Related Work	27
1.4	STS for Translation Quality Estimation	27
2	TransQuest: STS Architectures for QE	29
2.1	Introduction	29
2.2	Methodology	29
2.3	Results and Evaluation	29
	Bibliography	30

LIST OF TABLES

1.1	Example sentence pairs from the SICK dataset	6
1.2	Example sentence pairs from the STS2017 English dataset . .	8
1.3	Example sentence pairs from the Quora Question Pairs dataset	9

LIST OF FIGURES

LISTINGS

Part I

Semantic Textual Similarity

CHAPTER 1

INTRODUCTION

1.1 What is Semantic Textual Similarity?

1.2 Related Work

1.3 Datasets

We experimented with several datasets throughout the experiments in the Semantic Textual Similarity Section. In order to maintain the versatility of our methods we experimented with several English datasets as well as several non English datasets and several datasets from different domains which we will introduce in this section. All of the datasets which are described here are publicly available and can be considered as STS benchmarks.

1.3.1 English Datasets

1. **SICK dataset**¹ - The SICK data contains 9927 sentence pairs with a 5,000/4,927 training/test split which were employed in the SemEval 2014 Task1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual En-

¹The SICK dataset is available to download at <https://wiki.cimtec.unitn.it/tiki-index.php?page=CLIC>

1.3. DATASETS

tailment [1]. The dataset has two types of annotations: Semantic Relatedness and Textual Entailment. We only use Semantic Relatedness annotations in our research. SICK was built starting from two existing datasets: the 8K ImageFlickr data set² [2] and the SemEval-2012 STS MSR-Video Descriptions dataset³ [3]. The 8K ImageFlickr dataset is a dataset of images, where each image is associated with five descriptions. To derive SICK sentence pairs the organisers randomly selected 750 images and sampled two descriptions from each of them. The SemEval2012 STS MSR-Video Descriptions data set is a collection of sentence pairs sampled from the short video snippets which compose the Microsoft Research Video Description Corpus⁴. A subset of 750 sentence pairs have been randomly chosen from this data set to be used in SICK.

In order to generate SICK data from the 1,500 sentence pairs taken from the source data sets, a 3-step process has been applied to each sentence composing the pair, namely *(i) normalisation*, *(ii) expansion* and *(iii) pairing* [1]. The *normalisation* step has been carried out on the original sentences to exclude or simplify instances that contained lexical, syntactic or semantic phenomena such as named entities, dates,

²The 8K ImageFlickr data set is available at <http://hockenmaier.cs.illinois.edu/8k-pictures.html>

³The SemEval-2012 STS MSR-Video Descriptions dataset is available at <https://www.cs.york.ac.uk/semeval-2012/task6/index.html>

⁴The Microsoft Research Video Description Corpus is available to download at <https://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/>

numbers, multiword expressions etc. In the *expansion* step syntactic and lexical transformations with predictable effects have been applied to each normalized sentence, in order to obtain (i) a sentence with a similar meaning, (ii) a sentence with a logically contradictory or at least highly contrasting meaning, and (iii) a sentence that contains most of the same lexical items, but has a different meaning. Finally, in the *pairing* step each normalised sentence in the pair has been combined with all the sentences resulting from the expansion phase and with the other normalised sentence in the pair. Furthermore, a number of pairs composed of completely unrelated sentences have been added to the data set by randomly taking two sentences from two different pairs [1].

Each pair in the SICK dataset has been annotated to mark the degree to which the two sentence meanings are related (on a 5-point scale). The ratings have been collected through a large crowdsourcing study, where each pair has been evaluated by 10 different annotators. Once all the annotations were collected, the relatedness gold score has been computed for each pair as the average of the ten ratings assigned by the annotators [1]. Table 1.1 shows examples of sentence pairs with different degrees of semantic relatedness; gold relatedness scores are expressed on a 5-point rating scale. Given a test sentence pair the machine learning models require to predict a value between 0-5 which reflects the relatedness of the given sentence pair.

1.3. DATASETS

Sentence Pair	Relatedness
1. A little girl is looking at a woman in costume. 2. A young girl is looking at a woman in costume.	4.7
1. Nobody is pouring ingredients into a pot. 2. Someone is pouring ingredients into a pot.	3.5
1. Someone is pouring ingredients into a pot. 2. A man is removing vegetables from a pot.	2.8
1. A man is jumping into an empty pool. 2. There is no biker jumping in the air.	1.6

Table 1.1: Example sentence pairs from the SICK dataset with their gold relatedness scores (on a 5-point rating scale).

