Manual and Automatic Annotation of Meeting Reports with Young Offenders for Quality Assessment of Interventions

Pierre André Ménard[1], Sylvie Ratté[2], Geneviève Parent[3], Franck Barbedor[2]
Computer Research Institute of Montréal[1], École de Technologie Supérieure[2], Université du Québec en Outaouais[3]
pierre-andre.menard@crim.ca, sylvie.ratte@etsmtl.ca, genevieve.parent@uqo.ca, franck.barbedor.1@ens.etsmtl.ca

Résumé

We present an annotation project in criminology using meeting reports between clinicians and criminalized young offenders. The domain-specific goal is to assess the quality of the interventions versus the profile of criminal needs established for each offender. The project requires both the manual annotation of a significant number of reports by experts as well as the development of an automatic annotation process to classify the unannotated reports. Both annotation experiments help identify the needs and challenges of providing helpful linguistically relevant annotations to this type of task. Performances of a first classification effort is reported as well as the related manual process.

1. Introduction

Organizations often collect masses of textual information from their daily activities, storing and using them to help provide a better monitoring to the managers. But using this information to assess the quality of the activities is no small task, often requiring in-depth analysis, annotations and, depending on the quantity of data, machine learning methods in order to extract useful knowledge and give a clear view of the process and its output.

This is one such project, involving textual records of meetings between clinicians and teenagers convicted of various offences, who received a penal sentence and are doing mandatory follow-up meetings in relation to their convictions. These records contain much information about young offenders as well as the type and focus of interventions done by the clinicians.

From a criminology perspective, the goal of this research project is to validate if the interventions are relevant to the criminal profile of the young offenders (YLS/CMI from (Hoge and Andrews, 2010)). This is important for the quality of these activities as research shows (Baglivio et al., 2018; Bonta et al., 2008) that interventions targeting the relevant risk factors diminish the risk of reoffending of young offenders. On the other hand, doing interventions on irrelevant aspects is counterproductive and time-consuming.

From a computational linguistic perspective, the goal is to enable the automatic classification of thousands of reports using manual annotations from experts. Correctly classifying these reports will provide a better insight of the monitoring process of young offenders. Manual annotations also provide insights about which linguistic or semantic knowledge is relevant to identify the different types of interventions.

The following section will present a more in-depth context of the project, the analyzed data and the type of information sought by experts in the criminology field. Section 3. will present an overview of the manual annotation process. The fourth section provides insights about the requirements and challenges for this task, followed by classification experiments using the annotated data. The conclusion presents the current state of the project.

2. Context

In Canada, since the *Youth Criminal Justice Act* came into force in 2003, each time a teenager is convicted in court and receives a sentence, an organization responsible for youth protection takes action to protect the public and promote the rehabilitation and reintegration of the youth . To do so, many countries, including Canada, rely on the Risk-Need-Receptivity intervention model (well known as the RNR model) (Andrews and Bonta, 2010). The RNR model is one of the most effective, that is, the one most likely to reduce recidivism among juvenile offenders (Dowden and Andrews, 1999; Koehler et al., 2013; Lipsey, 2009). First, the clinician, responsible for the follow-up with the young offender, reduces recurring risks of future offences by establishing a risk profile for each teenager. This profile is built from interviews and questionnaires filled out with the offender, like the Youth Level of Service / Case Management Inventory (YLS/CMI) (Hoge and Andrews, 2010), to target his criminogenic needs.

Criminogenic needs are dynamic risk factors, like antisocial attitudes, associated to reoffending that can be reduced with clinical interventions. A clinical intervention (hereafter called intervention) is defined as any discussion, communication or interaction between a clinician and a young offender that aims to reduce a risk factor. The offender's criminogenic needs profile is usually done once after the sentence is received and every six months thereafter for cases that exceed this duration. Not all offenders have the same profile. It is therefore important to adapt the interventions according to the needs of each person.

For each sentence received by a young offender, a specific period is imposed by the judge during which he or she will meet with a clinician in order to help reduce the risk of reoffending. For each of these meetings, the clinician must detail in a report, where it happened, the interactions, the topics discussed, the exchange of information, news related to the sentence, etc. By reading a report, an expert should be able to judge if the described interventions are aligned with the profile of the offender or if it was on another topic.

