Final project for ANLP 21/22
SemEval 2022, Task 8: Multilingual News Article Similarity: Exploration of simplicity-performance trade-offs

Tamara Atanasoska
atanasoska@uni-potsdam.de

Abstract

This paper is a final project report for "Task 8: Multilingual News Article Similarity", part of the SemEval 2022 competition. Given mono- and multi-lingual news article pairs, the task requires the participants to find a more factual similarity between them, grounded in the location, time and named entities. However, the best-performing solutions were all different versions of fine-tuned transformer-based language models. Motivated by the diversifying nature of media and news reporting, the desire to include as many languages as possible and include low-resource settings, this project explores the simplicity-performance trade-offs. The results show that by using a modular approach and available pre-trained models we can get quite close to the performance of the best performing fine-tuned models while offering a solution that is flexible, widely applicable, easy to run and set up and accessible in terms of resources.

1 Introduction

Semantic textual similarity is a well researched natural language processing(NLP) task. The are various approaches and strategies used for assessing semantic similarity between two texts (Chandrasekaran and Mago (2021)). One of the most popular approaches for this task is to use embeddings. The performance of such approaches is commonly very good for most use cases, and performs well for short text, but longer text forms are still a challenge (Rawte et al. (2020)). Although performing the task of semantic textual similarity with the help of embeddings can give us a good assessment weather two documents are similar overall, does it perform well when we are only interested in specific aspects aspects of the text?

This is one of the main challenges of the topic of this project's task, "Task 8: Assessing multilingual news similarity" (Chen et al. (2022)), which was part of the SemEval 2022 competition[1]. The authors of the task called for solutions that "should also not overly rely on general language models" because the similarity should ideally be based on factual points: "what happened, where and when it happened, who was involved, as well as why and how it happened". The annotators of the task, besides grading the overall similarity of the articles, also focused specifically on aspects like time, location, entities and narrative.

The vast majority of the participants approached the task by fine-tuning a transformer-based language model. Although various attempts were made to enhance the language model measurement by adding named entities, images, miscellaneous metadata, the best performing model is still an optimised XLM-RoBERTa based model. The conclusions from the authors of the task are that "systems that fine-tuned or otherwise trained embeddings generally performed better than those that did not" and that the best performing systems generally combined multilingual embeddings and translation" (Chen et al. (2022)). The authors end on the note that besides those points, further research is necessary that there is no clear consensus as to what approach is the best.

My motivation to explore the simplicity-performance trade-offs came from several aspects. In a time where the media landscape is going through continuous alterations in its attempt to reach younger and new audiences (Newman and for the Study of Journalism (2022)), the formats of news and event reporting are diverse and ever changing. Fine-tuning a model on one set of articles locked in specific time that is now inevitably a few years in the past, limits the generalisability and applicability to a wider reach of relevant texts. Further more, a fine-tuning approach greatly limits the languages that can be included. Successfully fine-tuning a model requires a good amount of quality

[1] https://competitions.codalab.org/competitions/33835

annotated data, which is a challenge for the languages with less resources. Machine translation offers an option to mitigate this issue, but a modular, simpler approach might offer even more possibilities even if not top performing. Lastly, training large models is resource intensive. This makes the use of use methods out of reach for people who don't have access to these types of resources, and it also has climate impact. The main motivation question that I asked myself is: what can we achieve just by using readily available pre-trained models by creating modular pipelines?

My contributions consist of 1) exploring summarisation as a means to generalise different forms of text, shorten text without losing crucial information and summarisation as a way to prioritise the more meaningful named entities, 2) exploring named entity recognition and disambiguation using embeddings, 3) comparing the performance between different types of embeddings, 4) exploring cosine similarity between different parts of the texts(summaries, titles, named entities) and their combinations for the best results compared to the original annotator scores. The code compromising this modular pipeline, as well as a notebook demonstrating the usage can be found in a GitHub repository[2].

I find that:

- using pre-trained sentence embeddings for the English-English pairs performs almost as good as the fine-tuned models;
- the summarised text as input performs comparatively well to the full-text scores;
- assessing the similarity using sentence embeddings on the titles alone already produces good results, which invites for researching more techniques for reducing the input data;
- the best performing combination is full summary text combined with the title articles;
- using all named entities or just parts of them containing the relevant categories like time and location do not perform well on their own, but improve the similarity results when the score they produce is combined with just title similarity scores in a weighted average calculation.

2 Related work

2.1 Task 8: : Multilingual News Article Similarity

The results from the competition showed that the highest performing approaches combined machine translation and fine-tined models/embeddings, such as the very best performing group Xu et al. (2022). They picked XLM-RoBERTa (XLM-R) (Conneau et al. (2020)) as their foundational model and deployed multiple linguistically inspired strategies on top.

