Building A Corporate Corpus For Threads Constitution

Lionel Tadonfouet Tadjou, Fabrice Bourge, Tiphaine Marie, Laurent Romary, Eric Villemonte de La Clergerie

Building A Corporate Corpus For Threads Constitution

Lionel Tadonfouet Tadjou[1,2,3] **Fabrice Bourge**[1] **Tiphaine Marie**[1]
Laurent Romary[2] **Eric De La Clergerie**[2]
[1]Orange Caen, [2]Inria Paris, [3]Sorbonne University Paris, France
{lionel.tadjou,fabrice.bourge,tiphaine.marie}@orange.com
{laurent.romary,Eric.De_La_Clergerie}@inria.fr

Abstract

In this paper we describe the process of building a corporate corpus that will be used as a reference for modelling and computing threads from conversations generated using communication and collaboration tools. The overall goal of the reconstruction of threads is to be able to provide value to the collorator in various use cases, such as higlighting the important parts of a running discussion, reviewing the upcoming commitments or deadlines, etc.

Since, to our knowledge, there is no available corporate corpus for the French language which could allow us to address this problem of thread constitution, we present here a method for building such corpora including different aspects and steps which allowed the creation of a pipeline to pseudo-anonymise data. Such a pipeline is a response to the constraints induced by the General Data Protection Regulation GDPR [1] in Europe and the compliance to the secrecy of correspondence.

1 Introduction

Computer mediated communication (CMC) tools are used to produce and exchange an enormous amount of contents such as emails, chats and fora. But these contents are usually little or not structured, which makes the automatic extraction of knowledge difficult. In this paper, we tackle the issue of the constitution of threads in conversations within a specific company. This problem strongly relates to that of conversations disentanglement which have been addressed in several research works. Jiang et al. (2018) computed message similarity using language representation to address this problem. Such problems requires the existence of a corpus to analyse. It is why Kummerfeld et al. (2019) have built a large-scale corpus for conversations disentanglement, also the ENRON corpus was created by Klimt and Yang (2004) for email classification research. In the same way Chanier et al. (2014) built the CoMeRe corpus for French and Douglas et al. (2015) created the Avocado Research Email Collection. More recently Bevendorff et al. (2020) built the Webis Gmane Email Corpus from the archive of Gmane mailing list. Among these corpora, some are from ENRON and AVOCADO companies, but they are in English and their content have specific contexts different from those of the company of which we will create a corpus. Those that are in French have very few emails around 2030 (CoMeRe) or are mailing list conversations with themes and context that do not fit with corporate conversations. One approach could be transfer learning, (Bornea et al., 2020) use multilingual transfer learning for question answering problem. But the context of a company is very circumscribed and specific.

In this paper we describe a method for building a corporate corpus based on emails data. We specifically address the process of collecting, preprocessing and pseudo-anonymizing emails in order to meet legal, personal, acceptability and social constraints. This comprises the use of models for Named Entity Recognition (NER), rule-based systems and their evaluation. In the upcoming sections, we will present research works on corpus building and constraints related to it.

2 Related Work

Building a corpus is a prerequisite step to various research work in different fields such as discourse analysis, automatic translation and transcription, image recognition, emails segmentation, named entities recognition, etc. Chanier et al. (2014) built the CoMeRe corpus for discourse analysis and carried out linguistic studies of idiolects that appear

[1] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en

in different types of CMC. The CoMeRe corpus consists of heterogeneous CMC in French including 3 million chats, 44K SMS, 2,030 emails, 2,700 forum messages and 34k Tweets.

Automatically classifying emails in specific email directories and extracting information from them are issues that led to the creation of the ENRON Corpus by Klimt and Yang (2004), a corpus of emails in English from a private company. The ENRON corpus has been widely used: email formalities in work environments are studied by Peterson et al. (2011), also Chhaya et al. (2018) quantify the feelings and tone used in emails.

In order to disentangle conversations in the same message stream, Kummerfeld et al. (2019) developed a large corpus of 77,563 messages including 74,963 from Ubuntu's Internet Relay Chat (IRC) channel and 2,600 messages from Linux's that highly contribute to research in dialogue analysis.

Recently, Bevendorff et al. (2020) developed the largest corpus available named Webis Gmane Email Corpus 2019 consisting of 153 million emails extracted from 14,669 mailing lists. This corpus data exists in three major languages: English, German and French. These emails are segmented into 15 classes of emails by a neural model with a performance of 96%.

As a whole, the scarcity of reference data in general and in particular for the French language justifies our approach to build a corpus that covers the need of thread analysis in CMC context.