These reports and other activity entries (request for medical record access, adding documents, etc.) related to a case are stored in a secured database for five to seven years following the end of each sentence, at which time they are destroyed. This amounts to more than 150,000 reports for the entire follow-up period for our sample of 750 young offenders. Of these entries, about 30% are reports relevant to our project in which an intervention could take place.

2.1. Intervention types

Each intervention falls into one of the following categories :

- Administrative
- Antecedents
- Attitude
- Consumption
- Family/Couple
- Hobbies
- Peers
- Personality
- Occupational school/work

The *Administrative* category contains any discussion about the conditions of the penal case (contact restrictions, apology letters, etc), the mandatory community work, or the intervention plan. While these are not interventions aimed at lowering reoffence risk, it often takes up a large part of meetings, often in concurrence to useful interventions of other categories.

Antecedents relates to reinforcing believes or alternative behaviors which lower the chance of criminal reoffence when in the presence of a high risk situation.

Attitude regroups interventions that seek change in motivation, prosocial institution valorisation and value restructuration to recognize antisocial attitudes or criminal lifestyle and promote alternative prosocial identities and attitudes.

Consumption interventions favor the reduction of alcohol and drugs consumption or abuse, reduce the personal and interpersonal behavior that leads to consumption and develop new substitutes to these habits.

The *Family/couple* category contains interventions about developping or maintaining positive family relationships, respecting house rules and supervision as well as valuing couple relationship with a prosocial partner with a long term or positive outcome.

Any interventions that foster the participation or engagement in an organized prosocial activity like sports, gym, extracurricular or religious activities, fall into the *Hobbies* category.

Interventions about *Peers* target the reduction of interactions with criminals and the valorisation of relationships with prosocial persons.

Trying to help the young offender to cope with *Personality* issues can also reduce the risk of reoffence. This includes discussions about anger management, improving problem resolution skills, discourage manipulation of others or egocentrism.

Finally, *Occupational* interventions relate to school or work and can include helping and accompanying the offender through the subscription procedure for school, valuing active participation and attendance to either school or work, denoting positive rewards brought by an occupation and developing positive relationships with new colleagues or person of authority.

2.2. Criminology research objectives

As mentioned earlier, studies in criminology shows that interventions aligned with the criminogenic needs of a young offender reduce the chance of recidivism. Acting on this knowledge, the main goal of this project from a criminology standpoint is to validate if the interventions described in the reports are aligned with the profile of each individual.

For example, if a young offender is sentenced for theft, but his criminogenic needs profile reveals that he stole for his consumption habits, interventions should mainly target this last topic. Focusing the interventions on the first topic would thus be misaligned and much less effective, if at all.

In order to attain this goal, reports of each young offender must be annotated, either manually or automatically, to give an estimate of the number of interventions for each category. This will enable researchers to compare the estimated number of typed interventions with the main criminogenic risk of each young offender to validate if they match.

A secondary goal is to assess the intrinsic quality of the reports, validating if the interventions are clearly mentioned and explained. While this is not part of the current project involving natural language processing techniques, a qualitative evaluation will be done following the manual annotation effort to make recommendations to improve future reports. Providing a more detailed account of interventions might help the automatic annotation of future reports and thus enhance the performance of natural language processing tasks like classification or information extraction.

2.3. Computational linguistic aspects

In order to correctly assess the alignment between interventions and risk profile, each intervention reported by clinicians should be fully identified for each meeting regarding a case. We obtained a subset of more than 56,000 reports related to various cases, which is too much data for manual annotation alone. Also, as new meetings occurs every week, this would be an ongoing manual task that would require much effort.

All the reports are written in Canadian French as the organization targets mostly French speaking young offenders. While linguistic analysis tools exist for French, none are trained on the Canadian French variation and, more importantly, register which typically includes more anglicisms as well as different idiomatic expressions and semantic senses. In addition, most reports contains various amount of abbreviations, truncated words, missing words, missing letters, typos, domain-specific lingo, agglomerated words (missing space), implicit acronyms, colloquial terms, missing punctuation, anglicisms and spoken sentence formulation. As such, the reports can be viewed as noisy texts, which implies that usual natural language processing tools will fail to analyze them correctly.