Only one other group submitted a simpler approach to the competition, more in the lines of the approach described in this report. Wangsadirdja et al. (2022) created a system that "creates pair-wise cosine and arccosine sentence similarity matrices using multilingual sentence embeddings obtained from pre-trained SBERT and Universal Sentence Encoder models respectively". This approach takes advantage of the sentence embeddings by evaluating the similarity sentence-wise. When it comes to longer texts, such strategies are besides effective also necessary. Other popular methods when it comes to longer texts include dividing longer texts into smaller chunks/paragraphs and averaging the scores when comparing.

2.2 Parts of news articles with greater meaning

Not all parts of the text are equally meaningful when it comes to the information they hold and in this case, contributing to the similarity assessment. Starke et al. (2021) have asked 401 participants in their study which feature of an article contributes the most to their similarity judgment(images, textual parts etc.). They have found that users mostly indicated textual features like titles to be crucial for their assessment, with body text being the most important. Moreover, several articles have used news article titles to differentiate between real and fake news (Rai et al. (2022) and Horne and Adali (2017), as the title can hold crucial information, alone or in relation with the body text.

2.3 Summarisation

Summarisation is a task of extracting essential information from a text. Summarisation can be abstractive or extractive. Summarising news articles with transformer-based models is a well established practice (see: (Garg et al., 2021), S. et al. (2022), Yang et al. (2020) for examples of both extractive

[2] https://github.com/TamaraAtanasoska/SemEval-2022-Task-8-Multilingual-News-Article-Similarity

and abstractive methods). Besides the method of summarisation, the transformer-based models differ when it comes to architecture and the data used to train the model. Their differences make them better or worse for specific summarisation tasks regarding the length of the input articles and the desired size of the generated summaries. If the task requires summarising articles of various size, including potentially long articles it is important to carefully evaluate this dimension specifically ((Koh et al., 2022)).

2.4 Named entity recognition(NER) and named entity disambiguation(NED)

Named entity recognition is a task of searching and locating named entities in unstructured text. Named entity disambiguation, or also called named entity linking, is the task of assigning a unique identifier/identity to the named entities found in a text.

There are many strategies and techniques used for NER, although there is still space for improving the performance (Li et al. (2022), Wang et al. (2022)). Despite the methods used for a long time described in the review by Vychegzhanin and Kotelnikov (2019) specifically for their effectiveness in the news articles domain, in the recent years we have the rise of transformer-based models like the ones described in Vychegzhanin and Kotelnikov (2019). Often for the task of named entity recognition, especially in more niche domains, a custom approach using a fine-tuned model or custom knowledge-bases and/or dictionaries performs the best.

The same applies for the task of NED. However, interesting and promising results have been achieved just using word embeddings for the disambiguation (Almasian et al. (2019)), with other researchers exploring hybrid approaches that take advantage of both embeddings and knowledge bases (Shi et al. (2020)).

3 Task formalisation

The task formalisation calls for framing the problem formally, expressing it in a way that can be broken down to parts that can be handled by natural language processing methods. Below are the formalised problem, the tasks that constitute the solution and chosen methods for those tasks.

Problem: asses the similarity between two articles by prioritising the factual similarities like date, time, location and named entities.

Tasks and methods:

- Task: identify named entities;
  Method: NER;
- Task: disambiguate named entities:
  Method: NED;
- Task: find way to prioritise named entities that contribute the most to the similarity score;
  Method: summarisation;
- Task: compare similarity;
  Method: encode with word/sentence embeddings and compare with cosine similarity;
- Task: account for multiple languages;
  Method: machine translation or multilingual models *(only applicable in the multilingual version of the problem).*

4 Dataset

The data consists of mono- and multilingual news article pairs. For copyright reasons, the authors of the SemEval task were not able to provide the data, but instead provided files with links to the original articles. To facilitate the downloading process, they also provided a script[3], which first attempts to download the articles from the Internet Archive[4], and if they are not available, it attempts the original site of publication.

4.1 Provided training data

The distribution of languages and the percentage of monolingual pairs differs in the training and evaluation data. In the second batch of the training data which I used, there are 4918 pairs in total.

The language distribution of the training data look as follows: en-en: 1800, de-de: 857, de-en: 577, es-es: 570, tr-tr: 465, pl-pl: 349, ar-ar: 274, fr-fr: 72. Although these are the numbers representing the all possible pairs, due to articles quickly becoming unavailable, the number for the participants of the task varies.

The training data has been annotated for several categories. Each pair has 1-8 annotators, while most pairs have 1-3 annotators. The annotators were more than 20 students, compensated for their work. They recorded the similarity of the two news articles in the pair for 7 different categories, which

[3] https://github.com/euagendas/semeval_8_2022_ia_downloader

[4] https://archive.org/

I briefly outline below, summarising Chen et al. (2022).

- GEO: how similar is the geographical focus?
- ENT: how similar are the named entities, excluding the locations?
- TIME: are the articles based in or describing similar time periods?
- NAR: how similar is the narrative schema?
- STYLE: how similar is the writing style?
- TONE: how similar is the tone?
- Overall: overall similarity, excluding style and tone.

