Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2928–2939
Marseille, 20-25 June 2022

Evaluating Multilingual Sentence Representation Models in a Real Case Scenario

Rocco Tripodi[1], Rexhina Blloshmi[2], Simon Levis Sullam[3]
[1]Universisty of Bologna, [2]Sapienza University of Rome, [3]Ca' Foscari University of Venice
rocco.tripodi@unibo.it, blloshmi@di.uniroma1.it, levissmn@unive.it

Abstract

In this paper, we present an evaluation of sentence representation models on the paraphrase detection task. The evaluation is designed to simulate a real-world problem of plagiarism and is based on one of the most important cases of forgery in modern history: the so-called "Protocols of the Elders of Zion". The sentence pairs for the evaluation are taken from the infamous *forged* text "Protocols of the Elders of Zion" (*Protocols*) by unknown authors; and by "Dialogue in Hell between Machiavelli and Montesquieu" by Maurice Joly. Scholars have demonstrated that the first text plagiarizes from the second (Cohn, 1967; Taguieff, 2004), indicating all the forged parts on qualitative grounds. Following this evidence, we organized the rephrased texts and asked native speakers to quantify the level of similarity between each pair. We used this material to evaluate sentence representation models in two languages: English and French, and on three tasks: similarity correlation, paraphrase identification, and paraphrase retrieval. Our evaluation aims at encouraging the development of benchmarks based on real-world problems, as a means to prevent problems connected to AI hypes, and to use NLP technologies for social good. Through our evaluation, we are able to confirm that the infamous *Protocols* are actually a plagiarized text but, as we will show, we encounter several problems connected with the convoluted nature of the task, that is very different from the one reported in standard benchmarks of paraphrase detection and sentence similarity. Code and data available at `https://github.com/roccotrip/protocols`.

Keywords: sentence representation, sentence similarity, paraphrase detection, real-world evaluation

1. Introduction

Paraphrase detection is the task of analyzing two segments of text and determining if they have the same meaning despite the differences in structure and wording (Wieting et al., 2015). This definition, like that of *synonymity* for words, is hard to apply in real case scenarios, since each lexical form has a definite connotation (Bloomfield, 1933). To this end, in this work, we use the denotation by Bhagat and Hovy (2013) of (quasi-) paraphrases, i.e., *sentences or phrases that convey approximately the same meaning using different words*.

Paraphrase detection is an application of compositional semantics, a discipline that studies how lexical semantics units combine to generate complex thoughts, and therefore it is based on the principle of compositionality: the meaning of the whole is a function of the meaning of its parts (Frege, 2002). The first compositional semantics models were based on formal semantics (Montague, 2019), while the most recent models are based on distributional semantics (Mitchell and Lapata, 2008; Baroni and Lenci, 2010; Zanzotto et al., 2010; Ji and Eisenstein, 2013) and in particular on neural network models (Socher et al., 2011a; Yin et al., 2016; Liu et al., 2019a). Unfortunately, the evaluation of these models is conducted on datasets composed of short and grammatically simple sentences in most cases. This is because constructing more challenging benchmarks requires a vast amount of time. Therefore, semi-automatic procedures have been employed for creating these datasets. Furthermore, most of them are in English only, limiting the evaluation of models on other languages. Only recently, multilingual datasets have been released (Yang et al., 2019; Hu et al., 2020), which however, employ the characteristics of the existing datasets in English, i.e., are automatically constructed and contain short and highly overlapping sentence pairs.

In this work, we collected a small but challenging set of sentences for paraphrase detection based on a well-known forgery and plagiarism case. Historians (Cohn, 1967; Taguieff, 2004) have definitively established that several segments of the book "Dialogue in Hell between Machiavelli and Montesquieu" (henceforth, *Dialogue*) by Maurice Joly have been used to assemble the infamous booklet "The Protocols of the Elder of Zion" by unknown authors (henceforth, *Protocols*). Indeed, we choose this specific case study due to the fact that *Protocols* can be considered as one of the most famous – and pernicious – cases of plagiarism in modern history. For this reason, we decided to test whether an automatic analysis can effectively discover the plagiarized texts identified using the specialized knowledge of scholars in the field, aiming at developing technologies that can be used for social good.[1]

From a more technical point of view, our evaluation is tailored to alleviate some of the drawbacks observed in the current evaluation benchmarks, which: i. are mainly in English; ii. comprise short sentences; iii. consist of paraphrase pairs with high word overlap; iv. have simple

[1]We want to underline here that the publication and the spread of *Protocols* contributed to the rise of antisemitism during the first decades of Twentieth century and still today the text has a large circulation and following despite being a forgery (see Section 3.1 for a more detailed discussion on the origin of *Protocols*).

syntactic structures; v. are mostly formulated as binary classification; vi. do not provide retrieval tasks. The evaluation conducted in this paper is, in fact, conducted on two languages and involve long sentences with low word overlap. Indeed, many of our paraphrases are abstract and involve similarities at concept level that can be conveyed using elliptical constructions. We use this textual material to formulate three different tasks with increasing level of sophistication, in order to effectively simulate a real-world scenario.

Most importantly, the ultimate aim of this work is to encourage researchers to test models on highly challenging real case problems with a social impact.

2. Related Work

We divide this Section in two parts: Section 2.1 describes several datasets for paraphrase detection and sentence similarity; Section 2.2 introduces recent approaches to create sentence representations.