The granularity of annotation should also be tuned to fit the need for precision, either for the estimate number of interventions of different categories or for the exact expressions used to detail an intervention. As such, two natural language processing tasks can be devised in order to obtained the necessary information : multilabel report classification and intervention recognition and typing.

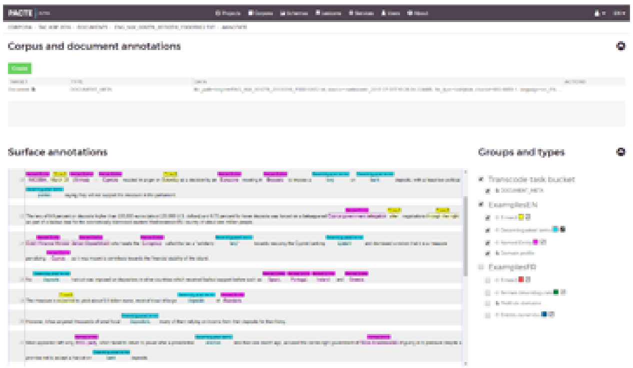

FIGURE 1 – PACTE manual annotation interface.

Annotation type	Training		Evaluation	
	Ann.	Doc.	Ann.	Doc.
Administrative	2,114	1,783	553	484
Antecedents	9	9	1	1
Attitude	79	76	6	5
Consumption	113	113	31	31
Family and couple	55	53	15	15
Hobbies	367	365	47	47
Occupational	1,817	1,597	584	492
Peers	77	74	17	16
Personality	333	302	49	44
Without annotations	—	4,720	—	1,637
Total	4,964	8,189	1,303	2,622

TABLE 1 – Training and evaluation sets distribution for annotations (Ann.) and documents (Doc.).

Multilabel report classification implies the automatic annotation of a report with all the categories corresponding to the detailed interventions mentioned in it. While it does not give an exact count when multiple annotations of the same category are found in one report, it still gives a good estimate as multiple annotation reports usually contain different categories instead of multiple instances of the same.
Intervention recognition and typing is akin to named entity recognition and typing tasks as relevant candidate expressions must first be identified in a text then classified into one of many type. As the exact expression is not needed, we view this as a sentence-level classification, as there is seldom more than one intervention in a single sentence.
As these reports are highly confidential, largely involving minors, no similar training data was available to help the classification process, thus prompting a manual annotation effort. As these data were obtained through an ethic committee and a court hearing, the data cannot be released publicly. The examples shown in this article have been redacted to remove any possibility of individual identification.

3. Annotation process

To simplify the interaction between the experts of each team from different institutions, we used the online text annotation platform PACTE (Ménard and Barrière, 2017) as the central repository for manual annotation, annotation curation and classification results for this project. PACTE enables an annotation project manager to import large text corpora, define custom annotation schemas, add participants to a project, define project's steps and allocate documents to be annotated by the project's participant as shown in Figure 1 (with unrelated text for confidentiality).

3.1. Dataset

In order to have a significant quantity of reports to train and evaluate the machine-learning algorithm, 10,811 single reports from randomly chosen young offenders' cases were selected. For each individual case, all the reports were taken, thus giving a full historical account of each case.
These reports were split into two sets for training and evaluation. The training set contains 8,189 randomly selected reports while the remaining 2,622 reports were kept aside for the evaluation set. A case-based random split was applied, which means that all the reports from one case are entirely found in either the training or evaluation set.
Both datasets were split into batches of 500 documents with a larger last batch for the remaining documents. This was done in order to better organize the work of annotators in the subsequent steps and provide them with a positive sense of advancement throughout the entire effort. They were then upload into PACTE as separate corpora but included in the same annotation project.
Looking at Table 1, we can readily see that the datasets are heavily unbalanced. For the training set, 79.2% of all annotations either comes from the *Administrative* or *Occupational* categories. *Antecedents* makes up less than 0.2% of the entire set. This will likely requires further manual annotation of this type to enrich the sample size.

