Machine Learning under the light of Phraseology expertise: use case of presidential speeches, De Gaulle -Hollande (1958-2016)

Mélanie Ducoffe, Damon Mayaffre, Frédéric Precioso, Frédéric Lavigne, Laurent Vanni, A Tre-Hardy

Machine Learning under the light of Phraseology expertise: use case of presidential speeches, De Gaulle - Hollande (1958-2016)

M. Ducoffe[1], D. Mayaffre[2], F. Precioso[1], F. Lavigne[2], L. Vanni[2], A. Tre-Hardy[1]

[1] Univ. Nice Sophia Antipolis - I3S, UMR UNS-CNRS 7271 06900 Sophia Antipolis, France - {ducoffe, precioso}@i3s.unice.fr

[2] Univ. Nice Sophia Antipolis - BCL, UMR UNS-CNRS 7320 - 06357 Nice CEDEX 4, France - {mayaffre, lavigne, lvanni}@unice.fr

Abstract

Author identification and text genesis have always been a hot topic for the statistical analysis of textual data community. Recent advances in machine learning have seen the emergence of machines competing state-of-the-art computational linguistic methods on specific natural language processing tasks (part-of-speech tagging, chunking and parsing, etc). In particular, Deep Linguistic Architectures are based on the knowledge of language specificities such as grammar or semantic structure. These models are considered as the most competitive thanks to their assumed ability to capture syntax. However if those methods have proven their efficiency, their underlying mechanisms, both from a theoretical and an empirical analysis point of view, remains hard both to explicit and to maintain stable, which restricts their area of applications. Our work is enlightening mechanisms involved in deep architectures when applied to Natural Language Processing (NLP) tasks. The Query-By-Dropout-Committee (QBDC) algorithm is an active learning technique we have designed for deep architectures: it selects iteratively the most relevant samples to be added to the training set so that the model is improved the most when built from the new training set. However in this article, we do not go into details of the QBDC algorithm - as it has already been studied in the original QBDC article - but we rather confront the relevance of the sentences chosen by our active strategy to state of the art phraseology techniques. We have thus conducted experiments on the presidential discourses from presidents C. De Gaulle, N. Sarkozy and F. Hollande in order to exhibit the interest of our active deep learning method in terms of discourse author identification and to analyze the extracted linguistic patterns by our artificial approach compared to standard phraseology techniques.

Résumé

L'identification de l'auteur et la genèse d'un texte ont toujours été une question de très grand intérêt pour la communauté de l'analyse statistique des données textuelles. Les récentes avancées dans le domaine de l'apprentissage machine ont permis l'émergence d'algorithmes concurrençant les méthodes de linguistique computationnelles de l'état de l'art pour des tâches spécifiques en traitement automatique du langage (étiquetage des parties du discours, segmentation et l'analyse du texte, etc). En particulier, les architectures profondes pour la linguistique sont fondées sur la connaissance des spécificités linguistiques telles que la grammaire ou la structure sémantique. Ces modèles sont considérés comme les plus compétitifs grâce à leur capacité supposée de capturer la syntaxe. Toutefois, si ces méthodes ont prouvé leur efficacité, leurs mécanismes sous-jacents, tant du point de vue théorique que du point de vue de l'analyse empirique, restent difficile à la fois à expliciter et à maintenir stables, ce qui limite leur domaine d'application. Notre article vise à mettre en lumière certains des mécanismes impliqués dans l'apprentissage profond lorsqu'il est appliqué à des tâches de traitement automatique du langage (TAL). L'algorithme Query-By-Dropout-Committee (QBDC) est une technique d'apprentissage actif, nous avons conçu pour les architectures profondes : il sélectionne itérativement les échantillons les plus pertinents pour être ajoutés à l'ensemble d'entrainement afin que le modèle soit amélioré de façon optimale lorsqu'on il est mis à jour à partir du nouvel ensemble d'entrainement. Cependant, dans cet article, nous ne détaillons pas l'algorithme QBDC - qui a déjà été étudié dans l'article original sur QBDC - mais nous confrontons plutôt la pertinence des phrases choisies par notre stratégie active aux techniques de l'état de l'art en phraséologie. Nous avons donc mené des expériences sur les discours présidentiels des présidents C. De Gaulle , N. Sarkozy et F. Hollande afin de présenter l'intérêt de notre méthode d'apprentissage profond actif en termes de d'identification de l'auteur d'un discours et pour analyser les motifs linguistiques extraits par notre approche artificielle par rapport aux techniques de phraséologie standard.

