De-identification of Emergency Medical Records in French: Survey and Comparison of State-of-the-Art Automated Systems

Loïck Bourdois[1], **Marta Avalos-Fernandez**[1,2], **Gabrielle Chenais**[1], **Frantz Thiessard**[1], **Philippe Revel**[1,3], **Cédric Gil-Jardiné**[1,3], **Emmanuel Lagarde**[1]

[1]University of Bordeaux, Bordeaux Population Health Research Center, UMR U1219, INSERM, F-33000, Bordeaux, France

[2]SISTM team Inria BSO, F-33405, Talence, France

[3]University Hospital of Bordeaux, Pole of Emergency Medicine, F-33000, Bordeaux, France

{first name.last name}@u-bordeaux.fr

Abstract

In France, structured data from emergency room (ER) visits are aggregated at the national level to build a syndromic surveillance system for several health events. For visits motivated by a traumatic event, information on the causes are stored in free-text clinical notes. To exploit these data, an automated de-identification system guaranteeing protection of privacy is required.

In this study we review available de-identification tools to de-identify free-text clinical documents in French. A key point is how to overcome the resource barrier that hampers NLP applications in languages other than English. We compare rule-based, named entity recognition, new Transformer-based deep learning and hybrid systems using, when required, a fine-tuning set of 30,000 unlabeled clinical notes. The evaluation is performed on a test set of 3,000 manually annotated notes. Hybrid systems, combining capabilities in complementary tasks, show the best performance. This work is a first step in the foundation of a national surveillance system based on the exhaustive collection of ER visits reports for automated trauma monitoring.

Introduction

Hospital emergency room (ER) data are one of the main data sources in syndromic surveillance. In France, the data transmitted in routine to local, regional and national health agencies concern demographic data (date of birth, gender, location), temporal data (date and time of arrival at the ER), medical diagnoses (coded using the 10th revision of the International Classification of Diseases diagnostic codes, ICD10) and outcome orientation (hospitalization or discharge). Surveillance of injuries is however unviable as it requires data on the circumstances and mechanisms: intentional/unintentional, self-inflicted/other-inflicted, etc.

Yet, the cause for the visit and injury mechanisms are fully described with free-text narratives stored in electronic health records. Narrative text fields from injury databases began to be used to extract useful epidemiological data more than two decades ago but only recently the operationality of the systems has become realistic (Marucci-Wellman, Corns, and Lehto 2017; Chen et al. 2018; Wang et al. 2020). In France, the potential exploitation by natural language processing (NLP) techniques of the more than twenty million per year unlabeled ER notes has been pointed out by a few research teams. Gerbier and colleagues proposed an automated extraction and encoding of information from the ER clinical notes for intra-hospital syndromic surveillance purposes (Gerbier et al. 2011). Metzger and colleagues evaluated the improvement in the estimation of the incidence rate of suicide attempts when using automated extraction and processing of computerized ER records compared with current manual coding by ER physicians (Metzger et al. 2017). Our team evaluated the feasibility of adapting a multi-purpose neural language model (NLM) to classify free-text ER notes (Xu et al. 2020), and showed that manual annotation requirements (generally time-consuming and prohibitevely expensive (Spasic and Nenadic 2020)) could be substantially reduced by unsupervised pre-training.

Unsupervised pre-training has recently achieved high levels of performance in the domain of NLMs by applying the concept of attention that consists in learning dependencies between words in a sentence without regard to their distances. This mechanism was implemented in a sequence to sequence neural network model, the Transformer architecture (Vaswani et al. 2017). However, the application of NLM to the classification of ER notes faces several challenges including the high number of required expert annoted samples, the limited clinical corpus in French, the scarcity of French-adapted NLP models and the key issue related to the protection of personal data.