4.2 Provided evaluation data

The evaluation data contains less monolingual pairs, and has languages that are not featured in the training data. Besides the languages not previously seen, the evaluation data also offers more language combinations of pairs that have two different languages compared to the training data. The number of pairs in the evaluation data is 4902.

The distribution of languages in the data is as follows: ar-ar: 298, de-de: 608, de-en: 185, de-fr: 116, de-pl: 35, en-en: 236, es-es: 243, es-en: 496, es-it: 320, fr-fr: 111, fr-pl: 11, it-it: 411, pl-pl: 224, pl-en: 64, ru-ru: 287, tr-tr: 275, zh-zh: 769, zh-en: 213.

4.3 Dataset as used in this report

The task as proposed by the authors contained a sub-task that was focusing only on the en-en (English-English) pairs. It wasn't my intention to pick the sub-task, but because of the issues I faced with translation described in section 5.1, combined with the time limitations stemming from doing this project individually rather than in a group, I decided to treat my project as a proof of concept and work only on the monolingual English data.

Thus, the dataset I used in this project is as follows:

- Training data: 1783 English-English pairs.
- Evaluation data: 236 English-English pairs.

The reduction of the dataset was achieved by extracting just the pairs that both had the 'en' value in the *url1_lang* and *url2_lang* columns.

5 Experiments

5.1 Multilingual task

The task of multilingual data can be approached in two ways: working directly with multilingual embeddings/models, or using machine translation to translate the data to English to make use of the most wide range of tools available for natural language processing.

The chosen initial approach was to try and use multilingual NER models, like "bert-base-multilingual-cased-ner-hrl"[5], and multilingual summarisation models like "mt5-base-multilingual-summarization-multilarge-cs"[6]. After running the training data through the models, it was evident that while they performed well for English, the rest of the performance was ranging from acceptable, to non-existent for some of the languages found in the data.

The task of obtaining reliable translation was much more arduous than my initial assumption. I spent a considerable amount of time on trying to achieve this goal, before abandoning it in the interest of time. Below I discuss two of the main machine translation approaches I took, the code for both provided in the repository.

5.1.1 Machine translation

The initial exploration of the task revealed a lot of possibilities from python packages, limited APIs, and pre-trained models. I attempted to use of the best machine translation models, M2M100 Fan et al. (2021), available to translate the data. After translating all the titles, whose translation could be evaluated as good enough, the model spent several days translating the body text. While the model performs great on shorter text, when it comes to larger text it suffers from repetitive sentences, which is a known issue[7]. This occurrence was too repetitive to make the data usable. Example in the appendix A.

Using the Google Translate API[8] is a popular way to translate articles. However, the free tier, especially if used outside of the company's cloud platform, has many limitations. The efforts spanning several days to go around those limitations resulted with a blocked IP address by the service.

[5] https://huggingface.co/Davlan/bert-base-multilingual-cased-ner-hrl

[6] https://huggingface.co/ctu-aic/mt5-base-multilingual-summarization-multilarge-cs

[7] https://github.com/huggingface/transformers/issues/19276

[8] https://cloud.google.com/translate

5.2 Monolingual task - only English-English pairs

The monolingual sub-task of the main task consisted of assessing the news articles similarity only between English-English pairs. While the training data had 1783 such pairs, the evaluation data had only 236. Working just with monolingual data allowed me to have more certainty about the effects of each method used, as the language variability was removed.

Because this project proposes a modular, flexible pipeline with simple components, in the subsections below I will go through all the methods used in detail, briefly mentioning the their specific contribution. In section 6 where the results are presented I will elaborate a bit more about the particular combination of these methods that led to that particular similarity measurement.

5.2.1 Summarisation

The role of summarisation in the project is threefold: to unify the representation of the text, shorten long text without losing crucial information for easier analysis and as a means to prioritise the important named entities in longer text to improve the chances of tangential similarity not taking over the score.

The model used to generate the summaries is "mrm8488/t5-base-finetuned-summarize-news"[9]. The base of the model is the T5 transformer(Raffel et al. (2020)) that is pre-trained for the summarisation task. The model was then fine-tuned on a dataset[10] composed of 4515 articles. The pretraining data consists of author name, url, summary and complete article. The complete articles were collected in the time range from February to August 2017, from the publications: Hindu, Indian Times and The Guardian. The summaries were collected from Inshorts[11].

I configured the model to produce generation of maximum 250 tokens. This particular configuration was picked after trying several different options, from the default being 512 tokens, to larger numbers like 750 tokens. The 250 tokens summary seemed to do the best with the texts of variable length. An example in the Appendix B. Some postprocessing applied as the model added a string of characters at the end as a form of padding.

The summarised articles made for easier encoding and similarity evaluations without losing much of the information that contributes to that evaluation. More on this in section 6.

