Some New Results in Epidemic Surveilance

Stephen Mutuvi and **Nhu Khoa Nguyen** (Main contributors)

\+ Adam Jatowt, Moses Odeo

\+ Emanuela Boros, Antoine Doucet

Gaël Lejeune (slide concatenation)

12 novembre 2020

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

Outline

General Context

Epidemic Surveillance

Mosts systems are designed for "clean and English data"

However, there are other needs :

- (I) Timeliness in event detection (multilinguality)
- (II) Ability to handle heterogeneous data (robustness to noise)

Three objectives to fulfill these needs :

- Improve and enrich reference datasets (I, II) [Mutuvi et al., 2020b]
- Compare SOA approach for multilingual text classification (I) [Mutuvi et al., 2020a]
- Evaluate different approaches on noisy documents (II) [Nguyen et al., 2020]

Challenges

Regarding classification

- Bender rule

Challenges

Regarding classification

- Bender rule
- Comparing the performances of multilingual classification methods
- Measuring the importance of news articles writing style and structure

Challenges

Regarding classification

- Bender rule
- Comparing the performances of multilingual classification methods
- Measuring the importance of news articles writing style and structure

Regarding noise

- Physical VS standard/digital-born formats

Challenges

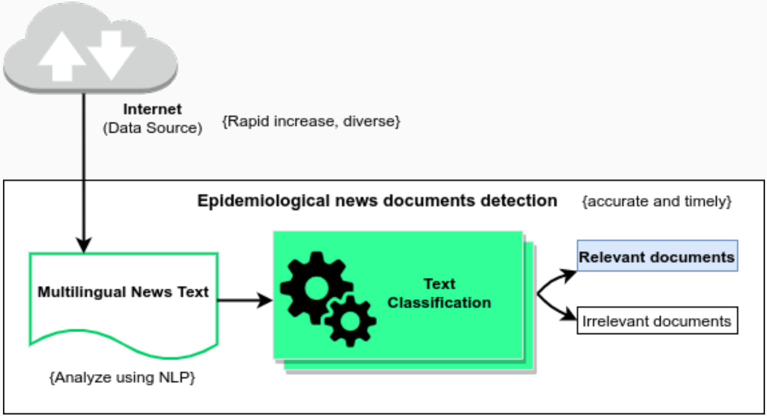
Regarding classification

- Bender rule
- Comparing the performances of multilingual classification methods
- Measuring the importance of news articles writing style and structure

Regarding noise

- Physical VS standard/digital-born formats
- Content alteration after digitization (archiving)
- Natural aging or damage in storage (encoding)
- Errors from conversion process (scraping)

General architecture



DAnIEL System

- **Document-level** analysis, exploiting **global structure of journalistic writing style**
- **Character-based algorithm**
- Crucial information in salient segments of text
- Early extraction of event in local language, thus timeliness in information deliveries

DAnIEL System

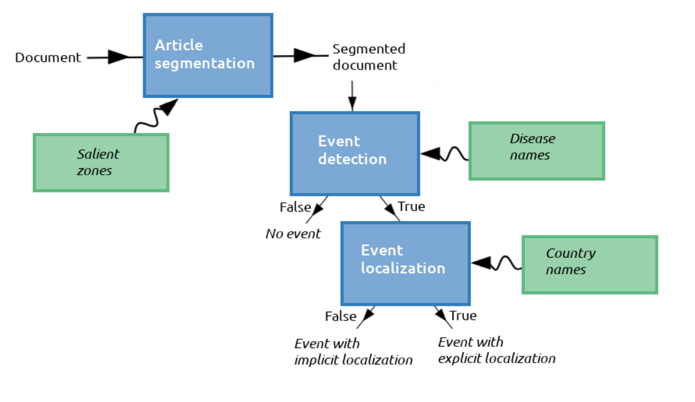


Figure 1 – Event Detection pipeline in DAnIEL.

