Character Based Pattern Mining for Neology Detection

Gaël Lejeune and **Emmanuel Cartier**
LIPN, Paris XIII University
99 avenue Jean-Baptiste Clément
93430 Villetaneuse FRANCE
firstname.lastname@lipn.univ-paris13.fr

Abstract

Detecting neologisms is essential in real-time natural language processing applications. Not only can it enable to follow the lexical evolution of languages, but it is also essential for updating linguistic resources and parsers. In this paper, neology detection is considered as a classification task where a system has to assess whether a given lexical item is an actual neologism or not. We propose a combination of an unsupervised data mining technique and a supervised machine learning approach. It is inspired by current researches in stylometry and on token-level and character-level patterns. We train and evaluate our system on a manually designed reference dataset in French and Russian. We show that this approach is able to outperform state-of-the-art neology detection systems. Furthermore, character-level patterns exhibit good properties for multilingual extensions of the system.

1 Introduction

This paper deals with automatic detection of formal neologisms in French and Russian, with a language-agnostic objective. Formal neologisms are composed of a new form linked to a new meaning, in opposition to semantic neologisms, composed of a new meaning with an existing form. Whereas formal neologisms represent a tiny part of lexical items in corpora, and thus are not yet attracting a lot of research, they are part of the living lexicon of a given language and notably the gate to understand the evolution of languages.

The remainder of the paper is organized as follows. Section 2 details related works on computational approaches to neology. Section 3 describes key aspects of our method and experiments for neology detection. Section 4 presents evaluation results for French and Russian. Finally, Section 5 summarizes the experiments and evokes future developments.

2 Previous work

The study of neology has not been a high level priority within computational linguistics for two reasons. First, large diachronic electronic corpora were scarcely available for different languages until recently. Second, novel lexical units represent less than 5 percent of lexical units in corpora, according to several studies (Renouf, 1993, e.g.). But, from a bird-eyes view, linguistic change is the complementary aspect of the synchronic structure, and every unit in every language is time-related and has a life-cycle.

As shown by (Lardilleux et al., 2011), new words and hapaxes are continuously appearing in textual data. Every lexical unit is subjected to time, form and meaning can change, due to socio-linguistic (*diastraty*) and geographical (*diatopy*) variations. The increasing availability of electronic (long or short-term) diachronic corpora, advances on word-formation theory and in machine learning techniques motivated the recent emergence of neology tracking systems (Cabré and De Yzaguirre, 1995; Kerremans et al., 2012; Gérard et al., 2014; Cartier, 2016). These tools have a two-fold objective: gaining a better overview on language lifecyle(s), and allow lexicographers and computational linguists to update lexicographic resources, language processing tools and re-

Proceedings of the First Workshop on Subword and Character Level Models in NLP, pages 25–30,
Copenhagen, Denmark, September 7, 2017.

sources.

From a NLP point of view, the main questions are : how can we automatically track neologisms, categorize them and follow their evolution, from their first appearance to their integration or disappearance? is it possible to induce neology-formation procedures from expert-curated examples and therefore predict new words formation?

The standard, and rather unique, approach to formal neology tracking consists in extracting novel forms from monitor corpora using lexicographic resources as a reference dictionary to induce unknown words. This is often call the "exclusion dictionary architecture" (EDA). The first system designed for English is due to Renouf (Renouf, 1993) : a monitor corpora and a reference dictionary from which unknown words can be derived. Further filters are then applied to eliminate spellings errors and proper nouns.

Four main difficulties arise from this approach. First, the design of a reference exclusion dictionary requires large machine-readable dictionaries: this entails specific procedures to apply this architecture to under-resourced languages, and an up-to-date dictionary for other languages. Second, the EDA architecture is not sufficient by itself : most of the unknown words are proper nouns, spelling errors or other cases derived from boilerplate removal: this entails a post-processing phase. Third, these systems do not take into account the sociological and diatopic aspects of neologism, as they limit their corpora to specific domains: an ideal system should be able to extend its monitoring to new corpora and maintain diastratic meta-data to characterize novel forms. Fourth, post-filtering has to be processed carefully. For instance, excluding all proper nouns makes it impossible to detect antonomasia (i.e. the fact that a proper noun is used as a common noun, for example "Is he a new kind of Kennedy?").

