Multilingual Email Zoning

Bruno Jardim
Cleverly, Lisbon, Portugal
NOVA-IMS, Lisbon, Portugal
`bjardim@novaims.unl.pt`

Ricardo Rei
NOVA-IMS, Lisbon, Portugal
Unbabel, Lisbon, Portugal
`rrei@novaims.unl.pt`

Mariana S.C. Almeida
Cleverly, Lisbon, Portugal
`mariana.almeida@cleverly.ai`

Abstract

The segmentation of emails into functional zones (also dubbed **email zoning**) is a relevant preprocessing step for most NLP tasks that deal with emails. However, despite the multilingual character of emails and their applications, previous literature regarding email zoning corpora and systems was developed essentially for English.

In this paper, we analyse the existing email zoning corpora and propose a new multilingual benchmark composed of 625 emails in Portuguese, Spanish and French. Moreover, we introduce OKAPI, the first multilingual email segmentation model based on a language agnostic sentence encoder. Besides generalizing well for unseen languages, our model is competitive with current English benchmarks, and reached new state-of-the-art performances for domain adaptation tasks in English.

1 Introduction

Worldwide, email is a predominant means of social and business communication. Its importance has attracted studies in areas of Machine Learning (ML) and Natural Language Processing (NLP), impacting a wide range of applications, from spam filtering (Qaroush et al., 2012) to network analysis (Christidis and Losada, 2019).

The email body is commonly perceived as unstructured textual data with multiple possible formats. However, it is possible to discern a level of formal organization in the way most emails are formed. Different functional parts can be identified such as greetings, signatures, quoted content, legal disclaimers, etc. The segmentation of email text into zones, also known as **email zoning** (Lampert et al., 2009), has since become a prevalent preprocessing task for a diversity of downstream applications, such as author profiling (Estival et al., 2007),
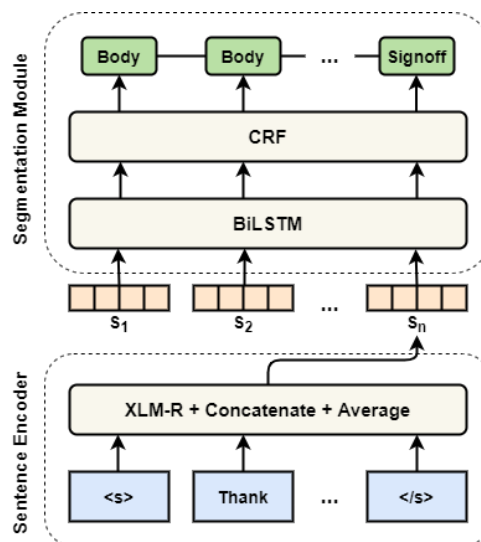

Figure 1: OKAPI is composed of two building blocks: 1) a multilingual sentence encoder (XLM-RoBERTa) to derive sentence embeddings; and 2) a segmentation module that uses a BiLSTM with a CRF on top to classify each sentence into an email zone.

request detection (Lampert et al., 2010), uncover of technical artifacts (Bettenburg et al., 2011), automated template induction (Proskurnia et al., 2017), email classification (Kocayusufoglu et al., 2019) or automated email response suggestion (Kannan et al., 2016; Chen et al., 2019).

Since email communication is a worldwide phenomenon, all previous applications are in fact highly multilingual. Despite this, email zoning literature remains English-centric and without a standardize zone taxonomy. To mitigate those problems, we make the following research contributions:

1. We discuss the existing zoning taxonomies and their limitations.

2. We release Cleverly zoning corpus, the first multilingual corpus for email zoning. This

corpus consists of 625 emails in 3 languages rather than English (Portuguese, Spanish and French), and encompasses 15 email zones as defined in (Bevendorff et al., 2020)

3. We introduce OKAPI, a multilingual email segmentation system built on top of XLM-RoBERTa (Conneau et al., 2020) that can be easily extended to 100 languages.

To the best of our knowledge, OKAPI is the first end-to-end multilingual system exploring pre-trained transformer models (Vaswani et al., 2017) to perform email zoning. Besides having multilingual capabilities, OKAPI is competitive with existing approaches for English email zoning, and attained state-of-the-art performance in domain adaptation tasks for English email zoning.