5.2.2 NER and NED

For the tasks of NER and NED, there are many choices. In order to keep the simplicity theme which allows for some performance sacrifices I made the choice to use the spaCy EntityRecognizer[12], specifically the package spacy-fishing[13] that takes advantage of the NER engine and performs disambiguation using Wikipedia concepts. Besides the unique identifiers, the library also adds a confidence score. Example in the appendix C.

The all spaCy EntityRecognizer available pretrained models use the OntoNotes Release 5.0 annotation scheme (Weischedel et al. (2022)). The scheme has 18 different entity labels. The relevant for the most interesting categories that the authors of the task hinted towards are: 'TIME', 'DATE', 'LOC' and 'GPE'. While 'TIME' and 'DATE' are self-explanatory, the entity 'GPE' stands for "geopolitical entities", differing from 'LOC' which stands for location, representing a physical place or area.

The categorisation of the named entities by assigning them a label and the disambiguation into Wikipedia concepts are not always correct. While the reported numbers for the accuracy of the NER system are quite high [14], it is worth mentioning that this is still a challenge.

The disambiguated concept identifiers at the end were not used. The two options I considered were creating new custom embeddings or utilising a knowledge base (Wang et al. (2019) as an example of such system). These efforts have a larger span that the time and focus of this project. Additionally, there is existing literature as referenced in section 2 that shows that contextual word/sentence embeddings aid named entity disambiguation (some more evidence found in: Gao et al. (2022), Jia et al. (2021)). Although the extended options would have been preferable, the connecting theme of the project is to test the simplicity-performance tradeoffs so utilising the embeddings for as much as possible aligns with it.

[9] https://huggingface.co/mrm8488/t5-base-finetuned-summarize-news

[10] https://www.kaggle.com/datasets/sunnysai12345/news-summary

[11] https://www.inshorts.com/

[12] https://spacy.io/api/entityrecognizer

[13] https://github.com/Lucaterre/spacyfishing

[14] https://spacy.io/usage/facts-figures

Similarity asssesment combination	Train data	Evaluation data
Summary body-text with SBERT	0.77	0.72
Summary titles with SBERT	0.74	0.74
Summary body-text and titles	**0.82**	**0.81**
NE with word embeddings	0.39	0.45
NE with SBERT	0.64	0.62
NE: only GPE, LOC, TIME, DATE	0.29	0.18
Only the rest of NE	0.56	0.55
Set NE: only GPE, LOC, TIME, DATE	0.28	0.1
A set of the rest of NE	0.56	0.55
Weighted average: 2*titles and summary NE	0.78	0.77

Table 1: A table that shows the modular pipeline combinations that led to the similarity metrics for both the train and the evaluation data. NE stands for "named entities" in the table.

5.2.3 Sentence and word embeddings

Utilising pre-trained embeddings is the main part of the project. For the experiments, both the titles, articles body texts and the various combinations of named entities were encoded with sentence and word embeddings.

The sentence embeddings were obtained through a pre-trained "sentence-transformers/paraphrase-MiniLM-L12-v2"[15] (Wang et al. (2020)) model. This model is smaller than the base model of the same type, but it is "5 times faster and offers good quality"[16]. I picked the paraphrasing option of the mini model because of the summarisation - this option searches for texts with identical or similar meaning. For the very condensed information in the summaries, that presumably reflects a factual, event based reflection of the news articles this would be a good fit.

The word embeddings appear only once in the experiments, and that when it comes encoding the named entities as a collection of words. For the word embeddings I used spaCy's word vectors[17], namely the large "en_core_web_lg"[18] model.

Contrasting between the word and sentence embedding was interesting for the experiment of comparing the similarity of the two articles based solely on the named entities, or any of their subsets. The named entities were joined in one big string with spaces between them. Although resembling one big sentence, these words did not necessarily belong together. Although the sentence embeddings performed really well, trying word embeddings for this particular case was interesting and relevant.

5.2.4 Cosine similarity

All the similarity scores were evaluated with cosine similarity.

For the sentence embeddings I used the implementation specified as suitable for the model used in the documentation[19].

For the word embeddings, I used the cosine similarity implementation found in the spaCy library[20] to keep the consistency.

6 Results

The overview of the results is presented in Table 1. The table represents all the different combinations of similarity measurements explored, for both the train and the evaluation data. Although I developed the ideas exclusively on the train data as I would do on a model and just executed the same process once on the evaluation data, the train data numbers

[15] https://www.sbert.net/examples/applications/paraphrase-mining/README.html#paraphrase-mining

[16] https://www.sbert.net/docs/pretrained_models.html#model-overview

[17] https://spacy.io/usage/spacy-101#vectors-similarity

[18] https://spacy.io/models/en#en_core_web_lg

[19] https://www.sbert.net/docs/package_reference/util.html #sentence_transformers.util.cos_sim

[20] https://spacy.io/api/token#similarity

are still relevant. The approach does not include any training nor the methods know anything about the training data specifically, so the numbers on this dataset many times the size of the evaluation dataset are still important for the analysis.