In many cases, the EDA technique is complemented by a human validation phase, in which experts have to assign each detected "neologism candidate" (NC) a label, either "excluded" or "neologism". This phase enables to complement the exclusion dictionary and to filter candidates to achieve a 100% precision for subsequent analysis. Usually, the guidelines for assessing the class of NCs are as follows : a formal neologism is defined as a word not yet pertaining to usage in the given language at assessment time[1]. A non-neologism is a word pertaining to one of the following categories : a spelling mistake, a boilerplate outcome, a word already in usage... With this procedure, Cartier (Cartier, 2016) evaluated on a one-year subset, that 59.87% of French NC were actual neologisms. In Russian, nevertheless, they evaluated that only 30% of NC were actual neologisms, mainly due to the fact that the EDA technique was in its early phases and that the POS-tagger and spell-checker were not accurate enough. Thus, this approach is not suitable for real time detection or multilingual extension.

In this paper, we advocate a new method to overcome the drawbacks of this method. It combines an unsupervised text mining component to retrieve salient features of positive and negative examples, and a supervised method using these features to automatically detect new neologisms from on-going texts.

3 Dataset and Methods

To the best of our knowledge, there are no existing NLP techniques that take advantage of text mining techniques for detecting neologisms. Intuitively and practically, formal neologisms, as new form-meaning pairs, appear in specific contexts, such as quotation marks(*c'est une véritable "trumperie"*[2]) or metalinguistic markers (*ce que nous pouvons appeler des catholibans*[3].The word-formation rules at stake (Schmid, 2015) involve affixation, composition and borrowings, each implying specific character-based features. From these intuition and analysis, we propose a novel method combining an unsupervised technique to retrieve the salient features of neologisms (internal structure and context), and a supervised machine learning approach to de-

[1]This definition is complemented by other clues like Google Ngrams Viewer statistics or reference dictionaries

[2]*It is a pure deception*, built from Trump + -erie suffix, phonetically near the French word for deception, tromperie.

[3]What we can call catholibans

	French	Russian
#Documents	15559	1750
#Candidates (occ.)	4321 (21511)	807(3563)
#Positives (occ.)	1903 (6339)	245 (715)
Positive ratio (Precision)	44.04%	30.3%

Table 1: Composition of the dataset for French and Russian

tect formal neologisms in on-going texts. In the following, we will first present our corpora and reference data and detail the algorithms used.

3.1 Corpora and Reference Data

As reference data, we use the evaluation data proposed by (Cartier, 2016). It contains a list of NCs and a label : excluded or neologism. In order to see the candidates in context we queried their website[4] to retrieve texts containing one or more NC occurrences. The dataset used here is then limited to NCs having at least one context available. Table 1 exhibits the statistics about this dataset[5]. One can see that the lack of experts for Russian has led to a much smaller dataset. Furthermore, the ratio of positive candidates is smaller in Russian due to a lower quality of the components.

3.2 Contextual character-level features for classification

The data mining component presented here aims to model the context of the candidates in order to classify them. It is an important tool to detect salient contextual and internal features of formal neologisms. Many Data Mining techniques have been used to deal with textual data (Borgelt, 2012), among them we chose an algorithm suitable for the particular type of patterns we wanted to compute (character-level patterns). Character-level analysis has received a growing attention from the scientific in recent years. This approach has proved its efficiency in various tasks (in particular in multilingual settings), among which Authorship Attribution (Brixtel, 2015), Information Extraction (Lejeune et al., 2015), Hashtags Prediction (Dhingra et al., 2016) or Terminology Extraction (Korenchuk, 2017). In this experiment, we mine closed frequent token and character sequences from the candidates contexts using the maximal repeated strings algorithm from Ukkonen (Ukkonen, 2009). These character level patterns (CLP) are computed in linear time thanks to augmented suffix arrays (Kärkkäinen et al., 2006). The CLP computed in this paper have two properties :

- they have a minimal frequency of 2 (in other words they are repeated);
- they are closed: CLP cannot be expanded to the left nor to the right without lowering the frequency.

Patterns are extracted by comparing the contexts of each occurrence of the candidates belonging to the training set. Two kinds of patterns are computed. First, we computed token-level patterns (TLP) which are words and punctuation marks. In some extent, the TLP method can be viewed as a variant of the Lesk algorithm (Lesk, 1986) where in addition to words unigrams there are n-grams mixing graphical words and punctuation. Second, character-level patterns (CLP) pattern which are sequences of characters without any filtering. With CLP, the objective is to represent different levels of linguistic description in the same time: morphology (prefixes, suffixes), lexicon (words or group of words) and style (punctuation and combinations between words and punctuation).

3.2.1 Patterns and contexts

For each attested neologisms found in our corpus, the start and end offsets of their occurrences in the corpus are computed. We model the context as a vector of CLP and TLP frequencies, afterwards we are able to compare the contexts of neologisms and compare them to the context of non-neologisms. Four types of contexts have been identified:

- Internal (resp. bilateral): n characters before the start offset of the NC and n characters after the end offset of the NC, including (resp. excluding) the NC itself
- Left (resp. right): n characters before (resp. after) the start offset (resp. end offset) of the NC plus the NC itself

[4] http://www.neoveille.org

[5] The precision is even worse than on the original data, due to the lack of contexts in the retrieval phase.