3 Constraints: Legal, personal and psychological

Certain conditions govern the exploitation of public data, including forum, mailing list or web site data. These types of conditions are generally displayed to and agreed upon by the user when registering for these forums and or mailing lists. Mentions contained in these terms and conditions often refer to the exploitation of user data for the purposes of training, service improvements and or advertising. Unlike this process of joining forums, mailing lists or websites, employees in corporate do not always have to explicitly validate such conditions, but are required to keep secret certain company information on the one hand and on the other hand the private information exchanged in corporate via CMC are governed by the secrecy of correspondence and the General Data Protection Regulation (GDPR) adopted in 2016 and became effective on 25 May 2018 in the European union.

Personal data protection concerns have been elicited since 1970 Hallinan and Zuiderveen Borgesius (2020), with Ann Cavoukain, the Privacy Commissioner of Ontario, Canada, co-author of the 1995 international report on Privacy Enhancing Technologies (PET) acting as a forerunner in the domain of data privacy in any technological design. Cavoukian (2010) further clarified that "the future of privacy cannot be assured solely by compliance with regulatory frameworks; rather, privacy assurance must ideally become an organization's default mode of operation".

With regard to the secrecy of correspondence, it would be a violation if a third party managed to access even a conversation of collaborators without their consent. Our goal of disentangling forum conversations, chats and corporate emails could only be carried out under the respect of the secrecy of correspondence. Thus obliging us to submit consent requests to the various collaborators involved in these conversations. This is a tedious task given the number of collaborators in the different teams or entities in a company.

In the same vein, the GDPR was established in the European Union territory to protect personal data. It applies to any organisation, public and private, that processes personal data on its behalf or not, provided that it is established in the territory of the European Union, or that its activity directly targets European residents. The GDPR strengthens the obligation of information and transparency towards the persons (collaborators in our case) whose data we process. Despite this obligation of transparency towards collaborators, they are able to exercise their rights of refusal, withdrawal (if committed beforehand) of exploitation of their data. Secure and substitute some information from this data are part of the conditions to be respected in order to be GDPR compliance. These conditions are binding for the data acquisition process because require to set up stable and streamlined processes that include administrative and material aspects.

The two previous forms of constraints put collaborators at the centre of any data acquisition process in a company. They need to know why and how their data will be used. And depending on the answers given to them, they are free to give their consent or not. Unwillingness to share their content which potentially includes personal or highly sensitive data, brings out another group of con-

straints that are personal, acceptability and even psychological.

All these constraints need to be addressed when dealing with business data acquisition and processing.

4 Methodology and Implementation

Our process is based upon three main stages: collecting conversations, preprocessing them and pseudo-anonymizing sensitive information in these conversations. For the global aim of conversation disentanglement, it is important to mention that the data will be further analysed by a limited number of people from two different teams, one within the company and another the associated public research team. Our goal is to process the data, with anonymization or pseudo-anonymization techniques in particular, so that it is not possible for any person outside that team to identify any participant in the conversations nor to be able to detect participants' personal data. In order to verify our compliance with GDPR and the secrecy of correspondences, our project has gone through an internal validation procedure supervised by a legal commission within the company on the basis of a data life-cycle and a completed risk analysis form. The data life-cycle describes in detail which operations would be performed by which person acting with what role. In particular it helped us identify the possible leaks in the process and their potential consequences. It also specifies that before the anonymization or pseudo-anonymization stage, only specific researcher, who bears the role of *technical data controller (TDC)* - can access the whole collected data on a secure storage device. The other collaborators in the team can only access the data collected from their own computer. This legal commission evaluated defined the process to follow.

This process consists first in obtaining a consent from the targeted collaborators. Then, once the data has been collected, any text string that directly or indirectly allows identification of a person or access to personal information must either be removed, modified or pseudo-anonymized. It is worth mentioning that the consent provided by the collaborators is valid for a maximum of one year at a time.

To obtain collaborators' consent, we produced media content (video and slides) explaining why and how the process of collecting and analysing data would take place. This is to provide transparency, as requested by GDPR. Obtaining a collaborator's consent means having him/her sign a document explaining why we want to collect emails he/she is involved in and under what conditions it is done. For COVID-19 reasons these documents were signed electronically and stored on a secure file system.

4.1 First stage: Collecting conversations

Since our objective was to collect a large corpus of French interactions occurring in the context of computer mediated exchanges in a business environment, we focused specifically on emails since a) these are still widely used in professional context and b) other sources appeared to be more fragmented and difficult to gather in a coherent way. Our main goal is to reconstruct threads from those conversations so that each thread only deals with a clear and well-circumscribed topic. After dealing with emails conversations, we will then explore other types of CMC in corporate such as forums, chats, IM, etc.

