Deconvolution for linguistic analysis

L. Vanni[1], V. Elango[2], C. Aguilar[1], D. Longrée[3], D. Mayaffre[1], F. Precioso[2], M. Ducoffe[2]
[1] Univ. Nice Sophia Antipolis - I3S, UMR UNS-CNRS 7271 06900 Sophia Antipolis, France
{lvanni, mayaffre}@unice.fr
[2] Univ. Nice Sophia Antipolis - BCL, UMR UNS-CNRS 7320 - 06357 Nice CEDEX 4, France
{ducoffe, precioso}@unice.fr - ecveer@gmail.com
[3] Univ. Liège - L.A.S.L.A, Bélgique
dominique.longree@uliege.be

Abstract

This document contains the instructions for preparing a paper submitted to COLING-2018 or accepted for publication in its proceedings. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut nec tellus at lectus suscipit porta. Sed ut aliquam tellus. Ut in arcu nec dui tincidunt suscipit in ultrices sem. Curabitur sed nibh quis est tincidunt aliquet. Nullam et nulla lorem. In hac habitasse platea dictumst. Nam aliquam nisi vel orci venenatis venenatis. Nunc aliquet nibh ut odio dapibus, sit amet convallis quam tincidunt. Suspendisse potenti. Vestibulum ut erat ac mauris imperdiet venenatis. Aenean porttitor mollis mi, eu placerat enim bibendum sodales. In vel ligula diam. Aenean faucibus lacinia rutrum. Aenean in risus neque.

2 Related work

Mauris ut magna ut diam hendrerit tincidunt. Duis turpis lacus, lacinia ut accumsan a, rutrum eu justo. Donec efficitur purus non leo iaculis elementum. Nulla pulvinar ligula ut pretium vulputate. Mauris non suscipit felis, ac molestie sem. Duis quis lacus sed massa pharetra eleifend non ut urna. Phasellus lobortis mattis pharetra. Phasellus mattis purus non quam molestie tincidunt.

Sed hendrerit at leo sed tristique. Vestibulum fringilla, nisi id rutrum congue, erat elit hendrerit mi, eget tempus erat nisl sit amet urna. Pellentesque ornare, nunc vel molestie scelerisque, ante augue condimentum orci, sollicitudin sodales odio neque at tortor. Mauris pellentesque ex neque, ut finibus ante pulvinar nec. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus et fermentum urna. Donec rutrum, ex vel interdum mattis, nibh augue mattis magna, vel varius ante ipsum sed diam. Sed dui risus, gravida sit amet ornare nec, lacinia id elit. Nunc consectetur commodo ante semper suscipit. Quisque ullamcorper mauris id arcu placerat pellentesque. Sed sit amet dolor metus.

3 model

We propose a deep neural model to capture linguistics partern of the text. This model is based on simple Convolutional Neural Network with an embedding layer for words representaion, one covolutional with pooling layer and finnaly one Dense layer. Figure 1 shows the global structure of our architecture. The input is a sequence of words $w_1, w_2 ... w_n$ and the output is class (for text classification). The embedding is build on top of a Word2Vec architecture trained on a Skip-gram model. Our text tokenizer keeps all the words, event if a words are rare, to make sure all linguistics material could be detected at the end by the model. This kind of embedding give us a good starting point for the word represention in vector, but we make it trainable also by the model to reach the best text-classification acuracy.

The Convolutional layer is based on 2 dimensional convolutions, the same as we use for pictures convolutions, but with a fixed width corresponding to the max width represented by the embedding size.

At the end we transform the 2 dimensional convolutions in 1 dimensional convolution. The common convolution used in general for the text. The only parameter we can modify here is the height of the filter corresponding to the number of words we want to put in the filter. The goal of this approach is to use the standard picture deconvolution (conv2D Transpose) for our model fit on text.

