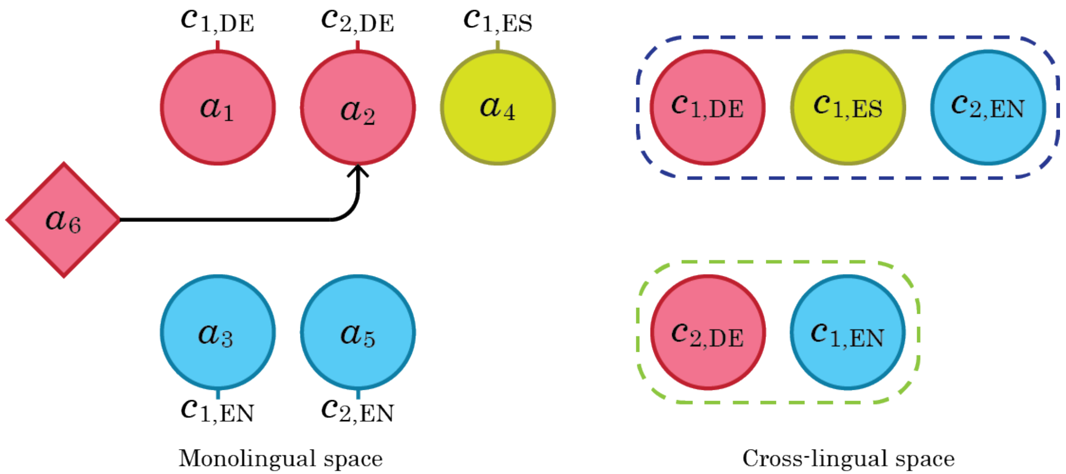
News Stream Clustering using Multilingual Language Models

Erik Novak
Jožef Stefan Institute
Jožef Stefan Institute Postgraduate School
Ljubljana, Slovenia

Motivation

- Online news is producing hundreds of thousands of articles per day
- News stream clustering algorithms are used to identify which news articles are about the same event
- The algorithms usually have two steps, both involving monolingual text features and advanced statistical or machine learning methods

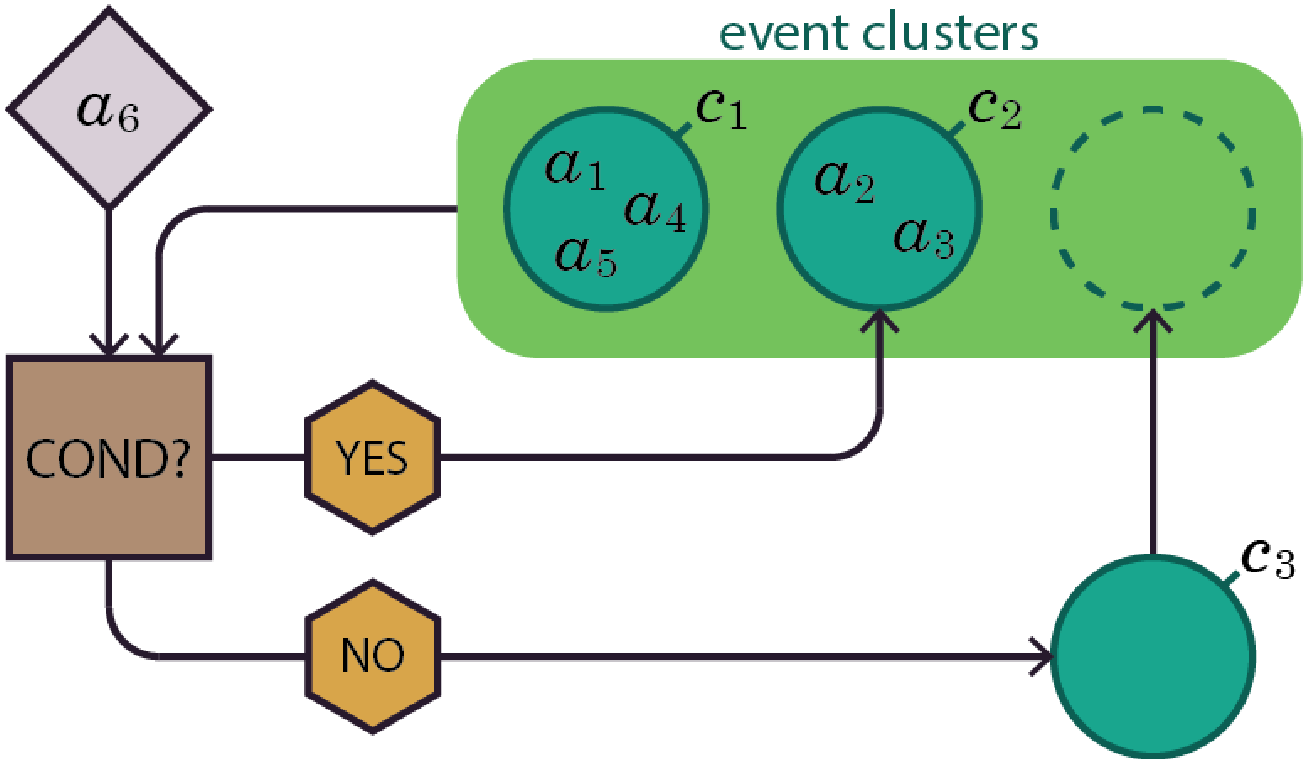
Is it possible to directly generate cross-lingual event clusters?