Still, the sole gathering of emails can appear to be a painful task. Beyond the difficulty of just gathering users' consents, the resulting issue is to be able to collect those messages or threads that only relate to the collaborators that have actually given their consent since obviously many threads involve several collaborators. Finally, the technical deployment of the mailing environment (in our case: Microsoft's Exchange server) can in itself be a hindrance to the precise querying and retrieval of the appropriate content.

As a consequence, we chose to collect emails directly from the collaborators' desktop computers. This choice was guided by two main reasons, namely:

- It is easier to obtain consents of some targeted collaborators. Aspects of security and confidence are in this way easily approached and collaborators have the full control of what is collected from their computer.
- It also allows us to limit the extraction to relevant emails, thus preventing that we come across, for instance, emails that the collaborator wants to keep confidential.

In order to build a corpus of significant size it is necessary to have the consent of a minimum number of collaborators. However targeting collaborators

at random proved not to be sufficient. Indeed since we can only collect emails for which all the collaborators involved gave their consent, we would then take the risk of collecting conversations with many missing emails. This would probably result in having a negative impact when extracting meaningful threads from these amputated conversations. Our approach was to try to maximise the number of emails/conversations to extract while minimising the number of collaborators to contact. In doing so we distinguished collaborators who gave their consent into two groups, namely: - *anchor*, are collaborators within our project and with whom we can perform manipulations on their workstation to extract data; - *participant*, the rest of those collaborators who gave their consent. The above approach is divided into two steps. The first one consists in selecting *participant* who are"close" to an *anchor*.

The output of this first step is a list of interlocutors that will be used as a guide when selecting potential *participant* or anchors on other *anchor*' workstations. The second step is focusing on the actual extraction of emails and conversations on *anchor's* workstations. During this emails/conversations extraction, we also made sure that we extrated metadata such as senders, receivers, emails and conversations identifiers, dates and times from emails headers. These metadata helped us later for the pseudo-anonymization step and to keep the real links between the different messages of a conversation. To bootstrap this phase we first contacted collaborators from a single team, then from a project, from a larger entity and so on until a substantial amount of data would be collected.

In a more practical way and because Microsoft Outlook is the application used in the company for emails exchanges, we developed a C# Windows Presentation Foundation (WPF)[2] application based on *Microsoft Office Interface*[3] that allows us to interact with Outlook application and also with backup files from Outlook. With this tool we can select and unselect folders and contacts from a collbaborator's Outlook mailbox. Once the tool is installed on a *anchor's* workstation, the *anchor* uses it with the support of one of the people who designed the tool. Thus it is easy for an *anchor* to choose the emails folders from which the emails may be extracted and to choose the collaborators that could be contacted to become participants. This gives the possibility to avoid folders with private or personal contents and contacts with whom he/she has confidential exchanges. It also helps to provide data in JSON and CSV formats for future processing.

Until now we have obtained 78 consents from close collaborators involved in shared projects. Based on these 78 consents and as a first stage of data collection, we succeeded in extracting 11K unique emails for a total of 1 023 736 tokens, from the workstations of four *anchors* (the TDC and three other collaborators of the same team). These emails included conversations dating back to 2013 which shows the extent and richness of the information collected. The current size of this extracted data is 194 MB. These extracted data are only a first part of our final corpus, as we will contact collaborators from other teams and involved in other projects to increase the size of this corpus.

4.2 Second stage: Preprocessing collected data

Emails are not structured due to the explosion of different email formats and styles, coupled with the ad hoc ways in which people vary the structure and layout of their messages Carvalho and Cohen (2004). Emails contain different zones that can be easily identified by humans. Estival et al. (2007) identify five categories of zone within emails, namely: *Author Text, Signature, Advertisement*(automatically appended advertising), *Quoted Text* (extended quotations), and *Reply Lines* (including forwarded and reply text). Three years later Carvalho and Cohen (2004) refined and extended to nine categories. Two of them caught our attention: there are **Quoted Conversation Zones** (reply and forwoard message) and **Boilerplate Zones** (*signature*, *advertising*, *disclaimer* and *attachment*).

