# Multilingual Email Zoning

Bruno Jardim, Ricardo Rei, Mariana S.C. Almeida

Cleverly

NOVA IMS Information Management School

Github

## Abstract


- We analyse the existing email zoning corpora and propose **Cleverly zoning corpus**, a new multilingual benchmark composed of 625 emails in Portuguese, Spanish and French.
- We introduce **OKAPI**, the first multilingual email segmentation model based on a language agnostic sentence encoder.
- Our model r: i) generalizes well for unseen languages, ii) is competitive with current English benchmarks, and iii) reached new state-of-the-art performances for domain adaptation tasks in English.


## Cleverly Zoning Corpus

- The first multilingual email zoning corpus.
- We used the Gmane raw corpus (Bevendorff et al., 2020).
- We followed the classification schema proposed by Bevendorff et al. (2020), with **15 annotated zones**.

| | PT | ES | FR |
|---|---|---|---|
| #zones | 15 | 14 | 14 |
| #emails | 210 | 200 | 215 |
| #lines | 12366 | 9824 | 6958 |

## References


Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein. *Crawling and preprocessing mailing lists at scale for dialog analysis.* ACL, 2020.

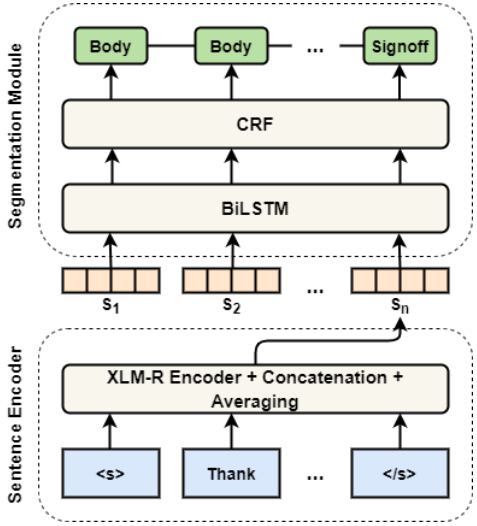
Tim Repke and Ralf Krestel. *Bringing back structure to free text email conversations with recurrent neural networks.* ECIR, 2018.


## OKAPI

**OKAPI** is composed of two building blocks:

1) a **multilingual sentence encoder** that extracts word-level embedding with XLM-RoBERTa (Conneau et al., 2020), then apply average pooling to the last 4 layers.

2) and a **segmentation module** that uses a BiLSTM with a CRF on top to classify each sentence into an email zone.




## Results

| Zone | PT | ES | FR |
|---|---|---|---|
| **All** | **0.91** | **0.93** | **0.93** |
| *Quotation* | 0.99 | 0.99 | 0.99 |
| *Paragraph* | 0.91 | 0.96 | 0.92 |
| *MUA Sig.* | 0.95 | 0.82 | 0.91 |
| *Pers. Sig.* | 0.81 | 0.87 | 0.79 |
| *Visual. Sep.* | 0.92 | 0.90 | 0.96 |

**Zero-shot OKAPI accuracy** on Ceverly Zoning corpus. OKAPI was only trained on Gmane English corpus (Bevendorff et al. (2020)).

**OKAPI** outperformes existing **monolingual methods** with various English corpora, zoning taxonomies, and tasks

| Model | Zones | Enron | ASF |
|---|---|---|---|
| *Jangada* | 2 | 0.88 | 0.97 |
| *Zebra* | 2 | 0.25 | 0.18 |
| *Quagga* | 2 | 0.98 | 0.98 |
| ***OKAPI*** | 2 | **0.99** | **0.99** |
| *Jangada* | 5 | 0.85 | 0.91 |
| *Zebra* | 5 | 0.24 | 0.20 |
| *Quagga* | 5 | 0.93 | **0.95** |
| ***OKAPI*** | 5 | **0.96** | **0.95** |

**Accuracy** on Repke and Krestel (2018) corpora

| Model | Zones | Gmane | Enron |
|---|---|---|---|
| *Tang et al.* | 15 | 0.86 | 0.73 |
| *Quagga* | 15 | 0.94 | 0.83 |
| *Chipmunk* | 15 | **0.96** | **0.88** |
| ***OKAPI*** | 15 | **0.96** | **0.88** |

**Accuracy** on Bevendorff et al. (2020) corpora

| Model | Train/Test | 2 Zones | 5 Zones |
|---|---|---|---|
| *Quagga* | Enron/ASF | 0.94 | 0.86 |
| ***OKAPI*** | Enron/ASF | **0.98** | **0.93** |
| *Quagga* | ASF/Enron | 0.86 | 0.80 |
| ***OKAPI*** | ASF/Enron | **0.97** | **0.88** |

**Accuracy** for **domain adaptation** tasks.

## Aknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 873904.

Horizon 2020 European Union Funding for Research & Innovation