Mots-clés: Deep learning, active learning, computational linguistic, presidential speeches

1. Introduction

Author identification and text classification have always been major concerns for statistical text analysis community. Recently, those fields have received more attention thanks to two major breakthroughs.

On the one hand, digital philology currently expands its research focuses to "new observable" linguistic objects Rastier (2011), either generated or unveiled by numerical processes (patterns, motifs, complex n-grams, morpho-syntaxical or semantical tags,...). On the other hand, the ongoing recent progress of Natural Language Processing (NLP) techniques leads to improvements and more complex textual data processings in particular involving bio-inspired models likely to widely extend the traditional "word bag" model considered in statistical text analysis since Guiraud (1954) or Muller (1968).

If Deep Learning has been so widely spread recently, it is mainly thanks to its ability to tackle difficult problems in mutliple areas from computer vision to speech recognition, being sometimes even more efficient than humans. For example, Convolutional Neural Networks (CNN), one of the different deep architectures, have outperformed human capabilities for category recognition task in the largest image dataset classification challenge He et al. (2015). For speech recognition also, deep learning methods have surpassed their main competitors, such as Gaussian Mixture Models and Hidden Markov Models Mohamed et al. (2012, 2011); Baker et al. (2009).

For NLP tasks also deep learning has shown impressive performances Collobert et al. (2011); Huang et al. (2012). In this context, deep learning is widely acknowledged for its ability to combine different granularity of the text (lexical and syntaxic Socher et al. (2010); Luong et al. (2013); Kalchbrenner and Blunsom (2013); Kalchbrenner et al. (2014)).

The computational analysis of political discourse such as sensed in France by M. Tournier Tournier (1986) has hence to benefit from these advances and our present contribution aims at analyzing a discriminative approach in order to classify the discourses from the french presidents, from De Gaulle to Hollande, with a deep learning architecture Ducoffe and Precioso (2015). On a corpus of 500 discourses and 2,500,000 word occurrences each described by 3 representations (raw format, lemma format, morphosyntactic attributes), we automatically extract prototypical sentences[1] of the presidents according to their notable recognition scores.

Therefore, our contribution is three-fold: First of all, we introduce the fundamentals of deep architectures Bengio (2009); Deng (2014), indeed deep architectures appeared to be particularly relevant for dealing with textual data, even though by construction the decision is implicit making thus the understanding of the language information involved in the final decision difficult to analyze; Secondly, we confront our automatic classification method to linguistic and phraseological knowledge on presidential discourses Mayaffre (2012a) both to complement our knowledge of presidential discourses, especially for a sentence structure point of view, and to improve, in a feedback process of this comparison analysis, the deeper understanding of our deep learning model; Finally and most of all, we aim at challenging the underlying learning mechanisms of deep architectures against linguistic knowledge relying our analysis on the recent QBDC work proposed in Ducoffe and Precioso (2015). We refer the readers to section 4 for further details about such a method.

2. Methodology

2.1. How to represent words

To apply deep learning techniques on a text input, we need to transform the sentences of the corpus into a more desirable format. Indeed we will associate a word and each of its lemma to a numerical representation in order to apply mathematical processing on it. This process is realised in two steps:

- we define a dictionary of pairs of words with successive indices. Although, only words which occur in the corpus more than a fixed threshold will have a specific index, others will all be gathered into a unique category called '*RARE*' word. We have a dictionnary for each level of abstraction.
- To obtain a more desirable output format, the indices will serve as input to a lookuptable containing embedding representations. We will have one lookuptable for each level of granularity considered.

[1] In this paper, we will not provide a linguistic definition of what a sentence is. The term "sentence" will refer to a 20-word long excerpt between two strong punctuation marks.

We learnt directly the lookup tables embeddings as parameters during the training of the whole model. The lookup tables embedding are concatenated to give a global representation of a word and its lemma. However one must now that, in our case of study, the values contained in the embeddings are not hand crafted, nor understandable in an intuitive way. It is but part of the learning process to learn the values of those embeddings to obtain a relevant representation of the words and lemma of our presidential corpus.

