Team Innovators at SemEval-2022 for Task 8: Multi-Task Training with Hyperpartisan and Semantic Relation for Multi-Lingual News Article Similarity

Nidhir Bhavsar*
Navrachana University
Vadodara, India
18103488@nuv.ac.in

Rishikesh Devanathan*
IIT Patna
Patna, India
rishikesh_2001cs85@iitp.ac.in

Aakash Bhatnagar*
Navrachana University
Vadodara, India
18124526@nuv.ac.in

Muskaan Singh, Petr Motlicek
Speech and Audio Processing Group,
IDIAP Research Institute, Switzerland
(msingh,petr.motlicek)@idiap.ch

Tirthankar Ghosal
ÚFAL, MFF
Charles University, Czech Republic
ghosal@ufal.mff.cuni.cz

Abstract

This work represents the system proposed by team Innovators for SemEval 2022 Task 8: Multilingual News Article Similarity (Chen et al., 2022). Similar multilingual news articles should match irrespective of the style of writing, the language of conveyance, and subjective decisions and biases induced by medium/outlet. The proposed architecture includes a machine translation system that translates multilingual news articles into English and presents a multitask learning model trained simultaneously on three distinct datasets. The system leverages the PageRank algorithm for Long-form text alignment. Multitask learning approach allows simultaneous training of multiple tasks while sharing the same encoder during training, facilitating knowledge transfer between tasks. Our best model is ranked 16 with a Pearson score of 0.733. We make our code accessible here[1]

1 Introduction

Over the last decade, English has been one of the most dominant languages on the internet. However, the number of non-English websites is rapidly increasing. From 2001-to 11, online use of English increased at a slower rate than that of Spanish, Chinese, etc. Approximately 4 billion people connect to the internet every day, but only half of them access web pages written in English (Pimienta, 2009). This creates a severe problem regarding verifying the integrity of the documentation because most systems in use are enhanced with English as the medium of information delivery.

SemEval has conducted similar tasks on sentence-level semantic textual similarity in the past. The SemEval 2017 Task 1 (Cer et al., 2017) dealt with finding the similarity between sentence pairs of both monolingual and cross-lingual nature. ECNU (Tian et al., 2017) was the best model that used feature engineering and deep averaging network (DAN) (Iyyer et al., 2015).

Furthermore, language-agnostic representation for sentence similarity described in Tiyajamorn et al. (2021) uses meaning embedding to estimate the cross-lingual sentence similarity without using human annotations. It improves the performance of any pre-trained multilingual sentence encoder, even in low-resource languages, with a few thousand parallel sentence pairs.

Inspired by Ham and Kim (2021), we explored the concept of semantically aligned multilingual sentence embedding. It covers the biases induced by monolingual similarity evaluation and multilingual sentence retrieval to generate language-aware embeddings. This method aligns semantic structures across different languages and uses a teacher network to distill the knowledge of pivot languages, thus achieving state-of-the-art STS 2017 multilingual corpora.

Our model is inspired by a multitask training approach using a BERT-based encoder. We use multiple subtasks to improve the overall accuracy score. Our model uses a machine translation model to cope with the articles' linguistic diversity. Additionally, since the average size of each article in the training data is close to 512 tokens, we use a text ranking algorithm that can capture the long-range sentence-level similarity between two documents. The next sections provide a more in-depth exposition.

[1] https://github.com/rdev12/Multilingual-News-Article-Similarity

* First three authors have equal contribution

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1163 - 1170

2 Task Description

The shared task emphasizes finding the similarity of multi-lingual news articles irrespective of the style of writing, political spin, tone, or any other more subjective "design decision" imposed by a medium/outlet. It gives participants access to a cross/multi-lingual dataset that spans over ten languages, including English, German, French, Italian, Polish, Russian, Chinese, Turkish, Arabic, and Spanish. The dataset includes an overall matching score between two different news articles. Additionally, the dataset consists of other dimensionality scores, such as Geo-location, Time, Shared Entities, and Shared Narratives (see table 1). These scores are based on a four-point scale ranging from the most to the least similar.