2.1. Datasets

Classification datasets. The most popular benchmark for sentence understanding tasks is GLUE (Wang et al., 2019b). It consists of nine datasets: i. Corpus of Linguistic Acceptability (Warstadt et al., 2018, CoLA); ii. Stanford Sentiment Treebank (Socher et al., 2013, SST-2); iii. Microsoft Research Paraphrase Corpus (Dolan et al., 2004, MRPC); iv. Quora Question Pairs; v. Semantic Textual Similarity Benchmark (Cer et al., 2017, STS); vi. The Stanford Question Answering Dataset (Rajpurkar et al., 2016, QNLI) vii. Multi-Genre NLI corpus (Williams et al., 2018, MNLI); viii. Recognizing Textual Entailment ix. Winograd Schema Challenge (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009, RTE); (Levesque et al., 2012, WNLI). The first two are single sentence tasks; the datasets iii, iv, v are sentence similarity and paraphrase detection tasks; the other are natural language inference tasks. Most of them are formulated as binary classification problems, except STS, which is evaluated using regression, and MNLI with three classes. More recently, Wang et al. (2019a) introduced SuperGLUE, a set of more challenging tasks than GLUE. It was necessary since recent pre-trained models proved to solve GLUE tasks easily.

Sentence similarity/relatedness datasets. Semantic *relatedness* tasks consist of learning to predict a score between two sentences ranging from 0 to 5. The benchmarks for this task are SICK (Marelli et al., 2014) and STS (Cer et al., 2017). Sentences in these datasets are concise and constructed following the same template. This might allow simple models, that deal only with the identification of common words, to obtain high results. For example, *The young boys are playing outdoors and the man is smiling nearby*, and *The kids are playing outdoors near a man with a smile* have a relatedness score of 4.7; *A group of friends are riding the current in a raft* and *This group of people is practicing water safety and wearing preservers* have a relatedness score of 3.1; *A person is wearing a hat and is sitting on the grass* and *A man is running in a field* have a relatedness score of 1.4. As we can see from all these examples, sentence pairs have similar length, high scores correspond to sentences that have key words replaced by synonyms, and low scores correspond to sentences that have words with contrasting meanings (running vs sitting).

The sentence similarity task consists in evaluating how a similarity measure between the representation of two sentences correlates with human judgments. Several datasets have been proposed in the SemEval shared task (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015; Agirre et al., 2016). Their limitations consist in the short length of the sentences, the limited vocabulary, the simplicity of the syntactic structure of the sentences, and the overlap of the words used in the sentence pairs.

Retrieval datasets. The task of finding comparable sentences in multiple corpora is gaining momentum, in particular, because of the success of transfer learning approaches in multilingual tasks (Ruder et al., 2019). Examples of such datasets include BUCC (Zweigenbaum et al., 2018), which has been released with the shared task organized in the workshop *Building and Using Parallel Corpora* consisting of parallel texts in 5 languages, and Tatoeba (Artetxe and Schwenk, 2019), which has been recently released for sentence retrieval, consisting of up to 1000 English sentence pairs aligned with 122 languages. While these datasets retrieve comparable sentences across languages, in our tackled task, we retrieve sentences written in the same language but which have been plagiarized.

Paraphrase detection datasets. The most popular benchmark for paraphrase detection is presented by Dolan et al. (2004, MRPC), which is organized as a binary classification task. Barrón-Cedeño et al. (2013) proposed a selection of MRPC annotated with paraphrases types. More recently, (Zhang et al., 2019) presented the PAWS dataset, which examples are generated through an automatic process of word scrambling and back translation via language models. It is designed for the paraphrases to have a high word overlap, and, similar to MRPC, it is a binary classification task. Yang et al. (2019, PAWS-X) instead, extend the PAWS dataset to six other languages, consisting of paraphrase pairs with high word overlap, similar to its English version. These features make the PAWS and PAWS-X dataset less challenging than our task, which, in contrast, consists of challenging sentence pairs that contain few overlapping words and are not automatically created. Furthermore, while these datasets might be suitable for the evaluation of a classifier, in real-world scenarios a paraphrase has to be found in a large document or a collection of documents, and it should produce a ranking of retrieved sentences, therefore identifying the degree of rephrasing.

Dataset	Sent. length # ↑	Word overlap % ↓
SemEval2017T1	8.7 ± 3.34	23.8 ± 11.8
SICK	9.6 ± 3.69	29.2 ± 12.1
SemEval2015T2	11.5 ± 6.38	15.0 ± 11.8
SemEval2016T1	14.3 ± 19.45	26.2 ± 11.9
MRPC	19.7 ± 16.03	28.0 ± 8.1
PAWS	21.5 ± 5.42	40.4 ± 4.7
Our work	$\mathbf{23.5 \pm 13.64}$	$\mathbf{10.3 \pm 6.4}$

Table 1: Dataset statistics: Average sentence length expressed as the average number of tokens; Average word overlap expressed as the % of common tokens. White space was used as token delimiter.

Miscellanea. The evaluation of sentence representation has also been conducted on downstream classification tasks including sentiment analysis (Pang and Lee, 2004; Socher et al., 2013), question type (Voorhees and Tice, 2000), product reviews (Hu and Liu, 2004), subjectivity/objectivity (Pang and Lee, 2004) and opinion polarity (Wiebe et al., 2005). In such case, it is evaluated not only the sentence representation model, but also the ability of the classifier to discriminate among features that belong to each class.

