News Stream Clustering using Multilingual Language Models

Erik Novak
erik.novak@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

In this paper, we propose a news stream clustering algorithm which directly outputs cross-lingual event clusters. It uses multilingual language models to generate cross-lingual article representations which enable a direct comparison of articles in different languages. The algorithm is evaluated using a cross-lingual news article data set and compared against a strong baseline algorithm. The experiment results show the algorithm has great promise, but requires additional modifications for improving its performance.

KEYWORDS

online news, event detection, news events, multilingual language model

1 INTRODUCTION

Online news is producing hundreds of thousands of articles per day reporting about any significant event that happened in the world. The articles cover various domains (such as politics, sports, and culture) and are written in different languages. In order to automatically identify these events, news stream clustering algorithms are used. These usually have the following steps: (1) they group articles written in the same language into monolingual clusters, and (2) form cross-lingual clusters by linking monolingual clusters that report on the same event. Both steps usually employ monolingual text features such as TF-IDF vectors; these do not allow cross-lingual comparison without using advanced statistical or machine learning methods.

In this paper, we propose a news stream clustering algorithm that directly generates cross-lingual event clusters. The algorithm uses multilingual language models for generating cross-lingual content embeddings and extracting named entities found in the articles. These are used to measure if an article should be assigned to an event. The algorithm is evaluated using a cross-lingual data set consisting of articles in English, Spanish, and German, and is compared against a strong baseline. While the experiment results look promising, there is still room for improving the algorithms performance.

The paper is structured as follows: Section 2 contains an overview of the related work on cross-lingual news stream clustering and multilingual language models. Next, we present the proposed clustering algorithm in Section 3, and describe the experiment setting in Section 4. The experiment results are found in Section 5. Finally, we conclude the paper and provide ideas for future work in Section 6.

Information Society 2021, 4 - 8 October 2021, Ljubljana, Slovenia

2 RELATED WORK

News Stream Clustering. The objective of news stream clustering is to group news articles that report about the same event that happened in the world. Grouping can be a difficult task, especially if the articles are written in multiple languages. To this end, various approaches were developed for cross-lingual event clustering. A statistical approach called Generalization of Canonical Correlation Analysis is used to compare news articles in different languages [9]. Information extraction techniques, such as named entity recognition and part-of-speech tagging, are also used for event detection [6]. With the increasing popularity of neural networks, more advanced approaches are used to link event clusters. The work in [3] uses word embeddings to compare and link monolingual event clusters into cross-lingual ones. Transformer-based language models are used for event sentence coreference identification [4], a task that links parts of articles to multiple events. However, the algorithm is performed only on a monolingual data set.

To the best of our knowledge, our work is the first that uses multilingual language models for grouping articles directly into cross-lingual events.

Multilingual Language Models. Since the introduction of the transformers [11], language model development has gained traction in the research community. One of the most well known language models, BERT [2], has improved the performance of various NLP tasks. By training it using multilingual documents, the multilingual BERT [5] enabled solving tasks that require cross-lingual text representations. While these models improved the performance of various NLP tasks, they do not provide good document embeddings for tasks like clustering. This changed with the introduction of Sentence-BERT [8], which generates monolingual sentence embeddings appropriate for measuring sentence similarity. A year later, an approach for making monolingual document representations cross-lingual [7] opened a way for using sentence embeddings for cross-lingual clustering.

In this work, we employ the multilingual Sentence-BERT model to generate cross-lingual embeddings used to group articles into events.

3 THE CLUSTERING ALGORITHM

We propose a news stream clustering algorithm that directly outputs cross-lingual events. It uses cross-lingual embeddings, named entities, and temporal features to measure if an article should be assigned to an event cluster. If none of the events are appropriate, a new cluster is created and the article is assigned to it. Figure 1 shows the algorithm's workflow diagram.

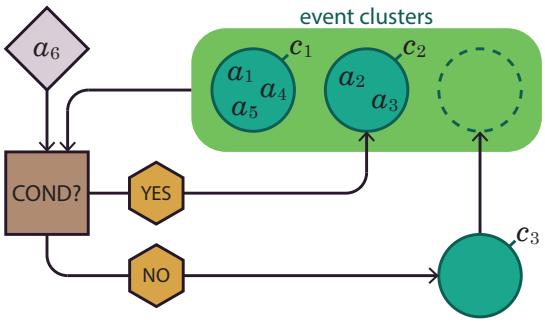

Figure 1: The algorithm's workflow diagram. The algorithm maintains a set of event clusters which are used when asessing if a new article (a_6) should be assigned to an existing event. If the conditions are met, the article is assigned to the most appropriate cluster (c_2). Otherwise, an empty event cluster is created (c_3), the article is assigned to it, and the newly created event is added to the cluster set.

