Multilingual Topic Detection in News Data

Nikolaos Pappas[♡] **João Lages**[♢] **Sebastião Miranda**[♢]

nikolaos.pappas@idiap.ch, joao.lages@priberam.com, sebastiao.miranda@priberam.com

Hierarchical Multi-label IPTC classification: Goals

- Develop models to predict IPTC Subject NewsCodes in news text
- Explore multilingual approaches while only having access to labelled data in portuguese

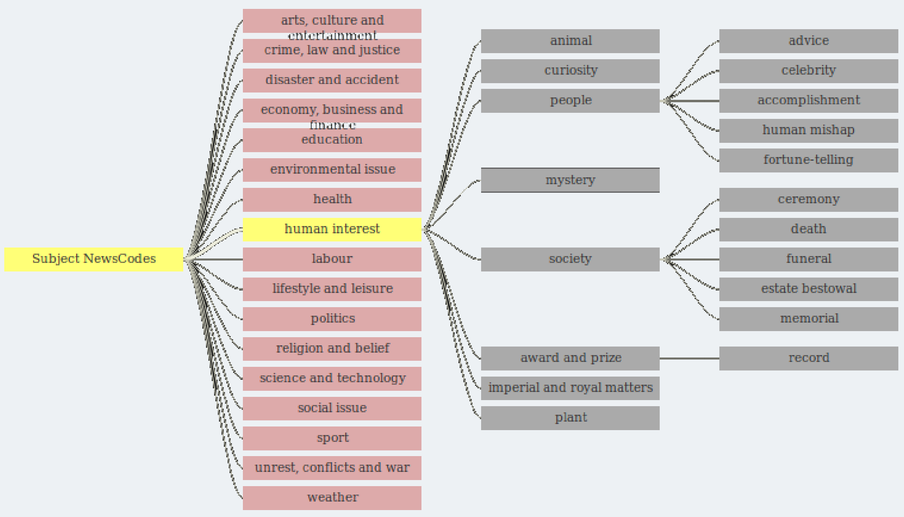


Figure. Tree-like visualization of some IPTC Subject NewsCodes.

Hierarchical Multi-label IPTC classification: Approach

- Build a hierarchical system for the 3 different levels of classification
- Use non-finetuned multilingual embeddings as a way of using the trained models on different languages

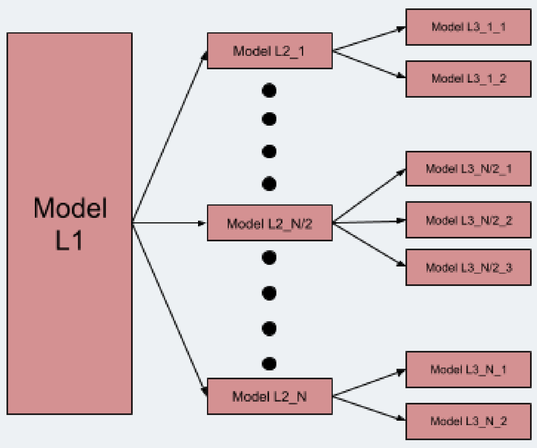


Figure. Overview of the implemented hierarchical system.

Hierarchical Multi-label IPTC classification: Results

LUSA corpus: 700k articles, 1400 terms structured in three levels

Models	L1 F_1	L2 F_1	L3 F_1
CNN	0.811	0.634	0.750
LSTM-CNN	0.832	0.675	0.761

Table. Results on the portuguese LUSA corpus. The columns are the micro F_1 scores for each level of the hierarchy. [a]

The aim of SUMMA is to significantly improve media monitoring by creating a platform to automate the analysis of media streams across many languages, to aggregate and distil the content, to automatically create rich knowledge bases, and to provide visualisations to cope with this deluge of data.
SUMMA integrates stream-based media processing tools (including speech recognition and machine translation) with deep language understanding capabilities (including named entity relation extraction and semantic parsing), for open-source applications and implemented in use cases at the BBC and Deutsche Welle.