3.2. Annotation schemas

Using the schema designer in PACTE, we defined nine different schemas corresponding to the nine categories at the top of Section 2.1.. For each of them we defined an attribute to specify the type of risk targeted by the intervention. For example, the *Consumption* schema has the *Reduction* and *Solutions* values for the type attribute while *Family/Couple* has *Relationship*, *Supervision* and *Couple* as the enumeration for the type attribute.
One exception is the *Occupational* category which has a type attribute with *School* and *Work* and a subtype attribute with *Help*, *Participation*, *Engagement*. *Satisfaction* and *Relationship*. All the type and subtype attributes were defined as mandatory when creating a new annotation in PACTE.
Finally, an additional *Comment* attribute was added in order for the annotators to provide additional information to the curator about uncertain annotation or edge cases.
We defined each schema as annotation targeting text surface (as opposed to document or corpus annotation). Text surface annotation schema in PACTE enables the annotators to create contiguous zones spanning from one letter to the whole document. It also enables them to create single annotation with multiple segments. This was quite useful as the reports often contains contextual information between parenthesis or apposition which are not part of the inter-

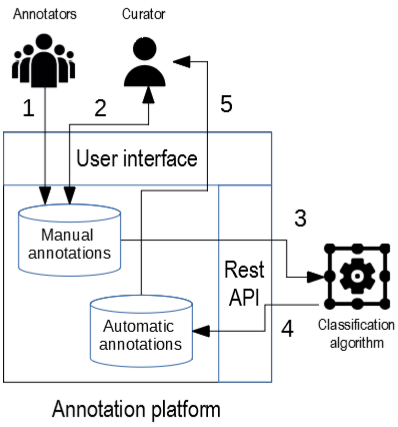

FIGURE 2 – Annotation process.

ventions. Using multipart annotations in this case provided a way to target precisely the sentence parts containing the relevant information.

3.3. Workflow

Once the reports and annotation schemas were imported and created in the platform, the annotation process could begin as shown in Figure 2. The first step (1) was done by two annotators who annotated collaboratively each batch of 500 reports with the web user interface. The curator was then able to review (2) and possibly correct the manual annotations for this batch of reports. After retrieving all the manual annotations available via an online web API, a machine-learning algorithm was used to train a classification model (3) with the N-1 batches and automatically annotate the last batch of reports (4) in a separate storage. Using the user interface, the curator could then (5) validate the performance of the classification and analyze potential issues.

The order of selected reports was randomized to minimize the chance of two contiguous reports on the same case during annotation session to minimize the prior knowledge issue where two reports of the same case with similar information would be annotated differently.

3.4. Effort metrics

Table 2 shows some statistics for the manual annotation effort of this project. The project in PACTE had 21 steps, one for each of the 16 batches for training and 5 for evaluation. This amounts to approximately 189 hours of annotation for each annotator, giving a total of 378 hours for this project. For each batch, half of the reports had annotations for an average of 274 annotations.

Still the unannotated reports took time to process as some had information that lead to a discussion about whether it was an intervention or not. Of course, some reports were very short (a few words like "*He didn't show up and we rescheduled for later*") while others were many paragraphs long.

Avg. time per batch	9 hours
Avg. number of annotations	274
Avg. time per document	1 min
Avg. annotation in single report	1.41
Avg. size of reports	19 words
Max. annotation in single report	5

TABLE 2 – Annotation effort statistics per batch.

4. Challenges

Automatic annotation of the reports in this project proved a challenge as there was no off-the-shelf tools suited to the enrichment of these texts. This section presents the analysis of some aspects of the data which represent a challenge for automatic processing.

4.1. Noisy data

As the processed documents in this study are internal reports, often hastily written at the end of the day, most of them contains misspellings, typos, phonological writing, structural inconsistencies and so on.

