Linguistic Evolution Analysis using "Le Temps" archives

Ivan Baeriswyl, Sylvain Beaud, Ismail Bouanani
ivan.baeriswyl@epfl.ch, sylvain.beaud@epfl.ch, ismail.bouanani@epfl.ch

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Introduction

Over time, semantics of the french language continuously varied. Words systematically acquire and lose different semantic content (meanings, synonyms...). For instance, the word "souris" (mouse), that went from referring only to the animal to having a meaning in the context of computers. Some words go through that evolution faster than others. Our work consists in analyzing the semantic changes of these words and finding a way to display them using the latest and most appropriate technique available. Our principal resource is the archives of the newspaper "Le Temps" for 200 years (1798-1998). We resorted to the word embedding Word2Vec methods [Hamilton et al., 2016] and t-SNE for the data visualization [Maaten and Hinton, 2008].

Pre-processing Methods

First, we applied some standard NLP preprocessing steps which are, in this case, extracting the content from the raw data, the tokenization of the text and the removal of stopwords. In addition to these standard procedures, we also needed to correct the errors.

The newspaper have been scanned and the text content has been retrieved using OCR (Optical Character Recognition). The tricky part about these kind of data is that the printed characters used in the beginning of the 19^{th} century were quite different thus most of the s characters are recognized as f and some words have now a different spelling.

So to correct the data, we created dictionaries from the raw data and performed some basic spell checking to remove some errors and to improve the accuracy of the spellchecker, we merged dictionaries together and removed the less used spelling which are often wrongly spelled words.

Conclusion

It is hard to get feedback and verification on natural languages that have two hundreds years. Due to the evolution of the written press, the distribution of the data is skewed since 2137 articles were written for the "Gazette de Lausanne" in 1805 and a bit more than 9000 in 1905 and 46833 in 1997. Hence, we tried to compare batches of data containing the same amount of articles rather than over the same period of time. It is also important to notice that the vocabulary used by the written press concerning everyday articles or politics is more or less constant through the years. The most significant changes are related to technological evolution or major political events such as the appearance of new countries or the falls of empires.

References

[Hamilton et al., 2016] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv:1605.09096*.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Used Methods (SGNS for Word2Vec)

The skip-Gram (SG) neural network architecture is used with Word2Vec to create a model that convert words into a multidimentional space (words are represented as vectors). The concept of negative sampling (NS) is implemented in the model by learning by modifying the weights of close (related) words only (the positive samples), while most of the other weights are untouched (the negatives). The SGNS is said to perform well on large datasets of embedded words [Hamilton et al., 2016]. On top of that, models are created for different periods of time, and compared to get an information on the semantic shift $\Delta_i^t = \text{cosdist}\ (\vec{w}_i^t, \vec{w}_i^{t+1})$. As the models created do not have the same vectors for the same words initially, an alignment of these to the same coordinate is operated by using the orthogonal Procrustes algorithm, giving a transformation matrix to go from one model to another with the error minimized.

Visualization of the results

Due to the high dimensionality of the embeddings, it is rather difficult to obtain an accurate representation of the data in a two-dimensional space. We thus reduce the number of dimensions from 300 to 2 using t-SNE [Maaten and Hinton, 2008]. But such a reduction is highly unstable and small changes in the model can change the appearance of the model.