Table 1 indicates two examples: the first pair of articles shows extreme similarity with a Pearson score of 1.25, and the second pair shows non-similar articles with a Pearson score of 4. The second pair of articles is a cross-lingual pair where one article is in English and another in German. The next section outlines the overview of the model proposed by our team.

3 System Overview

Our system pipeline can be decomposed into four modules i.e., extraction, translation, text ranking, and multitask training.

3.1 Extraction

We extract title, descriptions, meta-description, and text from the JSON files obtained by scraping the news articles from the URLs given in the dataset. In most instances, the description and meta-description are the same, so we merge them by creating an additional field in the dataset called "extra text." The intuition behind this is to provide more context, as the title and descriptions tend to convey the overall message of the news article. This led to an increase in Pearson score by 0.07.

3.2 Machine Translation

Our translation module is based on the OPUS-MT (Tiedemann and Thottingal, 2020), a transformer-based neural machine translation model. This model uses Marian-NMT (Junczys-Dowmunt et al., 2018), a stable production-ready neural machine translation toolbox with efficient training and decoding capabilities. It is pre-trained on freely available parallel corpora collected in the large bitext repository OPUS (Tiedemann, 2012). The pre-trained version of the OPUS-MT model has six self-attentive layers in both the encoder and decoder networks and eight attention heads in each layer. Also, to handle the long-form nature of text articles, we use the chunking technique to segment texts into various chunks and then concatenate these derived chunks with other characteristics.

3.3 Text Ranking

There are many instances where the combined token length for a pair of articles exceeded the length of 512 tokens. Often, there are many irrelevant sentences with no semantic significance. So, these sentences are eliminated since they contribute much less to the overall context of the article. To achieve this, we adopt sentence-level noise filtering approach similar to Pang et al. (2021) & Mihalcea and Tarau (2004). In this, we first concatenate the pair of texts d_a and d_b obtained in the previous sections and split them into their component sentences s_i as shown in equations 1 & 2.

$$d_a = \{s_1^1, s_2^1, s_3^1, ..., s_n^1\} \tag{1}$$

$$d_b = \{s_1^2, s_2^2, s_3^2, ..., s_m^2\} \tag{2}$$

Then we concatenate d_a and d_b into S as shown in equation 3 later we derive representation matrix by taking the mean of the component word embeddings.

$$S = (s_1^1, ..., s_n^1, s_1^2, ..., s_m^2) \tag{3}$$

We use fastText embeddings (Bojanowski et al., 2017) which construct a better node similarity matrix than traditional methods and generate the embeddings by capturing transitive relationships and utilizing very sparse random projections. To generate the sentence similarity graph, we calculate the pairwise similarity of the sentence embeddings. The sentence similarity is defined as the same as TextRank (Page et al., 1999) to measure the overlapping word ratio between two sentences:

$$Sim\,(s_i, s_j) = \frac{|\{w_k \mid w_k \in s_i, w_k \in s_j\}|}{\log\,(|s_i|) + \log\,(|s_j|)}, \quad s_i, s_j \in \mathcal{S} \tag{4}$$

Then, we apply the Page rank algorithm (Page et al., 1998) to calculate the sentence importance score of each s_i in S and sort in decreasing order of importance. Finally, we extract the top λ sentences from d_a and d_b separately such that it is less than 512 tokens when combined. If the two articles

		Geo	Entities	Time	Narrative	Overall	Style	Tone
Pair 1	India approves third moon mission, months after landing failure (link) India targets the new moon mission in 2020 (link)	1	1.25	1	1.25	1.25	1	1
Pair 2	Hong Kong exam question on China and Japan sparked outrage (link) Staatsanwalt wirft Reeder "inszenierte Machenschaften" vor (link)	4	4	3	4	4	1	4