In this section we describe how the algorithm represents the articles and events, and how it decides when to assign an article to the event cluster.

3.1 Article Representation

In this section we describe the different article representations used in the algorithm. Each article is assumed to have a title, body, and time attributes, which are used to (1) generate the content embedding and (2) extract its named entities.

Content Embedding. Each article is assigned an embedding that represents the article's content. Using multilingual Sentence-BERT[1], a language model designed for generating vectors used in cross-lingual clustering tasks, we get the content embedding by concatenating the article's title and body and inputing it into the language model. The output is a single 768 dimensional vector that captures the semantic meaning of the article.

Article Named Entities. For each article we extract the named entities that are mentioned in the article's body. To extract them, we developed a multilingual NER model using XLM-RoBERTa [1] and fine-tuned it using the CoNLL-2003 [10] data set.[2] Afterwards, we filter out the duplicates and store the remaining unique entities for later use.

3.2 Event Representations

An event is represented as an aggregate of its articles. This includes (1) the event centroid, (2) the named entities, and (3) the time statistics. In this section we describe how the aggregates are calculated and updated.

Event Centroid. The centroid represents the average content embedding of the articles assigned to the event. It is used to assess if an incoming article's content is similar enough to the event. Since the algorithm is intended to work on a news streams, we iteratively update the centroid with the newly assigned article's content embedding:

$$\vec{c_e}^{(0)} = \vec{0},$$
$$\vec{c_e}^{(k)} = \frac{(k-1)\cdot \vec{c_e}^{(k-1)} + \vec{c_{a_k}}}{k},$$

where $\vec{c_e}^{(k)}$ is the centroid calculated using the first k articles assigned to the event e, and $\vec{c_{a_k}}$ is the content embedding of the k-th article a_k.

Event Named Entities. Each event stores all of the unique named entities that are found in any of its articles. The named entities are used to identify if the incoming article mentions the event's entities. The event's named entities set is updated when a new article is assigned to the event:

$$r_e^{(0)} = \emptyset,$$
$$r_e^{(k)} = r_e^{(k-1)} \cup r_{a_k},$$

where $r_e^{(k)}$ is the set of named entities generated using the first k articles assigned to the event e, and r_{a_k} is the set of named entities of the k-th article a_k.

Time Statistics. The time statistics provide insights into the articles' temporal distribution. These are calculated using the articles' *time* attribute. In this experiment we measured the following statistics: the minimum, average, and maximum article timestamps. These are used to validate if an article was published at a time when it could still report about an existing event.

3.3 Assignment Condition

The most crucial part of the proposed algorithm is how to measure to which event should an article be assigned to, if any. We propose a condition that combines (1) the cosine similarity between the article's content embedding and the event's centroid, (2) the overlap between the article's and event's named entities, and (3) the time difference between the article's time and one of the event's time statistics.

Let $E = \{e_1, e_2, \ldots, e_j\}$ be the set of existing event clusters, where each event is represented with its centroid, named entities, and one of its time statistics $e_i = (\vec{c_{e_i}}, r_{e_i}, t_{e_i})$. Let the article be represented by its content embedding, named entities, and time attribute $a = (\vec{c_a}, r_a, t_a)$. We then check if the following conditions are met for each event:

$$\begin{aligned} \delta_c &= \frac{\langle \vec{c_{e_i}}, \vec{c_a} \rangle}{\|\vec{c_{e_i}}\|_2 \|\vec{c_a}\|_2} \geq \alpha, \\ \delta_r &= |r_{e_i} \cap r_a| \geq \beta, \\ \delta_t &= |t_{e_i} - t_a| \leq \tau, \end{aligned} \tag{1}$$

where α, β and τ are the thresholds corresponding to how similar the article's content must be to the event, the required amount of overlapping entities, and the time window in which an article has to be to be assigned to the event, respectively. Thus, $\delta_c, \delta_r, \delta_t$ correspond to the content similarity, entity overlap, and time window conditions, respectively.