Handling Noise

DAnIEL Dataset

- **DAnIEL Dataset** → DAnIEL event is defined at document-level
  - disease – location pairs
  - (listeria, U.S.)

Language	# Documents	# Relevant	# Sentences	# Tokens
French (fr)	2,733	340 (12.44%)	75,461	2,082,827
English (en)	475	31 (6.53%)	4,153	262,959
Chinese (zh)	446	16 (3.59%)	4,555	236,707
Russian (ru)	426	41 (9.62%)	6,865	133,905
Greek (el)	390	26 (6.67%)	3,743	198,077
Polish (pl)	352	30 (8.52%)	5,847	165,682
Total	4,822	489 (10.14%)	140,624	3,080,157

Data preparation

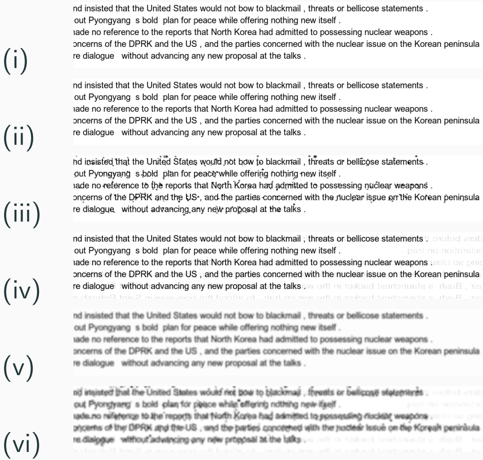


Figure 2 – (i) clean image, (ii) *Phantom Character*, (iii) *Character Degradation*, (iv) *Bleed Through*, (v) *Blur*, and (vi) all mixed together.

Clean Data : Evaluation of DAnIEL on the initial dataset for event identification and classification

Table 1 – Evaluation of DANIEL on the initial dataset for event identification (regardless of the types of the triggers).

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.6842	0.8	0.7115	0.641	0.592	0.4918	0.6052
	R	0.8667	1.0	0.9024	0.9259	0.9088	0.8571	0.9059
	F1	0.7647	0.8889	0.7957	0.7576	0.7169	0.625	0.7256
ratio=1.0	P	0.0	0.0	0.0	0.0	0.9155	0.0	0.9155
	R	0.0	0.0	0.0	0.0	0.5735	0.0	0.3988
	F1	0.0	0.0	0.0	0.0	0.7052	0.0	0.5556

Table 2 – Evaluation of DANIEL on the initial dataset for event classification.

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.3421	0.35	0.2692	0.4103	0.5211	0.2951	0.4645
	R	0.4	0.4118	0.3146	0.5079	0.5781	0.4737	0.5363
	F1	0.3688	0.3784	0.2902	0.4539	0.5481	0.3636	0.4978
ratio=1.0	P	0.0	0.0	0.0	0.0	0.7934	0.0	0.7934
	R	0.0	0.0	0.0	0.0	0.3592	0.0	0.2666
	F1	0.0	0.0	0.0	0.0	0.4945	0.0	0.3991