The numbers were obtained by calculating the Pearson correlation coefficient between the golden label of overall similarity in the evaluation and train data and the similarity measurements produced by the various combinations of the modular pipeline. This was the official method of comparison proposed by the authors of the competition, and how all solutions were evaluated.

The best performing method is adding the with sentence embeddings encoded tensors for summary body-text and titles. That method reached 0.81 Pearson correlation coefficient with the golden overall similarity labels, found in the evaluation data and 0.82 for the train data. Compared to the team Wangsadirdja et al. (2022) that did a similar simpler solution and achieved 0.85 on the English data only, and the best competition solution proposed by the team Xu et al. (2022) which achieved 0.87 on the same data, this solution is not far behind.

The winning competition solution requires advanced fine-tuning and involves a lot more information about the articles than just a title and a few sentences summarising the text with accuracy that does not trail far behind. It invites the question: *how much information suffices to make a good enough judgement about the similarity between two articles, and does involving more information benefits or hinders the results*? Another very interesting result that builds up on this question is the similarity achieved only by comparing the titles. For both the train and evaluation overall similarity golden labels, the Pearson correlation coefficient was 0.74. Compared to the combined accuracy of the summaries and titles, this is just 7% less and 8% less, with just one short sentence.

While combining the scores of both titles and summaries + all named entities doesn't have any a positive influence on the overall similarity score, if fact, it lowers it, combining the similarity based on just titles + all named entities produces an increased Pearson correlation coefficient to 0.77 for the evaluation and 0.78 for the train data. In extensive search of interesting combinations of scores, I performed many average and weighted average calculations that didn't produce results interesting enough to be featured in the table. However, combining the similarity scores from just the titles with the similarity scores of all the named entities extracted from the summaries, with a weight of 2 for the titles, produces a 3% increase for the evaluation and 4% increase for the train data. These results call for further attention. If better methods for named entities disambiguation are used that are based on more solid ground than just the similarities present in sentence embeddings, this combination might produce much better results with very little data required.

Using just the 'GPE', 'LOC', 'TIME', 'DATE' named entity labels is not enough information to produce meaningful similarity judgement using the method of encoding the named entities with sentence embeddings. Compared to the big jump to the accuracy of similarities based on all named entities and even just on the rest that are not these four, it is evident that the just the time, date and location do not suffice.

Interested in the question about the amount of information required to produce good enough judgement, I performed more experiments on the smallest bits of information I had in the form of the just the combined named entities with the labels 'GPE', 'LOC', 'TIME', 'DATE' and just the rest of the remaining named entities. In the longer articles, the named entities repeat a lot. If a set of the terms combined and encoded with sentence embeddings, from the table we can notice that there is a reduction in the accuracy of 1% the most.

Lastly, the word embeddings performed much worse than the sentence embeddings. Before each encoding, the named entities were extracted and then concatenated with spaces between in one long string. The word embeddings see each token as separate, so this particular representation did not influence the results. The sentence embeddings performed much better with a 0.62 Pearson correlation coefficient for the evaluation data, and 0.64 for the train data. In turn, the word embeddings had only 0.45 for the evaluation and 0.39 for the train data.

7 Conclusion

This project proposes a solution for the Task8: Multilingual news similarity entry for the SemEval 2022 competition. Exploring the simplicity-performance trade-offs when using a flexible, modular pipeline from just pre-trained sources, the project shows that the similarity assessment between two news articles does not lag far behind the winning complex and both time and resource intensive solutions.

Additionally, this project opens and attempts to answer some questions that need further research: how much information is enough to make a good enough similarity judgement?; does including more information in the the making of the decision improves or hinders it?; what are the key parts of the article that weigh the most into the similarity assessment?; do simpler solutions that do not impose as large environmental impact as retraining new models and are more accessible in terms of computational resources produce assessments that are on pair (or close) to the more complex solutions?; can we achieve language inclusion equity if we do not rely on large amounts of data to retrain models?

While all of these questions remain open, the results from the project show promising hints towards reducing the amount of data this type of factual similarity judgements are based upon, and making use of pre-trained models for it. In the case of language equity, the less/shorter text there is to translate, the more chances of inclusion for the low-resourced languages.

8 Limitations and ethical considerations

8.1 Limitations

The main limitation of the project is that it is focusing only on the English-English part of the data. If the same methods were used on texts that underwent machine translation, the results might differed greatly. The machine translation resources quality very for each language, and are lacking especially for the low-resourced languages.

Furthermore, the fields of NER and NED are still very much an open research question. While there is a some evidence as cited in several sections of this paper that sentence/word embeddings aid named entity disambiguation, it is not the best performing way to achieve it. Approaches using custom made dictionaries and knowledge bases, like the package used in the project called spacy-fishing, are a much better choice. The making of custom embedings or even better, a hybrid approach of combining the disasmbiguated concepts embeddings with the sentence/word embeddings are necessary to achieve high performance regarding a similarity assessment of two entities.

8.2 Ethical considerations

While there are not many ethical considerations with the data used in this project besides the evading of copyright by crawling the archive websites for the news articles, a lot of ethical considerations are woven into the solution.