2. **STS 2017 English Dataset**⁵ STS 2017 English Dataset was employed in SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation which is the most recent STS task in SemEval [4]. As the training data for the competition, participants were encouraged to make use of all existing data sets from prior STS evaluations including all previously released trial, training and evaluation data from SemEval 2012 - 2016 [3, 5, 6, 7, 8]. Once combined we had 8277 sentence pairs for training.

On the other hand, a fresh test set of 250 sentence pairs was provided by SemEval-2017 STS Task organisers [4]. The Stanford Natural Language Inference (SNLI) corpus [9] was the primary data source for this test set. Similar to the SICK dataset, Each pair in the STS 2017 English Test set has been annotated to mark the degree to which the two sentence meanings are related (on a 5-point scale). The ratings have

⁵The STS 2017 English Dataset is available to download at <http://ixa2.si.ehu.es/stswiki/>

been collected through crowdsourcing on Amazon Mechanical Turk⁶. Five annotations have been collected per pair and gold score has been computed for each pair as the average of the five ratings assigned by the annotators. However, unlike the SICK dataset, the organisers has a clear explanations for the score ranges. Table 1.2 shows some example sentence pairs from the dataset with the gold labels and their explanations. Similar to the SICK dataset, the machine learning models require to predict a value between 0-5 which reflects the similarity of the given sentence pair.

3. **Quora Question Pairs**⁷ The Quora Question Pairs dataset is a big dataset which was first released for a Kaggle Competition⁸. Quora is a question-and-answer website where questions are asked, answered, followed, and edited by internet users, either factually or in the form of opinions. If a particular new question has been asked before, users merge the new question to the original question flagging it as a duplicate. The organisers used this functionality to create the dataset and did not use a separate annotation process. Their original sampling method has returned an imbalanced dataset with many more

⁶Amazon Mechanical Turk is a crowdsourcing website for businesses to hire remotely located *crowd workers* to perform discrete on-demand tasks. It is available at <https://www.mturk.com/>

⁷The Quora Question Pairs Dataset is available to download at http://qm.fs.quoracdn.net/quora_duplicate_questions.tsv

⁸Kaggle is an online community of data scientists and machine learning practitioners that hosts machine learning competitions. The Quora Question Pairs competition is available on <https://www.kaggle.com/c/quora-question-pairs>

1.3. DATASETS

Sentence Pair	Relatedness
<i>The two sentences are completely equivalent as they mean the same thing.</i> 1. The bird is bathing in the sink. 2. Birdie is washing itself in the water basin.	5
<i>The two sentences are completely equivalent as they mean the same thing.</i> 1. The bird is bathing in the sink. 2. Birdie is washing itself in the water basin.	4
<i>The two sentences are roughly equivalent, but some important information differs/missing.</i> 1. John said he is considered a witness but not a suspect. 2. “He is not a suspect anymore.” John said.	3
<i>The two sentences are not equivalent, but share some details.</i> 1. They flew out of the nest in groups. 2. They flew into the nest together.	2
<i>The two sentences are not equivalent, but are on the same topic.</i> 1. The woman is playing the violin. 2. The young lady enjoys listening to the guitar.	1
<i>The two sentences are completely dissimilar</i> 1. The black dog is running through the snow. 2. A race car driver is driving his car through the mud.	0

Table 1.2: Example sentence pairs from the STS2017 English dataset with their gold relatedness scores (on a 5-point rating scale) and explanations.

true examples of duplicate pairs than non-duplicates. Therefore, the organisers have supplemented the dataset with negative examples. One source of negative examples have been pairs of *related question* which, although pertaining to similar topics, are not truly semantically equivalent.

The dataset has 400,000 question pairs and we used 4:1 split on that to

separate it into a training set and a test set resulting 320,000 questions pairs in the training set and 80,000 sentence pairs in the testing set. The machine learning models need to predict a value between 0 and 1 that reflects whether it is a duplicate question pair or not. 1 indicates that a certain question pair is a duplicate and 0 indicates it is not a duplicate.