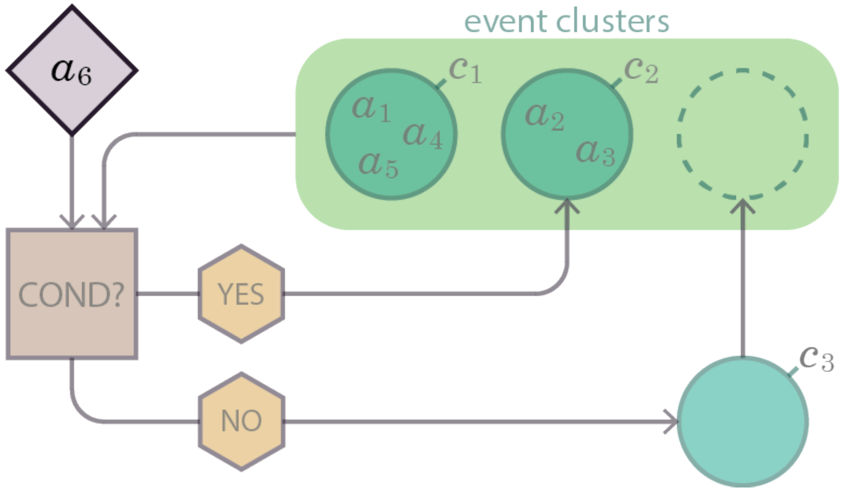
Outline

1. Clustering Algorithm
   a. Article Representation
   b. Event Representation
   c. Assignment Condition
2. Data Set
3. Results
   a. Evaluation
   b. Condition Analysis
4. Conclusion & Future Work

Clustering Algorithm



Article Representation
Clustering Algorithm



Each article is assumed to have a **title**, **body** and **time** attribute

Content Embedding

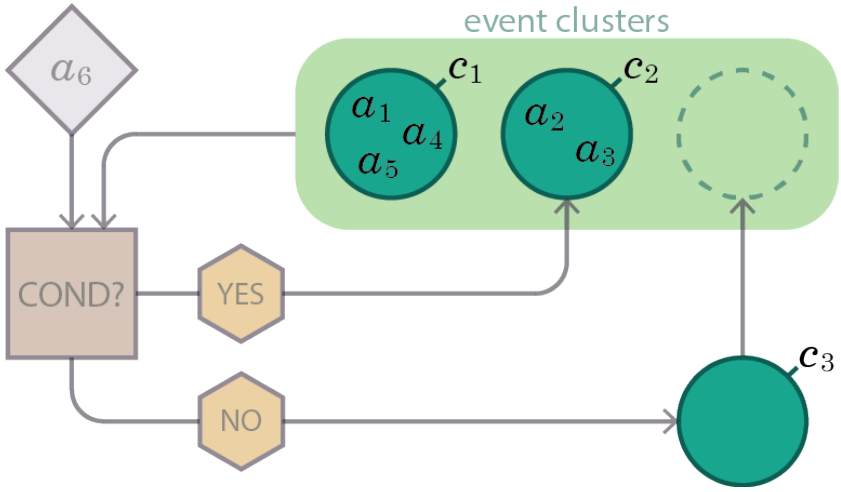
- Using Sentence-BERT; multilingual language models for generating vectors for cross-lingual clustering (INPUT LIMIT - 128 tokens)
- Articles title + body → vector representation

Article's Named Entities

- Extracted with a multilingual NER model using XLM-RoBERTa and fine-tuned on CoNLL-2003

Event Representation

Clustering Algorithm



Event representations are aggregates of its articles.
All representations are incrementally updated.