Dataset statistics. In the previous paragraphs, we often mentioned that datasets relevant to sentence representation evaluation contain short sentences or highly overlapping pairs. To quantify this statement, in Table 1, we show the statistics regarding the above-mentioned English test sets, and those related to the textual material used in this work. Specifically, we calculate two measures: the average sentence length and the average word overlap between sentences; the latter is computed as the number of tokens in common in each pair divided by the total number of words in the sentences. As one can see, the collected text for this study has both longer sentences and lower word overlap scores compared to all the other datasets.

2.2. Sentence Representation Approaches

Word embedding models based on deep neural networks have received significant attention, especially due to the success of distributional semantics models (Mikolov et al., 2013; Devlin et al., 2019), and are nowadays the main building block for NLP applications. Over the years, different approaches have been exploited to derive sentence embeddings starting from these word vectors, following the principle of compositionality. In fact, for many years, unsupervised models, reminiscent of the bag-of-words (BoW) approach and based on the construction of sentence vectors as a linear combination of word vectors, were very popular. Among them, we cite Smooth Inverse Frequency (Arora et al., 2017, SIF) that extends the BoW model and connects with the TF-IDF weighting schema. It uses a weighted average of the words composing a sentence, and then, modifies it using PCA. This simple model has demonstrated to be beneficial in many downstream tasks, providing an effective way of obtaining sentence representations compositionally.

Beside compositional models for sentence embeddings, several approaches have emerged to directly encode the semantics of a sentence based on: recurrent neural networks (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014); attention mechanisms (Bahdanau et al., 2015), where instead of using indiscriminately all the words in a sentence to construct its final representation, the relations among the words are weighted; recursive neural networks (Socher et al., 2011b), which use external knowledge about the structure of the sentence, e.g., syntactic structure; structured LSTM (Zhu et al., 2015); autoencoders (Socher et al., 2011c), to name a few.

In the recent years, several strong baseline models have been proposed, especially tailored for sentence representation learning and knowledge transfer to several downstream NLP tasks. Universal Sentence Encoder (Cer et al., 2018, USE) uses an attention mechanism to produce context-aware representations of words that are then averaged to obtain a sentence-level representation. LASER (Artetxe and Schwenk, 2019) is another language-agnostic sentence encoder trained in parallel sentences with the aim of producing similar representations for sentences expressing the same meaning in different languages. More recenlty, motivated by the performance of contextualized models (Devlin et al., 2019), Reimers and Gurevych (2019) developed a modification of the pre-trained BERT network for representing sentences (SBERT), using siamese and triplet network structures, which have proven to be more efficient and better performing in several sentence similarity tasks.

As regards formal semantics, there have been several attempts in representing text as a structured meaning representation. Abstract Meaning Representation (Banarescu et al., 2013, AMR) is a formalism for representing the meaning of sentences into semantic graphs, which has been previously exploited in the context of the paraphrase detection task (Issa et al., 2018) and multilingual sentence representation (Blloshmi et al., 2020; Procopio et al., 2021). AMR aims at abstracting away from its surface form, therefore, sentences expressing the same meaning should be represented by the same or close structures. Based on these features, AMR appears to be suitable for representing paraphrase sentences.

In this paper we collect and evaluate the performance of several approaches to sentence encoding in a challenging real case scenario of plagiarism, thus providing an exhaustive comparison of NLP tools.

3. The Textual Material

3.1. History of the Books

The *Protocols* were drafted in Russia at the beginning of the Twentieth century (De Michelis, 2004). They describe an imaginary meeting between a group of senior members of the Jewish community – represented as if it had really happened – discussing the conquest and ruling of international society, governments, and

Score	Label	Description
5	Very high	The two sentences are identical, or almost identical.
4	High	The key words are kept untouched (or using very close synonyms) in the paraphrase.
3.5	Medium-high	A few words are kept untouched, but the key words may have been replaced with synonyms or periphrasis.
3	Medium	The idea is similar but the terms and formulation are not obviously reused.
2	Low	It can be understood that the essence of both sentences is related upon a somewhat detailed reading
1	Very low	Based on the context or a subjective understanding, the two sentences seem to be somewhat related.

Table 2: Description of similarity scores.

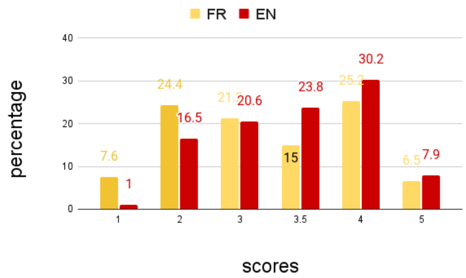

Figure 1: Distribution of similarity scores (%).

financial markets by the Jews – as a world domination project. The text was translated into the major European languages after the First World War and fed conspiracy theories about the Jews, especially within totalitarian regimes (particularly nazism), a period culminating in the Holocaust during the Second World War. Despite its apparent antisemitic contents and although it was unmasked as a forgery already in 1921, the circulation of the *Protocols* continued on a global scale in the second half of the Twentieth century, and it is still widespread today, in many languages, in print and online.

Numerous political and literary sources have been identified as models and subtexts to the *Protocols* (Cohn, 1967; Taguieff, 2004). The most relevant source is the treatise "Dialogue aux Enfers entre Machiavelli et Montesquieu" ("Dialogue in Hell betweeen Machiavelli and Montesquieu), published in Brussels in 1864 by Maurice Joly, which is a critique of the contemporary political misdeeds of Napoleon III. This work has no antisemitic content, but its critique of Napoleon's politics is turned into a handbook to the seizure of power and is represented as a conspiratorial Jewish project in the *Protocols*. Other indirect sources include the antisemitic booklet signed by Osman Bey, "La conquête du monde par les Juifs" ("The conquest of the world by Jews"), published in Bern in 1873, in French and in German, and the so-called "Rabbi's speech", a segment of the novel by Herman Goedsche, Biarritz (1868). *Protocols* remains, to this day, one of the major mediums of the global circulation of antisemitism and, more broadly, of conspiracy theories. For this reason, probing them as the fruit of plagiarism, and thus as a forgery, is especially relevant and one of the aims of this paper.