The rest of the paper is organized as follows: Section 2 presents an overview of the related literature. Section 3 provides a comprehensive review of existing email zoning corpora, and introduces Cleverly zoning corpus, our new multilingual email zoning corpus. Section 4 describes the OKAPI model architecture. Section 5 reports and discusses the results achieved. Finally, Section 6 concludes the paper.

2 Literature Review

Chen et al. (1999) were one of the pioneers in the topic of email segmentation. Looking at linguist and geometrical patterns, their work focuses on the identification of email signature. Similarly, Carvalho and Cohen (2004) developed JANGADA, a supervised learning system that classifies each line using a Conditional Random Field (CRF) (Lafferty et al., 2001) and a sequence-aware perceptron (Collins, 2002), that identifies signature blocks and quoted text from previous emails. Tang et al. (2005) proposed an email data cleansing system based on a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) model that aimed at filtering the non-textual noisy content from emails independently of downstream text mining applications, based on hand-coded features.

Estival et al. (2007) were the first to introduce a general segmentation schema for email text. Segmentation of emails is a crucial part on their work, which aims at identifying the author's basic demographic and psychometric traits. In that work, the authors compared a range of ML algorithms together with feature selection to classify email

segments into five functional parts, attaining improvements in the end task of auto profiling. Later, Lampert et al. (2009) formally defined the functional parts as email zones, describing the different segments inside email messages based on graphic, orthographic, and lexical features. Lampert et al. (2009) also proposed ZEBRA, an email zoning system based on a SVM. In a posterior work towards detecting emails containing requests for action, Lampert et al. (2010) used ZEBRA to "zone" emails, considering only the zones that had relevant patterns to increase the accuracy of their request detection task.

As email zoning surpassed its original purpose of signature identification and text cleansing into a more general task, Repke and Krestel (2018) extended its utility to thread reconstruction. Inspired by ZEBRA (Lampert et al., 2009), the authors proposed QUAGGA (Repke and Krestel, 2018), a neural system with a Convolutional Neural Network (CNN) (LeCun et al., 1989) to produce sentence representations followed by a Recurrent Neural Network (RNN) (Elman, 1990). QUAGGA was trained and evaluated on English emails from both the Enron (Klimt and Yang, 2004) corpus and the public mail archives of the Apache Software Foundation (ASF)[1], outperforming JANGADA and ZEBRA.

Until very recently, email zoning resorted to small samples of mailing lists or newsgroup corpus and was limited to the English language. Bevendorff et al. (2020) were the first to crawl email at scale, extracting 153 million emails from the Gmane email-to-newsgroup gateway[2] in different languages such as English, Spanish, French and Portuguese[3]. The authors annotated email zones for a subset of Gmane English emails and, due to the idiosyncratic characteristics of the corpora, they developed a more fine-grained zone classification schema with 15 zones. Moreover, Bevendorff et al. (2020) introduced an email zoning system, named CHIPMUNK, that combines a Bidirectional Gated Recurrent Unit (BiGRU) (Cho et al., 2014) with a CNN. When compared to other models in the literature, CHIPMUNK achieved better performance.

---

[1] http://mail-archives.apache.org/mod_mbox/
[2] https://news.gmane.io/
[3] https://webis.de/data.html?q=Webis-Gmane-19

Authors	Source	emails	zones	Language
(Carvalho and Cohen, 2004)[4]	20 Newsgroup[5]	617	2	English
(Estival et al., 2007)[6]	Donated[6]	9,836	5	English
(Lampert et al., 2009)[7]	Enron[8]	400	3/9	English
(Repke and Krestel, 2018)[9]	Enron[8]	800	2/5	English
	ASF[1]	500	2/5	English
Bevendorff et al. (2020)[10]	Gmane[3]	3,033	15	Multilingual*
	Enron[8]	300	15	English
Ours	Gmane[3]	625	15	Multilingual

Table 1: Summary of existing email zoning corpora. *Note that, although Bevendorff et al. (2020)'s Gmane corpus is technically multilingual, it only has 38 non-English test emails that are spread over 13 different languages.

