



**HAL**  
open science

## **Xart: Discovery of correlated arguments of n-ary relations in text**

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibia-Barthelemy, Mathieu Roche

► **To cite this version:**

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibia-Barthelemy, Mathieu Roche. Xart: Discovery of correlated arguments of n-ary relations in text. *Expert Systems with Applications*, 2017, 73, pp.115-124. 10.1016/j.eswa.2016.12.028 . hal-01508801

**HAL Id: hal-01508801**

**<https://agroparistech.hal.science/hal-01508801>**

Submitted on 18 Nov 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Xart: Discovery of correlated arguments of n-ary relations in text

Soumia Lilia Berrahou<sup>a,b</sup>, Patrice Buche<sup>\*a,b</sup>, Juliette Dibie<sup>c</sup>, Mathieu Roche<sup>a,d</sup>

<sup>a</sup>LIRMM - 860, rue de Saint Priest, 34095 Montpellier, FRANCE

<sup>b</sup>INRA - UMR IATE - 2, place Pierre Viala, 34060 Montpellier, FRANCE

<sup>c</sup>UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, FRANCE

<sup>d</sup>CIRAD - UMR TETIS - 500, rue J.F. Breton, 34093 Montpellier, FRANCE

---

## Abstract

Here we present the Xart system based on a three-step hybrid method using data mining approaches and syntactic analysis to automatically discover and extract relevant data modeled as n-ary relations in plain text. A n-ary relation links a studied object with its features considered as several arguments. We addressed the challenge of designing a novel method to handle the identification and extraction of heterogeneous arguments such as symbolic arguments, quantitative arguments composed of numbers and various measurement units. We thus developed the Xart system, which relies on a domain ontology for discovering patterns, in plain text, to identify arguments involved in n-ary relations. The discovered patterns take advantage of different ontological levels that facilitate identification of all arguments and pool them in the sought n-ary relation.

*Keywords:* Information extraction, N-ary relation, Ontology, Data mining, Sequential pattern, Quantitative data, Linguistic pattern

---

---

\*Corresponding author

*Email addresses:* [lilia.berrahou@gmail.com](mailto:lilia.berrahou@gmail.com) (Soumia Lilia Berrahou),  
[Patrice.Buche@inra.fr](mailto:Patrice.Buche@inra.fr) (Patrice Buche\*), [dibie@agroparistech.fr](mailto:dibie@agroparistech.fr) (Juliette Dibie),  
[mathieu.roche@cirad.fr](mailto:mathieu.roche@cirad.fr) (Mathieu Roche)

## Point-by-Point responses

---

---

### Reviewer #1

1. **Reviewer:** I suggest again to revise the style in the use of English. Moreover, there are yet some typos in the text.

**Response:** *Our paper has now been proofread by a native English speaker.*

2. **Reviewer:** I still think that the paper is too long. I suggest to make another simplification effort. For example, the second domain of application only appears from time to time. I think it could be safely removed (the authors could just mention in the conclusion that they have also tested the system in another domain and it is completely domain-independent).

**Response:** *We agree with this suggestion. In order to clarify our paper, the second domain application is now summarized in the Conclusion section. Moreover, it is important for us to consider our global system in order to highlight main characteristics of Xart and how each step is important for the final objective, i.e. extraction of n-ary relations. But we agree on significantly reducing the part devoted to data-mining (in particular the last step). So in the new version of our paper, we present the global system without the detailed description of the last step. More precisely, we removed section 6 (hybrid approach – p27-31) and associated experiments (p38-40).*

**Reviewer:** Section 2 could be divided in 2 subsections (binary relation extraction and annotated corpora), and its last paragraph on unrelated data mining techniques could be eliminated, along with its references.

**Response:** *we divided Section 2 into two subsections with another (and we hope better) organization.*

3. **Reviewer:** In 5.4 there are missing '(primes) in the definition of subsequence (for example, IS1 included in IS'j1).

**Response:** *This error has been fixed.*

4. **Reviewer:** Ex5 is a little bit confusing after the results of Ex4, since Packaging is linked to numvalthick and um, that are terms that did not appear in its 1-term neighborhood.