Various context sizes have been experimented, from 10 to 400 characters, in order to assess the influence of the window size on the classification results. The context size is always computed in characters in order to have the same data for computing CLP and TLP.

3.2.2 Learning Framework and Evaluation Metrics

Once the CLP are computed in all the training set, they are used as features to train classifiers. For each candidate, the value of each feature will be the frequency of the CLP in the given context (bilateral, internal, left or right). The training of the classifiers has been performed with Scikit-learn (Pedregosa et al., 2011). Various classifiers (decision trees, support vector machines, bayesian networks). 10-fold cross validation has been performed so that the figures presented here after are the mean of the results for each fold. In order to avoid learning biases, all the occurrences of a given candidate will be grouped in only one set per fold : the train set or the test set. Therefore, with TLP internal and bilateral contexts yield the same results : the NC itself can not be used by this method.

4 Results

Table 1 shows the results obtained with TLP for the French dataset with a SVM classifier (linear kernel) and C-parameter set at 1. We will only focus on SVM since this classifier outperformed Decision trees, random forests and bayesian networks . The results for the internal end bilateral context are the same because of the design of the train and test sets (see Section 3.2.2). Two results have to be highlighted here. First, the left context gave by far the best results, suggesting that there are clues announcing neologisms. Second, if we forget about left contexts, the results can be improved by expanding the windows size to 50 characters[6]. Our hypothesis is that expanding the context only improves the bad results and that expanding the left context mostly yields noise. With 72% F-measure in the best case, the TLP method was promising but it was quickly outperformed by the CLP method.

[6]More precisely, the best results for bilateral context are obtained with a window-size of 47.

	10	20	30	40	50
right	32.9	32.7	35.5	31.9	36.0
bilateral	48.3	54.2	56.8	60.3	61.4
left	72.1	70.6	65.6	66.0	66.1
internal	48.3	54.2	56.8	60.3	61.4

Table 1: French Data : F-measure for TLP according to the context length (10 to 50) and context types.

	10	20	30	40	50
right	84.2	81.2	82.2	82.7	81.7
bilateral	67.0	68.3	67.9	68.8	67.2
left	84.9	82.9	83.6	81.7	82.9
internal	82.4	81.6	80.9	80.2	80.4

Table 2: French Data : F-measure for CLP according to the context length (10 to 50) and context types.

On a first approach, we managed to tune the minimal ($minlen$) and maximal length ($maxlen$) of the CLP in order to reduce the search space because even in small windows there are a huge amount of CLP.

We first observed that the optimal F_1-measure scores were obtained with $minlen = 3$ and $maxlen = 7$. This result seemed to be consistent with what has been observed with comparable methodology used for the Authorship Attribution task (see for instance (Brixtel, 2015)). However, subsequent experiments with the same cross-validation method showed that removing these length constraints lead to similar results. Filtering patterns according to their support (relative frequency) has been tested as well but it gave instable results.

Finally, taking all the CLP appeared to

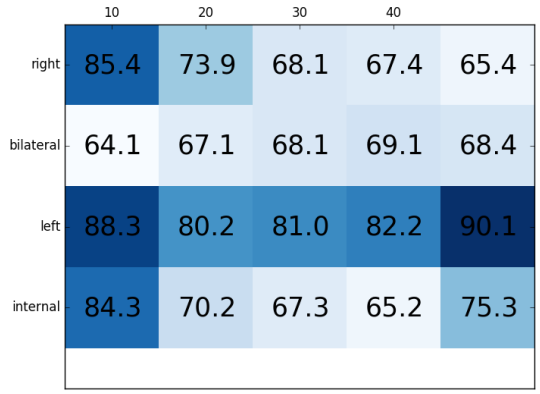

Table 3: Russian Data : F-measure for CLP according to the context length (10 to 50) and context types.

be the best configuration. These results are showed (Table 2). The CLP method takes advantage of internal properties of the candidates (prefixes and suffixes) and it allows us to get more clues in the immediate context of the NC. With a 84.9% F-measure, this method performs better than the 75%[7] presented in (Cartier, 2016). The bilateral context is the least efficient configuration. It shows that CLP including the candidate itself are very good features. Furthermore, it reduces the differences between the left and right contexts. The best results are still found in the immediate contexts but we do not find with CLP the same shift in the results when the context-size is modified.