There are also many truncations (i.e. "reso" for "résolution", "ds" for "dans" ["*in*"]), abbreviations and acronyms used across the reports, the amount used depending mostly on the author of the report. For acronyms, the implicit short forms (without explicitly linking to the long form) are often used as the report is intended for readers familiar with the domain of activity. This can hinder information extraction tasks applied to the dataset if no external reference list is used to explicitly link the two forms. It might also reduce the performance of the bag-of-word approach as concepts with multiple different surface forms will be separated in the td-idf processing.

4.2. Report versus reality

One key issue for annotation is trying to differentiate between the young offender simply relating an event or fact and the clinicians making an active intervention on the same subject. Because reporting this difference is not a requirement asked of clinicians, there is much variation in the ways it is expressed in the reports. As an example, one could only report that "He told me that he quit school" which does not count as an intervention.

On the other hand, if the previous sentence was followed by "I asked him what he intends to do next", this would be considered an active intervention and would have to be annotated as such. Then again, if there is discrepancies between what was said at the meeting and in the report, like missing information about the intervention, neither a human or a machine-learning agent could deduce what happened.

As there is no way of knowing, without recorded proof, exactly what was discussed and how during meetings, the manual annotation was done in an optimistic mindset. This implies that what may look like a young offender simply telling the clinician about something was annotated as an intervention. This will be taken into account when estimating the number of interventions of each type as the number will likely be inflated.

4.3. Expressing intervention

Without regard of what actually took place, the texts narrate the history of discussions, attempts, failures and commitments. Despite their simplicity, each snippet contains tacit knowledge and presents subtle characteristics. The analysis presented in the next subsections provides potential goals for automated annotation tools in order to help the detection of interventions in reports.

4.3.1. Speech acts in interactions

From the speech acts (Searle, 1969) perspective, the narratives contain many constative expressions that represent a state of things or the recollection of ascertainment by the clinician (e.g. expressions Add, Address, Announce, Discuss, Explore, Expose, Inform, Return, Read, Repeat, Talk, as shown in Table 3). They are pervasive in each category and reflect the continuing interaction and accompaniment of the young.

On the other hand, reinforcement expressions (e.g. Congrat, Reinforce, Underline) and commissive expressions (e.g. admit) are absent or nearly absent from *Antecedent* and *Family/Couple* categories. These ratios are understandable since both *Personality* and *Consumption* are part of a solution that can be within the control of the young offender, and thus, merit reinforcement and commitment, while *Antecedent* and *Family/Couple* are more or less likely to be solved directly by the young.

Directive expression (e.g. ask, explain, invite, question, respond) appear with high ratios in *Administrative* and *Occupational* categories as they consist of the clinician transferring administrative information to the young offender or the latter informing the clinician about his everyday activities such as school and work.

In terms of usage, groups of categories are positively correlated with the usage of these last three types of expressions (reinforcement, commissive and directive) : *Consumption* with *Personality* and *Hobbies*, *Antecedent* with *Family/Couple*, and *Attitude* with *Peers*.

4.3.2. Explicit versus implicit discussion

The first and third person pronouns are often used together, but third alone may express either passive event or interaction depending on the accompanying verb.

4.3.3. Implied third person

Since each snippet of text is intended to be read by people in the field, the young and the clinician are often mentioned implicitly (e.g. "J'aborde [avec lui]"). The same can also be noticed for specialized subject (e.g. "le positif du suivi") in which the intended reader understands without further explanation what is the implied meaning. In this specific example, "Le positif du suivi. Un endroit pour extérioriser ses émotions, pour ventiler." (*The positive of follow-up. A place to externalize his emotions, to breathe.* We can probably surmise that this was not a spontaneous expression of personality aspect, but derived from a discussion.

4.4. Structure of intervention

While most interventions are expressed in the same sentence fragment, some of them span multiple sentences. For example, the commissive-commissive-reinforce structure often unfold over three separate sentences. The first two commissive sentences bring contextual knowledge to the last reinforcement expressions. Using this type of structure and other relevant patterns across sentences could help to better identify important intervention and annotate them with the complete contextual knowledge.