**Response:** *We agree with this remark, but we assume that this pattern is not only based on the sentence (2) of the example 2 (data-mining applied on one sentence is really irrelevant) but this OSP was obtained using a large dataset. This has been specified.*

5. **Reviewer:** Some numbers in the textual description of Table2 do not match with those on the table.

**Response:** *These errors have been fixed.*

6. **Reviewer:** The caption of Table 1 should say that the best results (not recalls) are in bold.

**Response:** *This error has been fixed.*

7. **Reviewer:** In 7.1 the new paragraph before "Identification step" has a very bad redaction (e.g. "approach" 3 consecutive times, ":" at the end of a sentence, missing ")". It is really not very understandable. I suggest to remove it from here and move the discussion on the comparison between this new system and wrapper-based approaches to the conclusion.

**Response:** *As suggested, we changed this paragraph, this "discussion" has been moved to the conclusion section.*

8. **Reviewer:** Talking about the conclusion, I think that it has not been significantly improved, as I suggested. I still think it lacks a frank explanation of the limitations/weaknesses of the approach with respect to others, and maybe also a comment on the computational cost.

**Response:** *The conclusion has been changed according to the previous suggestions. We removed some parts. We added information about the genericity and different tests with another corpus (i.e. biorefinery domain).*

## 1. Introduction

Discovering and extracting information reported in textual documents is a crucial issue in several domains in order to be able to reuse, manage, exploit and analyze the information they contain, and use them for decision making purposes (Guillard et al., 2015). The proposed method addresses challenging issues related to n-ary relation identification and extraction in textual documents. More precisely, we aim to propose original patterns that could help domain experts in the difficult task of data annotation. Two examples of n-ary relations are given in sentences (1) and (2), which contain relevant information in two distinct domains, i.e. food packaging and civil aviation. In sentence (1), a studied object (i.e. *polypropylene* film) is analyzed according to different features represented by quantitative data, associated with their numerical value and unit (i.e. *thickness*, *oxygen permeability*, *temperature*, and *relative humidity (RH)*). In sentence (2), the studied object is a plane *A380-800* and its features associated with their numerical value and unit are *transport capacity*, *flying range*, *speed*.

(1) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64  $\mu\text{m}$  thickness polypropylene film with a permeability to oxygen of 110  $\text{cm}^3 \text{m}^{-2} \text{bar}^{-1} \text{day}^{-1}$  at 23  $^{\circ}\text{C}$  and 0 % RH

(2) The A380-800 has a 150 tons of transport capacity, a 15 400 kilometers of flying range that allow a non-stop New York-Hong Kong flight with a 900 km/h up to 1012 km/h of speed

The relevant information extracted from these two sentences can be considered as instances of n-ary relations that could help domain experts in decision making. Nevertheless, instances of n-ary relations are complicated to automatically identify and extract in text because the arguments are often separately expressed in several sentences, usually in implicit and various forms of expression. Moreover, the expression of quantitative arguments frequently varies with

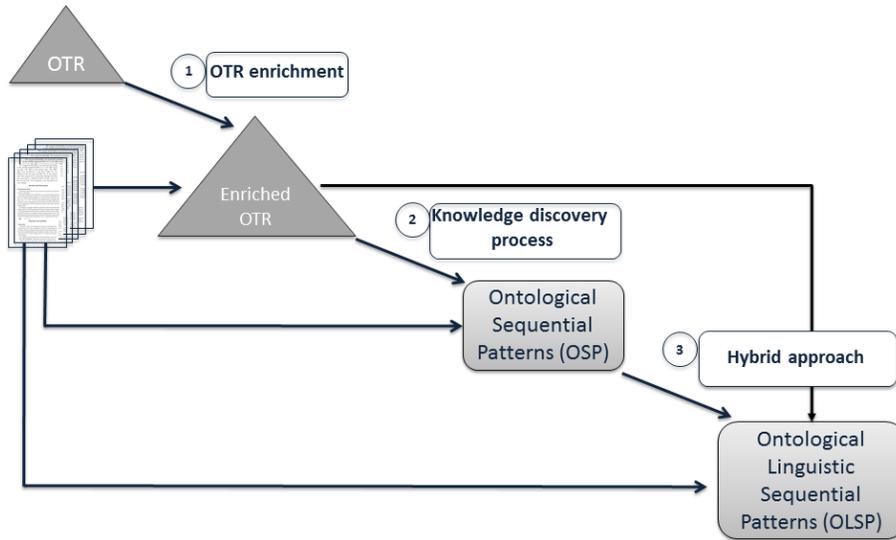


Figure 1: Xart system

30 regard to their attributes, i.e. the numerical value and measurement unit, be-  
 31 tween studied objects.