Preprocessing collected data initially consisted in deleting all duplicated emails, followed by identifying, dissociating or removing content from the quoted conversation and boiler-plates zones. To identify and extract quoted messages and signatures from an email, we used Talon[4], a python library. This library is inspired from the research work of Carvalho and Cohen (2004) and Joachims

[2] https://docs.microsoft.com/en-us/dotnet/desktop/wpf/overview/?view=netdesktop-5.0

[3] https://docs.microsoft.com/en-us/dotnet/api/microsoft.office.interop.outlook.application?view=outlook-pia

[4] https://github.com/mailgun/talon

(2001). As provided, this library works quite well for French e-mails both for the extraction of quoted messages and signatures with an average of predictions of 95% and 70%, respectively, despite the fact that the model within this library has been trained on ENRON corpus, a corpus in English.

Figure 1 shows the large gap between the number of tokens of an email with its quoted messages(in red on the figure) and without them (in blue). These quoted messages are actually the content of all or some previous emails in a conversation that are added to a new email.

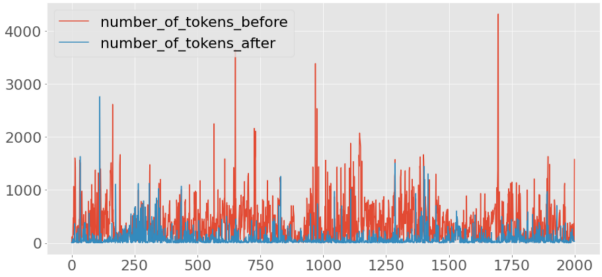

Figure 1. Number of tokens per email before and after quoted messages and signatures extraction for 2k emails sample

To comply with GDPR and the secrecy of correspondence, we started by replacing anything in the email body and subject that could indirectly lead to the identification of any person. Here, we were rather in an anonymisation process as we were trying to keep the semantic logic of sentences while substituting expressions. This meant replacing each email address, phone number, URL, path and any user's identifier respectively by *@EMAIL*, *@PHONE*, *@URL*, *@PATH* and *@ID* using python's internal *regular expression library*[5]. *@PATH* and *@ID* correspond respectively to any absolute or relative path of a folder or file and to sequences of text that uniquely identify the corporate collaborators. All these anonymized sequences do not have a great importance for the problem of thread reconstruction.

Within our collected data, emails are identified by two alphanumeric strings, namely:

- **ConversationID**: identifies a conversation, so it is the same for all emails within the same conversation
- **ConversationIndex**: is the unique identifier of an email, it contains its conversation identifier (*conversationID*). It helps to know what the position of an email is in the conversation tree.

We also replace those two strings for each email while keeping emails in their respective conversation and all relationships between emails.

This preprocessing step gives us an overview of the substitution method on strings easily identifiable. There are others whose identification is complex and we must find and substitute them. That is why we have conducted experiments of data annotation, Named Entity Recognition (NER) and disambiguation of collaborators names.

4.3 Third stage: Entities pseudo-anonymization

Besides indirect content which leads to the identification of a person in emails body, there are strings directly linked to collaborators such as their first names and last names. Emails body may also contain corporate sensitive data or intellectual property such as project names, tools, groups and/or entities names within it, conversation identifiers or any other kind of identifiers. Because of GDPR and secrecy of correspondence compliance, the identification and substitution of these data is necessary.

Identify direct strings listed above or sensitive data is become a general problem for a few decades called Named Entity Recognition (NER) in the field of NLP. A new state of art for French Named Entity Recognition has been recently established by Ortiz Suárez et al. (2020). However, some prior work has been carried out on NER in context of informal text such as emails. Minkov et al. (2005) propose a method based on Conditional Random Fields (CRF) and combined with rule-based system to extract personal names from emails. Zhang et al. (2018) use regular expressions to collect weak labels from web noisy data for the entity mentions and train a neural network to predict those RE-generated weak labels. These work show that regular expressions, rule-based system are still viable approach to Named Entity Recognition despite good results of recent transformers models which need some fine tuning. As our collected data could not be move on a cloud or GPU server for fine-tuning, we were constraint to approach the problem with a simple computer. We combine several approaches including annotation, CRF, fine-tuned **CamemBERT**[6] on NER task, regular

[5] https://docs.python.org/3/library/re.html

[6] https://camembert-model.fr/

expressions and rule-based system to achieve our goal of identifying and substitute information to be GDPR compliance.