When a secondary use of health data is planned, the protection of personal data has to be ensured according to the legislative framework (in our case, established by the European General Data Protection Regulation - GDPR, and the French data protection authority - CNIL). De-identification, consisting in separating and altering personal identifiers, is the main approach to protect privacy. This process involves two steps: detection of personal data and its replacement with surrogates or deletion. In the following we use the convention that the terms ***detection*** and ***anomymization*** indicate the first and the second tasks, respectively, and ***de-identification*** indicates the whole process.

As manual annotation, manual de-identification is time-consuming and costly, requiring automatic methods. The detection task can be viewed as a Named-Entity Recognition (NER) problem targeting personal data i.e. as the problem

of recognizing information units (like person and location names or date and telephone number numeric expressions) from free text, independently of the domain.

A benchmark of NER models on French commercial legal cases has been developed (Benesty 2019). The results encourage the use of the NER bi-directional long short term memory (Bi-LSTM) model by using the Flair library (Akbik, Blythe, and Vollgraf 2018) and the NER model of CamemBERT (Martin et al. 2020). CamemBERT is a French version of BERT –Bidirectional Encoder Representations from Transformers– (Devlin et al. 2018), which is itself based on the encoder part of the Transformer architecture (Vaswani et al. 2017). FlauBERT –French Language Understanding via Bidirectional Encoder Representations from Transformers– is another French version of BERT (Le et al. 2020), without an available pre-trained NER, yet. Finally, a BART –Bidirectional and Auto-Regressive Transformer – model (Lewis et al. 2020) for the French language (BARThez) has been freshly released (Eddine, Tixier, and Vazirgiannis 2020).

Regarding the techniques, automated de-identification systems can be rule-based, machine/deep learning-based, or hybrid (Khin, Burckhardt, and Padman ; Obeid et al. 2019; Trienes et al. 2020). Rule-based systems tend to perform better with personal data rarely mentioned in clinical texts, but are more difficult to generalize. In general, machine/deep learning-based systems tend to perform better, especially with terms not mentioned in the rules. Hybrid systems combine the advantages of both approaches. The self-attention mechanism has been newly applied to de-identification (Yang et al. 2019; Fu et al. 2020). Rule-based de-identification methods for clinical narratives previously developed in English have been adapted to French (Grouin and Névéol 2014; Névéol et al. 2018; Gaudet-Blavignac et al. 2018). Some of these works presume that hybrid systems would improve performance.

Objectives. We aim to compare the performances of several strategies to de-identify free-text clinical note, including rule-based, NER, deep learning (particularly Transformer-based) and hybrid systems available and adapted to the French clinical language used in the ER notes. The compared approaches perform the two de-identification tasks (detection of personal data and its replacement) either simultaneously or sequentially. The comparison criteria considered are recall, precision and specificity, while accounting for the reduction in time and cost.

Methods

The de-identification process which consists in identifying notes with personal data and anonymizing them was preceded by a preliminary task (figure 1) that consists in building a training set and a test set. The training set is used to fine-tune the Transformer-based FlauBERT model (Le et al. 2020) which is a French adaptation of the BERT model (Devlin et al. 2018). The test set is used to measure and compare performance of six selected strategies.