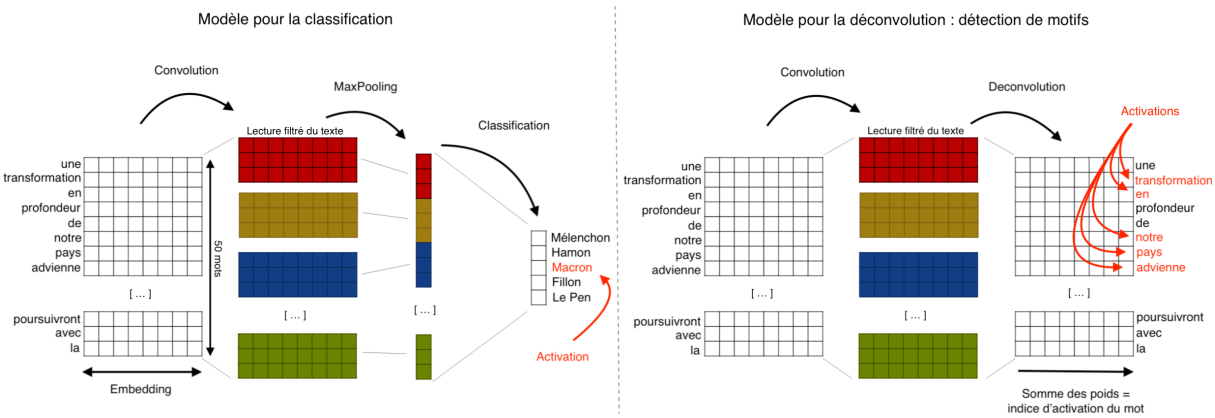

Figure 1: Deconvolution model

4 Experiments

4.1 Z-score Versus Activation-score

Ut est massa, rhoncus sagittis justo vel, cursus interdum odio. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Suspendisse elementum pretium sodales. Curabitur ultrices condimentum malesuada. Interdum et malesuada fames ac ante ipsum primis in faucibus. Duis non consectetur nisi, a viverra neque. Vestibulum sodales sed arcu non dictum.

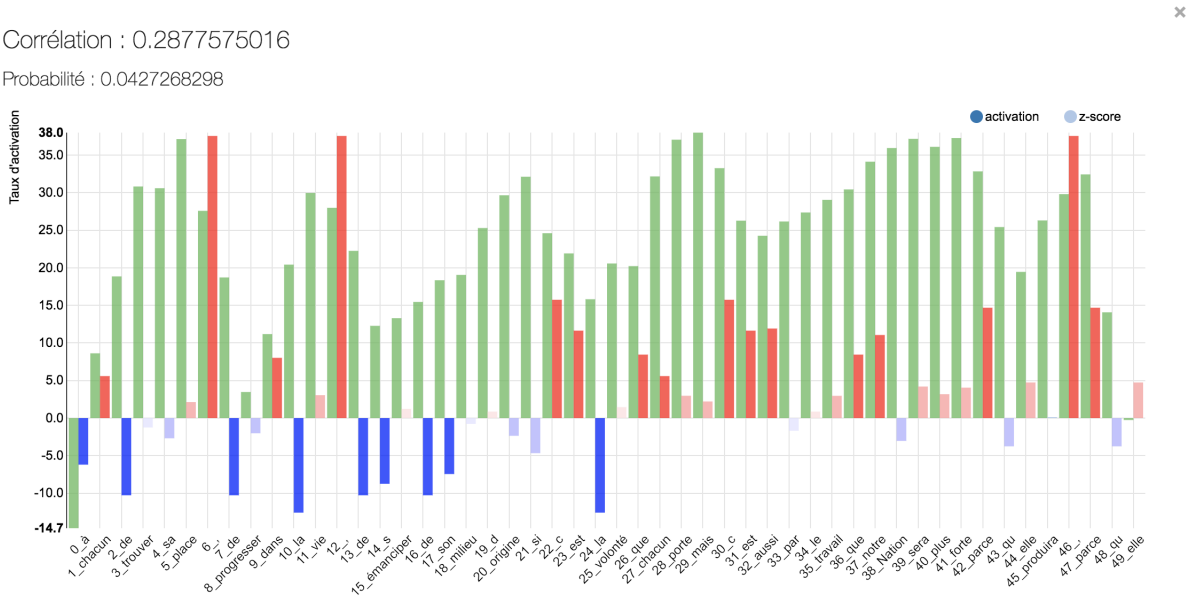

Figure 2: Z-score Versus Activation-score

4.2 Dataset : English

4.3 Dataset : French

" . . . notre pays advienne, à l'école pour nos enfants, au travail pour l'ensemble de nos concitoyens, pour le climat, pour le quotidien de chacune et chacun d' entre vous. Ces transformations profondes ont commencé et se poursuivront avec la même force, le même rythme, la même intensité"