32 Here we focused on discovering implicit relations in the expression of several  
 33 arguments. An implicit relation is seen as an informal textual expression of  
 34 arguments of the n-ary relation that is not predefined. If such relations exist,  
 35 they could facilitate argument identification in text and argument linkage in  
 36 the sought instance of n-ary relations. To this end, as shown in Figure 1, we  
 37 propose the Xart system based on three main steps driven by an Ontological  
 38 and Terminological Resource (OTR).

39 Since the second and third steps rely on the OTR, the first step consists in  
 40 enriching it with one relevant feature to identify quantitative arguments: the  
 41 measurement unit. The second step takes advantage of data mining approaches  
 42 for discovering correlated argument patterns in text using sequential pattern  
 43 mining. The third step proposes a hybrid approach that uses syntactic analysis  
 44 for constructing original argument identification patterns in text. This third  
 45 step of the Xart system is detailed in (Berrahou et al., 2016).

46

47 The paper is structured as follows. Section 2 presents related work on bi-  
48 nary and n-ary relation extraction fields. Section 3 presents the OTR and key  
49 definitions. Section 4 details the first step, which consists of enriching the OTR  
50 with measurement units that are located and identified with a new edit mea-  
51 sure in textual documents. Section 5 details the second step, which proposes a  
52 knowledge discovery process to extract Ontological Sequential Patterns (OSP).  
53 Section 6 presents the experiments and results. Section 7 concludes the paper.

## 54 2. Related work

55 In this section, we present and discuss related work on textual information  
56 extraction where relevant data are modeled as binary or n-ary relations.

57 **Binary relation extraction.** The approaches proposed to discover relations  
58 between entities as cooccurrences are essentially based on limited linguistic con-  
59 texts. Manually designed patterns are used to identify relevant information  
60 (Huang et al., 2004). In this context, linguistic or syntactic patterns are based  
61 on regular expressions constructed with terms and/or part-of-speech (POS) tags  
62 (Hawizy et al., 2011; Proux et al., 2000; Hao et al., 2005; Raja et al., 2013).  
63 Other approaches (Minard et al., 2011; Rosario & Hearst, 2005; Zhang et al.,  
64 2011; Miwa et al., 2009; Van Landeghem et al., 2009) are designed to resolve  
65 this issue by considering it as a classification problem. Entities are classified as  
66 part or not part of the sought relation. In our work, those methods cannot be  
67 efficiently applied because they rely on small linguistic contexts and require a  
68 large amount of annotated data for training, which usually takes a tremendous  
69 amount of human effort to build. Our approach aims to overcome those tasks  
70 with the hybrid approach that allows the construction of linguistic patterns  
71 based on sequential patterns of correlated arguments, i.e. from two to several  
72 arguments linked in the n-ary relation.

73 Several techniques are proposed, but the process of n-ary relation identification  
74 and extraction is generally based on three main steps: the first step consists