2.2. Modeling a deep architecture for text classification

We describe the whole process of learning to discriminate on our presidential corpus. Firstly, we took every sentences from the corpus, and apply padding and cuts to obtain fixed length windows of words and features. In a second step,we took sentences from different authors and

$$
\begin{aligned}
&\textit{The president speaks}\\
&\quad \textit{the president speak}\\
&\qquad \textit{Lookup Table } 1 = \{\ (\textit{RARE},0)\ (\textit{PADDING},1)\ (\textit{the},2)\ (\textit{president},3)\ (\textit{speaks},4)\ \} \quad (1)\\
&\qquad \textit{Lookup Table } 2 = \{\ (\textit{RARE},0)\ (\textit{PADDING},1)\ (\textit{the},2)\ (\textit{president},3)\ (\textit{speak},4)\ \}\\
&\qquad \textit{Lookup Table } 3 = \{\ (\textit{RARE},0)\ (\textit{PADDING},1)\ (\textit{Article},2)\ (\textit{Noun},3)\ (\textit{Verb},4)\ \}
\end{aligned}
$$

Table 1: Illustration of a dictionary for three levels of granularity. Words or lemma which are underrepresented among the corpus are gathered into a unique RARE tag. PADDING tags are repeated before and after each sentence so to obtain fixed length overlapping windows.

build a dictionary of index for every lookuptables. If the database contains the words and two level of features such as their radical and type, we obtain for example three dictionaries as illustration in 1.

The window is then converted into its equivalent numerical representation based on the previously built dictionaries. Why we do not take the full sentence as input for the network is a twofold reason:

1. It is possible to extend a non recursive deep architecture to unlimited size input along one dimension. In this purpose, we repeat the same operations on every units on that dimension and regroup the processed information into a fixed size using pooling so to feed the rest of the layers with a unique input size. This process is limited in that it reduces the size of the training set (*which needs to be important enough to train every parameters of the network*) and allows only one sample per minibatch so unables the parallel optmizations.
2. in that configuration, long sentence may be harder to train compared to short sentences while the contribution of their gradients remain unchanged.

Those windows are given as inputs to the lookup tables, then their embeddings are flattened to be fed to a neural network.

