Machine learning and natural language processing for the identification of synthesis parameters of NiMo sulfide catalysts

Aline Villarreal[1]*, Rodrigo Villarreal[2], Felipe Sánchez-Minero[1].

[1]*Unidad de Caracterización y Evaluación de Hidrocarburos, Dpto. de Ing. Quím. Petrolera, Esc. Nacional de Industrias Extractivas, IPN, Zacatenco, Alcaldía Gustavo A. Madero, C.P. 07738, CDMX.*

[2]*Departamento de Ing. Metalúrgica, Facultad de Química, UNAM, Circuito de la Investigación Científica, Ciudad Universitaria, Colonia Copilco Coyoacán, Alcaldía Coyoacán, C.P. 04510, CDMX*

**e-mail del autor de correspondencia: aline_vime@hotmail.com*

Introduction

To achieve a sustainable use of energy, new materials for energy storage and catalysis, among others, need to be develop at a much faster rate [1]. Catalysis is an interdisciplinary and complex field where several pieces of information must be put together to design a successful working catalyst. In recent years, theoreticians have contributed to accelerate the discovery of new catalytic materials by putting together information repositories like “Catalysis-Hub”, but often the models only address the molecular or the engineering aspect of the reaction [2].

Moreover, catalyst preparation is a trial an error process that relies heavily on experimentation due to the large number of parameters that need to be carefully controlled; and because small changes in these parameters can lead to huge variations in the final catalyst active sites [2]. Until now, the comparison in preparation methods is difficult even with materials prepared in the same laboratory.

One approach to accelerate materials discovery is the use of machine learning techniques to screen the existing literature and rationalize it [3]. However, a huge challenge to construct a “catalytic preparation procedures library” is that scientific articles use a free natural language that contains domain-specific terminology that lack a common accepted format to report the procedures and its outcomes. Fortunately, there are some natural language processing techniques tools that have been developed to extract information of a large number of scientific articles and retrieve the synthesis parameters such as “ChemDataExtractor” [3, 4]. While in some fields of chemistry these tools are more developed (i. e. organic synthesis, magnetic materials) this is, to our knowledge, the first work that aims to train a natural language processing tool to extract the synthesis parameters in heterogeneous catalysis; specifically those of NiMo sulfide catalysts whose activity varies widely depending on the preparation parameters.

Method

First, several articles in PDF format where downloaded. These articles where identified using a “Google Scholar” search using the keywords: “NiMo sulfide” + “supported” +“catalyst” + “preparation”. Then the text in these articles was converted to an XML file using CERMINE (a java library developed at the University of Warsaw). The paragraphs containing the catalyst preparation procedure were identified manually including catalyst preparation subsections. These paragraphs can be found at [5].

The *ChemDataExtractor* (developed at MIT) was used to parse and tag the retrieved catalyst preparation paragraphs. This toolkit is implemented in Python and allows to split each sentence in tokens (words) that are suitable for natural language processing. The algorithm contained in this toolkit detects and splits each sentence in an unsupervised yet reliable manner, this is largely because the information in scientific literature is orthographically correct and structured in a grammatically precise manner. Due to the precise nature of chemistry literature, the sentences must be parsed unambiguously in a form that leads to no confusion. This was achieved by coding multiple specialized grammar rules designed to extract the information shown in Table 1. This allowed us to extract the synthesis parameters and create a Python list for each relevant synthesis parameter (Support, Metal source, Additives, Impregnation, Drying, Calcination), each item of these lists represent information retrieved from one scientific article. Finally, we use this database and the *sci-kit* toolkit of Python to analyze this information. The chemical names in the figures are presented without Greek characters or subindexes because *sci-kit* does not support them.

Results

The information recovered from the articles showed that γ-Al_2O_3 is the preferred support for these catalysts, since it is used in 90% of the articles. Otherwise, there is a distinct preference for the metal source, for molybdenum the favorite precursor is MoO_3, followed by ammonium molybdate tetrahydrate ($(NH_4)_6Mo_7O_{24}\cdot 4H_2O$) while for nickel the top precursor is Ni $(NO_3)_2\cdot$ $6H_2O$. It was found that when MoO_3 is used the impregnation solution includes H_3PO_4.

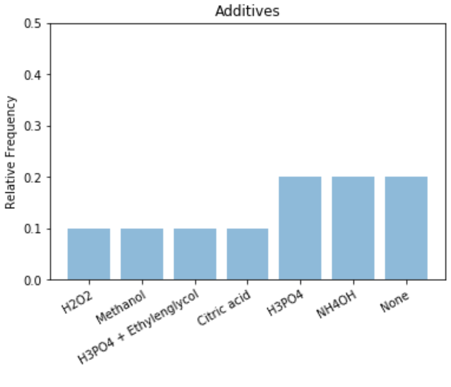

Figure 1. Common additives in the impregnation solutions.