Table 1: Two examples from the SemEval dataset. Pair 1 shows extreme similarity and pair 2 non similairty

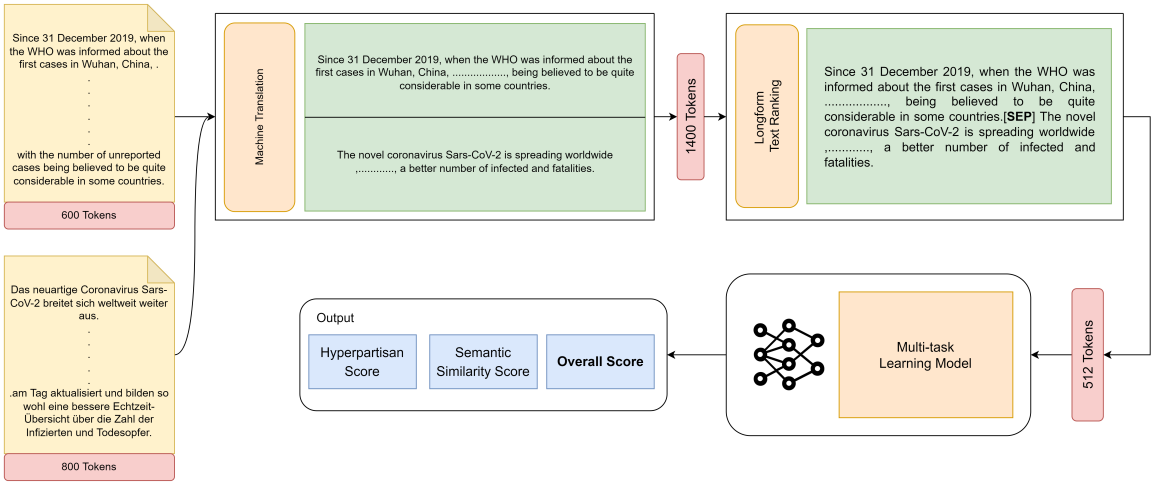

Figure 1: Main model pipeline

are similar, the sentences on the top of the ranked corpora reflect the same. As hypothesized, the title and *extra text* often appear on the top since they are rich in information.

3.4 Multi-task Training

Our multitask approach is based on the architecture proposed by Pruksachatkun et al. (2020). As shown in 2, the model consists of separate task-specific heads with their preprocessing methods but a shared encoder where the weights of all the tasks are updated simultaneously. This approach of introducing multiple subtasks supplements the main task. As evident from table 3, when more relevant tasks are added as task heads, the performance improves. In our best-performing model, the task head consists of an auxiliary semantic similarity task, a hyperpartisan identification task, and the main task. If required, the provided dataset and the hyperpartisan dataset are preprocessed by translation and text ranking modules. The sentences in the two ranked documents are combined, and a separator token separates the documents. This output is then fed into our multitasking model based on DeBERTa. The Loss during training was calculated using Mean Square Error (MSE) Loss function.

4 Experimental Setup

This section describes various hyper-parameters we use in data preprocessing and training. After scraping all the valid URLs, we could access 3651 pairs of articles for training, 408 for validation, and 4902 for a test. For training, we implement simple transformer models. We use DeBERTa (He et al., 2020) as the pre-trained language model with a batch size of 4. We found that the model performs better when the initial learning rate is 10^{-6} to 10^{-5}. We use the AdamW optimizer. We use Google Colab with a Tesla V100 GPU for various experiments. We apply various Python libraries like Pytorch, Transformers, Numpy, and Pandas to implement our multitask learning model.