Noisy Data : Evaluation of DAnIEL on the noisy dataset for event identification

		Orig	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.61	0.735 (+0.12)	0.755 (+0.14)	0.735 (+0.12)	0.74 (+0.13)	0.731 (+0.12)	0.758 (+0.14)
	R	0.91	0.859 (-0.05)	**0.674** (-0.23)	0.862 (-0.04)	0.857 (-0.05)	0.862 (-0.04)	0.718 (-0.19)
	F1	0.73	0.792 (+0.06)	0.712 (-0.01)	0.793 (+0.06)	0.794 (+0.06)	0.791 (+0.06)	0.737 (+0.00)
PL	P	0.68	0.643 (-0.03)	0.656 (-0.02)	0.658 (-0.02)	0.692 (+0.01)	0.643 (-0.03)	0.645 (-0.03)
	R	0.87	0.9 (+0.03)	0.7 (-0.17)	0.9 (+0.03)	0.9 (+0.03)	0.9 (+0.03)	0.667 (-0.20)
	F1	0.76	0.75 (-0.01)	0.677 (-0.08)	0.761 (+0.00)	0.783 (+0.02)	0.75 (-0.01)	**0.656** (-0.10)
ZH	P	0.8	0.882 (+0.08)	0.882 (+0.08)	0.789 (-0.01)	0.733 (-0.06)	0.789 (-0.01)	0.857 (+0.05)
	R	1.0	0.938 (-0.06)	0.938 (-0.06)	0.938 (-0.06)	0.917 (-0.08)	0.938 (-0.06)	0.75 (-0.25)
	F1	0.89	0.909 (+0.01)	0.909 (+0.01)	0.857 (-0.03)	0.815 (-0.07)	0.857 (-0.03)	**0.8** (-0.09)
RU	P	0.71	0.688 (-0.02)	0.691 (-0.01)	0.688 (-0.02)	0.705 (-0.00)	0.688 (-0.02)	0.727 (+0.01)
	R	0.9	0.805 (-0.09)	0.744 (-0.15)	0.846 (-0.05)	0.795 (-0.10)	0.846 (-0.05)	0.821 (-0.08)
	F1	0.8	0.742 (-0.05)	**0.716** (-0.08)	0.759 (-0.04)	0.747 (-0.05)	0.759 (-0.04)	0.771 (-0.02)
EL	P	0.64	0.59 (-0.05)	0.682 (+0.04)	0.59 (-0.05)	0.639 (-0.00)	0.59 (-0.05)	0.667 (+0.02)
	R	0.93	0.852 (-0.07)	0.556 (-0.37)	0.852 (-0.07)	0.852 (-0.07)	0.852 (-0.07)	0.518 (-0.41)
	F1	0.76	0.697 (-0.06)	0.612 (-0.14)	0.697 (-0.06)	0.73 (-0.03)	0.697 (-0.06)	**0.583** (-0.17)
FR	P	0.59	0.803 (+0.21)	0.828 (+0.23)	0.806 (+0.21)	0.801 (+0.21)	0.801 (+0.21)	0.816 (+0.22)
	R	0.91	0.849 (-0.06)	0.666 (-0.24)	0.849 (-0.06)	0.849 (-0.06)	0.849 (-0.06)	0.723 (-0.18)
	F1	0.72	0.826 (+0.10)	0.738 (+0.01)	0.827 (+0.10)	0.825 (+0.10)	0.825 (+0.10)	0.767 (+0.04)
EN	P	0.49	0.508 (+0.01)	0.458 (-0.03)	0.508 (+0.01)	0.516 (+0.02)	0.508 (+0.01)	0.52 (+0.03)
	R	0.86	0.943 (+0.08)	0.629 (-0.23)	0.943 (+0.08)	0.943 (+0.08)	0.943 (+0.08)	0.743 (-0.11)
	F1	0.62	0.66 (+0.04)	**0.53** (-0.09)	0.66 (+0.04)	0.667 (+0.04)	0.66 (+0.04)	0.612 (-0.00)