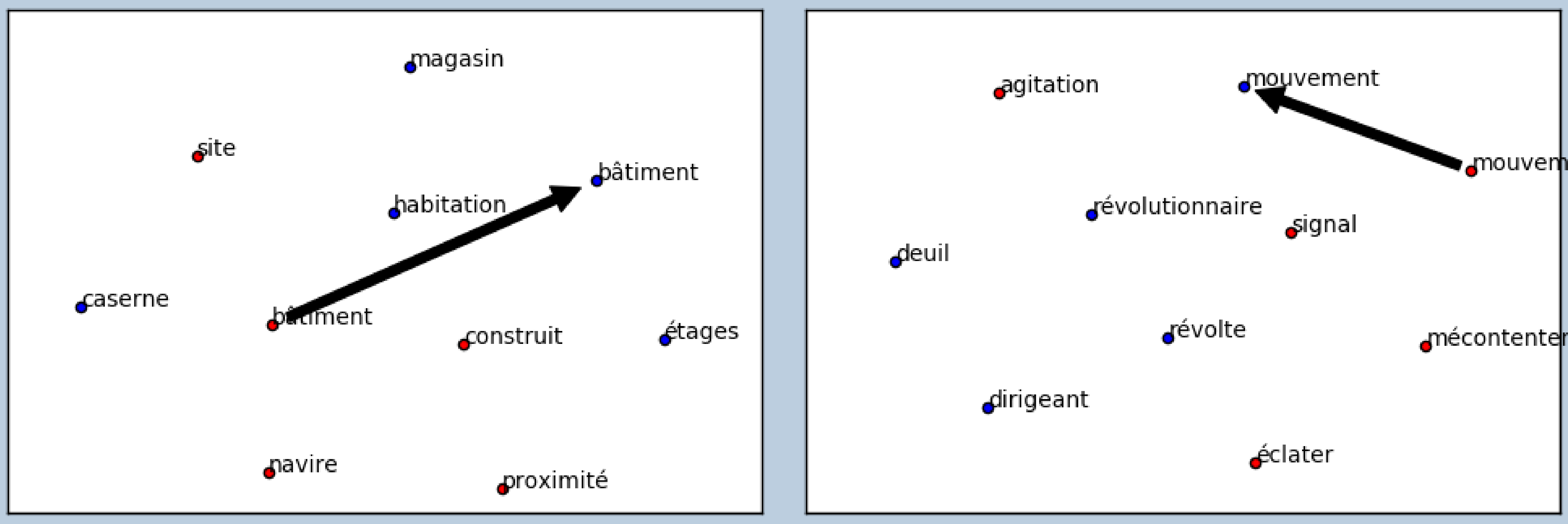

Figure 1: Evolution of "bâtiment" (left); Evolution of "mouvement" (right).

On Fig. 1, we can see the evolution of two words, the red dots represent the most similar words around 1830 and the blue dots are related to 1990. For example, the word "bâtiment" which has always been a homonym with both meanings "building" and "big ship". However, we can see a shift in the contexts where this word was used. In the 19^{th} century, it is related to the vessel and its context has shift towards the lexical field of the construction in the $20^{th} century$. The other word is "mouvement" which means "movement". We can see that in this case, the meaning of the word itself has not changed but the context in which it is used reflects new aspects.