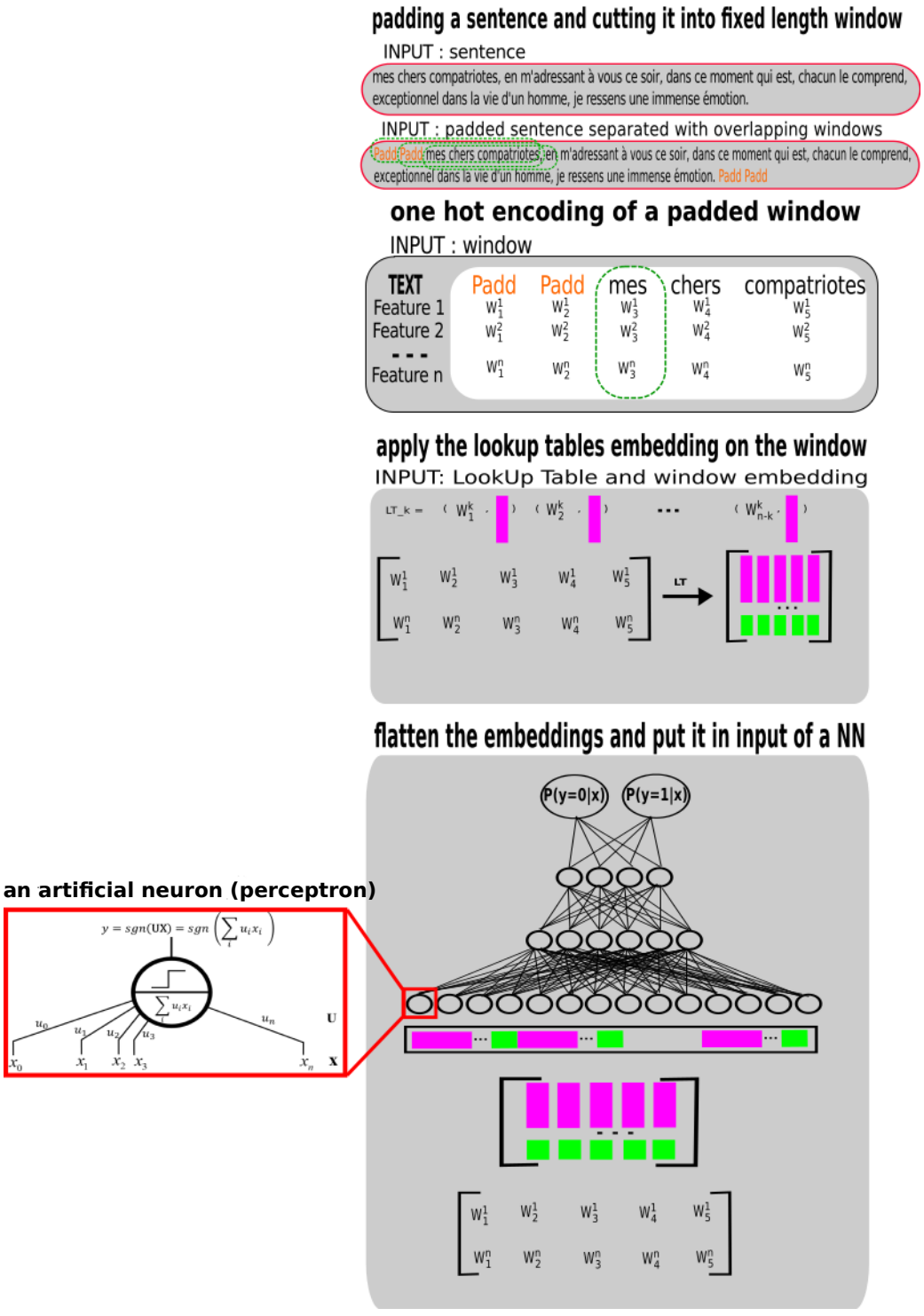

Figure 1: Technical diagram of every step for building the classification process

3. Experiments and results

3.1. Validation

During last decades, works such as Labbé (1990); Mayaffre (2012a) have been conducted towards a better comprehension of the corpus of french presidential discourses during the Fifth Republic. Although researchers have succeeded in providing an efficient speech-based characterisation of the presidents, relying on statistical text analysis methods such as distance between texts Luong et al. (2003); Mellet (2002), those techniques are largely considering words or lemma as independent information units. It should then be noticed here the strong advantage of deep learning versus state-of-the-art linguistic analysis since our model, presented in 2, is combining every level of granularity: short or long range extracted syntagmatic dimension, even or uneven distibution of linguistic entities (word or grammatical category), item distribution all along the text with respect to its global stucture, etc Mellet and Salem (2009).

In this article, we illustrate the deep learning ability to classify texts and authorship, and compare these results with state of the art linguistic assumptions.

Mayaffre et al., in Mayaffre (2012a), have highlighted a significant variation between discourses before and after the eighties based on two main factors:

- the themes, but also the lexicon used in the discourses, have of course evolved during the Fifth Republic, so as the presidential discourse style;
- Moreover, media coverage, which used to rely on radio and is now essentially based on internet and television, impacts also on the discourses.

The results obtained by our deep architecture on textual data are here presented in two use cases and compared with the linguistic literature.

	Recognition rate (%)
Hollande VS De Gaulle	85.08
Hollande VS Sarkozy	71.53

Table 2: Accuracy on two authorship classification tasks: Hollande versus De Gaulle and Hollande versus Sarkozy

After a training phase on a subset of the corpus, the automatic discrimination of sentences from De Gaulle versus Hollande provides very convincing results with 85% of classification rate, which confirms the aforementioned evolutions in the presidential discourses during the 80s Mayaffre (2012a). The machine learning algorithm achieves easily to catch all these differences and to assign the sentences to the right president.

The same process to automatically classify the sentences of Sarkozy versus Hollande results in a satisfying classification but a less striking discrimination. If both results are pretty satisfying, one would notice the decrease of accuracy when considering Hollande versus Sarkozy instead of versus De Gaulle. Indeed, as assessed by the linguistic analysis, we can observe in the

table 2 the gap in terms of accuracy when considering contemporary presidents. This result is interesting because it is in accordance with our work on the relative similarity of the discourses of Hollande and Sarkozy based on phraseological techniques, owing also to the predominance of the crisis theme and the economic vocabulary in their discourses Mayaffre (2012a). In this context, the machine learning approach still gets 71% of good classification rate.