4.3.1 Data annotation

Building up the ground truth for a downstream NLP task is a necessary step and usually corresponds to annotation which is time consuming especially when it is handcrafted. For NER case, tools have been built to accelerate annotations. **Prodigy**[7] and **INCEpTION**[8] are such tools. These are both accessible via web interfaces. In addition *Prodigy* offers command-line features. They are all based on *active learning (AL)* defined by Ren et al. (2020) as a method that aims to select the most useful samples from the unlabelled data set and hand it over to the "oracle" (e.g., human annotator) for labelling, so as to reduce the cost of labelling as much as possible while still maintaining performance. As *prodigy* is not free and *INCEpTION* needs some time to get hands on it, we used a tool built by a former intern of our team.

Tags of Named Entities	**To annotate**
PERSON (B-, I-)	first name, surname, full name and short forms for these
ORG (B-, I-)	known corporate
SUBGROUP(B-,I-)	company entities, departments, etc.
EMAIL	For email identification
PHONE(B-, I-)	phone number
PROJECT (B-, I-)	project names in the company
LOC (B-, I-)	any type of locations
ID	any identifier linked to a person
ROLE (B-, I-)	job names (Manager, Engineer, etc.)
UNCERTAIN	any identified entity that does not fall into the previous tags

Table 1. Tags used to annotate 1k emails

This tool is based on Conditional Random Fields (CRF) and active learning method and as well as being accessible via web interface. From the 11K extracted emails, we selected 1k emails containing at least 142 916 tokens. Three annotators did the annotation exercise on the 1k emails each with Named Entities(NE) tags we defined based on what we considered sensitive information to substitute. Those annotators come from the pool of *anchor* collaborators from whom we extracted the 11k emails and they annotated a total of 8689 tokens.. They are thus well aware of the general context of the task and therefore ended up being good raters for the task. Table 1 lists the various tags according to the BIO standard. Figure 2 shows the annotation statistics. We can see the very low rate of *ID*, *ROLE*, *LOC* and *ORG*, which is actually due to the pre-processing stage where we extracted signatures that usually contain identifiers (which have been annotated as *ID*), roles or functions, phone number, and addresses.

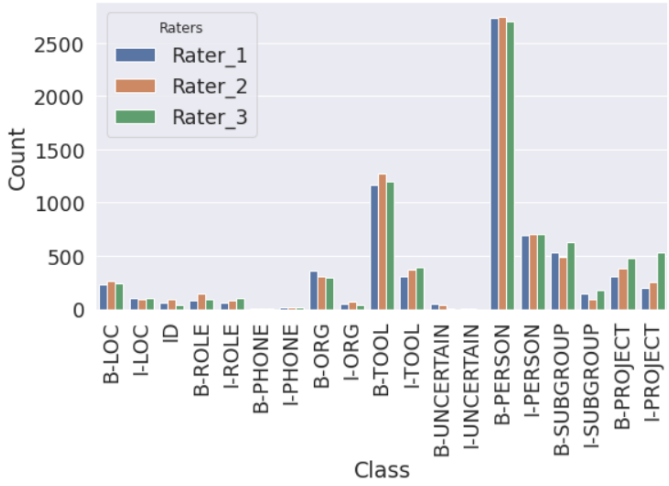

Figure 2. Statistics of annotations by three annotators

The very low rate of *phone number* is also related to the pre-processing stage described in section 4.2 with the use of regular expressions to identify and substitute *phone number*, *email address*, *url* and *path* (which sometimes contains the company ID of employees). Also *UNCERTAIN* is low because raters know well the context of emails that they annotated, otherwise we would have had a high number of *UNCERTAIN* tags. This shared knowledge of the context of annotated emails is also observed on several tags (*LOC*, *ID*, *TOOL*, *PERSON*) which are not very disproportionate for raters 1 and 2. Rater 3 annotated more *PROJECT* and *TOOL* than two others raters, this shows some ambiguity on strings related to those two tags. This effectively reflects the reality of distinction between project and tool because sometimes in a company, a tool developed in a project can bear the same name of the project. There is also such ambiguity between *SUBGROUP* (for departments or teams in the company) and *PROJECT*, due to the fact that sometimes a team name is identical to the product being developed. From the 8689 annotated tokens, we identified 236 different annotations among the three raters, in which the *SUBGROUP*, *PROJECT* and *TOOL* categories are predominant. These disparities show that certain concepts in corporate context are not accurate and may be subject to mis-

[7] https://prodi.gy/
[8] https://inception-project.github.io/

understanding.

