

# Scalable News Clustering and Topic Detection

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

<sup>1</sup>Idiap, <sup>2</sup>LETA, <sup>3</sup>Priberam, <sup>4</sup>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 → 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

Populism, 'fake news' set to dominate DW's 10th Global Media Forum	China tour agency stops North Korea trips	How climate change is increasing forest fires around the world	Rallies in Hamburg ahead of G20 summit focus on failed asylum seekers	Theresa May: EU citizens in UK must apply for 'settled status' to keep rights post-Brexit	France's Emmanuel Macron holds reins of a reform parliament
Colombia's ELN rebels announce release of kidnapped Dutch journalists	Helmut Kohl to be buried near symbolic Speyer Cathedral	Is Angela Merkel about to shift her party's position on gay marriage?	'Rugby mentality' - Low expects physical battle against Soccerroos	'A horror story': German clubs slam plan for Chinese U20 team to play in regional league	
Top US senator to block arms sales to Gulf states over Qatar crisis					
Combustible cladding on buildings similar to Grenfell Tower, says British PM Theresa May					

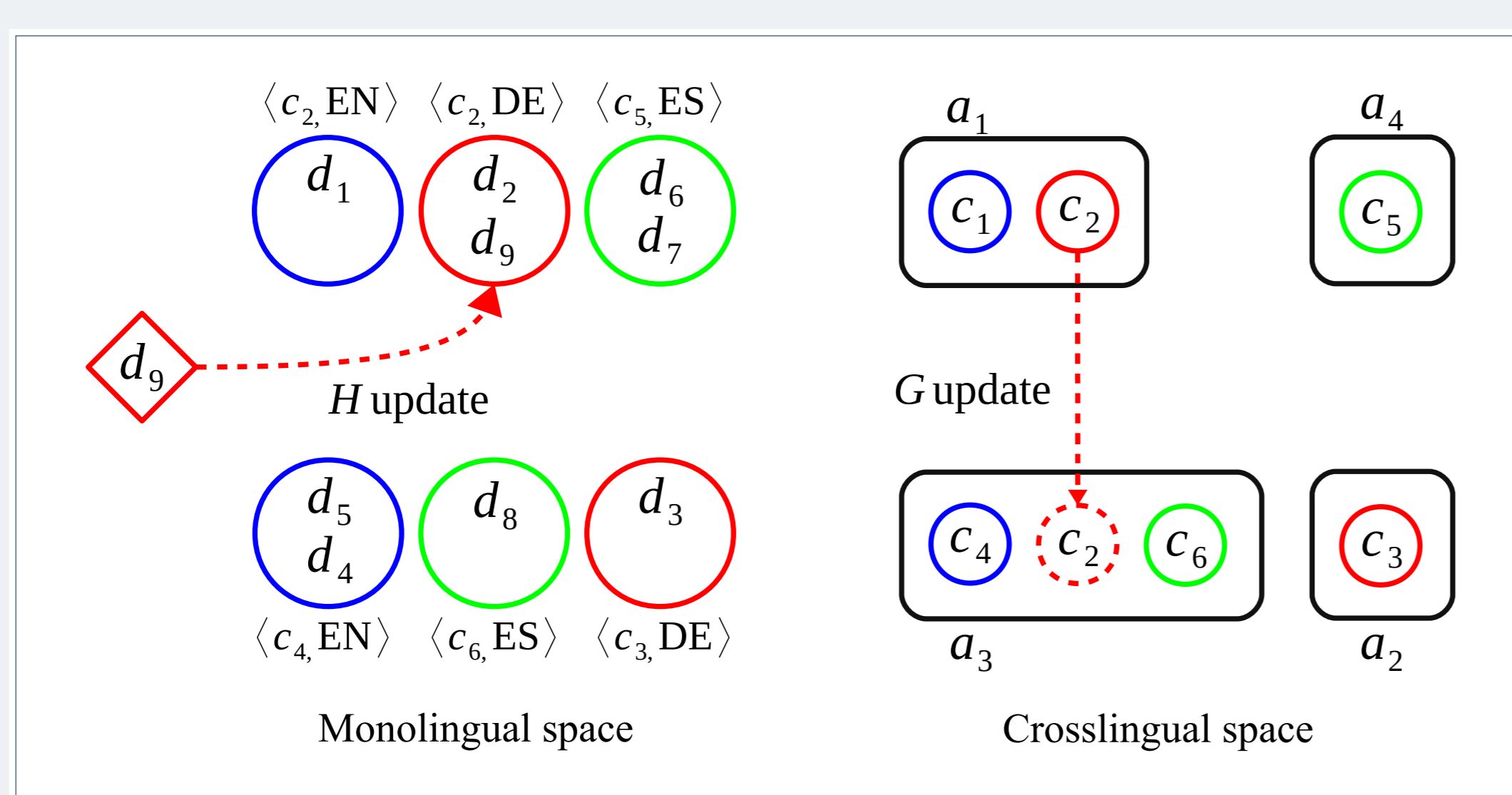
Each rectangle is a cluster of articles, represented by a story title