3.2. Extracting and prediction

We examine here two prototypical sentences among the thousands extracted by the algorithm, one correctly recognized as belonging to De Gaulle (versus Hollande), the other correctly recognized as belonging to Sarkozy (versus Hollande) in order to analyse the linguistic content.

"*...l'unité et l'intégrité, se pliant ensuite à une profonde transformation économique, technique et sociale, réformant voici...*"(from De Gaulle)

This first excerpt is particularly rich in terms of linguistic information for the language analyst in the sense that it gathers De Gaulle speech relevant features from a lexical, a morphosyntactic and a syntactic point of view. Thus the unveiled deep architecture is linguistically a concentrate of information which specifies our knowledge of the Gaullist discourse. The heuristic added value lies precisely in this concentrated or combined linguistic information that we are going to decompose now to support the demonstration:

- From a lexical point of view, most of the words of the excerpt belong de facto to Gaullist vocabulary such as statistical text analysis has allowed to describe it Mayaffre (2012a). For instance, "*unité*", "*transformation*" or "*profond*" (lemma-adjective), "*technique*", etc., are typical words both of the presidential discourse in the 1960s and also of a Gaullist style made of abstraction and concepts.
- From a morphosyntactical point of view, most of "parts of speech" and grammatical conventions of the excerpt belong also to the Gaullist rhetoric. The adjective or the noun for example, or also the coordinating conjunction ("*et*") mark a discourse first of all nominal (versus verbal for Hollande) and built or complex, as also highlighted in this excerpt by the weak punctuation (comma).
- From a syntactic point of view, basic structures of the excerpt such as Determiner + Noun + Coordinating conjunction + Determiner + Noun ("*l'unité et l'intégrité*", i.e. "*The unity and the integrity*"), or the sequence of several qualifying adjectives, (more specifically feminine, as in "*...la transformation économique, technique et sociale*", i.e. "*...the economic, technical and social transformation*") are also typical of the Gaullist phraseology.

The phraseologies of De Gaulle and Hollande are so different that the algorithm classifies correctly the sentences. It is however more difficult to discriminate the discourse of Hollande from the one of his immediate predecessor. However, the sentence on which it is the most confident are relevant for the linguist :

	unité	et	l'	RARE	,	se	RARE	ensuite	à	une		
	unité	et	le	RARE	,	se	RARE	ensuite	à	un		
	ncfs	cc	dafs	ncfs	ypw	px3	vpp	r	sp	da-fs-i		
RARE	transformation		économique	,	technique		et	sociale	,	RARE	voici	
profond	transformation		économique	,	technique		et	social	,	RARE	voici	
afpfs	ncfs		afpfs	ypw	afpfs		cc	afpfs	ypw	vpp	sp	

Table 3: decomposition of a sentence by our algorithm

"*...[au lieu de mettre de l'argent pour] que les gens restent chez eux à déprimer, on va mettre de l'argent pour que les gens trouvent une...*" (Sarkozy)

Yet here too, the prototypical excerpt from Sarkozy is a concentrate of information that sustains our in-depth knowledge of the discourse of Sarkozy Mayaffre (2012b). It is maybe from the semantics prevailing standpoint that this second excerpt is first remarkable in its characterization of Sarkozy's discourse: it deals indeed with the major theme of the Sarkozy presidential's five-year term, 2007-2012, that is to say the topic of labour promotion and the denouncement of the state handouts. But it is also from a lexical and grammatical point of view that the analyst will be able to recognize the phrasing of Sarkozy: "*les gens*" (and not "*le peuple*" or "*les ouvriers*" which can be found in other discourses) are its main operators; "l'argent" appears as a major concern; the indefinite pronoun "*on*" (i.e. one) is also a strong feature; The slack use of an immediate future by verbal form "*va mettre*" (i.e. to be going to) is another feature of an informal speech. The factorization of these linguistic elements, which are the basis of the architecture of the discourse, leads to predict precisely that this excerpt belongs to Sarkozy.