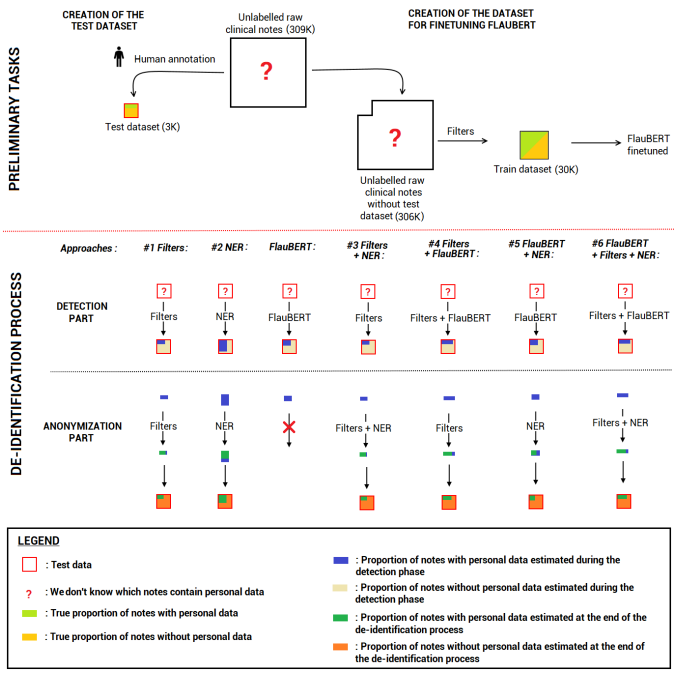

Figure 1: De-identification process scheme.

Preliminary tasks

We retrieved 309 380 unlabelled free-text clinical notes from the digital medical record system of the adult ER of the local University hospital, from 11-01-2012 to 10-16-2019.

Test data set. 3 000 clinical notes were randomly sampled from this data set to be annotated in order to become the labelled test data set. The manual annotation procedure consisted in a pilot study (that allowed to ensure that the annotators had a common understanding of the instructions and to refine the annotation grid) and subsequently, the simple annotation of the 3 000 sampled notes. A note was tagged as "with personal data", if a note contains either names, social security numbers, geographic addresses or telephone numbers (whether it concerned data relating to a patient or hospital staff). Otherwise it was tagged as "without personal data". A total of 414 clinical notes were identified with personal data.

Fine tuning data set. The FlauBERT model requires a fine tuning set. For this purpose, the remaining 306 380 notes were used as training sample. Since these notes were not annotated, we used filtering keywords to automatically build the required database. Clinical notes containing one or more of the predefined keywords were tagged in this fine-tuning dataset to be "with personal data". The list of predefined keywords is:

- *Docteur* (Doctor), *Professeur* (Professor), *Etudiant(e)* (Student)
  Healthcare students practicing at the University hospital often use an electronic signature for their reports in which their student status appear.

- *Mme* (Ms) and *Mr* (Mr)
  Spaces before the first letter and after the last one are informative (otherwise, words such as *"comme"*, meaning "as", would be falsely detected).
- Telephone numbers with the format "0000000000" or "00 00 00 00 00" or "00.00.00.00.00"

We didn't add to the list ambiguous abbreviations such as "Dr" (abbreviation of *Docteur*, Doctor, also used as an abbreviation of *droit*, right), "Pr" (abbreviation of *Professeur*, Professor, also used as an abbreviation of *pour*, for) to avoid false positive errors. To improve keywords matching, data were previously cleaned (also using filters): points that were not a period punctuating at the end of a sentence were removed. For example, "The patient saw Dr. X." became "The patient saw Dr X.". Nevertheless, to ensure fair comparisons of de-identification approaches, we used filters with discretion when cleaning the fine tuning data set. For instance, we didn't remove electronic signatures.

Filters detected $35\,991$ notes out of the $306\,380$ as including personal data. We randomly sampled $15\,000$ notes from this set of $35\,991$ notes "with personal data" and $15\,000$ notes from the remaining set of $270\,389$ notes "without personal data". The sample size of $15\,000$ was empirically determined as achieving a good trade off: less fine tuning examples involved poor coverage, on the other hand, more fine tuning examples involved high bias in the data since they were formed using keywords. Finally, we randomly mixed these $30\,000$ data.

De-identification process

Six de-identification approaches (see figure 1) were considered and evaluated on the test dataset. These approaches consist in applying filters (i.e. a series of rules defined one by one) and/or NER and/or a Transformer model and/or a combination of two or all of them.