Question Pair	is-duplicate
1. What are natural numbers? 2. What is a least natural number?	0
1. Which Pizzas are most popularly ordered in Dominos menu? 2. How many calories does a Dominos Pizza have?	0
1. How do you start a bakery? 2. How can one start a bakery business?	1
1. Should I learn Python or Java first? 2. If I had to choose between learning Java and Python what should I choose to learn first?	1

Table 1.3: Example sentence pairs from the Quora Question Pairs dataset with their gold is-duplicate value.

This is different to the previous datasets since it is not artificially created and use day to day language. Since it has more than 300,000 training instances deep learning systems will benefit more when used on this dataset.

1.3.2 Datasets on Other Languages

1. Spanish STS Dataset -

2. Arabic STS Dataset

1.3.3 Datasets on Different Domains

1. STS Dataset -

2. Arabic STS Dataset

1.4 Applications

CHAPTER 2

STATE OF THE ART METHODS

2.1 Introduction

[10]

2.2 Related Work

2.3 Improving State of the Art STS Methods

2.3.1 Portability to Other Languages

2.3.2 Portability to Other Domains

2.4 Conclusions

CHAPTER 3

SENTENCE ENCODERS

3.1 Introduction

[11]

3.2 Related Work

3.3 Exploring Sentence Encoders in English STS

3.4 Portability to Other Languages

3.5 Portability to Other Domains

3.6 Conclusions

CHAPTER 4

SIAMESE NEURAL NETWORKS

4.1 Introduction

[12]

4.2 Related Work

4.3 MAGRU: Improving Siamese Neural Networks

4.3.1 Portability to Other Languages

4.3.2 Portability to Other Domains

4.4 Conclusions

CHAPTER 5

TRANSFORMERS

5.1 Introduction

[13]

5.2 Related Work

5.3 Exploring Transformers in English STS

5.4 Exploring Transformers for STS in Other Languages

5.5 Exploring Transformers for STS in Other Domains

5.6 Conclusions

Part II

Applications - Translation Memories

CHAPTER 1

INTRODUCTION

1.1 What is Translation Memory?

[14]

1.2 Datasets

1.3 Related Work

1.4 STS for Translation Memories

CHAPTER 2

SENTENCE ENCODERS FOR TRANSLATION MEMORIES

2.1 Introduction

[15]

2.2 Methodology

2.3 Results and Evaluation

Part III

Applications - Translation Quality Estimation

CHAPTER 1

INTRODUCTION

1.1 What is Translation Quality Estimation?

1.2 Datasets

1.3 Related Work

[16]

1.4 STS for Translation Quality Estimation

CHAPTER 2

TRANSQUEST: STS ARCHITECTURES FOR QE

2.1 Introduction

[17]

2.2 Methodology

2.3 Results and Evaluation

BIBLIOGRAPHY

- [1] Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland, August 2014. Association for Computational Linguistics.
- [2] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon’s mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 139–147, Los Angeles, June 2010. Association for Computational Linguistics.
- [3] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

BIBLIOGRAPHY

- [4] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 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, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [5] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.
- [6] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, August 2014. Association for Computational Linguistics.
- [7] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 task 2: Semantic textual similarity, English,

BIBLIOGRAPHY

- Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June 2015. Association for Computational Linguistics.
- [8] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June 2016. Association for Computational Linguistics.
- [9] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
- [10] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. Enhancing unsupervised sentence similarity methods with deep contextualised word representations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 994–1003, Varna, Bulgaria, September 2019. INCOMA Ltd.
- [11] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and An-

BIBLIOGRAPHY

- toine Bordes. 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, September 2017. Association for Computational Linguistics.
- [12] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. Semantic textual similarity with Siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1004–1011, Varna, Bulgaria, September 2019. INCOMA Ltd.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [14] Peter J. Arthern. Machine translation and computerized terminology systems: A translator’s viewpoint. *Translating and the Computer, Proceedings of a Seminar, London 14th November 1978. Amsterdam: North-Holland Publishing Company*, pages 77–108, 1979.

BIBLIOGRAPHY

- [15] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. Intelligent translation memory matching and retrieval with sentence encoders. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 175–184, Lisboa, Portugal, November 2020. European Association for Machine Translation.
- [16] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy, July 2019. Association for Computational Linguistics.
- [17] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.