4.1 Selecting Loss Function

We tried various approaches for selecting the loss function for our model. Initially, we experimented with many different loss functions such as Mean Square Error (MSE), weighted MSE, and dice loss. Furthermore, we experimented with a multi-objective weighted loss function as the data had multiple features. This loss is calculated as the weighted loss of L_E, L_N, and L_O, which are the

https://huggingface.co/microsoft/deberta-base

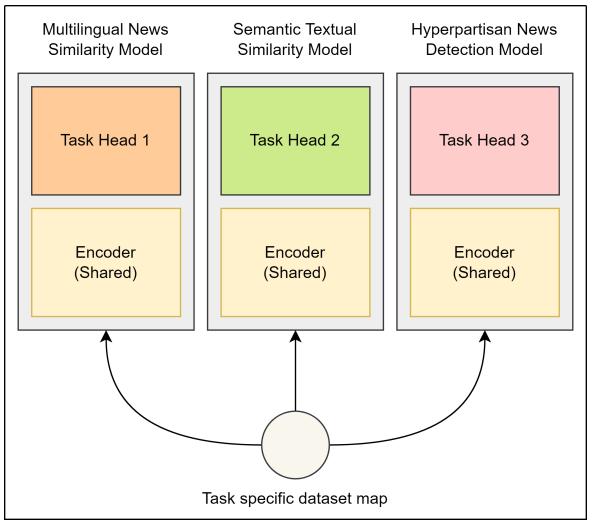

Figure 2: Multi-task Training Model

individual MSE Loss of entity, narrative, and overall similarity prediction. Multi-objective loss can be represented by the equation 5, where α, β, γ are the hyper-parameters decided while experimentation.

$$L = \alpha L_O + \beta L_E + \gamma L_N \tag{5}$$

5 Performance Analysis

In this section, we analyze the performance of several components of our system and compare different schemes for representing the document and the techniques used. Since the commencement of this task, Our team leaned toward using a multi-task learning strategy because it seemed to be the ideal fit for this problem. To do so, we ideated with several subtasks and experimented with our base encoder architecture. Table 3 shows all the different combinations of subtasks and base encoders used. Since the task prompts the participants to use multilingualism as a feature, our initial models utilized multilingual transformers such as XLM-Roberta (Conneau et al., 2019), and RemBERT (Chung et al., 2020), both of which are efficient with multiple languages. However, the primary issue we faced was the cross-linguality of the data; since the delivery style of information varies within every language, it becomes challenging to surface common information among pairs of text articles.

Thus, we emphasize more on translating the articles into English. We utilize a state-of-the-art publicly available model OPUS-MT. This translation module helped all articles to be represented in the same language. Doing so opens up the opportunity for us to choose from multiple encoders instead of our previous approach, where the number of encoders is limited.

After experimentation with various combinations of subtasks, as mentioned in table 3, we see that each of the subtasks has a unique attribute that can contribute to the overall performance of our model. After extensive experimentation, we chose two subtasks along with one main task.

5.1 Maintask - SemEval2022

As described earlier, our approach for generating the overall similarity is quite simple yet effective. The model uses an aggregate of title, text, and descriptions, which is then ranked based on the importance of each sentence in the article. The SemEval 2022 task-8 has various features to consider alongside the overall similarity score, rated between 1- and 4. During the task, the system pipeline supplies a pair of translated and ranked articles to the model, using the Deberta encoder to generate optimally aware document embedding. These embeddings are then supplemented with a set of linear layers to generate a final similarity score.

5.2 Subtask 1 - Hyperpartisan News Detection

As shown in Table 2 we used the Hyperpartisan News Detection dataset for detecting extreme sentences that may be biased towards a political group or a cause. The intuition behind this subtask was distinguishing news articles supporting extreme causes from more general articles. In sporadic cases, these two types of articles show similarities. This task also neglects political biases induced by media outlets and encourages us to treat each pair of articles impartially.

5.3 Subtask 2 - Semantic Textual Similarity

The semantic textual similarity closely relates to our main task since this involves finding the semantic proximity between pairs of texts. Because the STS-b dataset uses small sentence pairs with a max word limit of 40, it improved sentence-level similarity for the main task. The STS-b dataset covers a broad spectrum of sentences from news articles, image captions, and forums, which helps the model diversify across varied sentence-type situations.