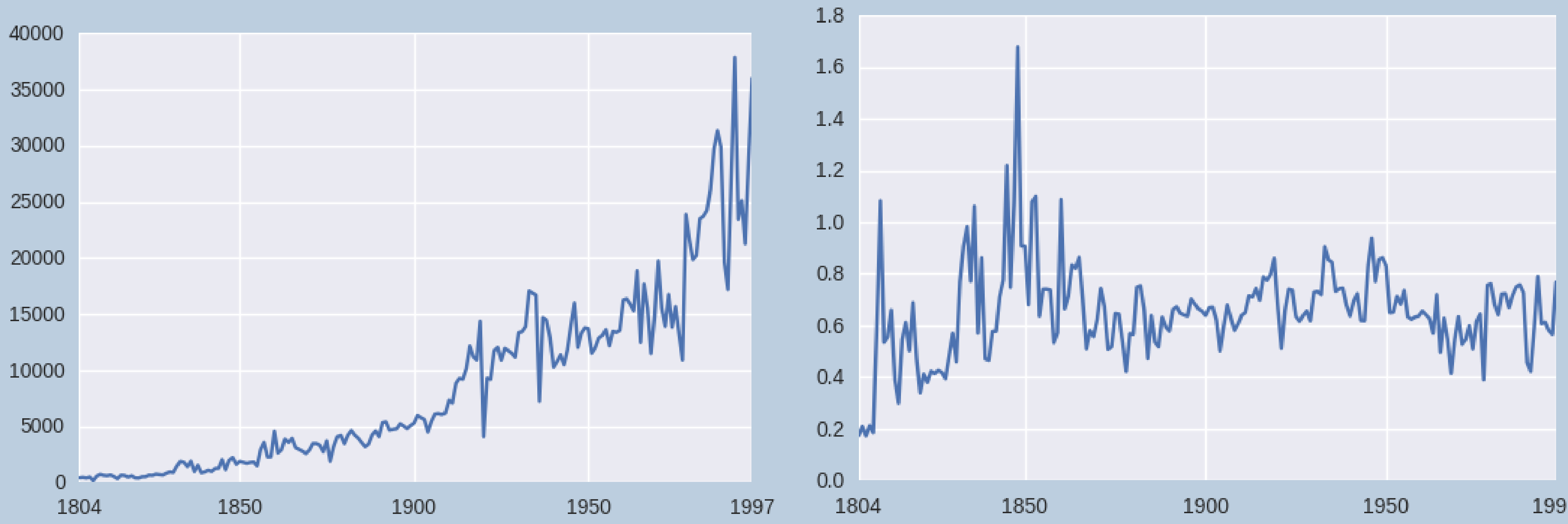

Figure 2: Frequency of occurrences of the word "suisse" (swiss) : Total number (left) and by article (right).

Another interesting point is to observe the evolution of the number of occurrences of a word over time. Due to the distribution of the number of articles, most words follow a trend similar to the one shown on the left in Fig.2. However, the plot of the occurrence's frequency of a word by article often reveals major events as the creation of the federal state in 1848 (Fig.2) or the two World Wars displayed on Fig.3.

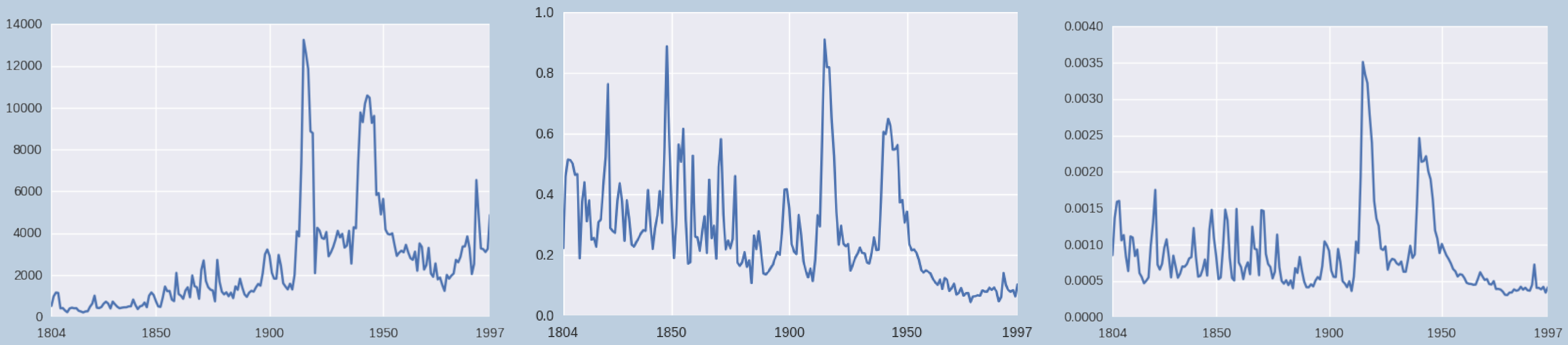

Figure 3: Frequency of occurrences of the word "guerre" (war) : Total number (left), by article (middle), with respect to all other words (right).