3.2. The Books in Numbers

Protocols is made up of 24 chapters, each introduced with a title and a summary. After discarding all summaries and titles, we collect the remaining 294 paragraphs, each of which comprising 3.5 sentences of 26.3 tokens, on average. Thus, *Protocols* consists of a total of 1031 sentences. Cohn (1967) and, more recently, Taguieff (2004) manually identified the pieces of text in *Protocols* that have been plagiarized from *Dialogue*. This collection of forged text covers 110 of the *Protocols* paragraphs, comprising 11731 tokens in total.

In our work, we collect and annotate the plagiarized texts of the books in French and English. As regards the French collection, similarly to Taguieff (2004), we used the 1921 edition of the *Protocols* (Anonymous, 1921) and the 1864 edition of *Dialogue* (Joly, 1864). Initially, a native French speaker manually selected the pieces of corresponding texts, starting from the *Protocols* sources, and judged the similarity between them using grades in a 1-5 range, following the scoring model in Table 2. Then, we asked three other native French speakers to annotate 50 randomly sampled sentence pairs from the resulting dataset each, with samples being different for every new annotator. This allows for computing the agreement among them and the first annotator using the Krippendorff's alpha (α) coefficient (Hayes and Krippendorff, 2007), which in turn ignores missing data entries, handles different sample sizes, and applies to any measurement type. Therefore, we believe it is an appropriate agreement measure for our case. Finally, the Krippendorff's alpha coefficient for the French dataset is 0.73. The same process has been applied for English, identifying the corresponding French sentences in the two respective translations of the books, i.e., Marsden (1934) for *Protocols* and Joly (2003) for *Dialogue*. Similarly to French, the annotator agreement for the English dataset is 0.71. Generally, values of α between 0.667 and 0.8 are considered acceptable.

In summary, the collected French text consists of 110 paragraphs comprising 353 sentence pairs, while the English one consists of 123 paragraphs consisting of 315 sentence pairs. In Table 3 we provide one example for each similarity score in our grading model (see Table 2). As seen from these examples, our chosen scoring range is necessary given the heterogeneity of paraphrases present in the collections, and to account for *quasi-paraphrases* (Bhagat and Hovy, 2013). Similarly to what was observed by Barrón-Cedeño et al. (2010) in different datasets, low similarity scores correspond to more abstract paraphrases in which a few concepts are maintained in the two sentences, while the highly scored pairs share the same meaning and also overlapping lexical units. The distributions of the dataset scores are presented in Figure 1.

4. Experimental Setup

4.1. Sentence Representations

We evaluate the performance of four different techniques to building sentence representations, based on:

S	L	*Protocols*	*Dialogue*
5	FR	Le résultat justifie les moyens.	La fin justifie les moyens.
	EN	The ends justify the means.	The result justifies the means.
4	FR	La liberté politique est une idée et non un fait.	La liberté politique n'est qu'une idée relative.
	EN	Political liberty is only a secondary idea.	Political freedom is an idea but not a fact.
3.5	FR	Il en est peu qui ne soient prêts à sacrifier les biens de tous pour atteindre leur propre bien.	Tous ou presque tous sont prêts à sacrifier les droits d'autrui à leurs intérêts.
	EN	and rare indeed are the men who would not be willing to sacrifice the welfare of all for the sake of securing their own welfare.	All, or nearly all, are ready to sacrifice another's rights to their own interests.
3	FR	Notre règne se signalera par un despotisme si majestueux qu'il sera en état, en tout temps et en tout lieu, de faire taire les chrétiens qui voudront nous faire de l'opposition et qui seront mécontents.	Dans un despotisme gigantesque, enfin, qui puisse frapper immédiatement, et à toute heure, tout ce qui résiste, tout ce qui se plaint.
	EN	These laws will withdraw one by one all the indulgences and liberties which have been permitted by the goyim, and our kingdom will be distinguished by a despotism of such magnificent proportions as to be at any moment and in every place in a position to wipe out any goyim who oppose us by deed or word.	It calls for a vast system of legislation that takes back bit by bit all the liberties that had been imprudently bestowed – in sum, a gigantic despotism that could strike immediately and at any time all who resist and complain.
2	FR	Pourquoi aurions-nous inventé et inspiré aux chrétiens toute cette politique, sans leur donner les moyens de la pénétrer, pourquoi, si ce n'est pour atteindre secrètement ce que notre race dispersée ne pouvait atteindre directement ?	À quoi servirait la politique, si l'on ne pouvait gagner, par des voies obliques, le but qui ne peut s'atteindre par la ligne droite ?
	EN	For what purpose then have we invented this whole policy and insinuated it into the minds of the goys without giving them any chance to examine its underlying meaning?	What's the use of political maneuvering if it can't attain the desired goal by devious ways, when straight ones are inadequate?
1	FR	L'idée de la liberté est irréalisable, parce que personne ne sait en user dans une juste mesure.	Sous certaines latitudes de l'Europe, il y a des peuples incapables de modération dans l'exercice de la liberté.
	EN	The idea of freedom is impossible of realization because no one knows how to use it with moderation.	In certain regions of Europe, there are people incapable of moderation in the exercise of liberty.