If an event meets the conditions described in Equation 1, the article is assigned to it. If multiple events are appropriate, the article is assigned to the event that has the greatest δ_c value. If none are appropriate, a new empty event cluster is created, the article is assigned to it, and the event representations are updated.

To compare the impact of the conditions, we implement multiple versions of the algorithm that use a different combination

[1]The model is available at https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1.

[2]The code of the model is available at https://github.com/ErikNovak/named-entity-recognition.

of δ_c, δ_r, and δ_t conditions. Table 1 shows all of the algorithm versions compared in the experiment.

Table 1: The list of algorithm versions. Each algorithm uses a different combination of conditions.

Algorithm	condition combination
CONTENT	δ_c
CONTENT + NE	δ_c *and* δ_r
CONTENT + TS	δ_c *and* δ_t
CONTENT + NE + TS	δ_c *and* δ_r *and* δ_t

4 EXPERIMENTS

We now present the experiment setting. We introduce the data set and how it is prepared for the experiment. Next, we present the evaluation metrics. Finally, the baseline algorithm is described.

4.1 Data Set

To compare the algorithm performances we use the news article data sets acquired via Event Registry and prepared by [3] for the purposes of news stream clustering. These data sets are in three different languages (English, German, and Spanish), and consist of articles containing the following attributes:

- *Title.* The title of the article.
- *Text.* The body of the article.
- *Lang.* The language of the article.
- *Date.* The datetime when the article was published.
- *Event ID.* The ID of the event the article is associated with. It is used to measure the performance of the algorithms.

For the experiment, we merge the three data sets together to create a single cross-lingual news article data set. We extract their content embeddings and named entities, and sort them in chronological order, i.e. from oldest to newest. Table 2 shows the data set statistics.

Table 2: Data set statistics. For each language data set we denote the number of documents in the data set (# docs), the average length of the documents (avg. length), the number of event clusters (# clusters) and the average number of documents in the clusters (avg. size).

Language	# docs	avg. length	# clusters	avg. size
English	8,726	537	238	37
German	2,101	450	122	17
Spanish	2,177	401	149	15
Together	13,004	500	427	30

4.2 Evaluation Metrics

For the evaluation we use the same metrics as [3]. Let tp be the number of correctly clustered-together article pairs, let fp be the number of incorrectly clustered-together article pairs, and let fn be the number of incorrectly not-clustered-together article pairs. Then we report precision as $P = \frac{\text{tp}}{\text{tp+fp}}$, recall as $R = \frac{\text{tp}}{\text{tp+fn}}$, and the balanced F-score as $F_1 = 2 \cdot \frac{P \cdot R}{P+R}$. While precision describes how homogenous are clusters the, recall tells us the amount of articles that should be together but are actually found in different clusters.

4.3 Baseline Algorithm

The baseline algorithm used in the experiment is presented in [3]. It performs cross-lingual news stream clustering by first generating monolingual event clusters using TF-IDF subvectors of words, word lemmas and named entities of the articles. Afterwards, it merges monolingual into cross-lingual clusters using cross-lingual word embeddings to represent the articles. The algorithm compares two approaches when performing cross-lingual clustering:

- *Global parameter.* Using a global parameter for measuring distances between all language articles for cross-lingual clustering decisions.
- *Pivot parameter.* Using a pivot parameter, where the distances between every other language are only compared to English, and cross-lingual clustering decisions are made only based on this distance.

Since the baseline algorithm was already evaluated using the cross-lingual data set we are using the the experiment, we only report their performances from the paper.

5 RESULTS

In this section we present the experiment results. For all experiments we fix the values $\beta = 1$ and $\tau = 3$ days, and evaluate the algorithms using different values of α. In addition, all experiments use the event's minimum time statistic when validating the time condition δ_t.

Baseline Comparison. Table 3 shows the experiment results of the best performing algorithm on the evaluation data set. We report the best performing CONTENT + NE + TS algorithm which uses the content similarity threshold $\alpha = 0.3$.

Table 3: The algorithm performances. The best reported algorithm uses all three asssignment conditions.