The usage of freely available pre-trained models reduces the ecological imprint that training new models for very specific purposes leaves. Adjacent to that, if we develop solutions that require less computational resources and produce good enough results, we open up the possibilities for more people to use them for purposes we have maybe not foreseen.

The questions of language equity and inclusion are a big problem with the wider natural language processing field. While the resources for the languages backed by the most economically powerful sources continue to grow, the languages with smaller speaker bases or coming from countries with less resources can not participate in the newer technological advancement as much. By keeping the pipelines simple are relying on as little data as possible like in the case of this project, we might be able to include more languages and communities in the efforts by default.

References

Satya Almasian, Andreas Spitz, and Michael Gertz. 2019. Word embeddings for entity-annotated texts. In *European Conference on Information Retrieval*.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2).

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott 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)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Lei Gao, Lijuan Zhang, Lei Zhang, and Jie Huang. 2022. Rsvn: A roberta sentence vector normalization scheme for short texts to extract semantic information. *Applied Sciences*, 12(21).

Apar Garg, Saiteja Adusumilli, Shanmukha Yenneti, Tapas Badal, Deepak Garg, Vivek Pandey, Abhishek Nigam, Yashu Kant Gupta, Gyan Mittal, and Rahul Agarwal. 2021. News article summarization with pretrained transformer. In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10*, pages 203–211. Springer.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

Bingjing Jia, Zhongli Wu, Pengpeng Zhou, and Bin Wu. 2021. Entity linking based on sentence representation. *Complex.*, 2021:8895742:1–8895742:9.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8).

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Trans. on Knowl. and Data Eng.*, 34(1):50–70.

Nic Newman and Reuters Institute for the Study of Journalism. 2022. *Reuters Institute Digital News Report 2022*. Reuters Institute for the Study of Journalism, [Oxford] :.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. 2022. Fake news classification using transformer based enhanced lstm and bert. *International Journal of Cognitive Computing in Engineering*, 3.

Vipula Rawte, Aparna Gupta, and Mohammed J. Zaki. 2020. A comparative analysis of temporal long text similarity: Application to financial documents. In *MIDAS@PKDD/ECML*.

Harivignesh S., Avinash S., Avinash V., and R. Kingsy Grace. 2022. Summarization of news articles using transformers. In *2022 5th International Conference on Advances in Science and Technology (ICAST)*, pages 159–163.

Wei Shi, Siyuan Zhang, Zhiwei Zhang, Hong Cheng, and Jeffrey Xu Yu. 2020. Joint embedding in named entity linking on sentence level. *ArXiv*, abs/2002.04936.

A.D. Starke, Sebastian Øverhaug Larsen, and Christoph Trattner. 2021. Predicting feature-based similarity in the news domain using human judgments. In *Proceedings of the 9th International Workshop on News Recommendation and Analytics (INRA 2021) colocated with 15th ACM Conference on Recommender Systems (RecSys 2021)*, volume 3143 of *CEUR Workshop Proceedings*. Rheinisch-Westfaelische Technische Hochschule Aachen. 15th ACM Conference on Recommender Systems, RecSys 2021 ; Conference date: 27-09-2021 Through 01-10-2021.

Sergey Vychegzhanin and Evgeny Kotelnikov. 2019. Comparison of named entity recognition tools applied to news articles. In *2019 Ivannikov Ispras Open Conference (ISPRAS)*, pages 72–77.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juan-Zi Li, and Jian Tang. 2019. Kepler: A unified model for knowledge embedding and pretrained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. Nested named entity recognition: A survey. *ACM Trans. Knowl. Discov. Data*, 16(6).

Dirk Wangsadirdja, Felix Heinickel, Simon Trapp, Albin Zehe, Konstantin Kobs, and Andreas Hotho. 2022. WueDevils at SemEval-2022 task 8: Multilingual news article similarity via pair-wise sentence similarity matrices. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1235–1243, Seattle, United States. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Hovy Eduard, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2022. OntoNotes Release 5.0.

Zihang Xu, Ziqing Yang, Yiming Cui, and Zhigang Chen. 2022. HFL at SemEval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1114–1120, Seattle, United States. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2020. Ted: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings*.

A M2M100 repetitive sentence example

Two examples of the repetitive sentences produced by the M2M100 model while summarising articles from the train data.

4936,['The United Nations (UN) announced that more than 2 thousand domestic migrants were damaged by the flood in Aden the United Nations (UN) announced that more than 2 thousand domestic migrants were damaged by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe caused by the flood catastrophe.']

4937,"['For example, if you want to find out how to find out what to do.']"

B Summarised text example

An example of the summarisation. This is data point 14 in the evaluation data.