5.4 Additional Models

We experimented with different combinations of Transformer based models and datasets. Table 4

Subtask	*Description*	*Dataset*
Semantic Textual Similarity	Determine how semantically similar two pieces of text are.	STS benchmark
Hyperpartisan detection	Given a news article, decide whether it follows a hyperpartisan argumentation, i.e., whether it exhibits blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person.	Hyperpartisan News Detection (Kiesel et al., 2019)
Stance detection	It involves estimating the relative perspective (or stance) of two pieces of text respective to a topic, claim or issue.	Fake News Challenge - 1 (Hanselowski et al., 2018)
Fake news inference detection	Fake news Detection using the Natural Language Inference. This entails categorizing a piece of text into categories such as "pants-on-fire", "false", "barely true", "half-true", "mostly true", and "true."	Fake news inference dataset
Language Inference	Determine whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise"	DocNLI Dataset (Yin et al., 2021)
Emotion Detection	Our Intuition behind this subtask was that if two articles are similar, they will exhibit same type of emotion. However, after experimentation we found this false as this subtask did not contribute towards the main task.	Go-Emotions Dataset (Demszky et al., 2020)
Paraphrase detection	Determine whether a particular sentence is a paraphrase of the original text.	Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005)

Table 2: Brief description of all different subtasks used for experimentation

Model subtasks	Model Type	Pearson Score (on validation set)
Stance detection	Translated text with RoBERTa	0.800
Stance detection	Untranslated text with XLM-RoBERTa	0.737
No Subtask (Multi-Objective Loss)	Weighted loss on Entity (0.15), Narrative (0.15) and Overall (0.7) Similarity (Sener and Koltun, 2018)	0.815
No Subtask (Multi-Objective Loss)	Weighted loss on Entity (0.2) and Overall (0.8) Similarity	0.811
Stance detection + Hyperpartisan detection	Translated text with RoBERTa	0.809
Hyperpartisan detection + Semantic Textual Similarity	**DeBERTa**	**0.835**

Table 3: Combination of all the subtasks and model types we used in experimentation. The results here are calculated on the validation set.

represents the score we achieved while developing the model. These scores are on the validation dataset that comprises 408 article pairs. Our model performed well during the developing stage. For instance, our best-performing model achieved a Pearson score of 0.835, and subsequently, we achieved a Pearson score greater than 0.8 in two other approaches. Table 3, however, shows the Pearson scores evaluated on the test dataset. We believe that the decline in performance was caused by the new languages (Chinese, Italian, Russian) introduced in the test dataset. These languages were not present in the training or validation set. This could be linked to the shift in writing style imposed on new languages, regardless of translation. According to our validation results, the model could not interpret the new languages and the unique cross-lingual pairs in the dataset.

6 Results

We use the Pearson score to evaluate SemEval 2022 task 8, which measures the linear relationship between the predicted and ground truth values. Like other correlation coefficients, the Pearson score varies between -1 and +1, with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. The dataset available in the evaluation phase of the task lacked features like Geography, Entity, etc., which were present in the training phase. Furthermore, the test dataset has an additional set of languages not present in the training dataset. This test dataset ensured that the model trained by the participants was an accurate multilingual model. Table 4 shows the models submitted by our team for the SemEval 2022 task-8; our best model achieves a Pearson score of 0.733, placing us amongst the top 15 of all participating teams.

As stated in the second row of table 3 we also experiment with a singleton multitask learning model separate from a machine translation system. The model, however, was unable to generalize adequately for each language in the dataset, and there appeared to be considerable sparsity between the predicted scores, particularly for low-resource languages like Turkish and Polish. Also, the lack of invariant multilingual data, specifically for each of the languages listed previously, was crucial in our decision to switch to a translation-based approach.

