Scalable News Clustering and Topic Detection

Andrei Popescu-Belis[1], Nikolaos Pappas[1], Guntis Barzdins[2], Roberts Dargis[2], Arturs Znotins[2], Sebastião Miranda[3], Afonso Mendes[3], João Prieto[3], Shay B. Cohen[4]

[1]Idiap, [2]LETA, [3]Priberam, [4]University of Edinburgh

{apbelis,npappas}@idiap.ch, {guntis.barzdins,roberts.dargis,arturs.znotins}@leta.lv,
{sebastiao.miranda,amm,jup}@priberam.pt, scohen@inf.ed.ac.uk

Objectives and Achievements

Group articles into storylines with topic labels

★ Monolingual Storyline Clustering Using BoW and TF*IDF
- Baseline clustering with cosine similarity, on English

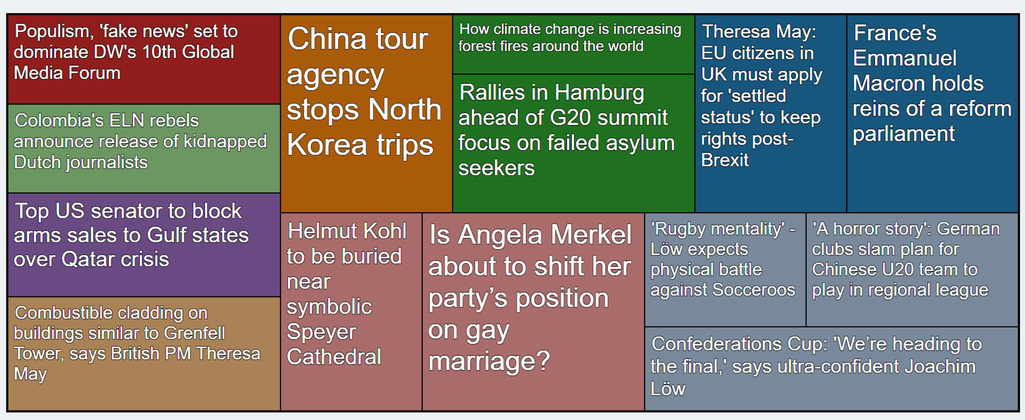
★ Multilingual Storyline Clustering Using Embeddings
- Groups stories from an incoming text stream into storylines

★ Deep Tagger for Topic Labeling
- Hierarchical neural network for multilingual document labelling with topics or keywords

Baseline Monolingual Storyline Clustering

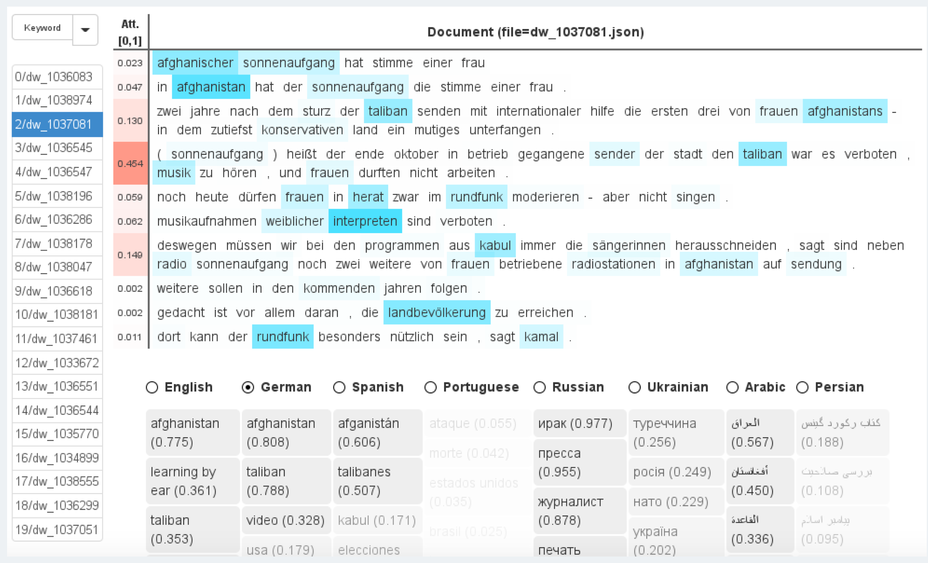
- Method
  - Bag-of-words model, top 100 keywords, using TF*IDF weighting
  - Cosine similarity above a threshold $\longrightarrow$ same cluster
- Evaluation method and data
  - True Positive (TP): articles about the same topic are in the same cluster
    True Negative (TN): articles about different topics are in different clusters
  - 9,664 articles in English, average length 534 words; 214 clusters
- Results
  - Precision $TP/(TP+FP)$: 92.4% | Recall $TP/(TP+FN)$: 86.2% | F1: 89.2%