For data annotation, we evaluated inter-annotator agreements (IAA) by means of the standard Cohen Kappa measure McHugh (2012) from the biomedical field. Cohen's Kappa measure is not the most relevant for NER as mentioned e.g. in Hripcsak and Rothschild (2005); Grouin et al. (2011), because it requires negative cases that do not exist for NER. Also Cohen's Kappa measure can not be used when there is more than two annotators. Since there are three raters for our data, we use Cohen's Kappa to evaluate pairs of raters (R_1, R_2), (R_1, R_3) and (R_2, R_3). To evaluate nominal inter-annotator agreements (IAA) with more than two annotators, Zapf et al. (2016) advise to use **Fleiss's Kappa** or **Krippendorff's alpha**. They study different cases and show that these measures are similar. Annotated data is very unbalanced because of the un-annotated tokens labelled with "O" which are highly represented compared to all existing tokens. As Brandsen et al. (2020) did, we computed all IAA scores in both cases with all tokens and only with annotated tokens. Table 2 shows different scores computed, and we observe that the scores with all tokens are quite high, but this is due to the bias of non-labelled tokens. Cohen's Kappa on annotated tokens only is **substantial** agreement with values in $[0.61 - 0.80]$ interval according to the interpretations in Viera and Garrett (2005) for three pairs of raters. Computed values for both Krippendorff's alpha and Fleiss's Kappa are identical 0.70 and belong to the same previous interval, interpret as substantial agreement. Based on these computed measures and their interpretation, we can find some correlations with diagram on the Figure 2. These annotations made it possible to build a repository of annotated tokens that we combined with CamemBERT-ner results and a rule-based system to pseudo-anonymize our entire data. This combination is what we call *data pseudo-anonymization chain* represented by Figure 3.

4.3.2 Data pseudo-anonymization chain

Pseudo-anonymising sensitive information starts with recognising them, and this is done by NER task. As we use a tool based on CRF coupled with active learning to annotate our data, we test the resulting model on unseen emails and the result was very bad compared to the test performed with **CamemBERT-ner**[9], a transformers based model.

[9] https://huggingface.co/Jean-Baptiste/camembert-ner

	(R_1,R_2)	(R_1,R_3)	(R_2,R_3)
Cohen's Kappa*	0.8879	0.8450	0.8344
Cohen's Kappa #	0.7832	0.6959	0.6690
Krippendorff's alpha*	0.8554		
Krippendorff's alpha #	0.7155		
Fleiss's Kappa*	0.8554		
Fleiss's Kappa #	0.7158		

Table 2. Inter-annotator agreement measures on 1k emails with Cohen's Kappa, Krippendorff's alpha and Fleiss's Kappa. R_i *stands for* $Rater_i$, $i \in \{1, 2, 3\}$; * For all tokens and # For annotated tokens only

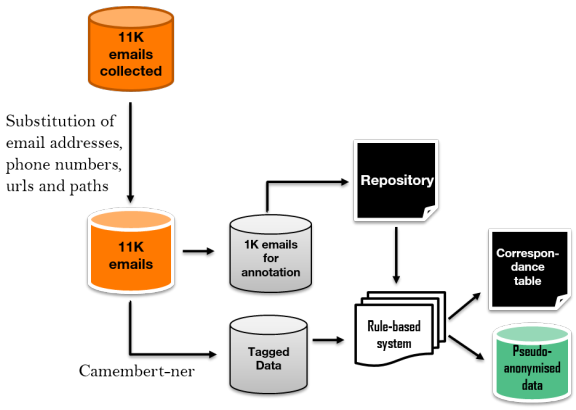

Figure 3. Data pseudo-anonymization chain

Figure 4 shows statistics of *CamemBERT-ner* model on the bunch of 1k emails data with a total number of 6952 identified strings of which 3915 are classified in the miscellaneous category (MISC).

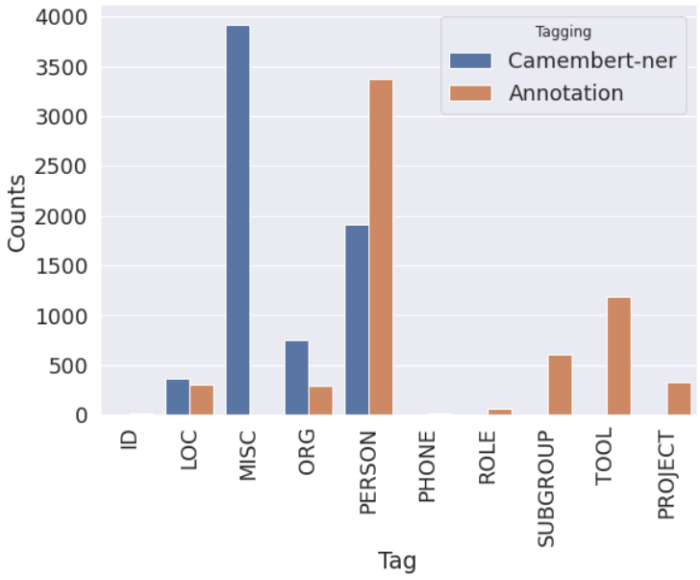

Figure 4. Number of elements identified and classified by CamemBERT-ner and tags identified by annotators on 1K emails

