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.

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

List of Tables

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 re 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-

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

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 set [2] [2] and the SemEval-2012 STS MSR-Video Descriptions dataset [3] [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 Corpus [4]. 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,

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

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

[4] The 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.

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. More information about the datasets used to build the training set is available in Table 1.2.

   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 En-

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

glish 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 been collected through crowdsourcing on Amazon Mechanical Turk[6]. 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 1.3 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 Competition[8]. 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

[6] Amazon 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/

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

[8] Kaggle 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

Year	**Dataset**	**Pairs**	**Source**
2012 [3]	MSRpar	1500	newswire
2012 [3]	MSRvid	1500	videos
2012 [3]	OnWN	750	glosses
2012 [3]	SMTnews	750	WMT eval.
2012 [3]	SMTeuroparl	750	WMT eval.
2013 [5]	HDL	750	newswire
2013 [5]	FNWN	189	glosses
2013 [5]	OnWN	561	glosses
2013 [5]	SMT	750	MT eval.
2014 [6]	HDL	750	newswire headlines
2014 [6]	OnWN	750	glosses
2014 [6]	Deft-forum	450	forum posts
2014 [6]	Deft-news	300	news summary
2014 [6]	Images	750	image descriptions
2014 [6]	Tweet-news	750	tweet-news pairs
2015 [7]	HDL	750	newswire headlines
2015 [7]	Images	750	image descriptions
2015 [7]	Ans.-student	750	student answers
2015 [7]	Ans.-forum	375	Q&A forum answers
2015 [7]	Belief	375	committed belief
2016 [8]	HDL	249	newswire headlines
2016 [8]	Plagiarism	230	short-answer plag.
2016 [8]	post-editing	244	MT postedits
2016 [8]	Ans.-Ans.	254	Q&A forum answers
2016 [8]	Quest.-Quest.	209	Q&A forum questions
2017 [4]	Trial	23	Mixed STS 2016

Table 1.2: Information about the datasets used to build the English STS 2017 training set. The **Year** column shows the year of the SemEval competition that the dataset got released. **Dataset** column expresses the acronym used describe a dataset in that year. **Pairs** is the number of sentence pairs in that particular dataset and **Source** shows the source of the sentence pairs.

and did not use a separate annotation process. Their original sampling method has returned an imbalanced dataset with many more true examples of duplicate pairs than non-duplicates. Therefore, the organisers have supplemented the dataset with negative examples. One

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

Table 1.3: Example sentence pairs from the STS2017 English dataset with their gold relatedness scores (on a 5-point rating scale) and explanations.

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 1.4: Example question pairs from the Quora Question Pairs dataset with their gold is-duplicate value. **Question Pair** column shows the two questions and **is-duplicated** column denotes whether it is a duplicated pair or not.

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

In order to evaluate the potability of our STS methods we used several non-English datasets through out the experiments which we describe below.

1. **Spanish STS Dataset** [9] - Spanish STS dataset that we used was employed for Spanish STS subtask in SemEval 2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation [4]. The training set has 1250 sentence pairs annotated with a relatedness score between 0 and 4. The training set combined several datasets from previous SemEval STS shared tasks also[4]. Table 1.5 shows more information about the trainin set. There were two sources for test set - Spanish news and Spanish Wikipedia dump having 500 and 250 sentence pairs respectively [4]. Both datasets were annotated with a relatedness score between 0 and 4. Table 1.6 shows few pairs of sentences with their similarity score.

Year	Dataset	Pairs	Source
2014 [6]	Trial	56	NR
2014 [6]	Wiki	324	Spanish Wikipedia
2014 [6]	News	480	Newswire
2015 [6]	Wiki	251	Spanish Wikipedia
2015 [7]	News	500	Sewswire
2017 [4]	Trial	23	Mixed STS 2016

Table 1.5: Information about the datasets used to build the Spanish STS training set. The **Year** column shows the year of the SemEval competition that the dataset got released. **Dataset** column expresses the acronym used describe a dataset in that year. **Pairs** is the number of sentence pairs in that particular dataset and **Source** shows the source of the sentence pairs.