3 Email Zoning Corpora

Several corpora and zoning schemes have been proposed in the literature under different contexts. This section provides an overview of the existing corpora, hoping to make it easier to develop and compare new email zoning methods in the future.

Table 1 compiles the information of existing email zoning corpus. To the best of our knowledge, Carvalho and Cohen (2004) released the first email zoning corpus. The corpus consists of 617 emails[4] from the 20 Newsgroup corpus[5] identified with two zones: *signature* and *quotation*. Despite the usefulness of identifying those zones for email cleansing, this level of detail is still insufficient for a general email segmentation.

Estival et al. (2007) released a corpus of 9,836 recruited respondents donated email messages[6] and introduced a wider annotation schema focusing on more email parts: *author text*, *signature*, *advertisement*, *quoted text*, and *reply* lines. However, Estival et al. (2007) still did not divide the email text into some other relevant zones, such as greetings, closings nor identify attachments and code lines.

Lampert et al. (2009) were arguably the first to conceptualize the email zoning task and fully define the characteristics of each identified zone, as well as dividing the authored text into different zones. They annotated 400 English emails[7] from the Enron email corpus database dump, identifying 3 email zones: *sender*, *quoted conversation* and *boilerplate* zones, each containing a different set of sub-zones, within a total of 9 sub-zones.

Repke and Krestel (2018) also resorted to the Enron database[8], annotating a total of 800 emails[9]. Reconsidering the task as thread reconstruction, they produced a new annotation schema, considering a 2-level and a 5-level approach (the latter being a refinement of the 2-level segmentation). Repke and Krestel (2018) also annotated 500 ASF emails[7] using both the 2-level and 5-level taxonomies. Their 5-level annotation schema consists of segmenting emails into: *body* (typically comprising $\sim 80\%$ of the lines), *header*, *signoff*, *signature* and *greetings*.

Bevendorff et al. (2020) introduced the Gmane corpus for email zoning[10]. Even though the corpus is composed of 31 languages, the annotated emails are mostly in English, and their test set only contains a residual number of non-English emails (38 emails covering 13 different languages), which is insufficient for a consistent multilingual evaluation. Due to the richness of the Gmane conversations on technical topics, Bevendorff et al. (2020) developed a more fine grained classification schema, considering the segmentation of blocks of code, log data and technical data. Whilst also preserving most of the common zones introduced in previous works, they ended up with a total of 15 zones: *closing*, *inline headers*, *log data*, *MUA signature*, *paragraph*, *patch*, *personal signature*, *quotation*, *quotation marker*, *raw code*, *salutation*, *section heading*, *tabular*, *technical*, *visual separator*. Following the same zone taxonomy they also released a set of 300 English emails from the Enron database dump. In both Enron and Gmane

---

[4] http://www.cs.cmu.edu/~vitor/codeAndData.html
[5] http://qwone.com/~jason/20Newsgroups/
[6] available upon contact with the authors.
[7] http://zebra.thoughtlets.org/

[8] http://www.cs.cmu.edu/~enron/
[9] https://github.com/HPI-Information-Systems/Quagga
[10] https://github.com/webis-de/acl20-crawling-mailing-lists

emails, the majority of the email segments belong to the *paragraph* and *quotation* zones. This being said, Gmane has much more lines of *quotation* than *paragraph*, while Enron is the other way around.

Overall, email zoning corpora show a great variability of zone taxonomies and most works have introduced new zones to face the nature of each email source or downstream task. The Enron database dump has been the most used source to retrieve emails to build new corpus. On the other hand, the recent Gmane raw dump of emails is multilingual and it contains various functional zones, which opens the door to new challenges in email zoning and multilingual methodologies.

3.1 Cleverly Zoning Corpus

	pt	es	fr
# zones	15	14	14
# emails	210	200	215
# lines	12366	9824	6958
# lines / email	58.9	49.1	32.4
# zones /email	8.6	6.5	5.9
# unique zones / email	5.8	5.1	4.9

Table 2: Some statistics of the Cleverly zoning corpus.

This section presents Cleverly zoning corpus, the first multilingual email zoning corpus. To create the corpus, we searched the Gmane raw corpus (Bevendorff et al., 2020) for Portuguese (pt), Spanish (es) and French (fr) emails. Then, following the classification schema proposed by Bevendorff et al. (2020), we produced a total of 625 annotated emails.