Original text:

INN turned to experts in the market to get a sense of what's in store for investors looking to plug into the gaming industry in the new year. Gaming and the esports industry — characterized by massive video game tournaments — saw unprecedented growth in 2019 as investors recognized opportunity. Now, heading into 2020, the gaming sector is poised to further shift into a lucrative market as increased interest supports the growth of players in the space. Here the Investing News Network (INN) turns to experts in the market to get a sense of what's in store for investors looking to plug into the esports industry in the new year. Why are Google and Apple are investing heavily in the gaming market? Read our FREE outlook report on the gaming market! Give me my free report! Gaming outlook 2020: A closer look at the numbers The esports industry is looking to capitalize on a successful 2019 moving into 2020 and beyond. In its 2019 global games market report, gaming research firm Newzoo states that the international gaming market could be worth US$196 billion by 2022. Mobile games in particular are projected to generate US$95.4 billion by 2022, driven primarily by smartphone activity. Cross-platform titles such as Mario Kart Tour on iOS and Android, a general increase in smartphone users and improvements to hardware and infrastructure are factors that will lead to the projected growth, according to the firm. The console segment, on the other hand, is set to reach US$61.1 billion in 2022, Newzoo estimates, boosted by the upcoming generation of Xbox and PlayStation systems, created by Microsoft (NASDAQ:MSFT) and Sony (NYSE:SNE,TSE:6758), respectively. PC gaming will be a bit slower in the coming years, according to the researchers, though it is gaining traction and could be valued at US$39.5 billion over the next two year period. In a previous interview with INN, Erik Dekker, senior vice president and portfolio manager with Dekker Hewett Group at Canaccord Genuity, said, "Gaming is bigger than film box offices, bigger than television, bigger than digital music. I think it's about a US$116 billion market at this time." Gaming outlook 2020: The growth of team organizations In esports, organizations like Team Liquid and OG are big generators of cash for the industry. Team Liquid, which has teams in many of the top esports leagues, currently has a merchandise collaboration with Marvel, and OG just announced a Counter Strike: Global Offensive team following its big Dota 2 win. Team support and attachment have been reflected in the sheer amount of interest in gaming tournaments, another important revenue stream for esport companies. In August, the Dota 2 The International 9 tournament had the world's largest prize pool in history at a whopping US$34.3 million in winnings, and OG took home the top prize. But organizations are evolving *to rely on more than just their athletes and tournaments — instead, they're taking a holistic approach to branding and maintaining a public image. Why are Google and Apple are investing heavily in the gaming market? Read our FREE outlook report on the gaming market! Give me my free report! "It's going to ultimately come down to teams that can create brand lifestyles, create great content programming and diversify the revenue streams to find more profitability," Daniel Mitre, CEO of New Wave Esports (CSE:NWES), told INN via email. Mitre said the impressive growth of the sector has already led heavy hitters in gaming to throw their hats into the ring. The executive explained that companies like Activision Blizzard (NASDAQ:ATVI) and Riot Games, a subsidiary of Chinese conglomerate Tencent (HKEX:0700), have carefully crafted plans attached to their esports franchises and leagues. "None of this movement around big games, AAA publishers and developers is coming to a slow," said Mitre. "In fact, it's just going to start picking up further." This sentiment was shared by another expert in gaming who is optimistic about the market's trajectory. Haywood Securities investment banker Sean MacGillis said the skill of the gaming industry is in the flexibility of its revenue channels, which include everything from merchandise and ticket sales to advertising and sponsorship revenue. "The ability to monetize games is an order of magnitude larger than the ability to monetize traditional forms of content," MacGillis explained. Gaming outlook 2020: Possible concerns in the new year Some experts have noted possible pitfalls for the industry, despite its recent levels of success. For Nick Mersch, associate portfolio manager at Purpose Investments, the biggest risks for the space could come with the hypothetical collapse of the Overwatch League, one of the biggest leagues in the sector. Mersch said a downturn for the league could turn off existing and prospective investors. "A lot of traditional sports owners wrote huge checks into teams in this league, and if they get burned here it will be difficult to attract capital back to the space," Mersch told INN. According to a report from Variety, viewers for the Overwatch League Grand Finals grew by 16 percent in 2019, with the event reaching 1.12 million average viewers. Mersch said he's also keeping his eye on the overvaluation that has plagued esports, pointing to a recent Forbes list ranking "the most valuable companies in esports" that still seems to exaggerate the values of some organizations on it, according to the expert. Why are Google and Apple are investing heavily in the gaming market? Read our FREE outlook report on the gaming market! Give me my free report! Other concerns have larger political implications. The ongoing US-China trade war dampened the growth of the sector, according to Trevor Doerksen, CEO of ePlay Digital (CSE:EPY), and could continue to do so into 2020, especially given China's place as one of the world's largest hubs for gaming. In 2019, it was projected that US$36.5 billion was spent on gaming in China, according to Newzoo. "Geopolitical forces and the US-China trade war muted global collaboration and access to markets, and important data that would help better understand and grow esports in 2019," said Doerksen. The executive said there could be better access to content, games and data across both markets if the relationship between the two countries improves in 2020. Gaming outlook 2020: Investor takeaway Mersch said moving forward, the most important factor for gaming will be the work done by data firm Nielsen, owned by The Nielsen Company (NYSE:NLSN), in its esports division, which works to get accurate information on engagement and viewership data. In his view, Nielsen's work could help close the existing gap between how the industry is growing and how it's monetizing that growth, Mersch told INN. "To close this gap, Fortune 500 companies need actionable return of investment metrics in order to justify more (spending) in the space. Nielsen and other data analytic companies can provide this." In terms of the overall health within the sector, Mitre told INN that the increasing popularity of gaming and its viewership will help in building a strong foundation for an interesting investment market. "We've always had this propensity to watch people play better than us, whether it's standing at the arcade and watching someone play Street Fighter, or the emergence of a Twitch and YouTube influencer space watching other people play video games," said Mitre. Don't forget to follow us INN_Technology for real-time news updates! Securities Disclosure: I, Danielle Edwards, hold no direct investment interest in any company mentioned in this article. Editorial Disclosure: The Investing News Network does not guarantee the accuracy or thoroughness of the information reported in the interviews it conducts. The opinions expressed in these interviews do not reflect the opinions of the Investing News Network and do not constitute investment advice. All readers are encouraged to perform their own due diligence.*