Table 3: Examples of *quasi*-paraphrases for each score included in the French and English text collections.

i) static word embeddings, ii) contextualized embeddings, iii) sentence embeddings, and iv) explicit semantic representations.

Static word embedding models. We use the pre-trained word embeddings by Mikolov et al. (2013, word2vec) obtained via the skip-gram algorithm (SG) and the ConceptNet Numberbatch (Speer et al., 2017a) embeddings (V. 17.06), for English. For French instead, we train the word2vec model via skip-gram on a corpus of 30.813 French books (Levis Sullam et al., 2021).
We create sentence representations as a combination of embeddings of words appearing in the sentence, by using the unweighted mean of their vectors, and also the weighted mean according to the TF-IDF and SIF weighting schemata. Throughout our experiments, we denote these combinations as $model_{\text{AVG}}$, $model_{\text{TF-IDF}}$ and $model_{\text{SIF}}$, respectively, where $model$ is a variable.

Contextualized word embedding models. We use BERT[2] (Devlin et al., 2019) for English and CamemBERT (Martin et al., 2019) for French, a model based on the RoBERTa (Liu et al., 2019b) architecture trained on French texts. We also employ two multilingual models: the multilingual version of BERT (BERT-M) and XLM-RoBERTa (Conneau et al., 2020, XLM-R) for encoding both English and French sentences. Similarly to when using static word embeddings, we build sentence representations from the contextualized word embeddings using the TF-IDF weighting schema, i.e., we first weight each token of a sentence using TF-IDF, and then, average all the word vectors.[3]

[2]All the models in this Section have been obtained using the Transformers library (Wolf et al., 2019).

[3]We noticed that this approach yields better results compared to other weighting schemata in preliminary experiments.

Sentence embedding models. To directly create latent sentence representations, we use the following pre-trained sentence encoders: i. LASER (Artetxe and Schwenk, 2019) multilingual model, trained on parallel corpora, for both English and French; ii. USE for English and its multilingual extension (mUSE) for both languages[4]; and iii. multilingual SBERT (Reimers and Gurevych, 2019) for both English and French.

Explicit Semantic Representation As we mention in Section 2.2, the ability to abstract away from the sentence makes AMR adequate for representing paraphrase sentences. We follow (Issa et al., 2018), and use a pre-trained AMR parsing model, i.e., AMREager (Damonte et al., 2016), to represent the English sentences as structured representations.

4.2. Tasks

In this Section we present a set of experiments on three different tasks. As a first task we evaluate the correlation between the paraphrase similarity scores calculated using different representation models, and that assigned by human annotators. We formulate the second task as a retrieval task, i.e., given an input sentence from the *Protocols* marked as a paraphrase by the annotators, find the corresponding sentence in the full text of the *Dialogue* book. The third task consists of using the full text of both books to find how many paraphrases a model is able to detect, based on various similarity thresholds. We use the cosine similarity among all the dense representations, while for AMR, we measure the similarity between sentence AMR graphs via the Smatch (Cai and Knight, 2013) metric.

[4]We use the models available through TensorFlow Hub.

	SG_{AVG}	SG_{SIF}	camemBERT$_{TF\text{-}IDF}$	BERT-M$_{TF\text{-}IDF}$	XLM-R$_{TF\text{-}IDF}$	LASER	mUSE	SBERT
ρ	0.30	0.55	0.44	0.38	0.28	0.57	**0.66**	0.59
r_s	0.33	0.54	0.45	0.40	0.19	0.58	**0.66**	0.60

Table 4: Pearson (ρ) and Spearman (r_s) correlation for each model on the French dataset.

	ConceptNet$_{SIF}$	SG_{SIF}	BERT$_{TF\text{-}IDF}$	XLM-R$_{TF\text{-}IDF}$	LASER	USE	mUSE	SBERT	AMR
ρ	0.54	0.51	0.35	0.26	0.41	**0.64**	0.56	0.62	0.50
r_s	0.52	0.48	0.38	0.27	0.49	**0.61**	0.54	0.59	0.44

Table 5: Pearson (ρ) and Spearman (r_s) correlation for each model on English texts.

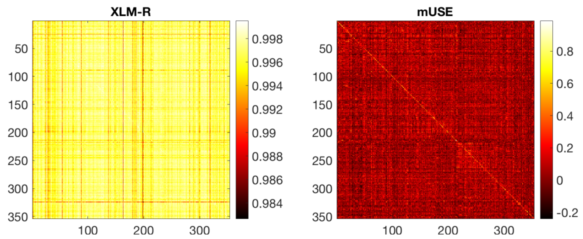

Figure 2: Heatmaps of the similarity matrices obtained using XLM-R (left) and mUSE (right).

4.2.1. Correlation Experiment

We compute the Pearson (ρ) and Spearman (r_s) correlation between the similarity scores given by human annotators and those we compute automatically via the above-listed representation models.