Noisy Data : Evaluation of DAnIEL on the noisy dataset for event classification

		Orig	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.46	0.552 (+0.09)	0.548 (+0.08)	0.549 (+0.08)	0.558 (+0.09)	0.548 (+0.08)	0.547 (+0.08)
	R	0.54	0.497 (-0.04)	**0.377** (-0.16)	0.496 (-0.04)	0.497 (-0.04)	0.498 (-0.04)	0.4 (-0.14)
	F1	0.5	0.523 (+0.02)	0.447 (-0.05)	0.521 (+0.02)	0.526 (+0.02)	0.521 (+0.02)	0.462 (-0.03)
PL	P	0.34	0.333 (-0.00)	0.328 (-0.01)	0.342 (+0.00)	0.359 (+0.01)	0.333 (-0.00)	0.274 (-0.06)
	R	0.4	0.431 (+0.03)	0.323 (-0.07)	0.431 (+0.03)	0.431 (+0.03)	0.431 (+0.03)	0.262 (-0.13)
	F1	0.37	0.376 (+0.00)	0.326 (-0.04)	0.381 (+0.01)	0.392 (+0.02)	0.376 (+0.00)	**0.268** (-0.10)
ZH	P	0.35	0.412 (+0.06)	0.353 (+0.00)	0.342 (-0.00)	0.367 (+0.01)	0.342 (-0.00)	0.464 (+0.11)
	R	0.41	0.412 (+0.00)	0.353 (-0.05)	0.382 (-0.02)	0.423 (+0.01)	0.382 (-0.02)	0.382 (-0.02)
	F1	0.38	0.412 (+0.03)	**0.353** (-0.02)	0.361 (-0.01)	0.393 (+0.01)	0.361 (-0.01)	0.419 (+0.03)
RU	P	0.27	0.302 (+0.03)	0.312 (+0.04)	0.302 (+0.03)	0.295 (+0.02)	0.302 (+0.03)	0.273 (+0.00)
	R	0.31	0.326 (+0.01)	0.357 (+0.04)	0.341 (+0.03)	0.306 (-0.00)	0.341 (+0.03)	0.282 (-0.02)
	F1	0.31	0.314 (+0.00)	0.333 (+0.02)	0.32 (+0.01)	0.301 (-0.00)	0.32 (+0.01)	**0.278** (-0.03)
EL	P	0.41	0.333 (-0.07)	0.341 (-0.06)	0.333 (-0.07)	0.361 (-0.04)	0.333 (-0.07)	0.357 (-0.05)
	R	0.51	0.413 (-0.09)	0.238 (-0.27)	0.413 (-0.09)	0.413 (-0.09)	0.413 (-0.09)	0.238 (-0.27)
	F1	0.45	0.369 (-0.08)	**0.28** (-0.17)	0.369 (-0.08)	0.385 (-0.06)	0.369 (-0.08)	0.286 (-0.16)
FR	P	0.47	0.691 (+0.22)	0.693 (+0.22)	0.69 (+0.22)	0.689 (+0.21)	0.689 (+0.21)	0.675 (+0.20)
	R	0.51	0.527 (+0.01)	0.402 (-0.10)	0.524 (+0.01)	0.527 (+0.01)	0.527 (+0.01)	0.431 (-0.07)
	F1	0.49	0.598 (+0.10)	0.509 (+0.01)	0.596 (+0.10)	0.597 (+0.10)	0.597 (+0.10)	0.526 (+0.03)
EN	P	0.47	0.292 (-0.17)	0.26 (-0.21)	0.292 (-0.17)	0.297 (-0.17)	0.292 (-0.17)	0.31 (-0.16)
	R	0.51	0.5 (-0.01)	0.329 (-0.18)	0.5 (-0.01)	0.5 (-0.01)	0.5 (-0.01)	0.408 (-0.10)
	F1	0.49	0.369 (-0.12)	**0.291** (-0.19)	0.369 (-0.12)	0.372 (-0.11)	0.369 (-0.12)	0.352 (-0.13)

Handling noise : conclusion

- Event extraction is **prone to digitization errors**
- In general, image degradation has **little impact** on the performance of DAnIEL
- **Interesting** : Precision, recall, and F1 can increase, despite the noise $\rightarrow$ noise helps disambiguating false positives, hence more distinguishable
- **Realistic scenarios** might generate other undiscovered issues in the digitization process

Handling Multilinguality

Experiments : Dataset

- **Extended dataset**[Mutuvi et al., 2020b]
- **Language families** : Germanic (English), Hellenic (Greek), Romance (French), Slavic (Russian and Polish), Sino-Tibetan (Chinese)