Cet extrait issu du discours de vœux d'Emmanuel Macron (31 décembre 2017) est mal attribué par l'ADT (Z-score) qui le rapproche statistiquement de De Gaulle, et bien attribué par le Deep learning qui reconnait du Macron. L'erreur d'attribution statistique s'explique par une phraséologie très gaullienne et la multiplication de marqueurs linguistiques fortement indicés chez de Gaulle : par exemple, de Gaulle avait pour caractéristique de faire des phrases longues et littéraires articulées autour de conjonctions de coordination comme " et " (z-score = 28 chez de Gaulle, 2 occurrences dans l'extrait). Son discours était également plus conceptuel que la moyenne, et cela se traduisait par une sur-utilisation des articles définis (le, la, l', les) très nombreux dans l'extrait (7 occurrences) ; particulièrement au féminin singulier (" la " République, " la " liberté, " la " nation ", " la " guerre, etc ; ici " la " même force, " la " même intensité).

Les meilleures performances du deep learning interrogent quant à elle le linguiste et épouse parfaitement ce que l'on sait socio-linguistiquement du discours dynamique de Macron. La zone d'activation la plus importante de l'extrait concerne le syntagme nominal " transformation profonde ". Pris séparément, aucun des deux mots du syntagme n'est très macronien d'un point de vue statistique (" transformation " = XXX ; " profonde " = YYY). Mieux : le syntagme lui-même n'est pas attesté dans le corpus d'apprentissage du Président (0 occurrence). Cependant, on peut constater que la co-occurrence de " transformation " et de " profonde " s'élève à +XXX chez Macron : ce n'est donc pas l'occurrence d'un mot seul, ou de l'autre, qui est macronienne mais l'apparition simultanée des deux dans une même fenêtre. Pour autant, la cooccurrence de " transformation " et de " profonde " ne saurait être suffisante pour caractériser Macron, notamment parce que la cooccurrence des deux mots est plus fréquente encore chez Pompidou par exemple ; d'autres indices additionnés sont nécessaires à l'attribution. Les zones d'activation secondes et complémentaires de l'extrait concernent ainsi les deux verbes " advienne " et " poursuivront ". D'un point de vue sémantique, les deux verbes conspirent parfaitement, après le syntagme " transformation profonde ", à donner la dynamique nécessaire à un discours qui prône le changement. Mais c'est les temps verbaux (portés par la morphologie des verbes) qui apparaissent déterminants dans l'analyse. Le calcul des codes grammaticaux co-occurrents au mot " transformation " indique ainsi que les verbes au subjonctif et les verbes au futur (et également les noms) sont les codes privilégies chez Macron. (GRAPH XXX) Plus précisément encore, l'algorithme indique que, chez Macron, lorsque " transformation " est associée à un verbe au subjonctif (ici "advienne "), alors il existe le plus souvent un verbe au futur co-présent (ici " poursuivront "). " Transformation profonde ", " advenir " au subjonctif, " poursuivre " au futur : tous ces éléments signent, ensemble, un discours fait de promesse d'action, dans la bouche d'un président jeune et dynamique. Enfin, le graphique indique que " transformation " est surtout associée aux noms chez le Président : dans une concentration extraordinaire, l'extrait en recense ainsi 11 (" pays ", " école ", " enfants ", " travail ", " concitoyens ", " climat ", " quotidien ", " transformations ", " force ", " rythme ", " intensité ").

4.4 Dataset : Latin

5 Conclusion

Mauris ut magna ut diam hendrerit tincidunt. Duis turpis lacus, lacinia ut accumsan a, rutrum eu justo. Donec efficitur purus non leo iaculis elementum. Nulla pulvinar ligula ut pretium vulputate. Mauris non suscipit felis, ac molestie sem. Duis quis lacus sed massa pharetra eleifend non ut urna. Phasellus lobortis mattis pharetra. Phasellus mattis purus non quam molestie tincidunt.

Sed hendrerit at leo sed tristique. Vestibulum fringilla, nisi id rutrum congue, erat elit hendrerit mi, eget tempus erat nisl sit amet urna. Pellentesque ornare, nunc vel molestie scelerisque, ante augue condimentum orci, sollicitudin sodales odio neque at tortor. Mauris pellentesque ex neque, ut finibus ante pulvinar nec. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus et fermentum urna.

Donec rutrum, ex vel interdum mattis, nibh augue mattis magna, vel varius ante ipsum sed diam. Sed dui risus, gravida sit amet ornare nec, lacinia id elit. Nunc consectetur commodo ante semper suscipit. Quisque ullamcorper mauris id arcu placerat pellentesque. Sed sit amet dolor metus.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.