Table 2 compiles a brief description of the email statistics for each of the languages. While French is the language with more emails, Portuguese and Spanish emails tend to be longer, resulting in a greater amount of lines and an overall higher number of zones per email. The distribution of zones is similar between the three languages, as detailed in Table 3.

The annotation was carried out by two annotators. The first annotator was a native Portuguese speaker and the second annotator a native Spanish speaker, both with academical background in French and fluent in the third language. Each email was annotated by both annotators using the tagtog[11] annotation tool.

---

[11] https://www.tagtog.net

Zone	pt (%)	es (%)	fr (%)
Quotation	52.43	59.02	46.20
Paragraph	16.33	17.36	27.61
MUA Sig.	12.04	3.84	9.04
Personal Sig.	3.93	4.47	2.00
Visual Sep.	2.94	2.29	2,60
Quot. Mark.	2.72	1.54	2.10
Closing	2.63	2.00	3.73
Log Data	1.04	3.79	1.82
Raw Code	1.28	2.45	2.07
Inl. Head.	2.96	0,82	1.33
Salutation	0,96	0.81	1,35
Tabular	0.32	0.42	0.27
Technical	0.30	1.00	0.38
Patch	0.02	0.20	0.02
Sec. Head.	0.15	0.04	0.03

Table 3: Distribution, for each language, of the number of lines per zone in the Cleverly zoning corpus. The distributions were obtained by averaging statistics from both annotators.

measure	pt	es	fr
accuracy	0.93	0.92	0.96
F_1 A_1A_2	0.93	0.92	0.96
F_1 A_2A_1	0.94	0.92	0.96
k	0.90	0.87	0.94

Table 4: Inter-annotator agreement for each language in the Cleverly zoning corpus, using Cohen's kappa (k), accuracy and F_1 between annotators A_1 and A_2.

Table 4 shows the inter-annotator agreement scores for each language using the Cohen's kappa coefficient (k) (McHugh, 2012), accuracy and F_1 of one annotator versus the other. All annotations and required information to compile the original emails are freely available at https://github.com/cleverly-ai/multilingual-email-zoning.

4 OKAPI Architecture

We propose OKAPI, an email segmentation model composed of two building blocks: a multilingual sentence encoder and a segmentation module. Figure 1 shows the OKAPI architecture.

4.1 Multilingual Sentence Encoder

To address the multilingual nature of emails we developed a language agnostic sentence encoder that turns each email line into an embedding. Figure 2 illustrates the encoding process.
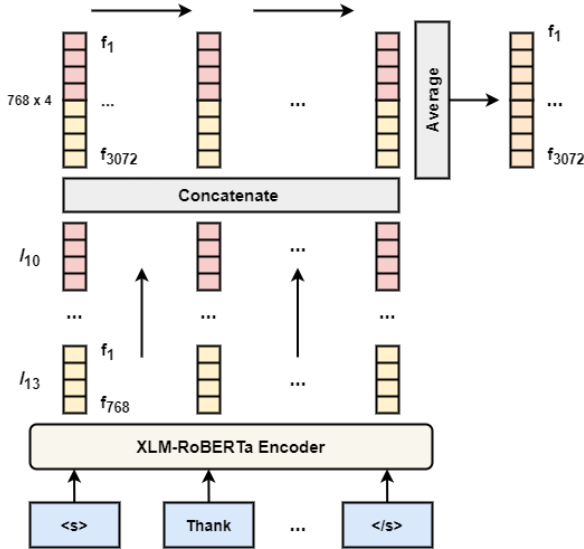
Figure 2: To derive a cross-lingual line embedding we use XLM-RoBERTa (Conneau et al., 2020) to extract word-level embeddings, and then we apply average pooling to the last 4 layers. This leads to a final 3072 features embedding.

Given an email line $x = [x_0, x_1, ..., x_n]$, our encoder module uses XLM-RoBERTa (base) (Conneau et al., 2020) to produce an embedding $e_j^{(\ell)}$ for each token x_j and each layer $\ell \in \{0, 1, ..., 13\}$. Since it has been shown that BERT-like models capture within the network layers diverse linguistic information, and, particularly, the last layers preserve most of the semantic information (Tenney et al., 2019), we keep, for each sentence, only the word embeddings from the last 4 layers. Lastly, as in (Reimers and Gurevych, 2019), these word embedding are turned into a 3072 sentence embedding s_k by averaging the concatenation of the 4 word layer embeddings.