Expression	Examples
Add	"Nous ajoutons dans son CV" (*We add to his resume*)
Address	"J'aborde la révision..." (*I address the revision*) "On aborde chacun des point" (*We address each point*)
Admit	"Admet impulsivité." (*Admits impulsivity*)
Announce	"Je lui annonce" (*I announce him*)
Ask	"Je lui demande si" (*"I ask him if"*)
Congratulate	"Le félicite d'emblée pr..." (*"I readily congratulate him for"*)
Discuss	"Discutons du plan..." (*Discussing the plan*) "Discussion sur " (*"Discussion on"*) "nous avons discuté des ..." (*We have discussed about the*)
Do	"Nous faisons une première ébauche" (*We do a first draft*) "On fait ensemble son devoir" (*We do her homework together*)
Explore	"Nous tentons d'explorer ses pensées..." (*We try to explore his toughts*)
Expose	"Je lui expose la situation" (*I expose him the situation*)
Explain	"M'explqieu qu'entre chaque cours" (*(He) explain that between each course*)
Inform	"Informons que nous avons" (*We inform that we have*)
Invite	"Je l'invite à faire les bons choix " (*I invite him to make the right choices*)
Mention	"je lui mentionne que" (*I mention him that*)
Question	"Questionne à savoir où il se trouve" (*Question to know where he is*)
Respond	"Je lui répond que..." (*I answer him that*)
Return	"Retour sur la révision" (*Return on the revision*)
Read	"Je lui lit les conditions " (*I read him the conditions*)
Reinforce	"Je le renforce en le félicitant " (*"I validate him by congratulating him"*)
Repeat	"Nous devons lui faire répéter certians propos" (*We must make him repeat some points*)
Talk	"Nous parlons de ses travaux" (*We talk about his work*) "Lui parlons du ..." (*Talk to him about*)
Try	"On tente de mêtre en place" (*We try to put in place*)
Underline	"Je lui souligne aussi que" (*"I also point out to him him that"*)
Understand	"Il semble comprendre" (*He seems to understand*)

TABLE 3 – Samples (verbatim, underlined noise) of expressions used for interventions.

Algorithm	Recall	Precision	F1
Complement Naive Bayes	0.8160	0.6301	0.7116
Naive Bayes network	0.5801	0.5663	0.5731
Random Forest	0.3196	0.6105	0.4195
REPTree	0.4258	0.6184	0.5043
SimpleCART	0.7013	0.4846	0.5731
J48 Consolidated	0.5478	0.6250	0.6000

TABLE 4 – Average performances on all types for sentence-level classification task.

5. Classification results

We present in this section some of the classification experiments done for the second task of sentence-level classification. Using the manual annotations from the training set, a prediction model was built and then applied on the evaluation set to assess the performances. As some reports have no annotations at all, a null class was added to the nine relevant classes listed in section 2.. As shown in Table 1, the dataset is unbalanced mostly in favour of the *Administrative* and *Occupational* categories and to a lesser extent *Hobbies* and *Personality* which predict better results on the modal classes and lower scores on the less represented ones.

5.1. Preprocessing

Using a tokenizer and sentence splitter, each report was broken down as a single instance per sentence in the dataset. For simplicity, the few consecutive sentences that were spanned by the same annotation were kept together and tagged with the annotation type. The stop words were removed with the exception of personal pronouns as they can be helpful, as explained in the last section.

The baseline uses a bag-of-word approach with tf-idf vector build using generated ngrams from 1 to 5 words long. A subset composed of the most discriminating 2,000 features was kept for training and evaluation. Table 5 shows a sample of ngrams generated for the *Occupational* category with stop words left in place for clarity.

5.2. Performances

We applied Naive Bayes network (Friedman et al., 1997), Complement Naive Bayes (Rennie et al., 2003), Random Forest (Breiman, 2001), SimpleCART (Breiman et al., 1984), J48 consolidated (Pérez et al., 2007) and REPTree (Quinlan, 1987) on the current data to compare performances on classic machine-learning algorithms from different types (rule-based, decision tree, function). While these approaches are not cutting edge, they provide a quick view to assess the potential performance of current data.

The scores shown in Table 4 are averages combining performances on all ten categories (the nine basic ones plus the null class). We can see that the complement naive bayes outperforms the others, as it was specifically designed to overcome the challenge of text classification. The less frequent classes like *Antecedents* and *Family/Couple* had 0% score as none were correctly classified in the evaluation set.