From the data annotation stage (carried out on the same 1k emails), there are 6219 identified strings that have been similarly annotated by the three annotators or by two of them. The difference of about 733 entities identified by CamemBERT-ner compared to manual annotations shows its performance in identifying elements that it cannot classify. This comes from the fact that CamemBERT-ner has been fine-tuned from CamemBERT on the wikiner-fr data set ($\sim$ 170 634 sentences). We anticipate that we will have to compare these results with those obtained with fine tuning *CamemBERT* on our own data in order to recognise additional entities (tool, project, ID, subgroup, etc.) used by raters during the annotation process.

As can be seen in figure 4, the main challenge is to be able to reclassify on these categories those from MISC. Also the organisations recognised by CamemBERT-ner are more than double of those from the annotation and need to be reclassified into other categories like subgroup or project. A rule-based system was used to refine category of entities from MISC. During the annotation process, a repository was built, containing a list of string with their respective tags (tags from table 1) without duplicates. The rule-based system consists of comparing strings identified as *miscellaneous* by *CamemBERT-ner* with same from the annotation repository and finally use his tag (defined by raters) to refine *miscellaneous* category. This Refining stage helps the rule-based system in creating semantic code for the substitution of those strings from emails content ans subjects. Created codes look like *user_xxxxxxxx*, *org_xxxxxxxx*, *tool_xxxxxxxx* respectively for identified collaborators, organisations, tools and this same coding pattern was used for all other entities classes. When substituting collaborators first names or last names or both in an email, the rule-based system checks sender or recipients of the email to keep the email and conversation thread context. While running, the rule-based system builds a new repository containing all pairs of substituted string and their respective code. This repository is call **correspondence table** and could be use later to rebuild emails and subjects without codes but strings which was substituted before. Below is an example of a paragraph before and after pseudo-anonymisation.

Les espaces de co-working sont plutôt traités dans le ***Research Paper*** *en cours de rédaction par* ***Pierre***, ***Paul***, ***Louise***. *L'étude de* ***Louise*** *(seule) devait porter initialement sur la valeur pour* ***NomEntreprise*** *des tierslieux (fablabs, espaces de coworking) pour ce qui concerne l'apprentissage et la transmission de connaissances à distance. A part le recentrage sur une population interne* ***NomEntreprise***, *l'idée générale est globalement conservée.*

The above paragraph contains bold text segments that are names of collaborators and companies that have been replaced by code as in the pseudo-anonymised text below.

Les espaces de coworking sont plutôt traités dans le ***misc_55e6a*** *en cours de rédaction par* ***user_8e47d***, ***user_c6f1d***, ***user_5ff59***. *L'étude de* ***user_5ff59*** *(seule) devait porter initialement sur la valeur pour* ***org_252f2a*** *des tierslieux (fablabs, espaces de co-working) pour ce qui concerne l'apprentissage et la transmission de connaissances à distance. A part le recentrage sur une population interne* ***org_252f2a*** *, l'idée générale est globalement conservée.*

Everything that has been done so far has allowed to set up a data processing pipeline that takes data collected in CSV format as input and produces pseudo-anonymized data as output in the same format.

5 Future work

After obtaining pseudo-anonymised data, we will analyze this data in order to approach the problem of threads constitution with its inherent problems including conversation disentanglement, dialog acts identification and text segmentations. To reconstruct conversation threads, Domeniconi et al. (2016) propose an approach that combines similarity calculations of 8 *features* built from each email. These features actually bring out the context of an email. Taking into account the context of messages or emails allowed researchers of *Yahoo* and *Amazon* Avigdor-Elgrabli et al. (2018) to automatically evaluate the semantic relationship between messages within a mailbox.

Regarding conversation disentanglement, Elsner and Charniak (2010) and Elsner and Charniak (2011) approach chat disentanglement by first using binary classifier and local Coherence Models one year later. Jiang et al. (2018) take advantage of learning language representations to disentangle conversations. compute similarities between messages using a model they name ***Siamese Hierarchical Convolutional Neural Network (SHCNN)*** which is a Siamese hierarchical convolutional network.

SegBot was developed by Li et al. (2018) to propose a solution to the problem of text segmentation.