In Russian we observed the same phenomena with TLP. Therefore, we only present here the results for the CLP method (Table 3). Here, the results are even better than the results for the French dataset with more than 90% F-measure with a left context of length 50. The main difference is that there is more instability when the size and the types of contexts changes. This instability may come from the size of the dataset and the subsequent lower number of features.

There may be room for improvement for the bilateral and internal configurations by taking into account the relative position of the pattern (e.g. if the pattern has been found on the left side, right side or both sides of the candidate) and not only its number of occurences.

[7] 60% precision and probably a recall close to 100%

Finally, among the classifiers we tested, SVM with linear kernels offers the best results. This is a result we expected since it is consistent with state-of-the-art results in stylometry (Sun et al., 2012). Decision trees perform a bit worse and, interestingly, random forests offer very little added-value. We plan to experiment Conditional Random Fields in order to take advantage of the sequential aspect of our input data.

According to the data we collected, the EDA approach shows a precision around 44 % (61% F-measure) for French and 30 % (46% F-measure). Even if it is difficult to precisely assess recall, we can only say that the method presented here shows a real improvement : 82% for French (84.9% F-Measure) and 87% for Russian (90.1% F-measure) in terms of precision.

5 Discussion and Perspectives

The preliminary study we have conducted demonstrates that a combination of unsupervised data mining and supervised Machine learning techniques can largely outperform the EDA approach used to detect formal neologisms. Moreover, this technique does not need any NLP pre-processing (tokenization, lemmatization, POS tagging. . .) of the textual data, which is a great advantage for poorly endowed languages. It reduces the marginal cost for processing new languages.

We plan additional experiments to back the legitimacy of the approach :

- experiment on other languages : we are currently collecting data Chinese, Czech and Portuguese;
- compare with other machine learning techniques, especially CRF, which have proved good accuracy in sequence labelling;

Additionally, we want to experiment the model to detect not only neologisms as a unique category, but categories of neologisms, as affixation, composition and borrowing are likely to retain specific and discriminative features that could be exploited in the detection process.

References

Christian Borgelt. 2012. Frequent item set mining. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2(6)*. pages 437–456.

Romain Brixtel. 2015. Maximal repeats enhance substring-based authorship attribution. In *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*. pages 63–71.

M. Teresa Cabré and Luis De Yzaguirre. 1995. Stratégie pour la détection semi-automatique des néologismes de presse. *TTR : traduction, terminologie, rédaction, 8 (2), p. 89-100.* .

Emmanuel Cartier. 2016. Néoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica, 10, Revue internationale de néologie, p.101-131* .

Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William Cohen. 2016. Tweet2vec: Character-based distributed representations for social media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 269–274. http://anthology.aclweb.org/P16-2044.

Christophe Gérard, Ingrid Falk, and Delphine Bernhard. 2014. Traitement automatisé de la néologie : pourquoi et comment intégrer l'analyse thématique ? *Actes du 4e Congrès mondial de linguistique française (CMLF 2014), Berlin, p. 2627-2646.* .

Juha Kärkkäinen, Peter Sanders, and Stefan Burkhardt. 2006. Linear work suffix array construction. *Journal of the ACM* 53(6):918–936.

Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. 2012. The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring on-going change. *Kathryn Allan and Justyna A. Robinson, eds., Current methods in historical semantics, Berlin etc.: de Gruyter Mouton, 59-96.* .

Yuliya Korenchuk. 2017. *Méthode d'enrichissement et d'élargissement d'une ontologie à partir de corpus de spécialité multilingues*. Ph.D. thesis, Université de Strasbourg.

Adrien Lardilleux, Yves Lepage, and François Yvon. 2011. The Contribution of Low Frequencies to Multilingual Sub-sentential Alignment: a Differential Associative Approach. *International Journal of Advanced Intelligence* 3(2):189–217.

Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. 2015. Multilingual event extraction for epidemic detection. *Artificial Intelligence in Medicine* Doi: 10.1016/j.artmed.2015.06.005.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*. ACM, New York, NY, USA, SIGDOC '86, pages 24–26. https://doi.org/10.1145/318723.318728.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Antoinette Renouf. 1993. Sticking to the Text : a corpus linguist's view of language. *ASLIB Proceedings, 45 (5), p. 131-136* .

Hans-Jörg Schmid. 2015. The scope of word-formation research. *Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen and Franz Rainer, eds., Word-Formation. An International Handbook of the Languages of Europe. Vol. I.* .

Jianwen Sun, Zongkai Yang, Sanya Liu, and Pei Wang. 2012. Applying stylometric analysis techniques to counter anonymity in cyberspace. *JNW* 7(2):259–266. https://doi.org/10.4304/jnw.7.2.259-266.

Esko Ukkonen. 2009. Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theoretical Computer Science* 410(43):4341–4349.