75 in identifying entities (or arguments) using resources such as ontologies or dic-  
76 tionaries; the second step involves identifying the trigger word of the relation  
77 using dictionary-based methods or rule-based approaches to construct patterns  
78 from dependency parse results (Le Minh et al., 2011), or using machine learn-  
79 ing methods (Buyko et al., 2009; Bui & Sloot, 2011; Björne et al., 2009; Zhou  
80 et al., 2014) for predicting which word of the sentence is the trigger word of  
81 the relation; and the third step involves constructing a set of binary relations  
82 using the trigger word, with a given argument being classified as part or not  
83 part of the n-ary relation using machine learning methods. Unfortunately, de-  
84 composing the problem of n-ary relation extraction in extracting several binary  
85 relations results in lower performance. Our approach relies on the knowledge  
86 discovery process using domain knowledge for representing relevant data and  
87 for discovering sequential patterns, including several correlated arguments and  
88 the trigger word of the relation. The trigger word discovered in the patterns  
89 allows all other arguments to be gathered in the sought n-ary relation.  
90 Data mining approaches are used in (Di-Jorio et al., 2008) for enriching on-  
91 tologies with new concepts, in (Béchet et al., 2012; Cellier et al., 2015) for dis-  
92 covering linguistic patterns without external resources, and in (Qiu, 2007) for  
93 adding more semantics and drawing up enhanced association rules. Moreover,  
94 in (Jaillet et al., 2006), the authors use association rules and sequential patterns  
95 to propose comprehensive and reusable text categorization rules. Those tech-  
96 niques have already been successfully used for processing textual data. In line  
97 with these authors, we propose to take advantage of data mining approaches to  
98 discover sequential patterns of several correlated arguments in text.

99 **Available annotated Corpora.** As cited in (Zhou et al., 2014), sev-  
100 eral corpora have been designed for binary relation extraction, e.g. GENIA<sup>1</sup>,  
101 LLL05<sup>2</sup>, AIMed<sup>3</sup>. Those corpora essentially contain sentences with interactions

---

<sup>1</sup><http://www.nactem.ac.uk/genia/genia-corpus>

<sup>2</sup><http://genome.jouy.inra.fr/texte/LLLchallenge/>

<sup>3</sup><ftp://ftp.cs.utexas.edu/pub/mooney/bio-data>

102 between proteins. Other corpora such as LDC2014T27<sup>4</sup> contain benchmarks for  
103 open relation extraction, including binary and n-ary relations, according to sen-  
104 tences extracted and annotated from the New York Times and the Treebank-3.  
105 While we looked for a standard evaluation dataset to assess our approaches,  
106 those corpora do not concern us since we focus on quantitative data involving  
107 numerical values and measurement units in n-ary relations. The aforementioned  
108 corpora are designed for binary or n-ary relations involving essentially named  
109 entities (e.g. proteins, locations, organisations). To the best of our knowledge,  
110 open corpora involving quantitative data do not exist. We chose to build our  
111 own corpus with complete articles from online databases (e.g. Wiley, Elsevier,  
112 Springer) with expert validation for the assessment task.

113

### 114 **3. Xart system key elements**

115 In this section, we present key elements of the Xart system involving a hy-  
116 brid approach to extract correlated arguments of n-ary relations from text. The  
117 Xart system relies on an Ontological and Terminological Resource (OTR). The  
118 OTR is a relevant semantic support for the Xart system, which enables termi-  
119 nology associated with n-ary relations in text to be represented with different  
120 conceptual levels.

#### 121 *3.1. An ontology for n-ary relation representation*

122 In our work, relevant data are represented as n-ary relations where a stud-  
123 ied object is modeled as a symbolic argument and its features as quantitative  
124 arguments associated with their attributes, i.e. the numerical value and mea-  
125 surement unit. Our representation of n-ary relations is that of the naRyQ (n-ary  
126 Relations between Quantitative data) OTR (Touhami et al., 2011; Buche et al.,  
127 2013b). naRyQ contains two components, i.e. a terminological component and  
128 a conceptual component. The conceptual component of naRyQ is composed of

