Automatic language register prediction of digital writings in French

Xingyu Liu

January 2021

Abstract

This article presents a semi-supervised approach to train a tweet register classifier with 500 tweets labeled and 20000 tweets unlabeled. This approach is based on an initial and limited set of expert data. We apply this approach to formal, neutral, informal and trash registers. At the end of the process, the labeled corpus gathers 20000 texts, and the classifier, a neural network, has an accuracy of 64%.

1 Introduction

Language register is the level of formality with which we speak. Different situations and people call for different registers.This stylistic dimension has a strong informational value. However, it is still little studied in Natural Language Processing (NLP). With some tweet data composed of different language registers, labeled and unlabeled, we will extract some linguistic features and apply a semi-supervised approach on the corpus.

2 Literature review

2.1 Language register

In every communication situation, there is a specific set of linguistic signs that defines the communication register (Biber 1995). Each register (slang, public speaking, etc.) corresponds to a different way of expressing the same intentions and ideas and has a linguistic system that fits the communication situation. A register is a set of linguistic variations that are context-dependent (Eckert and Rickford 2001). The terms 'register', 'genre', and 'modality' –although not all authors agree on their usage (Grimshaw 2003) – nevertheless allow us to define

the complex framework in which discourse is produced: all three involve a linguistic variation based on context. 'Genres' are different types of productions, as defined by their function, e.g. narratives, letters, recipes, manuals. 'Modality' is defined by the characteristics of on-line production processes: presence/absence of an interlocutor, duration of the language signal, user's degree of control over the linguistic production(Volckaert-Legrier, Bernicot, and Bert-Erboul 2009). A 'register' permits the expression of social dimensions like power,authority, politeness, and familiarity. Halliday (1964) defined 'register' as a set of variations according to use in the sense that each speaker has a range of varieties and chooses between them at different times.

The concept of register, in combination with those of genre and modality, can be applied to tweets by hypothesizing that this setting is defined by a specific set of linguistic signs that differ from those used in standard writing.

In this article, we adopt a vision with a division into 3 registers: informal, neutral and formal (which correspond respectively familier, courant and soutenu in French). This choice is above all motivated by pragmatism, this division being indeed relatively consensual to ambiguity for the manual labeling of an initial data set, while not prohibiting possible refinements for the future. Besides, due to the quality of our data set, some tweets are too short, or simply unreadable, this kind of tweets are annotated as trash (poubelle in French).

2.2 Language register in NLP

To our knowledge, few French-language works have studied the language registers in natural language processing. There is a work of language registers in tweets presented by Lecorve and al (Lecorvé et al. 2018). In their article, they propose a semi-supervised approach which jointly builds a corpus of texts labeled in registers and an associated classifier. This approach is based on an initial and limited set of expert data. Using an massive automatically retrieved collection of web pages, it iteratively proceeds by alternating the learning of an intermediate classifier and the annotation of new texts to augment the labeled corpus.The registers chosen are formal, neutral, and informal registers.

3 Method

Labeled data is often sparse in common learning scenarios, either because it is too time consuming or too expensive to obtain, while unlabeled data is almost always plentiful. This asymmetry is exacerbated in multi-label learning, where the labeling process is more complex than in the single label case. It is thus important to consider semi-supervised methods for multi-label learning. Semi-Supervised Learning describes the setting where we have a small amount of labeled data and a large, usually absolutely massive, set of unlabeled data. (Note that one tweet could be annotated with more than one labels, which means it mixes two or even three registers. Our classification is then multi-label classification.)

As illustrated on figure 1, the process is initiated by a model training of seed corpus which has been labeled. The training is based on linguistic features extracted from normalized corpus. After that the classifier is constructed, we use this initial classifier to predict registers' probability of the corpus unlabeled. If the probability of a number of corpus reaches our criteria, this part of corpus will be considered as reliable text and then be added in training corpus (see figure 2).