FlauBERT
Detection. We applied the uncased version with 138M parameters available on the Transformers Hugging Face library (Wolf et al.). We used the pretrained weights provided by (Le et al. 2020). We then ran it on the $30\,000$ fine tuning data on 1 epoch with a learning rate of 0.005 on a single Nvidia® GeForce GTX 1080 Ti with 11GB of VRAM. Results are based on a majority vote carried out on 5 executions. The default threshold 0.5 was used for classification decision. This language model allows only detection and needs to be combined with another strategy to complete the anonymization procedure.

$\#1$ **Filters**
Detection. We used the same filters that were used to create the fine tuning dataset with one exception: we added the abbreviations "dr" and "pr".
Anonymization. The first word following all keywords were removed as well as electronic signatures of hospital staff.

$\#2$ **NER**
Detection. NER is particularly well-adapted to handle proper names. We used the NER model of the Flair library trained on the WikiNER base (aij-wikiner-en-wp3) (Nothman et al. 2013) using the Fasttext embedding (Grave et al. 2018). Directly available in a exceptionally user-friendly pipeline, we preferred Flair to CamemBERT which is still very new, and therefore improvable.
Anonymization. Words detected as personal data were removed by the following tags of the Flair NER model: <B-PER>, <I-PER>, <E-PER> and <S-PER>.

$\#3$ **Filters+NER**
Detection. Filters of approach $\#1$ were applied for the detection part. *Anonymization*. We first applied filters as in the anonymization step of approach $\#1$. Then the task was completed by removing the tags generated by the NER algorithm.

$\#4$ **Filters+FlauBERT**
Detection. We applied filters to clean the notes (following the same procedure as for the creation of the data set used to fine tune FlauBERT). FlauBERT was then applied to detect note with personal data.
Anonymization. We applied the same procedure using filters as in the anonymization step of approach $\#1$.

$\#5$ **FlauBERT+NER**
Detection. FlauBERT was applied as in approach $\#4$.
Anonymization. We applied NER as in the anonymization step of approach $\#2$.

$\#6$ **FlauBERT+Filters+NER**
Detection. This step was performed as in approach $\#4$.
Anonymization. This step was performed as in approach $\#3$.

Performance criteria

Performance criteria measured on the test set included recall (fraction of notes detected as including personal data from all notes with personal data and fraction of fully de-identified notes at the end of the process from all notes originally with personal data, in the detection step and at the end of the de-identification process, respectively), precision (fraction of notes with personal data among those detected as including personal data and fraction of notes fully de-identified at the end of the process among those detected as including personal data in the detection step and at the end of the de-identification process, respectively), and specificity (fraction of notes predicted as without personal data from all notes without personal data and fraction of notes that did not need to be de-identified from the process in the detection step and at the end of the de-identification process, respectively).

Recall is frequently computed as the number of correctly removed personal data with respect to the total number of personal data items, as denominator. However, this criterion doesn't resolve if a note is fully de-identified, our target.

Results

Detection and de-identification performances of the compared approaches are showed in figure (2). Circles, squares and diamonds represent, respectively, precision, specificity and recall.

Approach $\#1$ was simple to implement, fast and gave very satisfactory results. 93% of the reports to be de-identified were detected. Only six false positive predictions were made. They corresponded to notes in which the