Koshorek et al. (2018) address text segmentation as a supervised learning problem and present a large dataset for this problem. Dialog acts identification could be seen as determining interlocutors intentions within a conversation. Wang et al. (2019) study the identification of intentions in emails in a workplace situation.

All these work give us some leads of experience to be carried out very soon with our data in order to approach our problem of threads constitution.

6 Conclusion

In this paper, we presented a method for building corporate corpora including different aspects and steps which allowed the creation of a pipeline to pseudo-anonymise data. Such a pipeline is a response to the constraints induced by the GDPR and the secrecy of correspondence compliance.

The process we described consists of several steps: first we prospected with our collaborators to obtain their consent agreement, this in order to be GDPR and secrecy of correspondence compliant; second we collected and pre-processed emails from Outlook mailboxes; the third step dealt with manual annotation and Named Entities Recognition; fourth, we performed data pseudo-anonymization. During the second step, we developed a tool called OutlookScraping that allowed us to collect a first batch of 11k emails on the workstations of 4 close collaborators out of 78 collaborators who gave their consent agreement.

All these steps contributed to the production of pseudo-anonymized data that we will use for our future work of thread constitution. For our constitution of threads problem, we will focus on conversations disentanglement, dialog acts identification and text segmentation.

References

Noa Avigdor-Elgrabli, Roei Gelbhart, Irena Grabovitch-Zuyev, and Ariel Raviv. 2018. More than threads: Identifying related email messages. CIKM '18, page 1711–1714, New York, NY, USA. Association for Computing Machinery.

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.

Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. 2020. Multilingual transfer learning for qa using translation as data augmentation.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Vitor R. Carvalho and William W. Cohen. 2004. Learning to extract signature and reply lines from email. In *IN PROCEEDINGS OF THE CONFERENCE ON EMAIL AND ANTI-SPAM*.

Ann Cavoukian. 2010. Privacy by design: The definitive workshop. a foreword by ann cavoukian, ph.d. *Identity in the Information Society*, 3(2):247–251.

Thierry Chanier, Céline Poudat, Benoît Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for language technology and computational linguistics*, 29(2):1–30. Final version to Special Issue of JLCL (Journal of Language Technology and Computational Linguistics (JLCL, http://jlcl.org/): BUILDING AND ANNOTATING CORPORA OF COMPUTER-MEDIATED DISCOURSE: Issues and Challenges at the Interface of Corpus and Computational Linguistics (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel).

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Giacomo Domeniconi, Konstantinos Semertzidis, Vanessa Lopez, Elizabeth M. Daly, Spyros Kotoulas, and Gianluca Moro. 2016. A novel method for unsupervised and supervised conversational message thread detection. In *Proceedings of the 5th International Conference on Data Management Technologies and Applications*, DATA 2016, page 43–54, Setubal, PRT. SCITEPRESS - Science and Technology Publications, Lda.

Oard Douglas, Webber William, A. Kirsch David, and Golitsynskiy Sergey. 2015. Avocado research email collection. Linguistic Data Consortium (LDC).

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*, 36(3):389–409.

Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1179–1189, Portland, Oregon, USA. Association for Computational Linguistics.

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

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

Dara Hallinan and Frederik Zuiderveen Borgesius. 2020. Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle. *International Data Privacy Law*, 10(1):1–10.

G. Hripcsak and A. S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc*, 12(3):296–298.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to disentangle interleaved conversational threads with a Siamese hierarchical network and similarity ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana. Association for Computational Linguistics.

Thorsten Joachims. 2001. A statistical learning learning model of text classification for support vector machines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 128–136, 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 *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg. Springer Berlin Heidelberg.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4166–4172. International Joint Conferences on Artificial Intelligence Organization.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282.

Einat Minkov, Richard C. Wang, and William W. Cohen. 2005. Extracting personal names from email: Applying named entity recognition to informal text. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 443–450, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. 2020. Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4631–4638, Marseille, France. European Language Resources Association.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon. Association for Computational Linguistics.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2020. A survey of deep active learning.

A. Viera and J. Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Family medicine*, 37 5:360–3.

Wei Wang, Saghar Hosseini, Ahmed Hassan Awadallah, Paul N. Bennett, and Chris Quirk. 2019. Context-aware intent identification in email conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 585–594, New York, NY, USA. Association for Computing Machinery.

Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. 2016. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16(1).

Shanshan Zhang, Lihong He, Slobodan Vucetic, and Eduard Dragut. 2018. Regular expression guided entity mention mining from noisy web data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1991–2000, Brussels, Belgium. Association for Computational Linguistics.