The use of additives reflects the wide variety of preparations that are used in heterogeneous catalysis, the most common additives in the impregnation solution are shown in Figure 1.

Table 1. Relevant synthesis information and common phrases used in the description of the preparation method.

Materials
• Support [Compound (Brand), SurfaceArea (Value, Units)]
Example: "Al2O3 was obtained by annealing (500 °C, 5 h under static air) commercial Pural SB..."
• MetalSource [Compound (Brand)]
Example: "…an impregnation solution containing ammonium heptamolybdate tetrahydrate (Sigma-Aldrich), cobalt (II) carbonate hydrate (Sigma-Aldrich) and citric acid monohydrate (Sigma-Aldrich, 99 %) in deionized water was prepared."
• Additives [Compound (Brand)]
Example: "…an impregnation solution containing ammonium heptamolybdate tetrahydrate (Sigma-Aldrich), cobalt (II) carbonate hydrate (Sigma-Aldrich) and citric acid monohydrate (Sigma-Aldrich, 99 %) in deionized water was prepared."
Method
• Impregnation [Type, Conc (Value, Units)]
Example: "The g-alumina-supported 6 wt. % CoO, 24 wt. % MoO3 catalyst precursors were prepared by wet co-impregnation."
• Drying [Temperature, Time]
Example: "Subsequently, the sample was dried at 110 C for 6 h…"
• Calcination [Type, Temperature, Time, HeatingRate]
Examples: "This solid was dried at 120 8C (2 h), calcining being avoided.", "… the catalyst …and then calcined at the temperature program 823 K (6 h) at a heating rate of 1.7 K/min."

There is also a clear trend in the type of impregnation, around 70 % of the retrieved articles mention the catalyst by pore volume simultaneous impregnation. It was found that only 3 out of 10 articles report the impregnation solution concentration. Figure 2 shows the combinations of temperature and time used during the calcination and drying steps. Most articles report a drying period of 12 h at 120 °C and a calcination stage of between 2 and 6 h at 440 °C on average. Although in this work does not retrieve the final properties of the catalysts, it seems that some of the variations in drying and calcination time and temperature are established empirically.

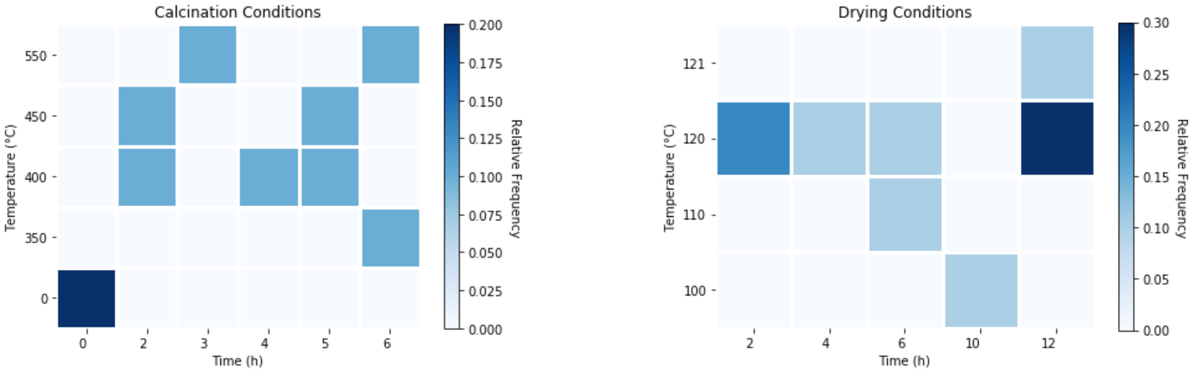

Figure 2. Temperature and time used during the calcination and drying steps of the preparation of NiMo sulfide catalysts.

Conclusions

In this work, synthesis parameters for NiMo sulfide catalysts were extracted from existing literature adding specific algorithms (parsers) to the ChemDataExtractor tool. It was shown that natural language processing techniques can be used to extract information and gain knowledge from a great number of systems and allow to find hidden or misregarded links between preparation conditions. However, some problems happened during data extraction. First, since the methods are written in a free form, the parsers did not retrieve information from all the articles. Second, many articles leave data out, such as the concentration in the impregnation solutions, which leads to gaps in the information. We believe that these problems will be addressed as more researchers are aware (and use) these tools.

Acknowledgements: A. V. work is supported by CONACyT-SENER Hidrocarburos postdoctoral grant.

References

1. Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. J. Cheminform. 2011, 3 (1), 1–13.
2. German Catalysis Society. Whitepaper: The Digitalization of Catalysis-Related Sciences. DECHEMA, 2019.
3. Swain, M. C.; Cole, J. M. J. Chem. Inf. Model. 2016, 56 (10), 1894–1904.
4. Kim, E.; Huang, K.; Tomala, A.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, Sci. Data 2017, 4, 170127.
5. https://github.com/AlineVime/ChemDataExtractor