Event Centroid

- The average content embedding of the articles in the event

$$\vec{c_e}^{(0)} = \vec{0},$$

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

Event's Named Entities

- The set of all unique named entities that are found in the event's articles

$$r_e^{(0)} = \emptyset,$$

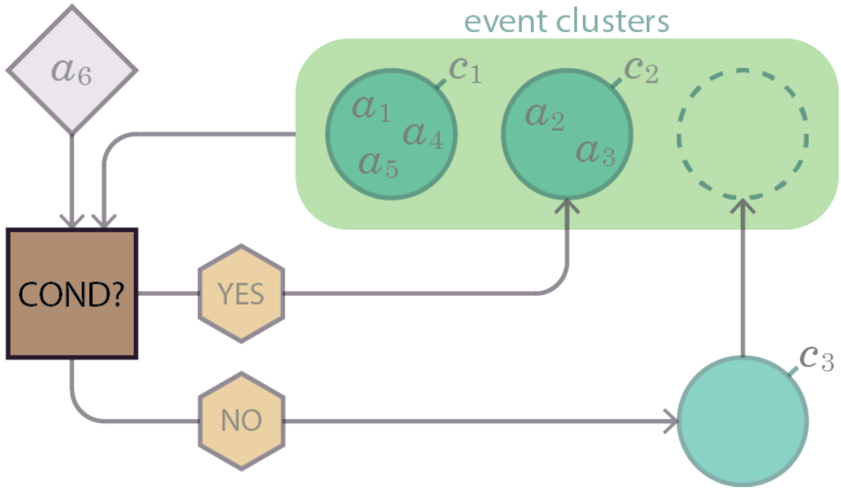
$$r_e^{(k)} = r_e^{(k-1)} \cup r_{a_k}$$

Time Statistics

- The minimum, average and maximum article's time articles

Assignment Condition

Clustering Algorithm



Assigning the article to an event based on

a. The content similarity between the article and the event

$$\delta_c = \frac{\langle \vec{c_{e_i}}, \vec{c_a} \rangle}{\|\vec{c_{e_i}}\|_2 \|\vec{c_a}\|_2} \geq \alpha$$

b. The overlap of the article's and event's named entities

$$\delta_r = |r_{e_i} \cap r_a| \geq \beta$$

c. Time difference between the article's time and the event's minimum time statistic

$$\delta_t = |t_{e_i} - t_a| \leq \tau$$

Using different combination of conditions to compare their impact on the algorithms performance

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

Data Set

- News article data set acquired via Event Registry and prepared for news stream clustering
- The data sets are in three different languages: English, German and Spanish
- Each article consists of its **title**, **body**, **language**, **date** of publish, and **event ID**.

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

Evaluation
Results

- Baseline model performs cross-lingual news stream clustering in two steps; uses **word embeddings** for merging monolingual event clusters into cross-lingual ones
  - Baseline (global). Using a global parameter for measuring distances between all language articles
  - Baseline (pivot). Using a pivot parameter, where the distances between every other language are only compared to English

- Fixed thresholds
  - content similarity ($\alpha = 0.3$)
  - entities overlap ($\beta = 1$)
  - time window ($\tau = 3$)

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

Condition Analysis

Results

- Evaluated how the content similarity condition effects the algorithms performance

- **Increasing α increases precision, decreases recall, and generates a larger number of clusters**

- **Algorithms with more conditions can achieve better performance**

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

Conclusion

- We propose a news stream clustering algorithm that generates cross-lingual event clusters
- Evaluated on a news article data set and compared to a strong baseline
- The algorithm results look promising, still room for improvement

It is possible to directly generate cross-lingual event clusters

Future Work

1. Modify the assignment conditions and learn its thresholds
2. Using language models that accept longer inputs
3. Learning the rates at which articles of a specific topic (sports, politics, etc.) are published and using them
4. Use a gold-standard data set for the evaluation