Towards the extraction of partial instances of N-Ary relations in textual data

Martin Lentschat[1,2,4,], Patrice Buche[2,], Juliette Dibie-Barthelemy[3,], Mathieu Roche[4,5]

[1] I2S Graduate School Montpellier University, France
martin.lentschat@umontpellier.fr

[2] Équipe ICO - UMR IATE Montpellier SuppAgro, Montpellier, France
[3] UMR MIA-Paris AgroParisTech, INRA, University of Paris-Saclay, 75005 Paris, France
[4] UMR TETIS AgroParisTech, Cirad, CNRS, Irstea, Montpellier University, Montpellier, France
[5] CIRAD TETIS, Montpellier, France

Abstract :
This paper presents a generic approach in order to extract experimental information from scientific documents in specialized domains. We are here describing the first phase of our work: the use of an *Ontological and Terminological Resource* (*OTR*) to research *partial instances* of *N-Ary relations*. The *OTR* drives the extraction with its domain vocabulary and guides the creation of a *Representation* of the results. Our contributions in this process is the identification of terminological variations and acronymic forms recognition to increase the coverage of the *OTR* vocabulary. The recognition and classification of text *Segments* (ie. document sections, figures...) is also used to contextualize the extraction of the *partial instances* of *N-Ary relations*.

Mots-clés : N-Ary relations, Ontological and Terminological Resource, Information Extraction in Specialized Domain, Terms Variation, Text Segments

1 Context

The ARTEXT4LOD project (n-ARy relaTions EXTraction for Linked Open Data — ARTEXT4LOD) aims is to develop a method enabling the extraction of experimental data from a corpus of scientific documents in a specialized domain relying on an *Ontological and Terminological Resource* (*OTR*). Our generic method is an extension of (Berrahou *et al.*, 2017).

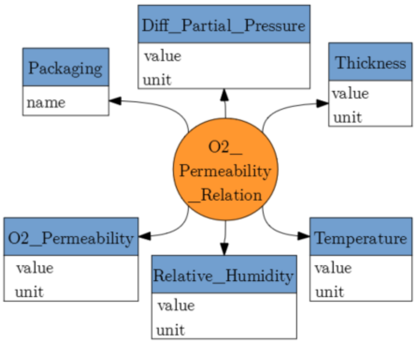

Figure 1: Structure and arguments of an N-Ary relation (food packaging O^2 permeability)

The *OTR* we use is a structured representation of a specific domain, its different concepts and their relations. It defines the information we seek to extract as *N-Ary relations*, with a specific structure and *arguments*, and drives the different steps of the process. This resource also comes with a vocabulary associated with each of its concepts, useful to identify the terms of interest in the documents and their corresponding *arguments*.

One of the biggest hurdles we face is that the *arguments* of *N-Ary relations* are scattered throughout the document. Considering for example the *N-Ary relation* in Figure 1, we may encounter the *argument* concerning the *packaging* name in the *Introduction* section, its *thickness* in *Materials and Methods* and the actual O^2 *permeability* value in *Results and Discussion*.

This involves two necessary phases: **(I)** the extraction of *partial instances* of *N-Ary relations* and **(II)** the reconstitution of *complete instances*.

This paper focuses on the first phase **(I)**, its general process and the contributions we are offering.

2 General Process

Experimental information are characterized by the presence of unit of measure. The extraction of *partial instances* of *N-Ary relations* **(I)** then start with the identification of *Measure Units* as defined in the *OTR*. We consider them as *pivot terms*, defining a context favorable to the presence of *arguments* in textual *Windows* formed by ± 1 sentences around *pivot terms*. This choice of *pivot term* offers the best agreement between quantity and quality of the results and is supported by previous research (Berrahou *et al.*, 2017). To identify and extract the *arguments* we use *features*, associations of *OTR* vocabulary and concepts, to produce *Representations* (in example 1 the term *"polyethylene"* corresponds to the *feature* $< packaging >$).

However, no *OTR* can be absolutely comprehensive and we set up two resources to create an expansion of its vocabulary. Identifying the variations of the terms contained in the *OTR* vocabulary is a simple way to allow the recognition of more *arguments* in the *Representation*. For that, we pre-processed the corpus and the *OTR* vocabulary with **FASTR** (Bourigault & Jacquemin, 1999), a tool used to extract term variations. Scientific publications use acronyms to represent important terms and being able to identify them is essential. We adapt an acronym recognition/disambiguation method (Okazaki & Ananiadou, 2006) by driving it with the *OTR*. Term variations and acronyms are afterwards added to the *features* used to build the *Representations* and should increase its coverage.

Example 1
Window *[...]OP of the polyethylenimine film at 50% RH were* $0.60 * 10^{-18}$ $m^3.mm^2.Pa$*[...]*

Representation $< quantity\_OPacro > < packaging_{termVar} > < numval > < measure Unit\_RH > < quantity\_RHacro > < numval > < measureUnit\_OP >$

In addition, scientific publications are structured documents where the information is contextualized in text *Segments* (ie. sections, figures ...) (Shah *et al.*, 2003). With a recognition and classification of these *Segments* (Hofmann *et al.*, 2009) we can use terms frequency measures to associate a *confidence score*, based on *arguments* and *Segment* associations, to *Representations*. Thresholds will later be applied on *confidence scores* to filter the partial instances before phase (II).

3 Conclusion

This first phase highlights two contributions in the process of *N-Ary relations* extraction: the extension of an *OTR* concepts vocabulary with terms variation and acronyms recognition, and the consideration of the context in which the *arguments* are expressed. Our expectations are that these tasks will respectively improve **recall** and **precision** of the results: by an increasing number of *arguments* identified and the selection of the best *partial instances* before the reconstitution of complete *N-Ary relations* in phase **(II)**.

References

BERRAHOU S. L., BUCHE P., DIBIE J. & ROCHE M. (2017). Xart: Discovery of correlated arguments of n-ary relations in text. *Expert Systems with Applications*, **73**, 115–124.

BOURIGAULT D. & JACQUEMIN C. (1999). Term extraction-i-term clustering: An integrated platform for computer-aided terminology. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

HOFMANN K., TSAGKIAS M., MEIJ E. & DE RIJKE M. (2009). The impact of document structure on keyphrase extraction. In *Proceedings of the 18th ACM conference on Information and knowledge management*, p. 1725–1728: ACM.

OKAZAKI N. & ANANIADOU S. (2006). A term recognition approach to acronym recognition. In *Proceedings of the COLING/ACL on Main conference poster sessions*, p. 643–650: Association for Computational Linguistics.

SHAH P. K., PEREZ-IRATXETA C., BORK P. & ANDRADE M. A. (2003). Information extraction from full text scientific articles: where are the keywords? *BMC bioinformatics*, **4**(1), 20.