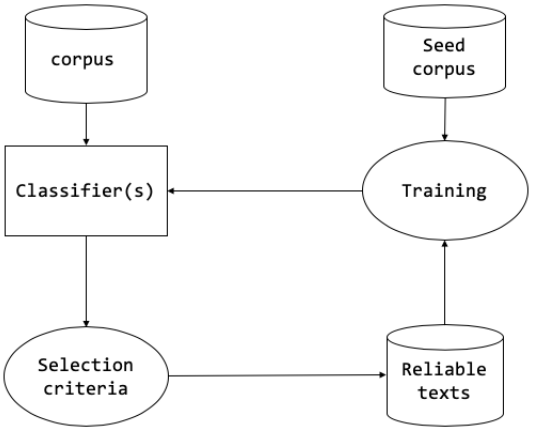

Figure 1: Iterative process of the semi-supervised model

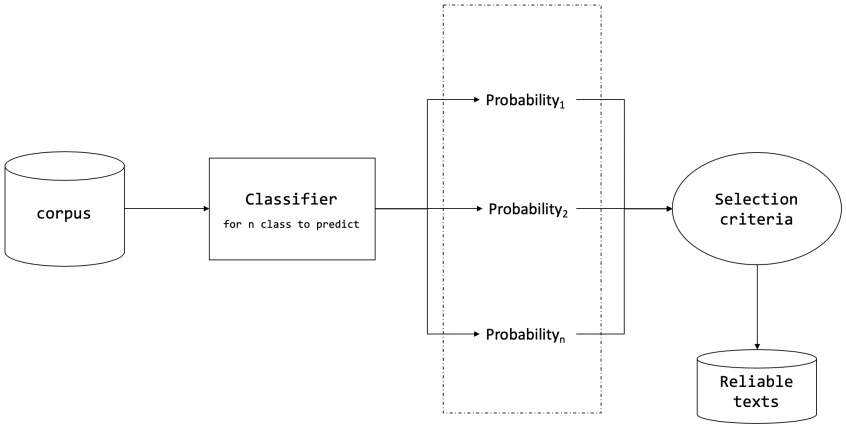

Figure 2: Selection of reliable texts according to a threshold set by the user

4 Dataset

In our work, the seed corpus contains 500 tweets annotated with one or two labels, whereas the corpus unlabeled has 20000 tweets.

From unorganised text file, we write tweet texts and their ids in a data frame and then apply normalized functions on tweet texts. Based on normalized data, we transform the data frame to format conllu containing tokens, lemmas and syntactic information between tokens.

5 Linguistic features

Based on the characteristics of the tweet text, we chose 11 linguistic features:

(1) sms: number of sms terms like abbreviations
(2) ça: if there *ça* is in text
(3) cela: if there is *cela* in text
(4) emoji: if there is emoji in text
(5) proportion_lemma: diversity of lemma
(6) verb_tense: diversity of verb tense
(7) frequency_letter: mean of letter's frequency
(8) proportion_head: how many tokens take ROOT as dependency
(9) fr_spelling_error
(10) fr_accord_number_error
(11) fr_accord_gender_error

6 Classifier training

Our classifier is based on 11 linguistic features mentionned above. They cover multiple levels of abstraction of language, including aspects related to lemma, morphosyntax and syntax. These descriptors are all global relative frequencies to each text. The spelling and grammatical analyses were produced using the language_tool_python tool. The rest of the work is done by a set of Python scripts.

The classifier is a multi-layered neural network. This choice is justifie above all by the current ease to build neural networks thanks to the multiple toolboxes available. In addition, the possibilities of interconnections between neurons and the multiple existing activation functions make it possible to model by neural networks other techniques. The neural network we consider takes as input the vector of 11 values representing a text. The output values are the probabilities of belonging to each register. All the layers of the network are dense layers. The first 2 are composed with a leaky ReLU activation function. The last layer is composed with a softmax function to produce a probability distribution. The dropout mechanism is introduced when learning the neural network to lead it to predictions based on a broader spectrum of information.

Table 1 records the main experiences in our training process. When the threshold is 0.5, which means more texts, even not that "reliable" will be included in next training. It led us to an accuracy "well performed", i.e. overfitting.

If we compare the metrics of same threshold 0.7, when we launch 30 epochs, the accuracy is better than 20 and 40 epochs. Mae and mse are also relatively low.