[9] The Spanish STS dataset can be downloaded at http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

Sentence Pair	**Similarity**
1. Amás, los misioneros apunten que los númberos d'infectaos puen ser shasta dos o hasta cuatro veces más grandess que los oficiales. *(Furthermore, missionaries point out that the numbers of infected can be up to two or up to four times larger than the official ones.)* 2. Los cadáveres de personas fallecidas pueden ser hasta diez veces más contagiosos que los infectados vivos. *(The corpses of deceased people can be up to ten times more contagious than those infected alive.)*	0.6
1. La policía abatió a un caníbal cuando devoraba a una mujer Matthew Williams, de 34 años, fue sorprendido en la madrugada mordiendo el rostro de una joven a la que había invitado a su hotel. *(Police killed a cannibal while devouring a woman Matthew Williams, 34, was caught early in the morning biting the face of a young woman he had invited to his hotel.)* 2. La policía de Gales del Sur mató a un caníbal cuando se estaba comiendo la cara de una mujer de 22 años en la habitación de un hotel. *(South Wales police killed a cannibal when he was eating the face of a 22-year-old woman in a hotel room.)*	2
1. Ollanta Humala se reúne mañana con el Papa Francisco. *(Ollanta Humala meets tomorrow with Pope Francis.)* 2. El Papa Francisco mantuvo hoy una audiencia privada con el presidente Ollanta Humala, en el Vaticano. *(Pope Francis held a private audience today with President Ollanta Humala, at the Vatican.)*	3

Table 1.6: Example sentence pairs from the Spanish STS dataset. **Sentence Pair** column shows the two sentences. We also included their translations in the table. The translations were done by a native Spanish speaker. **Similarity** column indicates the annotated similarity of the two sentences.

2. **Arabic STS Dataset** [10] The Arabic STS dataset we selected was also used for the Arabic STS subtask in SemEval 2017 Task 1: Semantic

[10] The Arabic STS dataset can be downloaded at http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

Textual Similarity Multilingual and Cross-lingual Focused Evaluation [4]. Unlike Spanish, no data from previous SemEval competitions were available since this was the first time an Arabic STS task was organised in SemEval. More information about the extracted sentences will be shown in the Table 1.7.

Dataset	**Pairs**	**Source**
Trial	23	Mixed STS 2016
MSRpar	510	newswire
MSRvid	368	videos
SMTeuroparl	203	WMT eval.

Table 1.7: Information about the datasets used to build the Arabic STS training set. **Dataset** column expresses the acronym used describe the dataset. **Pairs** is the number of sentence pairs in that particular dataset and **Source** shows the source of the sentence pairs.

To prepare the annotated instances, a subset of the English STS 2017 dataset has been selected and human translated into Arabic. Sentences have been translated independently from their pairs. Arabic translation has been provided by native Arabic speakers with strong English skills in Carnegie Mellon University in Qatar . Translators have been given an English sentence and its Arabic machine translation5 where they have performed post-editing to correct errors. STS labels have been then transferred to the translated pairs. Therefore, annotation guidelines and the template will be similar to the English STS 2017 dataset. 1103 sentence pairs were available for training and 250 sentence pairs were available in the test set. Table 1.8 shows few pairs of sentences with their similarity score.

Sentence Pair	Similarity
1. أحدهم يقلي لحما. *Someone is frying meat.* 2. أحدهم يعزف البيانو. *Someone plays the piano.*	0.250
1. أمرأة تظيف المكونات في الإناء. *A woman cleaning ingredients in the bowl.* 2. إمرأة تكسر ثلاثة بيضات في الإناء. *A woman breaks three eggs in a bowl.*	1.750
1. طفلة تعزف القيثارة. *A Child is playing harp.* 2. رجل يعزف القيثارة . *A man plays the harp.*	2.250
1. المرأة تقطع البصل الأخضر. *The woman chops green onions.* 2. إمرأة تقشر بصلة. *A woman peeling an onion.*	3.250
1. رجل يرقص على صقف الغرفة. *A man dancing on the roof of the room.* 2. رجل يرقص رأسا على عقب على السقف. *A man is dancing upside down on the ceiling.*	4.600

Table 1.8: Example question pairs from the Arabic STS dataset. **Sentence Pair** column shows the two sentences. We also included their translations in the table. The translations were done by a native Arabic speaker. **Similarity** column indicates the annotated similarity of the two sentences.

1.3.3 Datasets on Different Domains

1. **Bio-medical STS Dataset** - [10]

2. **Clinical STS Dataset** [11]

1.4 Applications

Chapter 2

State of the Art Methods

2.1 Introduction

[12]

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

[13]

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

[14]

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

[15]

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?

[16]

1.2 Datasets

1.3 Related Work

1.4 STS for Translation Memories

Chapter 2

Sentence Encoders for Translation Memories

2.1 Introduction

[17]

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

[18]

1.4 STS for Translation Quality Estimation

Chapter 2

TransQuest: STS Architectures for QE

2.1 Introduction

[19]

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.

[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,

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] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 07 2017.

[11] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. Medsts: a resource for

clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72, Mar 2020.

[12] 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.

[13] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine 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.

[14] 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.

[15] 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 Amer-*

ican 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.

[16] 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.

[17] 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.

[18] 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.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual rep-

resentation 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.