---

<sup>4</sup><https://catalog.ldc.upenn.edu/LDC2014T27>

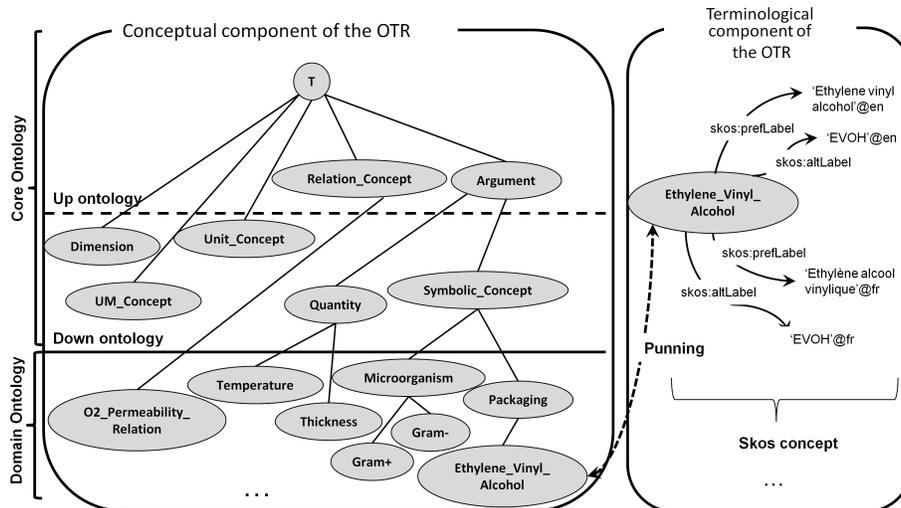


Figure 2: An excerpt of the naRyQ concept hierarchy in the food packaging domain

129 a core ontology to represent n-ary relations and a domain ontology to represent  
 130 specific concepts of a given application domain. Note that each step of the Xart  
 131 system relies on the core ontology, which is domain independant.

132 Figure 2 illustrates an application of naRyQ in the food packaging domain.  
 133 In the up core ontology, generic concepts Relation\_Concept and Argument re-  
 134 spectively represent n-ary relations and their arguments. In the down core on-  
 135 tology, generic concepts Dimension, UM\_Concept, Unit\_Concept and Quantity  
 136 allow the management of quantities and their associated measurement units.  
 137 Note that the measurement units are represented by instances of the generic con-  
 138 cept Unit\_Concept. The subconcepts of the generic concept Symbolic\_Concept  
 139 represent the non-numerical arguments of n-ary relations. The domain ontology  
 140 contains specific concepts of a given application domain. They appear in naRyQ  
 141 as subconcepts of the generic concepts of the core ontology. The terminological  
 142 component of naRyQ contains the set of terms describing the studied domain.  
 143 naRyQ presented in (Touhami et al., 2011; Buche et al., 2013b) may be formally  
 144 define as follows.

145 **Definition 1.**

146 An Ontological and Terminological Ressource is a sextuple  $OTR = \langle C_{OTR}; R; I; V; \leq_o$   
 147  $; W_{oi} \rangle$  where:

- 148 •  $C_{OTR}$  is a set of conceptual components of the OTR,
- 149 •  $C_{OTR} = C_{Rel} \cup C_{Qty} \cup C_{Symb}$  with  $C_{Rel}$  the set of n-ary relations,  $C_{Qty}$   
 150 the set of quantities,  $C_{Symb}$  the set of symbolic concepts;
- 151 •  $R$  is a set of relations in  $C_{OTR} \times C_{OTR}$ ;
- 152 •  $I$  is a set of instances with  $I_{UM} \subset I$ , i.e. the subset of instances which  
 153 represents measurement units;
- 154 •  $V$  is a set of values;
- 155 •  $\leq_o$  is a specialisation relation in  $(C_{OTR} \times C_{OTR}) \cup (R \times R)$ ;
- 156 •  $W_{oi}$  is a set of terms in the terminological component of the OTR, where  
 157 all terms  $w_i \in W_{oi}$  denote either a concept  $c \in C_{OTR}$  or a measurement  
 158 unit  $u \in I_{UM}$ .

159 A n-ary relation is represented by a concept which is linked to its arguments  
 160 by binary relations such that none of these arguments has a specific role (e.g.  
 161 subject or object). A formal definition of the representation of n-ary relations  
 162 between quantitative data is given below.