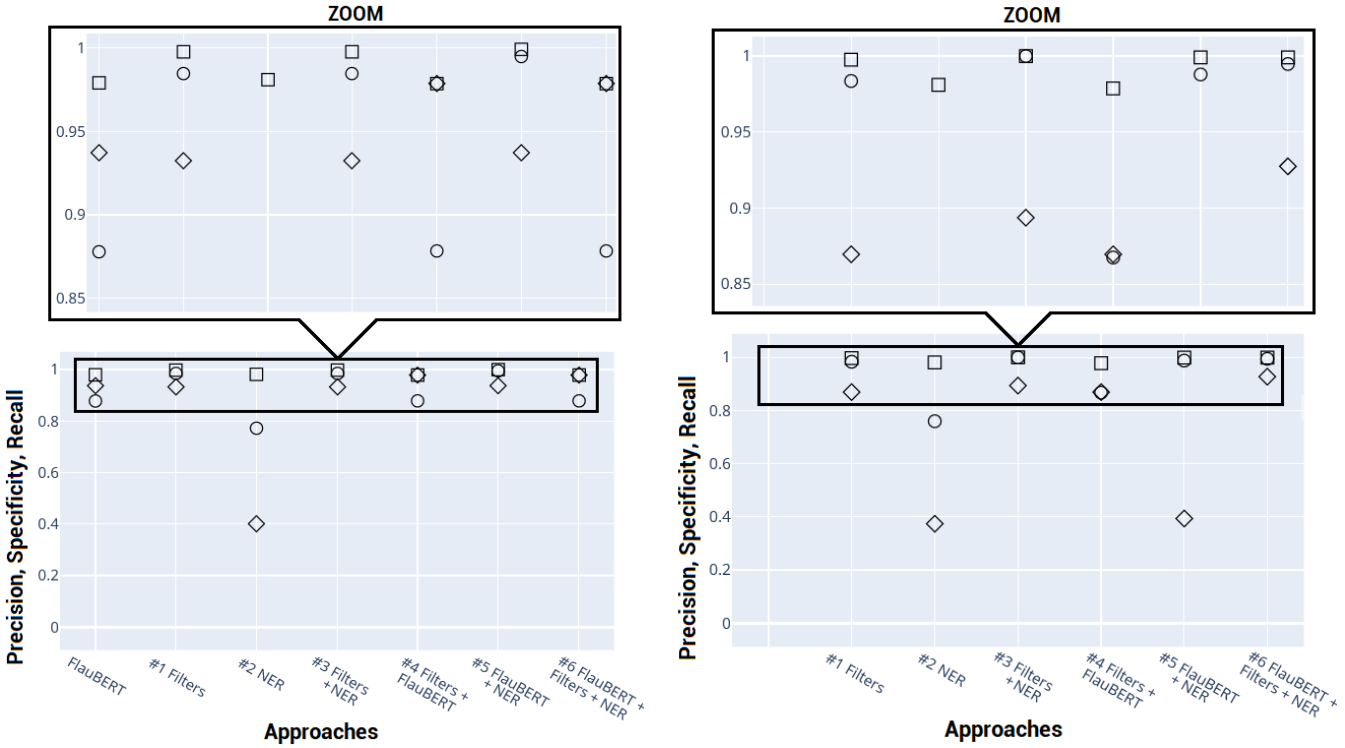

Figure 2: Precision (circles), specificity (squares) and recall (diamonds) following detection (left) and anonymization (right).

abbreviations "mr" and "mme" are generically used for "monsieur" (mister) or "madame" (madam), without being followed by a proper name. Unsurprisingly, "dr" abbreviations for "droit" (right) instead of "docteur" (doctor) leaded to errors. Approach $\#1$ anonymized 93% of the detected notes, leading to an overall 87% of all clinical notes to be anonymized. While this naive approach works for forms of type "Dr LAST_NAME", it doesn't for forms of type "Dr FIRST_NAME LAST_NAME" or "Dr COMPOUND_LAST_NAMES". This feature was present in 12 of the 386 detected notes.

Approach $\#2$ was also simple to implement and fast once the library and model weights were downloaded. However its performance proved to be poor. NER alone detected only 40% of the clinical notes to be de-identified. As a reference, when applying a single filter to remove staff's electronic signatures, 66% of the clinical notes to be de-identified were detected. In addition, approach $\#2$ led to many false positives, mostly hospital's names also corresponding to person's names and, illnesses, syndromes, diseases or body parts named after the people who discovered them.

Applying NER after the filters (approach $\#3$) improved recall. In addition, the false positive rate was dropped to 0. Indeed, when applying NER to the 9 incorrectly detected notes, no <B-PER>, <I-PER>, <E-PER> or <S-PER> tags were generated. At the end of the process, these 9 notes were thus considered as de-identified, which is their true class.