Algorithm	F_1	P	R
Baseline (global)	72.7	89.8	61.0
Baseline (pivot)	84.0	83.0	85.0
CONTENT + NE + TS	72.2	79.7	66.0

While the proposed algorithm does not perform better than any of the baselines with respect to the F_1 score, our algorithm still shows promising results. Its performance is comparable to the baseline using the global parameter and also outperforms the baseline (global) recall by 5%, showing it is better at grouping articles.

Condition Analysis. We have analyzed the impact the conditions have on the algorithm's performance. For each algorithm version we run the experiments using different values of $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, and measure the balanced F-score, precision, and recall, as well as the number of clusters it generated. Table 4 shows the condition analysis results. By analysing the results we come to two conclusions:

Increasing α increases precision, decreases recall, and generates a larger number of clusters. When α is bigger, the content condition δ_c requires the articles to be more similar to the event. This condition is met when the article's content embedding is close to the event's centroid. Since this has to hold for all articles in the event, then the articles that have high similarity are clustered together, increasing the algorithm's precision.

Table 4: The condition analysis results. The bold values represent the best performances on the data set.

Algorithm	α	# clusters	F_1	P	R
CONTENT	0.3	46	29.6	19.7	59.8
	0.4	234	51.6	46.2	58.4
	0.5	849	57.7	67.7	50.3
	0.6	1762	45.3	73.1	32.8
	0.7	3185	26.0	81.9	15.5
CONTENT	0.3	279	43.7	33.3	63.8
+ NE	0.4	648	52.9	55.8	50.3
	0.5	1168	56.5	67.4	48.6
	0.6	1939	45.1	73.6	32.5
	0.7	3254	25.9	82.3	15.4
CONTENT	0.3	344	58.8	63.2	55.0
+ TS	0.4	806	64.1	76.5	55.2
	0.5	1346	58.8	83.4	45.4
	0.6	2068	47.1	81.7	33.1
	0.7	3356	25.2	**84.8**	14.7
CONTENT	0.3	925	**72.2**	79.7	**66.0**
+ NE	0.4	1221	**72.2**	80.5	65.5
+ TS	0.5	1554	54.0	81.9	40.2
	0.6	2174	46.7	80.7	32.9
	0.7	3403	25.0	**84.8**	14.7

However, if the α is to large then the condition is too strong, thus similar articles can be split into multiple clusters, consequently decreasing recall and increasing the number of clusters the algorithm generates.

Algorithms with more conditions can achieve better performance. The algorithm's performance is increasing with added conditions. While the worst performance is achieved when only the content condition δ_c is used (CONTENT algorithm), the best is reached when all three conditions are used (CONTENT + NE + TS algorithm). The most significant contribution is provided by the time condition δ_t which drastically improves the F_1 score.

6 CONCLUSION

We propose a news stream clustering algorithm that directly generates cross-lingual event clusters. It uses multilingual language models to generate cross-lingual article representations which are used to compare with and generate cross-lingual event clusters. The algorithm was evaluated on a news article data set and compared to a strong baseline. The experiment results look promising, but there is still room for improvement.

In the future, we intend to modify the assignment condition and learn the condition parameters instead of manually setting them. Modifying the language models to accept longer inputs could better capture the articles semantic meaning. In addition, events from different domains are reported with different rates. Learning these rates and including them in the algorithm could improve its performance.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the Humane AI Net European Unions Horizon 2020 project under grant agreement No 952026.

REFERENCES

[1] 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*, 8440–8451.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.

[3] Sebastião Miranda, Artūrs Znotiņš, Shay B Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium.

[4] Faik Kerem Örs, Süveyda Yeniterzi, and Reyyan Yeniterzi. 2020. Event clustering within news articles. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, 63–68.

[5] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4996–5001.

[6] Xiaoting Qu, Juan Yang, Bin Wu, and Haiming Xin. 2016. A news event detection algorithm based on key elements recognition. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*. (June 2016), 394–399.

[7] Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 4512–4525.

[8] Nils Reimers and Iryna Gurevych. 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)*. Association for Computational Linguistics, 3982–3992.

[9] Jan Rupnik, Andrej Muhic, Gregor Leban, Primoz Skraba, Blaz Fortuna, and Marko Grobelnik. 2016. News across languages - Cross-Lingual document similarity and event tracking. en. *J. Artif. Intell. Res.*, 55, (January 2016), 283–316.

[10] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.