Summarisation:

Investing News Network has asked experts in the gaming industry what's in store for investors looking to invest in the esports industry in 2019. The global gaming market is estimated to be worth US$196 billion by 2022, while mobile games are projected to generate US$95.4 billion by 2022. The esports industry saw unprecedented growth in 2019 as investors recognized opportunity. The industry is characterized by massive video game tournaments. The esports industry is also expected to grow by 2022. Gaming is bigger than film box

C NER and NED with spacy-fishing

The named entity recognition and disambiguation results of the spacy-fishing library with the summary body text as input. This is datapoint 15 of the evaluation data.

Summary text:

Fort Lauderdale officials have provided a New Year's Day update on the sewage main issues plaguing residents in the area. The city has reported that two crews were working on the latest spills and are doing cleanup and remediation work on the damaged pipes. A total of six sewage main breaks have happened in Fort Lauderdale since the beginning of December. Restoration and remediation work is expected to continue this month.

A tuple of (named entities, labels, unique Wikipedia identifiers, confidence scores):

[('Fort Lauderdale', 'GPE', 'Q165972', 0.4622), ("New Year's Day", 'EVENT', 'Q196627', 0.4635), ('two', 'CARDINAL', None, None), ('six', 'CARDINAL', None, None), ('Fort Lauderdale', 'GPE', 'Q165972', 0.4622), ('the beginning of December', 'DATE', None, None), ('this month', 'DATE', None, None)]

D Reproducibility

All the code required for this project can be found in the main repository of the project[21]. This repository was started in the early winter of 2022, when the official guidelines did not require the usage of GitUp, so the project still lives there.

The many docstrings accompanying all the functions compromising the project API are extracted in comprehensive docs[22]. The functions have longer, expressive names describing their function, so it the case where they are are self-explanatory the docstrings are missing.

In a notebook[23], the full setup of the project including the usage of all the API points is demonstrated. The notebook follows the whole process of obtaining every result that is in the table featured in this paper for the evaluation data, but the same can be easily and exactly repeated for the train data just by replacing the data sources.

E My contributions

All the work in this project was done by me as the sole contributor of the project. The learning objectives that I expressed in the planning paper we not very relevant a year after. While I mention that I would like to get a chance to do another bigger NLP project from scratch and from start to finish, in the meantime I have finished several bigger projects of the sort, including my Individual Module and all Project Modules. The other objective noted there was to get more experience with academic writing. This still remains an area where I welcome every opportunity to practice.

[21] https://github.com/TamaraAtanasoska/SemEval-2022-Task-8-Multilingual-News-Article-Similarity

[22] https://tamaraatanasoska.github.io/SemEval-2022-Task-8-Multilingual-News-Article-Similarity/

[23] https://colab.research.google.com/drive/1k-2Pq858ADX6-A1j1oWpCIRPHs3zmj9K?usp=sharing

From the perspective of one year later, the biggest impact of this project was deep diving into the topics of named entity recognition and named entity disambiguation. I was not aware before attempting the project that many aspects of these tasks remain open questions and that the best solutions for performing them are still very resource intensive and require a custom base. Additionally, finding the similarities between a group of named entities poses another layer of complexity. I spent weeks reading papers and trying to come up with a simple solution that would fit the theme of the project and perform decent. News articles are a special challenge because they often talk about what is happening now, and that is ever changing, with equally ever changing actors in the spotlight. Big newspaper houses like "The Guardian" keep extensive tech teams and large knowledge bases, and the open sourcing of such material might be just what the field of named entity recognition for the domain of news articles needs.

There is far more work that went into this project that didn't produce results interesting enough to make it into this paper, but I am very grateful for the opportunity to spend the time on it. This was a very interesting project and I have learned a lot.