One epoch of FlauBERT' fine tuning on our dataset took about one hour and a half. Few additional notes were detected ($< 0.5\%$) when compared to approach $\#1$. FlauBERT's specificity was 10% lower.

Approach $\#4$ provided the best true positive rate in detection. However, it also presented the worst false positive rate.

Approach $\#5$ kept the recall performance of FlauBERT in detecting notes with personal data. In addition, specificity was improved, for the same reasons as approach $\#3$: when applying NER to the 54 incorrectly detected notes, <B-PER>, <I-PER>, <E-PER> or <S-PER> tags were generated for only two notes (two hospitals having person's names). The other 52 reports were therefore not anonymized. The de-identification rate remained lower than that resulting from just removing staff's electronic signatures.

Approach $\#6$ benefited from both, filters combined with FlauBERT's performance in the detection step and filters combined with NER's performance in the anonymization step. Therefore, this approach showed the highest precision, and the highest recall in the detection step as well as in the full de-identification process. Also, its specificity was among the best ones. In other words, this approach presented the lowest false positive rate (2/2586) and the highest true positive rate in detection (400/414) as well as in de-identification (387/414) (figure 3).

Figure 3 presents the distribution of the detected notes following anonymization in terms of fully, partially or not anonymized. A partially anonymized note is obtained when

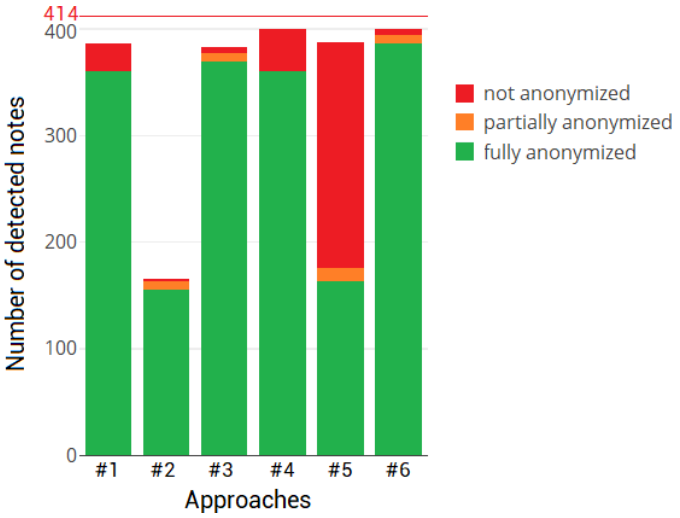

Figure 3: Distribution of the detected notes following anonymization.

several names of patients or doctors are present in the text and the corresponding approach succeeded to delete one of them but not all of them, or when a person is designated by his or her first and last name and only one of them was successfully removed.

FlauBERT's recall could be improved by varying the default probability threshold. However, when reducing the number of undetected notes from 14 to 0, the number of wrongly detected notes increased from 55 to 2100. Nevertheless, the anonymization step based on NER considerably reduced this latter amount: from 2100 to 47 (against 2, with the default 0.5 threshold). Sources of errors in these 47 wrongly anonymized notes were: commercial drug names, hospitals bearing person's names, names of diseases bearing person's names, and diseases or parts of the body not bearing person's names. With the new threshold, 45 additional notes lost substantial information leading to the de-identification of only 5 additional texts. Changing the threshold did not appear to be beneficial.

Conclusion

Because manual de-identification is time-consuming and costly, automatic methods based on well-developed clinical corpus and adapted language models are needed. We performed a survey of the available approaches to de-identify medical records in French and we compared them using ER data. We used a training set of more than 300K unlabeled clinical notes to fine-tune NLP models and a test set of 3 000 manually annotated notes to measure performance.