French. In Table 4, we report the correlation scores on the French dataset. The highest correlation scores are reached by the multilingual sentence encoders, i.e., LASER, mUSE, and SBERT, with mUSE performing best as it surpasses the other models by more than 0.06 points. Furthermore, we observe the remarkable performance of SG_{SIF}, which, despite its simplicity, can compete with more sophisticated systems such as LASER. This result suggests that the SIF weighting schema, originally developed and tested only in English, can also be used in other languages. Moreover, compared to SG_{AVG}, SG_{SIF} obtains a significantly higher performance, confirming the importance of weighting each word's contribution in the sentence. Surprisingly, the performance of the multilingual contextualized embedding models (BERT-M and XLM-R) is relatively low. This might be due to the tendency of these models to produce representations with high similarity. In fact, the starting (sub) token embeddings are anisotropic, occupying a narrow cone in the vector space (Ethayarajh, 2019). These aspects are particularly evident from the heatmaps of the similarity matrices between the sentence representations produced with XLM-R and mUSE, that we show in Figure 2. Evidently, higher similarity values appear mostly on the main diagonal of the mUSE matrix, while values in the XLM-R matrix are very close to 1. However, there is a noticeable performance gap between the multilingual contextualized models, i.e., BERT-M and XLM-R, and monolingual model, i.e., camemBERT. Indeed, the fact that it has been trained to model French text only, might be the reason to its significantly better performance than the multilingual models.

English. In Table 5 we show the correlation results in the English dataset. Similarly to French, the models that achieve higher performances are the sentence encoders, except for LASER, which is not among the top-ranked models. In fact, LASER degrades with more than 0.1 points in English when compared to French. The best performing model is the monolingual USE model, which outperforms even its multilingual version by more than 0.07 points. SBERT instead, has stable performances across languages and is the second best model across the board. The performance of contextualized embedding is low even in English, from which the monolingual BERT model obtains the highest results. Moreover, the static word embeddings aggregated using SIF obtain competitive results, especially the ConceptNet embeddings. Indeed, the latter encode also the word meanings derived from the ConceptNet semantic network (Speer et al., 2017b), which might motivate their competitve performance. Finally, representing sentences with AMR, achieves a performance that is in line with static embeddings weighted using SIF, despite the drastically different approach, i.e., distributional versus explicit semantics.

4.2.2. Retrieval Using an Input Sentence

The experiments in this Section consist in analyzing the ability of different sentence embedding models to identify paraphrases of specific target sentences. We use the paraphrased sentences in *Protocols* as the source and the *Dialogue* full text as target sentences and compute precision at *k* (P@k) with $k \in \{1, 5, 10\}$ to evaluate the performance of the models.

French. We present the retrieval performance of models in French in Figure 3. The patterns that emerge from these results are: i. no system is able to achieve a P@1 higher than 0.4 if we consider the performance on all the sentences, regardless of their gold annotation; ii. the best performing models are the sentence encoders, specifically LASER, mUSE, and SBERT; iii. only SBERT can detect all the paraphrases with similarity score 5 (starting from P@5); iv. the gap between precision computed on 5 relatedness scores and those computed on lower scores is large, suggesting that the models detect

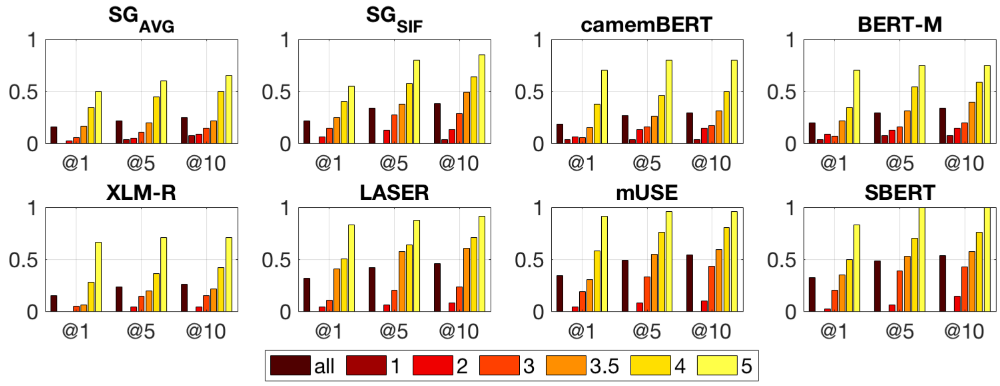

Figure 3: Retrieval precision on French computed at three different points (1, 5 and 10), on the whole dataset (all) and on single relatedness scores (1, 2, 3, 3.5, 4, 5).

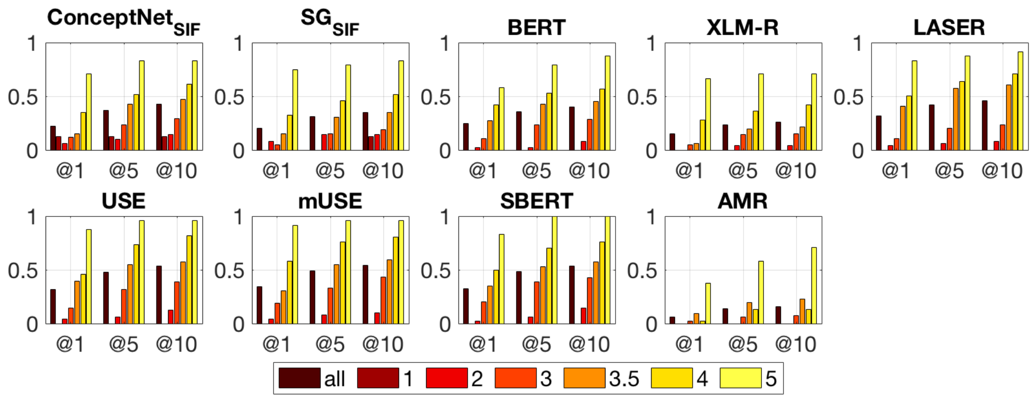

Figure 4: Retrieval precision on English computed at three different points (1, 5 and 10), on the entire dataset (all) and on single scores (1, 2, 3, 3.5, 4, 5).

well highly similar pairs, but fail when the rephrasing is more abstract; v. P@5 and P@10 do not change much, suggesting that some paraphrases are very difficult to be discovered, and even more sophisticated models would not be able to handle these cases in a real-world application; vi. only a few models (camemBERT and BERT-M) can detect paraphrases with score 1 when evaluated with P@1; vii. sentence encoders perform well on sentences with scores above 3, which might be because they have been trained on parallel sentences with high similarity; viii. sentence encoders do not retrieve paraphrases with similarity score 1, while contextualized word embedding models detect some of them.