	All	**Polish**	**Chinese**	**Russian**	**Greek**	**French**	**English**
Train	5,074 (10.8)	241 (7.4)	300 (2.6)	296 (9.45)	253 (6.7)	1,593 (10.9)	2,365 (11.7)
Validation	1,250 (10.9)	54 (7.4)	71 (2.8)	60 (10.0)	68 (10.2)	388 (13.4)	583 (12.6)
Test	1,250 (10.5)	46 (13.0)	75 (6)	70 (10.0)	63 (4.7)	434 (12.4)	614 (12.8)

Experiments : Dataset

- **Extended dataset**[Mutuvi et al., 2020b]
- **Language families** : Germanic (English), Hellenic (Greek), Romance (French), Slavic (Russian and Polish), Sino-Tibetan (Chinese)

	All	**Polish**	**Chinese**	**Russian**	**Greek**	**French**	**English**
Train	5,074 (10.8)	241 (7.4)	300 (2.6)	296 (9.45)	253 (6.7)	1,593 (10.9)	2,365 (11.7)
Validation	1,250 (10.9)	54 (7.4)	71 (2.8)	60 (10.0)	68 (10.2)	388 (13.4)	583 (12.6)
Test	1,250 (10.5)	46 (13.0)	75 (6)	70 (10.0)	63 (4.7)	434 (12.4)	614 (12.8)

Unfortunately we were not able to use all the data :

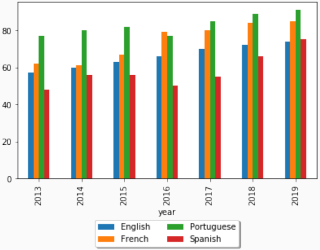


Figure 3 – Percentage of ProMED sources accessible by year.

Experiments

- Baseline model : **DAnIEL**
  - does not rely on any language-specific grammar analysis and considers text as a sequence of strings instead of words.
- Machine learning models :
  - **Logistic regression**, **Random forests**, **Support vector machines**
- Deep learning models :
  - **CNN** and **BiLSTM** (with FastText embeddings)
- **BERT**-based architectures :
  - `bert-base-multilingual-cased`
  - `bert-base-multilingual-uncased`)

Results : all models

Models	Precision %	Recall %	F1 %
DANIEL	33.9	60.61	43.48
LR	93.81	68.94	79.48
RF	**95.70**	67.42	79.11
SVM	91.26	71.21	80
CNN+FastTtext	86.11	70.45	77.5
BiLSTM+FastTtext	77.44	78.03	77.74
BERT (cased)†	88.62	82.58	85.49
CNN+BERT (cased)†	88.79	71.97	79.5
BiLSTM+BERT (cased)†	90.20	69.70	78.63
BERT (uncased)†	84.67	87.88	**86.25**
CNN+BERT (uncased)†	82.14	87.12	84.56
BiLSTM+BERT (uncased)†	83.72	81.82	82.76
BERT (cased)	80.71	85.61	83.09
CNN+BERT (cased)	86.67	78.79	82.54
BiLSTM+BERT (cased)	75.95	**90.91**	82.76
BERT (uncased)	88.52	81.82	85.04
CNN+BERT (uncased)	86.07	79.55	82.68
BiLSTM+BERT (uncased)	81.51	73.48	77.29
VGCN+BERT	87.18	77.27	81.93