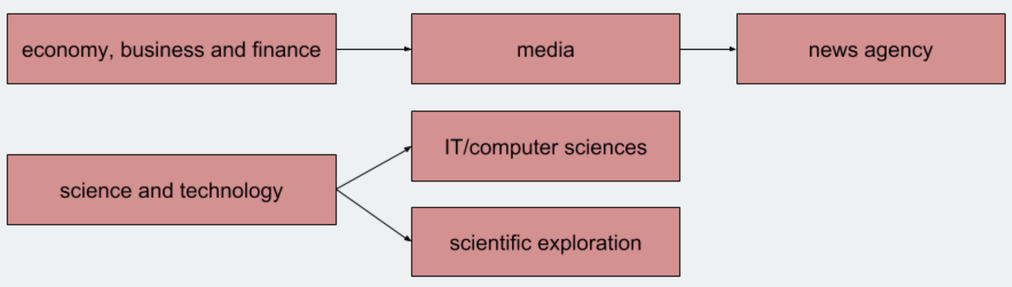
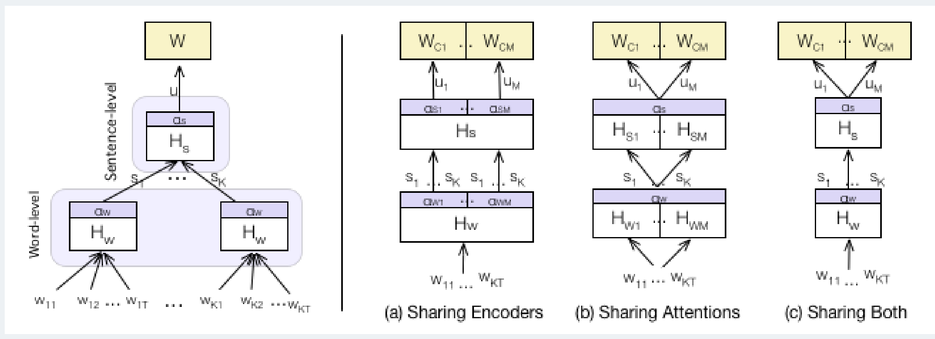


Figure. IPTC topics predicted for a given news text.

[a] Evaluation in other languages is still in progress.

Deep tagging: Model

- Goal: Predict multilingual tags assigned by journalists
- Multilingual hierarchical attention networks [IJCNLP 2017]
  1. Share parameters of the encoders across languages
  2. Outperform monolingual models while having fewer parameters than them
- Joint input-label multilingual embedding [TACL 2018[a]]
  1. Leverage multilingual word similarities to better understand target labels
  2. Share parameters of the encoders and output layers across languages
  3. Improve monolingual and multilingual models even further



[a] Currently under review.

Deep tagging: Evaluation

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

	Models		General labels		Specific labels	
	abbrev.	# lang.	n_l	f_l	n_l	f_l
[PB17]	HAN	1	50K	77.41	90K	44.90
[PB17]	MHAN	2	40K	78.30	80K	45.72
[PB17]	MHAN	8	32K	77.91	72K	45.82
Ours	GO-HAN	1	50K	**79.12**	90K	**45.90**
Ours	GO-MHAN	2	40K	**79.68**	80K	**46.49**
Ours	GO-MHAN	8	32K	**79.48**	72K	**46.32**

Table. Multilingual learning results. The columns are the average number of parameters per language (n_l), average F_1 per language (f_l).

→ Multilingual models improve over strong monolingual ones

→ Sharing attention mechanisms is the optimal configuration

→ Joint input-label embedding (GO) further improves all models

Deep tagging: Example

Figure. Multilingual tags predicted for a given document. Important sentences and words according to model's attention are marked with red and blue respectively.

Code

http://github.com/idiap/mhan

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