The low performance of all the models, when evaluated for the pairs with low relatedness scores, suggests that it is challenging to find abstract similarities. For example, the sentences with gold similarity score of 1, *In certain regions of Europe, there are people incapable of moderation in the exercise of liberty* and *The idea of freedom is impossible of realization because no one knows how to use it with moderation*, are not detected even when evaluated with P@10.

English. In Figure 4 we show the results in English. The patterns that emerge from these results are the following: i. the best performing models are the sentence embedding models; ii. the multilingual sentence embedding models perform similarly to the monolingual sentence embedding model and, in fact, USE and mUSE achieve similar P@10 performances of 0.53 and 0.54, respectively. iii. only SBERT can detect all the paraphrases with score 5 (@5), similarly to the French case; iv. contextualized embedding models (BERT, XLM-R) perform poorly, which, as observed in the correlation experiment, tend to produce very similar sentence representations making the selection of candidates challenging; v. sentence embedding models are very good at finding paraphrases with high gold similarity scores, but struggle on low similarity scores, similarly to the French case; vi. only SG and ConceptNet can detect paraphrases with relatedness score 1, with ConceptNet detecting some of them even when evaluated with P@1.

Summary. Overall, we observe better performance of the models in English than in French, with the best model (SBERT) achieving an overall P@10 of 0.54 and 0.51, respectively. However, the search space for English is larger than for French, i.e., *Dialogue* contains 4684 sentences in English and 3348 in the French version. While this should make the task in English more

SG_{AVG}	SG_{SIF}	camemBERT	BERT-M	$XLM\text{-}R_{TF\text{-}IDF}$	LASER	mUSE	SBERT
0.55	0.54	0.55	0.59	0.55	**0.65**	0.63	0.61

Table 6: AUC measure for each model on the French dataset.

$ConceptNet_{SIF}$	SG_{SIF}	$BERT_{TF\text{-}IDF}$	$XLM\text{-}R_{TF\text{-}IDF}$	LASER	USE	mUSE	SBERT	AMR
0.62	0.59	0.61	0.55	0.61	0.63	**0.65**	0.63	0.52

Table 7: AUC measure for each model on the English dataset.

difficult, on the other hand, the English sentences are shorter than the French ones. Therefore, it might be easier to embed their content, motivating the better results in English.

4.2.3. Paraphrase Identification in the Full Texts

The final task consists of using the models to embed the full text of the two books and to evaluate how many paraphrases it is possible to find with each approach. This task does not consider the degree of similarity between the paraphrases. Since we do not have a development or training set to tune the threshold parameter above which two sentences are considered paraphrases, we used the Area Under the Receiver Operating Characteristics (AUC) curve, i.e., true-positive rate against the false-positive rate at various threshold settings. AUC is widely used in classification problems to evaluate the ability of a model to rank a random positive example higher than a random negative one, and evaluates the quality of predictions at different thresholds.

French. In Table 6 we show the results in French retrieval task. The best performing model is LASER with an AUC of 0.65. This means that LASER is more suited to discriminate among different similarity scores. It is followed by two other sentence embedding models, i.e., mUSE and SBERT, with slightly lower scores. Instead, contextualized and static word embedding models achieve significantly lower results, i.e., around 0.1 point lower than the sentence encoders, making them inappropriate for resolving this task.

English. Similar to the French case, the best performing models in English are the sentence encoders. Overall, results in Table 7 show that the models cannot achieve an AUC higher than 0.65. ConceptNet performs relatively well considering its simplicity, also confirming the observation in the correlation task (Section 4.2.1), where it is competitive with more complex sentence encoders.

Summary. In general, considering the relatively small search space, the models we analyze perform poorly in both languages. This is because the range of paraphrases is wide, with low scored pairs being significantly more abstract than high scored pairs. For this reason, it is difficult to find a *paraphrase* threshold which would includes paraphrases of different grades.

5. Conclusions

In this paper we presented an evaluation of sentence embedding models on a small but challenging setting for paraphrase detection, based on a real case of plagiarism.

We compared automatic systems for paraphrase detection with the historical analyses that have identified the Protocols as a forgery and the linguistic knowledge of the annotators that scored the sentence pairs. We conducted the evaluation on two languages, and we plan to extend it to German, Italian, and Spanish. Among other insights, we showed that current approaches are good at identifying paraphrases when the sentences are almost identical, and share common words. However, they struggle to detect paraphrases when periphrases are introduced, making the relationship among sentences more abstract, i.e., quasi-paraphrases. Indeed, even if the search space in which we searched for paraphrases is small, the analyzed systems did not achieve good performances. These observations suggest that it would be difficult to use the existing models in real-world scenarios, since they can mainly detect highly similar paraphrase sentences. Furthermore, as much as our presented evaluation represents a real task, it is rather simplified. This is because we already presented to the systems the two books in which the similarities actually exist. In a real case scenario instead, the search space could be broader including an extensive collection of heterogeneous texts to search from. Indeed, this urges the research community to develop more sophisticated tools to deal with pressing issues in modern societies.