Original	Translation
cumulé autre absence	cumulate other absence
démarches	actions
retour sur ses absences scolaires	feedback on his school absences
va toujours à l'école	still going to school
il est encore suspendu	he was suspended again
imprime des copies de son cv	print copies of his resume
donne ses preuves d'emploi	give his proof of his employment status
toujours pas d'emploi	still no job
jeune dit aimer	youngster said he likes

TABLE 5 – Some frequent ngrams from Occupational manual annotations.

5.3. Performance and error analysis

As the datasets are created based on noisy data, one issue is the frequency restriction on ngrams used as features. In order to lower the number of features generated, we used a cut-off frequency of 2, so that any ngram occurring only once was not used as a feature. This means that relevant but uniquely miswritten words are eliminated from the datasets which impacts the representation power of features, especially for categories with few instances.

The same sentence can be classified in different categories depending on the surroundings. This is not captured by the current model of vectorizing single sentences without contextual information. Thus a short sentence like *"We discuss his continuing effort"* creates confusion for the prediction process as they were manually classified in different categories in two separate instances.

The unbalanced nature of the datasets, both training and evaluation, also influence the results. For example, as shown in Table 1, the *Attitude* has a single instance in the evaluation set and only nine for training the model.

We can see that performance is not yet useful to provide an adequate estimation of intervention numbers and types in reports for this corpus. Taking into account the task of assigning one of ten classes to a sentence (the nine categories plus the null class), it is still far better than an average 0.1 performance provided by a random baseline.

6. Conclusion

We presented a first set of experiments and analysis using reports relating meetings with young offenders. While the analysis provided in Section 4.3. is in a preliminary stage, it will be further explored to evaluate its potential as a helping linguistic annotation for the automatic detection and classification of interventions for young offenders' reports. The next step in this project is to address the issue with noisy data to single out expressions detailing interventions. If the number of raw reports allows it, an approach using neural network will be applied to profit from the manual annotation while being able to use the noisy text from the whole corpus.

7. References

Andrews, D. A. and Bonta, J. (2010). *The psychology of criminal conduct (5e ed.)*. Lexis Nexis.

Baglivio, M. T., Wolff, K. T., Howell, J. C., Jackowski, K., and Greenwald, M. A. (2018). The search for the holy grail : Criminogenic needs matching, intervention dosage, and subsequent recidivism among serious juvenile offenders in residential placement. *Journal of Criminal Justice*, 55 :46 – 57.

Bonta, J., Rugge, T., Scott, T.-L., Bourgon, G., and K.Yessine, A. (2008). Exploring the black box of community supervision. *Journal of Offender Rehabilitation*, 47(3) :248–270.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1) :5–32, Oct.

Dowden, C. and Andrews, D. A. (1999). What works in young offender treatment : A meta-analysis. In *Forum on Corrections Research*, volume 1, pages 21–24.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Mach. Learn.*, 29(2-3) :131–163, November.

Hoge, R. D. and Andrews, D. A., (2010). *Youth Level of Service/Case Management Inventory 2.0 (YLS/CMI 2.0) : User's Manual*.

Koehler, J. A., Lösel, F., Akoensi, T. D., and Humphreys, D. K. (2013). A systematic review and meta-analysis on the effects of young offender treatment programs in europe. *Journal of Experimental Criminology*, 9(1) :19–43, Mar.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders : A meta-analytic overview. *Victims & Offenders*, 4(2) :124–147.

Ménard, P. A. and Barrière, C. (2017). Pacte : a collaborative platform for textual annotation. In *Proc of the 12th International Conference on Computational Semantics*.

Pérez, J. M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., and Martín, J. I. (2007). Combining multiple class distribution modified subsamples in a single tree. *Pattern Recognition Letters*, 28(4) :414–422.

Quinlan, J. R. (1987). Simplifying decision trees. *Int. J. Man-Mach. Stud.*, 27(3) :221–234, September.

Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 616–623. AAAI Press.

Searle, J. R. (1969). *Speech Acts : An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London.