A hybrid system combining filters, the Transformer-based FlauBERT and NER achieved the best performance in terms of recall. It also obtained good performance in terms of specificity and precision. This confirms the assumption that hybrid systems, combining advantages, may outperform other methods (Grouin and Névéol 2014; Névéol et al. 2018; Gaudet-Blavignac et al. 2018).

These results will be instrumental in protecting personal data in secondary use of health data. In particular, to build surveillance systems based on electronic health records generated by visits to ER. Some challenges still remain.

First, it is possible to apply filters to automatically classify notes as with or without personal data to build a fine-tuning data set. Classification errors are thus likely and may have an impact on detection performance. A large dataset with manual annotations would have been of course preferable. A single annotation for only 3 000 notes was available for the present work and was used only to measure performance. A more extensive dataset, double-annotated (with a review of the team manager acting as a tie-breaker, when two reviewers disagree) to be used as test set but also for the fine tuning of FlauBERT is planned.

Second, our objective is to reach 100% in detecting notes with personal data and 100% in de-identifying them. It is inconceivable to develop a National injury observatory based on hospital ER data if notes coming from hospitals around the territory are not properly de-identified. Even if we change the classification threshold to reduce the number of incorrectly not detected notes, 22 notes remain not de-identified at the end of the process. A common practice consists in replacing personal data with realistic surrogates (Trienes et al. 2020). Also, the anonymization part may be improved by using other models. For example, a French adaptation of ELMo –Embeddings from Language Models– representations (Peters et al. 2018), trained on the OSCAR dataset (Ortiz Suárez, Sagot, and Romary 2019), has been proposed (Ortiz Suárez et al. 2020). Using NER with the OSCAR corpus rather than Wikipedia corpus has been shown to be beneficial (Martin et al. 2020). It would be also interesting to assess the system performance using a clinical corpus.

Third, while NLP applications require language resources, annotated clinical corpora in French are scarce. International ontologies are not fully translated (Névéol et al. 2018), existing corpora remain inaccessible outside the setting of the hospital that provided the data for annotation, and self-development costs are very high. This problem is shared by languages other than English.

Forth, we focused on protected information relative to names, social security numbers, geographic addresses and telephone numbers. Although unlikely, other identifiers could be present in data coming from other hospitals and should be considered.

Finally, the objective of 100% recall may involve a higher number of incorrectly detected notes. Applying de-identification systems to notes that didn't need to be de-identified could deteriorate the information. For example, when applying filters removing the word after "Mister", "Monsieur chute dans les escaliers" (Mister falls down the stairs) leads to "Monsieur dans les escaliers" (Mister down the stairs). The information on the trauma mechanism is thus lost.

Nevertheless, advances in the field are coming rapidly, opening new perspectives, both in terms of NLP technologies and in terms of appropriate data sets for pre-training in more specialized semantic fields.

References

Akbik, A.; Blythe, D.; and Vollgraf, R. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on COLING*, 1638–49.

Benesty, M. 2019. NER algo benchmark: spaCy, Flair, m-BERT and camemBERT on anonymizing French commercial legal cases.

Chen, X.; Xie, H.; Wang, F. L.; Liu, Z.; Xu, J.; and Hao, T. 2018. A bibliometric analysis of natural language processing in medical research. *BMC Med Inform Decis Mak.* 18(S1):14.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

Eddine, M. K.; Tixier, A. J. P.; and Vazirgiannis, M. 2020. BARThez: a skilled pretrained French sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Fu, S.; Chen, D.; He, H.; Liu, S.; Moon, S.; Peterson, K. J.; Shen, F.; Wang, L.; Wang, Y.; Wen, A.; Zhao, Y.; Sohn, S.; and Liu, H. 2020. Clinical concept extraction: A methodology review. *J Biomed Inform.* 109:103526.

Gaudet-Blavignac, C.; Foufi, V.; Wehrli, E.; and Lovis, C. 2018. De-identification of French medical narratives. *Swiss Med Informatics.* 1–3.