Now if we keep 30 epochs and raise the threshold at 0.8, because of the lack of corpus that reaches this threshold, few text will be added in training model. The performance is less satisfying than the threshold of 0.7.

classifier	iter	threshold	epoch	nbr_folds	accuracy	mse_final	mae_final
RN	8	0.5	20	5	0.99	0.16	0.18
RN	8	0.6	20	5	0.97	0.14	0.15
RN	8	0.7	20	5	0.60	0.22	0.24
RN	8	0.7	30	5	0.64	0.22	0.26
RN	8	0.7	40	5	0.56	0.29	0.29
RN	8	0.8	30	5	0.48	0.35	0.34

Table 1: list of experiences

Figure 3 illustrate the evolution of performance over iterations with the parameters kept (threshold=0.7, epoch=30)

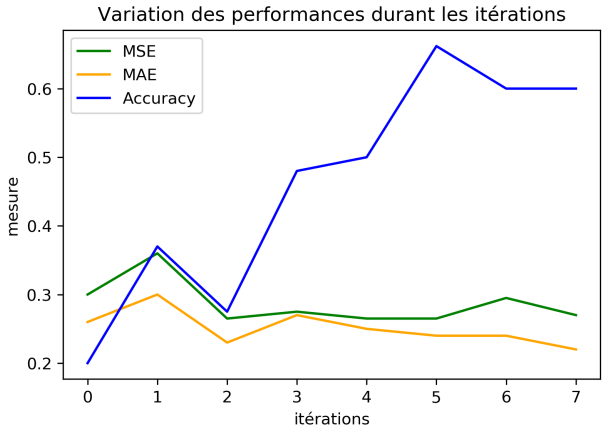

Figure 3: Evolution of performance over iteration

Now we apply the trained model on the corpus of 20000 tweets and predict their register, as shown in figure 4, there are 8857 tweets labeled as informal, 10017 as neutral, 1126 as formal and 0 trash.

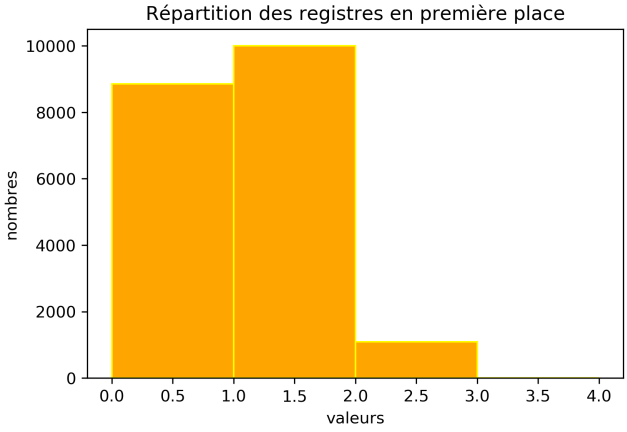

Figure 4: Distribution of registers

7 Conclusion

In this article, we presented a semi-supervised approach to train a tweet register classifier with 500 tweets labeled and 20000 tweets unlabeled. 11 linguistic features were proposed for the model training. At the end of the process, the labeled corpus gathers 20000 texts, and the classifier, a neural network, has an accuracy of 64%. Future perspectives could be to introduce word embedding features in current classifier and apply a more structured method in commissioning parameters.

References

Biber, Douglas (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Eckert, Penelope and John R Rickford (2001). *Style and sociolinguistic variation*. Cambridge University Press.

Grimshaw, Allen D (2003). "Genres, registers, and contexts of discourse". In: *Handbook of discourse processes*, pp. 25–82.

Lecorvé, Gwénolé et al. (2018). "Construction conjointe d'un corpus et d'un classifieur pour les registres de langue en français". In: *Traitement automatique du langage naturel (TALN)*.

Volckaert-Legrier, Olga, Josie Bernicot, and Alain Bert-Erboul (2009). "Electronic mail, a new written-language register: A study with French-speaking adolescents". In: *British Journal of Developmental Psychology* 27.1, pp. 163–181.