4.2 Segmentation Module

After passing each email line into the previous sentence encoder we get a cross-lingual line embedding s_k. After that, we pass all line embeddings of an email into a Bidirectional Long Short-Term Memory (BiLSTM) (Graves and Schmidhuber, 2005), with 1 layer and 64 hidden units, to derive compact line representations that encompass information from the entire structure of the email. Finally, as in Huang et al. (2015), we use a CRF output layer to predict the zone of each line in the document. Preliminary experiments showed that not using CRF either slightly deteriorates model

performance or does not have an impact on the results.

4.3 Training setup

During training, XLM-RoBERTa's weights were kept frozen and only the BiLSTM and CRF layers were updated. We experimented BiLSTM with 16, 32, 64, 128, 256 and 512 hidden units and more layers, but in the end, having a small segmentation module, with 64 hidden units and 1 layer, generically yielded the best performances in the validation splits. We used a dropout layer of value 0.25 between the BiLSTM and the CRF, and the RMSprop optimizer with a fixed learning rate of 0.001.

5 Results and Discussion

In this section, we analyse both multilingual and monolingual capabilities of OKAPI, considering various zoning corpora and annotation schemas.

5.1 Multilingual Email Zoning

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
Personal Sig.	0.81	0.87	0.79
Visual Sep.	0.92	0.90	0.96
Quot. Mark.	0.55	0.97	0.97
Closing	0.59	0.58	0.69
Log Data	0.56	0.53	0.57
Raw Code	0.54	0.74	0.84
Inl. Head.	0.78	0.77	0.58
Salutation	0.65	0.69	0.89
Tabular	0.30	0.00	0.60
Technical	0.67	0.56	0.48
Patch	0.00	0.00	0.00
Sec. Head.	0.34	0.00	0.00

Table 5: Multilingual zero-shot evaluation of OKAPI, using Cleverly zoning corpus. Global accuracy and recall of each zone, computed by averaging the scores regarding both annotators.

We evaluate the multilingual capabilities of OKAPI in a zero-shot fashion. For that, we trained the model with the Gmane English corpus released by Bevendorff et al. (2020), and tested it with the Cleverly multilingual corpus that we annotated for Portuguese, Spanish and French.

Table 5 presents the performances of OKAPI in our multilingual corpus for each zone. Comparing with the typical performance of email zoning and the Gmane corpus (see next Tables), OKAPI achieves quite reasonable performances, confirming its multilingual character. As expected, zone recall seems to be dependent on the total number of lines per zone.

5.2 English Email Zoning

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**

Table 6: Email zoning accuracy of various models, for the corpus of Repke and Krestel (2018).

Model	Zones	Gmane	Enron
Tang et al. (2005)	15	0.80	0.73
QUAGGA	15	0.94	0.83
CHIPMUNCK	15	**0.96**	**0.88**
OKAPI	15	**0.96**	**0.88**

Table 7: Zoning accuracy of various models, under the 15-level zoning schema of Bevendorff et al. (2020).

Resorting to the numbers reported in the literature for email zoning, we compared OKAPI with existing monolingual methods using various English corpora and zoning taxonomies. In particular, Table 6 compares OKAPI with other zoning systems on the corpora annotated by Repke and Krestel (2018) with 2 and 5 types of zones; and Table 7 shows the results obtained with the most recent and fine-grained annotation schema with 15 zones proposed by Bevendorff et al. (2020). For all those combination of corpora and zoning strategies, OKAPI achieved competitive, and sometimes better results when compared with state-of-the-art methods for English email zoning, being simultaneously able to perform well on different languages.

Finally, we analyse how OKAPI adapts to new domains. For that, Table 8 shows the performance of both OKAPI and QUAGGA (Repke and Krestel, 2018), when evaluated in a different corpus then

Model	Corpus Train/Test	Accuracy 2 zones	Accuracy 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**

Table 8: Comparison between OKAPI and QUAGGA for domain adaptation, considering Repke and Krestel (2018) 2 and 5 zoning schema.

the one they were trained on. In these experiments, OKAPI clearly outperformed QUAGGA, indicating a superior ability to generalize to unseen domains.