Next, table 5 showcases two instances where our model is successful and two instances where it fails to predict the correct value. We believe that this behavior is due to the machine translation module in our pipeline. Our model performs poorly for the following language pairs: de-en, ar-en, pl-en, and zh-en. This can be closely connected to the OPUS-MT model's BLEU score values, which are much lower for the language mentioned above pairs. This

Model Type	Model Subtasks	Pearson Score (on test set)
RoBERTa	Semantic Textual Similarity + Stance Detection	0.724
	Phrase Detection + Stance Detection	0.730
DeBERTa	**Semantic Textual Similarity + Hyperpartisan Detection**	**0.733**

Table 4: Models submitted by our team for the SemEval 2022 Task-8

Article 1	Article 2	Predicted Similarity Score	Actual Similarity Score	Language Pairs
Paraguay: Presidente promulga ley contra el “dinero sucio” en campañas (link)	Paraguay: Deputies approve the law on "dirty money" in political campaigns (link)	1.045	1	de-de
Conductores chocan por detenerse a ver accidente (link)	Paraguay: Deputies approve law on "dirty money" in political campaigns (link)	3.8	4	de-de
Schlag gegen den rechtsnationalen Flügel (link)	AfD-Rechtsaußen unter Druck (link)	1.02	3	es-es
Neue Debatte um Steuern: Millionen Arbeitnehmer zahlen Höchstsatz (link)	Norbert Walter-Borjans wants a higher top tax rate from 76,000 euros (link)	3.8	1	es-es

Table 5: First two rows of this table represents the instances where our model performed up to the mark and the last two rows represent the cases where our model failed to predict the right values

	Pearson Score
Ours	**0.733**
Average	0.624
Best in all teams	0.818

Table 6: Comparison of ours result with the best and the average pearson score

indicates that the model cannot comprehend these languages adequately, resulting in unclear results. Additionally, our system performs exceptionally well for data in French and Spanish, where the BLEU score values were 59.66 and 57.5, respectively. Furthermore, the results on en-en pairs are very accurate since they are not translated, thus retaining the writing style of the editor/outlet. Also, We encountered numerous challenges while scraping the data; not all websites were fully accessible, others only included photos or titles, and a handful swapped the URL with that of the main website's landing page. Next, the scrapper made numerous mistakes in discovering and assigning relevant tags to the articles' various subsections. A common erratum was combining description and title and attributing the title of the continuing piece to some random advertisement or supporting material.

7 Conclusion & Future Work

Our model performs well on English, Spanish, and French data while falling short on German. Even though the German data is the second largest, the biggest problem is that there is a lot of data to consider. The model could not correctly address each attribute of the article, resulting in underperformance. Some of the future work we anticipate to do which can increase the performance of our model are stated below:

1. Better Neural Machine Translation system which can effectively produce English sentences. We do suggest using AWS Translate since it makes more accurate predictions.

2. Using a finely trained multilingual model, enhanced explicitly for dealing with documents. DocMT5 (Lee et al., 2021) is one considered model; however, due to its public unavailability at the time of writing this paper, made us unable to use it in our approach.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.

Xi Chen, Ali Zeynali, Chico Q. Camargo, Fabian Flock, Devin Gaffney, Przemyslaw A. Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 Task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Jiyeon Ham and Eun-Sol Kim. 2021. Semantic alignment with calibrated similarity for multilingual sentence embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1781–1791, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar, and Melvin Johnson. 2021. Docmt5: Document-level pretraining of multilingual language models. *CoRR*, abs/2112.08709.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Larry Page, Sergey Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Liang Pang, Yanyan Lan, and Xueqi Cheng. 2021. Match-ignition: Plugging pagerank into transformer for long-form text matching. *CoRR*, abs/2101.06423.

Daniel Pimienta. 2009. Twelve years of measuring linguistic diversity in the internet: balance and perspectives.

Yada Pruksachatkun, Philip Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. *CoRR*, abs/2003.02249.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *CoRR*, abs/1810.04650.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 191–197, Vancouver, Canada. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of*

the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenpeng Yin, Dragomir R. Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. *CoRR*, abs/2106.09449.