que	les	gens	RARE	RARE	eux	à	RARE	,	on	
que	le	gens	rester	RARE	lui	à	RARE	,	on	
cs	da	nc	vmip3p	sp	pp3mpd	sp	vm	ypw	pp3	
va	mettre	de	l'	argent	pour	que	les	gens	RARE	
aller	mettre	de	le	argent	pour	que	le	gens	trouver	
vmip3s	vmip	sp	da	ncms	sp	cs	da	ncms	vmip3p	

Table 4: decomposition of a sentence by our algorithm

4. Going further by challenging QBDC and linguistic

4.1. Introducing QBDC: active learning for deep architectures

Among the corpus, we might suspect that some sentences will decrease the accuracy of the system because of several reasons:

- doppelganger: sentences repeated several times by one author. As our learning model is not taking in account windows overlapping along two sentences, the repetition will never

be a factor to help the classification. So in this case keeping track of the all the repetitions will only unbalanced the training set.

- outliers: sentences repeated several times but by different authors: this can happen for basic sentences and more commonly for cut sentences as in our experiments. In this case it is best for the convergence of the learning to keep a single version of the sentence.
- untrainable data: some sentence, based on the current knowledge the network has acquired, will never have another prediction labels but the one the network will attribute. In this case to reduce the computation cost of learning, it is best to discard that kind of data.

However we do not know what kind of *knowledge* the network is relying on: syntaxic, ontology, semantic or another non linguistic intuitive information. In order to have a glance at what kind of information the network is working on, we propose to select iteratively the samples which are the most helping the network to improve its accuracy and then to acquire knowledge. How to collect this data is done by a machine learning technique called active learning.

Active learning is a special case of learning when the model restricts its learning knowledge to a subset of the data and may gather more data in an online fashion: the model is able to interactively query new data which may be (or not) annotated by a human annotator (or an automatic oracle) and then added to the current set of training data. The main reason to be for active learning is the difficulty in gathering annotated data, especially when it recquires experts. The fieldhas been investigated through several solutions (uncertainty sampling, query by committee, variance reduction,...see Settles (2010) for a survey of the existing methods). Among all the proposed solutions (...), few of them are convenient for deep learning architectures. Indeed previously active learning approach techniques have a quadratic complexity based on the number of parameters in the network. Recently we proposed an active learning method suitable for deep learning architectures. It is a query by committee based approach which consists in building a set of models trained on the same current labeled database and make each instance vote on the output of queried elements. Eventually the score of an element is determined by the disagreement it provokes among the members of the committee.

Among such methods, QBDC Ducoffe and Precioso (2015) is an active learning technique designed to build a committee of deep architectures with a low computation cost.

4.2. Understanding the underlying mechanisms

The text analysis with our machine learning approach proceeds through active learning stages by selecting new samples at each iteration to be added to the training set. The linguistic analysis of these special samples helps to understand the processes at work.

- First active learning phase: the selected sentences are indeed ambiguous for the linguist. In addition to too short sentences whose length seems to prevent a correct assignment, we find for example these two excerpts:
"*c' est cela que les événements m' ont amené à représenter à travers toutes les tempêtes*"
"*je transmettrai ma charge officielle à celui que vous aurez élu pour l'assumer après moi*"

In these two examples, one can foresee a contradictory linguistic characterization between the lexical level and the grammatical level. The lexical composition would be rather Gaullist with a vocabulary recherché ("*tempêtes*", "*événements*", "*assumer*", "*charge officielle*"). The grammatical structure is rather associated to Hollande with the use of the first person ("m"', "je", "ma", "moi") and a verbal tone (lots of verbs). In the end, at this stage, the analyst may therefore not be more sure of the paternity of these excerpts than the algorithm.

- In the later active learning phase. After several iterations, selected sentences are gradually refined and disambiguated. After three active learning selection, for example, the algorithm remains indeterminate on the following excerpt:
"*cela dit, l'apparition de l'Algérie dans la situation d'un Etat indépendant coopérant organiquement avec la France.*"
The analyst recognizes without difficulty the phraseology, the lexicon and the concerns of De Gaulle period (the issue of "*Algérie*" and "*France*", the nominal tone). However, we can assume that the introductory words "*cela*" and "*dit*" scramble the classification since they do not belong to the phraseology of De Gaulle.