## Visualisation of Multilingual Topic Labeling

Keyword	Att. [0,1]	Document (file=dw_1037081.json)
afghanischer sonnenaufgang	0.023	afghanischer sonnenaufgang hat stimme einer frau
in afghanistan	0.047	in afghanistan hat der sonnenaufgang die stimme einer frau
zwei jahre nach dem sturz der taliban	0.130	zwei jahre nach dem sturz der taliban senden mit internationaler hilfe die ersten drei von frauen afghanistans in dem zuletzt konservativen land ein mutiges unterfangen
(sonnenaufgang) heißt der ende oktober in betrieb gegangene sender der stadt den taliban	0.054	(sonnenaufgang) heißt der ende oktober in betrieb gegangene sender der stadt den taliban war es verboten musik zu hören, und frauen durften nicht arbeiten
noch heute dürfen frauen in herat zwar im rundfunk moderieren - aber nicht singen	0.059	noch heute dürfen frauen in herat zwar im rundfunk moderieren - aber nicht singen
musikaufnahmen weiblicher interpreten sind verboten	0.062	musikaufnahmen weiblicher interpreten sind verboten
deswegen müssen wir bei den programmen aus kabul immer die sängerinnen herausschneiden, sagt sind neben radio sonnenaufgang noch zwei weitere von frauen betriebene radiostationen in afghanistan auf sendung	0.140	deswegen müssen wir bei den programmen aus kabul immer die sängerinnen herausschneiden, sagt sind neben radio sonnenaufgang noch zwei weitere von frauen betriebene radiostationen in afghanistan auf sendung
weitere sollen in den kommenden jahren folgen	0.002	weitere sollen in den kommenden jahren folgen
gedacht ist vor allem daran, die landbevölkerung zu erreichen	0.002	gedacht ist vor allem daran, die landbevölkerung zu erreichen
dort kann der rundfunk besonders nützlich sein, sagt kamal	0.011	dort kann der rundfunk besonders nützlich sein, sagt kamal

## 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

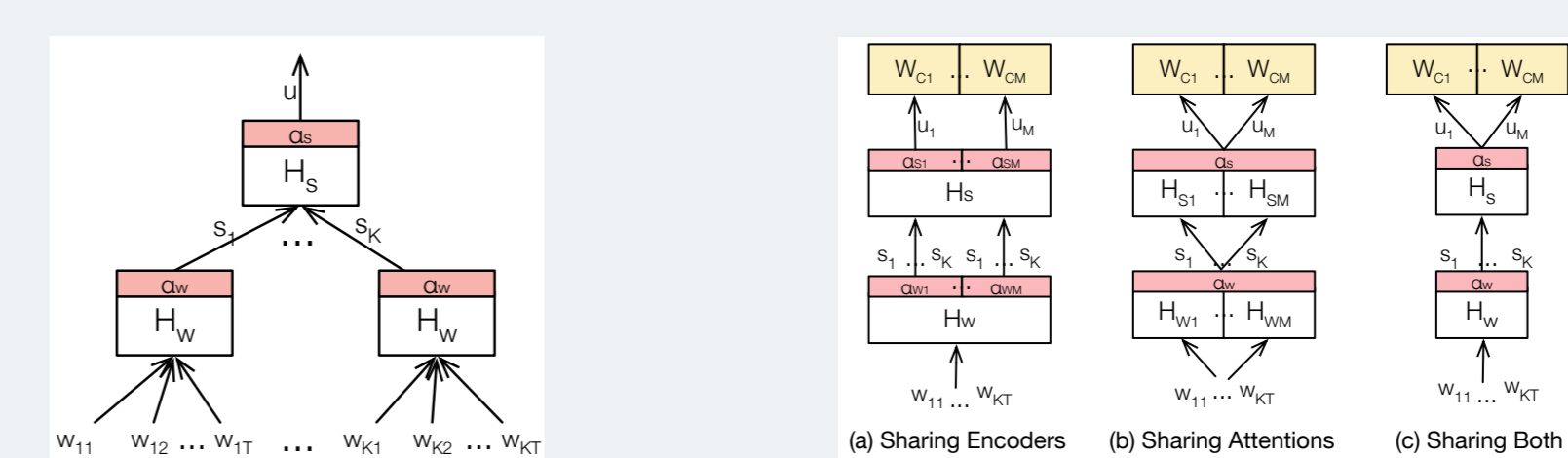
	Algorithm	F <sub>1</sub>	P	R
English	CluStream (IBM)	79.0	98.6	65.9
	TOKENS+LEMMAS+ENTITIES	<b>92.7</b>	92.9	92.5
	+TIMESTAMPS	82.8	96.5	72.4
German	CluStream (IBM)	88.2	99.8	79.05
	TOKENS	88.2	99.9	79.0
	+TIMESTAMPS	<b>96.5</b>	99.9	93.4
Spanish	CluStream (IBM)	78.1	73.4	83.5
	TOKENS+LEMMAS+ENTITIES	88.8	95.9	82.7
	+TIMESTAMPS	<b>94.2</b>	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,  $w_{w,s}$ : 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