6 Conclusion

To overcome the English-centric email zoning literature we propose OKAPI. Besides having multilingual capabilities, the proposed model is competitive with existing approaches for English email zoning, and attained state-of-the-art performance in domain adaptation tasks of English email zoning. Futhermore, to evaluate our model and to foster future research into multilingual email zoning, we release Cleverly zoning corpus – a corpus with 625 emails annotated in Portuguese, Spanish and French.

7 Acknowledgments

References

Nicolas Bettenburg, Bram Adams, Ahmed E. Hassan, and Michel Smidt. 2011. A lightweight approach to uncover technical artifacts in unstructured data. In International Conference on Program Comprehension, pages 185–188, Los Alamitos, CA, USA. IEEE Computer Society.

Janek Bevendorff, Khalid Al Khatib, Martin Potthast, and Benno Stein. 2020. Crawling and preprocessing mailing lists at scale for dialog analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1151–1158, Online. Association for Computational Linguistics.

Vitor R. Carvalho and William W. Cohen. 2004. Learning to extract signature and reply lines from email. In CEAS - 2004 (Conference on Email and Anti-Spam), Mountain View, CA, USA.

Hao Chen, Jianying Hu, and Richard W. Sproat. 1999. Integrating geometrical and linguistic analysis for email signature block parsing. ACM Transactions on Information Systems, 17(4):343–366.

Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail smart compose: Real-time assisted writing. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, page 2287–2295, New York, NY, USA. Association for Computing Machinery.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Panayotis Christidis and Álvaro G. Losada. 2019. Email based institutional network analysis: Applications and risks. The Social Sciences, 8(11):306.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, page 1–8, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. Machine Learning, 20:273–297.

Jeffrey L. Elman. 1990. Finding structure in time. Cognitive Science, 14(2):179 – 211.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for english emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 263–272.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5):602–610.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. CoRR, abs/1508.01991.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart reply: Automated response suggestion for email. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 955–964, New York, NY, USA. Association for Computing Machinery.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In Proceedings of the 15th European Conference on Machine Learning, ECML'04, page 217–226, Berlin, Heidelberg. Springer-Verlag.

Furkan Kocayusufoglu, Ying Sheng, Nguyen Vo, James Wendt, Qi Zhao, Sandeep Tata, and Marc Najork. 2019. Riser: Learning better representations for richly structured emails. In The World Wide Web Conference, WWW '19, page 886–895, New York, NY, USA. Association for Computing Machinery.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Andrew Lampert, Robert Dale, and Cécile Paris. 2009. Segmenting email message text into zones. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, EMNLP 2009, pages 919–928. Association for Computational Linguistics (ACL).

Andrew Lampert, Robert Dale, and Cecile Paris. 2010. Detecting emails containing requests for action. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 984–992, Los Angeles, California. Association for Computational Linguistics.

Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551.

M. McHugh. 2012. Interrater reliability: the kappa statistic. Biochemia Medica, 22:276 – 282.

Julia Proskurnia, Marc-Allen Cartright, Lluis Garcia-Pueyo, Ivo Krka, James B. Wendt, Tobias Kaufmann, and Balint Miklos. 2017. Template induction over unstructured email corpora. In Proceedings of

the 26th International Conference on World Wide Web, WWW '17, page 1521–1530, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Aziz Qaroush, Ismail M. Khater, and Mahdi Washaha. 2012. Identifying spam e-mail based-on statistical header features and sender behavior. In Proceedings of the CUBE International Information Technology Conference, page 771–778, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tim Repke and Ralf Krestel. 2018. Bringing back structure to free text email conversations with recurrent neural networks. In European Conference on Information Retrieval, pages 114–126. Springer.

Jie Tang, Hang Li, Yunbo Cao, and Zhaohui Tang. 2005. Email data cleaning. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, page 489–498, New York, NY, USA. Association for Computing Machinery.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30, pages 5998–6008. Curran Associates, Inc.