Apart from the contribution of providing an evaluation of sentence representations in a real-world scenario, the more noticeable impact of this paper is to encourage the use of language technology in different fields with social impact. We aim to highlight the need to develop more efficient technologies to solve pressing issues, such as plagiarism detection or the spread of misinformation. As the diffusion of the *Protocols* has had tragic consequences in the past and remains deplorable today, the risk that similar cases emerge and propagate is very high and can be amplified by digital technologies. Finally, empirically verifying the “Protocols of the Elders of Zion” as a case of plagiarism and thus a forgery, remains relevant to fight against the spread of antisemitism, religious hatred, and conspiracy theories: together with the scientific experiments and the proposed technological solution, this has been the ethical aim of this paper.

6. Bibliographical References

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). 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. Association for Computational Linguistics.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). *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. Association for Computational Linguistics.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). 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. Association for Computational Linguistics.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). 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. Association for Computational Linguistics.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). 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. Association for Computational Linguistics.

Anonymous. (1921). *Protocols des Sages de Sion*. Bernard Grasset.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *LAW@ACL*.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., and Szpektor, I. (2006). The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Barrón-Cedeño, A., Potthast, M., Rosso, P., and Stein, B. (2010). Corpus and evaluation measures for automatic plagiarism detection. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Languages Resources Association (ELRA).

Barrón-Cedeño, A., Vila, M., Martí, M. A., and Rosso, P. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947.

Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D. (2009). The fifth pascal recognizing textual entailment challenge. In *TAC*.

Bhagat, R. and Hovy, E. (2013). Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Blloshmi, R., Tripodi, R., and Navigli, R. (2020). XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500, Online, November. Association for Computational Linguistics.

Bloomfield, L. (1933). Language. 1933. *New York: Holt*.

Cai, S. and Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In *ACL*.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (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, Vancouver, Canada, August. Association for Computational Linguistics.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Cohn, N. (1967). *Warrant for Genocide: The Myth of the Jewish World-conspiracy and the Protocols of the Elders of Zion*. Eyre & Spottiswoode London.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). 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. Association for Computational Linguistics.

Dagan, I., Glickman, O., and Magnini, B. (2005). The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Damonte, M., Cohen, S. B., and Satta, G. (2016). An incremental parser for abstract meaning representation. In *EACL*.

De Michelis, C. (2004). *The Non-Existent Manuscript. A Study of the" Protocols of the Sages of Zion"*. Nebraska University Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Dolan, B., Quirk, C., and Brockett, C. (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, Geneva, Switzerland, aug 23–aug 27. COLING.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November. Association for Computational Linguistics.

Frege, G. (2002). *Funktion, Begriff, Bedeutung*, volume 4. Vandenhoeck & Ruprecht.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080.

Issa, F., Damonte, M., Cohen, S. B., Yan, X., and Chang, Y. (2018). Abstract meaning representation for paraphrase detection. In *NAACL-HLT*.

Ji, Y. and Eisenstein, J. (2013). Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896, Seattle, Washington, USA, October. Association for Computational Linguistics.

Joly, M. (1864). *Dialogue aux enfers entre Machiavel et Montesquieu*. A. Mertens et fils.

Joly, M. (2003). *The Dialogue in Hell Between Machiavelli and Montesquieu: Humanitarian Despotism and the Conditions of Modern Tyranny*. Lexington Books.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October. Association for Computational Linguistics.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Levis Sullam, S., Minello, G., Tripodi, R., and Warglien, M. (2021). Representation of jews and anti-jewish bias in 19th century french public discourse: Distant and close reading. *Frontiers Big Data*, 4:723043.

Liu, X., He, P., Chen, W., and Gao, J. (2019a). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 216–223, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

Marsden, V. E. (1934). *The Protocols of the Meetings of the Learned Elders of Zion: With Preface and Explanatory Notes by Henry Ford and Others*. University Press of the Pacific.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2019). Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

Montague, R., (2019). *English as a formal language*, pages 94–121. De Gruyter Mouton.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain, July.

Procopio, L., Tripodi, R., and Navigli, R. (2021). SGL: Speaking the graph languages of semantic parsing via multilingual translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 325–337, Online, June. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (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, Austin, Texas, November. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011a). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 801–809, Red Hook, NY, USA. Curran Associates Inc.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011b). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 801–809.

Socher, R., Lin, C. C., Ng, A. Y., and Manning, C. D. (2011c). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 129–136.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Speer, R., Chin, J., and Havasi, C. (2017a). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Speer, R., Chin, J., and Havasi, C. (2017b). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Taguieff, P.-A. (2004). *Les Protocoles des sages de Sion: faux et usages d'un faux.* Fayard.

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207. ACM.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019a). Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537.*

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Seventh International Conference on Learning Representations.*

Warstadt, A., Singh, A., and Bowman, S. R. (2018). Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471.*

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Williams, A., Nangia, N., and Bowman, S. (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, New Orleans, Louisiana, June. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019). PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3685–3690, Hong Kong, China, November. Association for Computational Linguistics.

Yin, W., Schütze, H., Xiang, B., and Zhou, B. (2016). ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Zanzotto, F. M., Korkontzelos, I., Fallucchi, F., and Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1263–1271, Beijing, China, August. Coling 2010 Organizing Committee.

Zhang, Y., Baldridge, J., and He, L. (2019). PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zhu, X., Sobhani, P., and Guo, H. (2015). Long short-term memory over recursive structures. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1604–1612.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2018). A multilingual dataset for evaluating parallel sentence extraction from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).