Gerbier, S.; Yarovaya, O.; Gicquel, Q.; Millet, A. L.; Smaldore, V.; Pagliaroli, V.; Darmoni, S.; and Metzger, M. H. 2011. Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. *BMC Med Inform Decis Mak.* 11:50.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning word vectors for 157 languages. In *Proceedings of the 11th International Conference on LREC*.

Grouin, C., and Névéol, A. 2014. De-identification of clinical notes in French: towards a protocol for reference corpus development. *J Biomed Inform.* 50:151–161.

Khin, K.; Burckhardt, P.; and Padman, R. A deep learning architecture for de-identification of patient notes: Implementation and evaluation. *arXiv preprint arXiv:1810.01570*.

Le, H.; Vial, L.; Frej, J.; Segonne, V.; Coavoux, M.; Lecouteux, B.; Allauzen, A.; Crabbé, B.; Besacier, L.; and Schwab, D. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th International Conference on LREC*, 2479–90.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the ACL*, 7871–80.

Martin, L.; Muller, B.; Ortiz Suárez, P. J.; Dupont, Y.; Romary, L.; de la Clergerie, É.; Seddah, D.; and Sagot, B. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the ACL*, 7203–19.

Marucci-Wellman, H. R.; Corns, H. L.; and Lehto, M. R. 2017. Classifying injury narratives of large administrative databases for surveillance-a practical approach combining machine learning ensembles and human review. *Accid Anal Prev.* 98:359–371.

Metzger, M. H.; Tvardik, N.; Gicquel, Q.; Bouvry, C.; Poulet, E.; and Potinet-Pagliaroli, V. 2017. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res.* 26(2):e1522.

Nothman, J.; Ringland, N.; Radford, W.; Murphy, T.; and Curran, J. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 194:151–175.

Névéol, A.; Dalianis, H.; Velupillai, S.; Savova, G.; and Zweigenbaum, P. 2018. Clinical natural language processing in languages other than English: opportunities and challenges. *J Biomed Semantics.* 9(1):12.

Obeid, J. S.; Heider, P. M.; Weeda, E. R.; Matuskowitz, A. J.; Carr, C. M.; Gagnon, K.; Crawford, T.; and Meystre, S. M. 2019. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Stud Health Technol Inform.* 264:283–287.

Ortiz Suárez, P. J.; Dupont, Y.; Muller, B.; Romary, L.; and Sagot, B. 2020. Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the 12th International Conference on LREC*, 4631–38.

Ortiz Suárez, P. J.; Sagot, B.; and Romary, L. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Proceedings of the 7th Workshop on the CMLC*.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *CoRR* abs/1802.05365.

Spasic, I., and Nenadic, G. 2020. Clinical text data in machine learning: Systematic review. *JMIR medical informatics* 8(3):e17984.

Trienes, J.; Trieschnigg, D.; Seifert, C.; and D., H. 2020. Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records. In *Proceedings of the ACM HSDM Workshop*, 3–11.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 30th conference on NeurIPS*. 5998–6008.

Wang, J.; Deng, H.; Liu, B.; Hu, A.; Liang, J.; Fan, L.; Zheng, X.; Wang, T.; and Lei, J. 2020. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on pubmed. *J Med Internet Res.* 22(1):e16816.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; and Funtowicz, M .and Brew, J. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xu, B.; Gil-Jardiné, C.; Thiessard, F.; Tellier, E.; Avalos, M.; and Lagarde, E. 2020. Pre-training a neural language model improves the sample efficiency of an emergency room classification model. In *Proceedings of the 33rd International FLAIRS Conference, AAAI Press*.

Yang, X.; Lyu, T.; Li, Q.; Lee, C. Y.; Bian, J.; Hogan, W. R.; and Wu, Y. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Med Inform Decis Mak.* 19(Suppl 5):232.