5. Discussion-Conclusion

Deep architectures have demonstrated compelling potential for a better sampling of the target manifold Bengio et al. (2007) thanks to their " *expressive power* " Bengio and Delalleau (2011). However, the lack of comprehensive understanding (both on a theoretical or a practical aspect) of their underlying mechanisms hampers their wider application to intricate linguistic tasks. In this article we have made a step towards understanding the shared linguistic knowledge entailed in both machine and human analysis processes. We have indeed analyzed the ability of deep learning approaches to cross the different levels of text granularity, vocabulary granularity, and morphosyntactic structure granularity, so as to encompass all the linguistic knowledge at once. Furthermore, we have shed a light on the persistent intricacy of the predictive process even for relatively simple classification task from a linguist's point of view.

Acknowledgements

This work benefited from using Inria Sophia Antipolis - Méditerranée, Computation cluster Nef

References

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., and O'Shghughnessy, D. (2009). Research developments and directions in speech recognition and understanding, part 1. *IEEE Signal Processing Magazine*, 26(3):75–80.

Bengio, Y. (2009). Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1):1–127.

Bengio, Y. and Delalleau, O. (2011). *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, chapter On the Expressive Power of Deep Architectures, pages 18–36. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. pages 153–160.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*.

Ducoffe, M. and Precioso, F. (2015). QBDC: query by dropout committee for training deep supervised architecture. *CoRR*, abs/1511.06412.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: essai de méthodologie*. Presses universitaires de France.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *The IEEE International Conference on Computer Vision (ICCV)*.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent convolutional neural networks for discourse compositionality. *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality*.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Labbé, D. (1990). Le vocabulaire de françois mitterrand.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *CoNLL*.

Luong, X. et al. (2003). La distance intertextuelle. *Corpus*, (2).

Mayaffre, D. (2012a). *Le discours présidentiel sous la Ve république: Chirac, Mitterrand, Giscard, Pompidou, de Gaulle*. Collection académique. Presses de la fondation nationale des sciences politiques.

Mayaffre, D. (2012b). *Mesure et démesure du discours. Nicolas Sarkozy (2007-2012)*. Fait politique. Presses de Sciences-Po.

Mellet, S. (2002). Corpus et recherches linguistiques. introduction. *Corpus*, 1.

Mellet, S. and Salem, A., editors (2009). *Topographie et topologie textuelles*, volume 7 of *Lexicometrica*.

Mohamed, A.-r., Hinton, G., and Penn, G. (2012). Understanding how deep belief networks perform acoustic modelling. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4273–4276. IEEE.

Mohamed, A.-r., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., and Picheny, M. A. (2011). Deep belief networks using discriminative features for phone recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5060–5063. IEEE.

Muller, C. (1968). *Initiation à la statistique linguistique*. Collection Langue et langage. Larousse.

Rastier, F. (2011). La mesure et le grain. sémantique de corpus. volume 10 of *Collection Lettres numériques*, page 280 pages. Paris, éditions honoré champion edition.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Socher, R., Manning, C. D., and Ng, A. Y. (2010). Learning continuous phrase representations and syntactic parsing with recursive neural networks. *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pages 1–9.

Tournier, M. (1986). La lexicométrie socio-politique. In *Le Courrier du CNRS*, number 65, pages 24–32.

Appendix

French nomination	English equivalent nomination
afps	SingularFeminineAdjective
cc	ConjunctionCoordination
cs	ConjunctionCoordination
da	ArticlePlural
da	ArticleMasculineSingular
dafs	SingularArticle FeminineDefinite
da-fs-i	SingularFeminineIndefiniteArticle
nc	PluralCommonNoun
ncfs	SingularFeminineCommonNoun
ncms	MasculineSingularCommonNoun
pp3	PronounNoReflexive
pp3mpd	PronounThirdPersonNonReflexivePlural
px3	ThirdPersonReflexivePronoun
r	Adverbe
sp	Preposition
vm	VerbInfinitive
vmip3p	VerbThirdPersonPlural
vmip3p	VerbPresent
vmip3s	VerbPresentThirdPersonSingular
vpp	VerbPresentParticiple
ypw	PunctuationComma

Table 5: lemma definitions