Visualisation of Multilingual Clustering



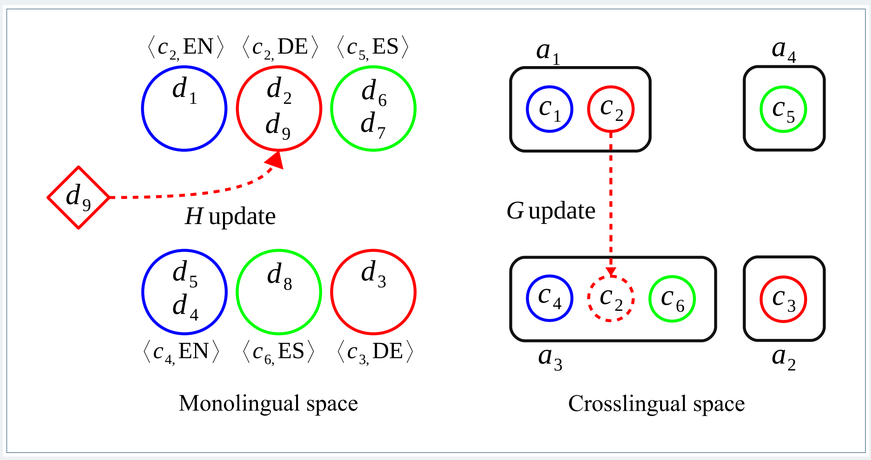
Each rectangle is a cluster of articles, represented by a story title

Visualisation of Multilingual Topic Labeling



Multilingual Storyline Clustering Using Embeddings

- Method based on crosslingual embeddings and timestamps
- Scalable production-ready online implementation
  - Experimented with English, German, Spanish and Portuguese
  - Demo: https://api.priberam.com/StorylineClustering/



Novel Interlaced Monolingual and Crosslingual Clustering

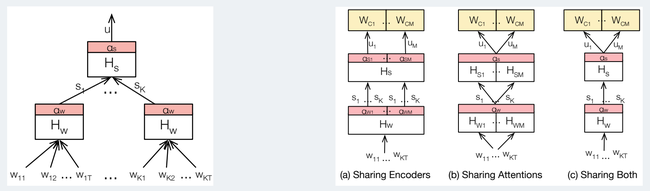
	Algorithm	F_1	P	R
English	CluStream (IBM)	79.0	98.6	65.9
	Tokens+Lemmas+Entities	**92.7**	92.9	92.5
	+Timestamps	82.8	96.5	72.4
German	CluStream (IBM)	88.2	99.8	79.05
	Tokens	88.2	99.9	79.0
	+Timestamps	**96.5**	99.9	93.4
Spanish	CluStream (IBM)	78.1	73.4	83.5
	Tokens+Lemmas+Entities	88.8	95.9	82.7
	+Timestamps	**94.2**	97.0	91.6

State-of-the-art results for all three languages

Deep Tagger for Topic Labeling

- **Goal:** identify topics assigned by journalists
- **Approach:** Multilingual Hierarchical Attention Networks
  - Learning shared structures across languages
- **Advantages** with respect to monolingual methods
  - Sub-linear parameter growth, cross-lingual knowledge transfer

Hierarchical Modeling + Multilingual Configurations



Notations. w_{ij}: word embeddings, H_w, H_s: encoder functions, a_w, a_s: attention mechanisms, $_w$,$_s$: word and sentence levels, W_{ci}: classifier for language i.

Results

Deutsche Welle corpus: 8 languages, 600k articles, 1240 general / 4397 specific labels

Word embeddings	$\|L\|$	$Y_{general}$		$Y_{specific}$	
Aligned	1	50K –	77.41 –	90K –	44.90 –
	2	40K ↓	78.30 ↑	80K ↓	45.72 ↑
	8	32K ↓	77.91 ↑	72K↓	45.82 ↑
Non-aligned	8	32K ↓	71.23 ↓	72K ↓	33.41 ↓

Parameters per language, average F_1 over 8 languages, and variation with the n. of languages $|L|$.

→ Multilingual models improve over strong monolingual ones

→ Sharing attention mechanisms is the optimal configuration

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n. 688139 (SUMMA).