Result : per-language evaluation

Models	Polish	Chinese	Russian	Greek	French	English
DAnIEL	40	80	33.33	33.33	71.43	32.23
LR	0	0	66.67	66.67	84.21	80
RF	0	0	40	66.67	86.84	78.83
SVM	0	0	33.33	0	87.18	81.38
CNN+FastText	0	0	0	0	84.21	81.88
BiLSTM+FastText	0	0	0	0	73.12	85.71
BERT (cased)†	50	80	66.67	66.67	**94.12**	82.89
CNN+BERT (cased)†	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (cased)†	0	80	40.00	66.67	87.36	86.27
BERT (uncased)†	57.14	80	50	**100**	91.95	86.08
CNN+BERT (uncased)†	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (uncased)†	0	80	40	66.67	87.36	86.27
BERT (cased)	33.33	80	50	66.67	87.50	85.54
CNN+BERT (cased)	0	0	40	66.67	83.33	86.45
BiLSTM+BERT (cased)	0	80	22.22	28.57	85.11	**88.37**
BERT (uncased)	0	66.67	85.71	66.67	87.18	86.25
CNN+BERT (uncased)	0	50	40	66.67	82.35	86.45
BiLSTM+BERT (uncased)	0	0	33.33	0	72.94	84.42
VGCN+BERT	**71.43**	**88.89**	**88.89**	80	87.80	78.26

Results : Zero-shot transfer learning

Train \ Test	Polish	Chinese	Russian	Greek	French	English
Polish	40	-	66.67	66.67	76.92	85.71
Chinese	-	80	60	-	70.97	81.08
Russian	33.33	-	33.33	66.67	62.86	88.61
Greek	-	-	-	66.67	-	63.05
French	-	66.67	57.14	-	91.95	85.90
English	50	-	33.33	66.67	39.29	84.35

- **English** and **French** documents have consistently higher F-score than when the model is trained on the other languages

Results : effect of article structure

Text Position	Models	Precision %	Recall %	F1 %
Beginning	VGCN+BERT	**87.18**	77.27	81.93
	BERT (uncased)†	84.67	87.88	86.25
Body	VGCN+BERT	79.83	71.97	75.70
	BERT (uncased)†	75.71	80.30	77.94
End	VGCN+BERT	72.93	73.48	73.21
	BERT (uncased)†	76.12	77.27	76.69
Beginning+End	VGCN+BERT	86.61	83.33	84.94
	BERT (uncased)†	85.61	**90.15**	**87.82**

- The combination of **the beginning and the ending text** in the news documents provided the best features

Discussion

- Deep learning (**BERT-based**) approaches performed better
    - capable of much deeper and complex representation
- Low-resource languages :
    - **CNN** and **BiLSTM**-based models were not able to distinguish relevant documents : size of training data

Handling Multilinguality : conclusions

- A detailed study of the performance of different methods on the task of epidemiological news report detection
- Models based on **fine-tuned language models** achieve good performance on the classification of multilingual epidemiological text
- **Transfer learning** can benefit the process of extracting useful information from multilingual epidemiological text

Perspectives : dataset ?

Dataset in LREC paper

The sources from experts from ProMED :

Language	#Documents	#Sentences	#Words
English (en)	19,149	558,448	53,325,455
French (fr)	1,849	28,823	5,593,184
Spanish (es)	3,453	27,918	4,458,533
Portuguese (pt)	3,451	48,591	5,994,583

Table 3 – Statistics for Retrieved ProMED Documents

Dataset in LREC paper

The sources from experts from ProMED :

Language	#Documents	#Sentences	#Words
English (en)	19,149	558,448	53,325,455
French (fr)	1,849	28,823	5,593,184
Spanish (es)	3,453	27,918	4,458,533
Portuguese (pt)	3,451	48,591	5,994,583

Table 3 – Statistics for Retrieved ProMED Documents

- Has new languages
- Curated by experts
- How can we get True Negatives ?

Références

References I

Mutuvi, S., Boros, E., Doucet, A., Jatowt, A., Lejeune, G., and Odeo, M. (2020a).
Multilingual epidemiological text classification : A comparative study.
In *COLING 2020*, page to appear.

Mutuvi, S., Doucet, A., Lejeune, G., and Odeo, M. (2020b).
A Dataset for Multilingual Epidemiological Event Extraction.
Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4139–4144.

Nguyen, N. K., Boros, E., Lejeune, G., and Doucet, A. (2020).
Impact analysis of document digitization on event extraction.
In *Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI), 19th International Conference of the Italian Association for Artificial Intelligence*, page to appear, Roma, Italy. -.