163 **Definition 2.**

164 Let us consider  $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$  of Definition 1. A **n-ary**  
 165 **relation concept**  $rel \in C_{OTR}$ ,  $\leq_o(rel, Relation\_Concept)$ , is defined in OTR  
 166 by the set of binary relations  $r_j \in R$  which link the n-ary relation  $rel$  with its  
 167 arguments, with this set being composed of at least two binary relations:

$$Def(rel) = \{r_j(rel, a_j) \mid r_j \in R, \\ (a_j \in C_{OTR} \wedge \leq_o(a_j, Argument))\}, \\ \text{such that } |Def(rel)| \geq 2$$

168 A n-ary relation is characterized by its signature, i.e. the set of its arguments.

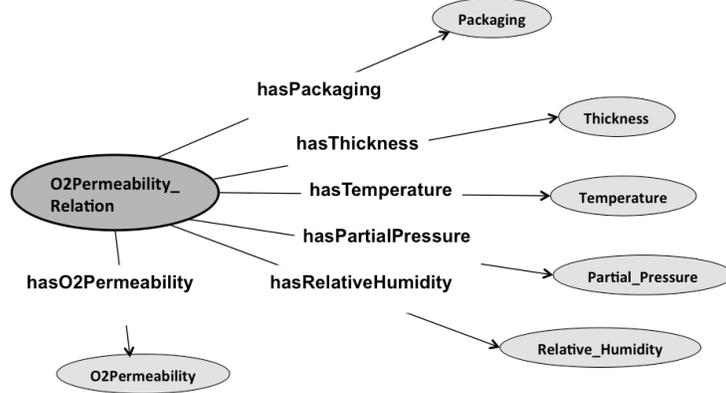


Figure 3: The n-ary relation O2Permeability\_Relation

169 **Definition 3.**

170 *Let us consider  $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$  of Definition 1 and a n-ary*  
 171 *relation concept  $rel \in C_{OTR}$  as defined in Definition 2. The **signature** signa-*  
 172 *ture  $R: C_{OTR} \rightarrow 2^{C_{OTR}}$  of the n-ary relation concept  $rel$  is:*

$$signatureR(rel) = \{(a_j \in C_{OTR} \wedge \leq_o(a_j, Argument)) \mid r_j(rel, a_j) \in Def(rel)\}$$

173 An example of a n-ary relation concept is given in Figure 3 and represents  
 174 the relation O2Permeability\_relation in the naRyQ\_pack OTR (food packaging  
 175 domain OTR). The signature of the n-ary relation O2Permeability\_Relation is:  
 176 signatureR(O2Permeability\_Relation)= {Packaging, Thickness, Temperature,  
 177 Partial\_Pressure, Relative\_Humidity, O2Permeability}.

178

179 *3.2. Xart textual context*

180 The following hypothesis underlies the Xart system: **measurement units**  
 181 **associated with quantitative arguments are considered as relevant fea-**  
 182 **tures in text to define an optimal context for discovering the sought**

183 **arguments.** From this hypothesis, we propose two relevant textual search con-  
184 texts: pivot sentence and textual window defined as follows:

185

186 **Definition 4.** (*Pivot sentence*)

187 *A pivot sentence is defined as the sentence where at least one unit referenced in*  
188 *the OTR is identified*

189 **Definition 5.** (*Textual window*)

190 *A textual window denoted  $f_{sn}$  is defined as the set of sentences composed of the*  
191 *pivot sentence and the  $n$  previous sentences, and/or the  $n$  subsequent sentences,*  
192 *where  $n$  is the window dimension. The search direction in sentences, denoted  $s$ ,*  
193 *is represented with  $-$  considering previous sentences,  $+$  considering subsequent*  
194 *sentences or  $\pm$  considering previous and subsequent sentences*

195 The textual window is a relevant textual context for the discovery of infor-  
196 mation about n-ary relations over the three steps of the Xart system.

#### 197 **4. The Xart first step: Enrichment of the OTR with measurement** 198 **units**

199 In this section, we present the first step of the Xart system based on Def-  
200 initions 4 and 5 which consists of locating and identifying measurement units  
201 in text in order to enrich the OTR. Those tasks are difficult because the units  
202 are hampered by a wide range of typographic variations in text (e.g.  $cm^3 m^{-2}$   
203  $bar^{-1} day^{-1}$  or  $cm^3/m^2/bar/day$ ) and a wide range of combinations between  
204 subunits to express a complex unit (e.g. unit of permeabilities). In this context,  
205 we cannot apply predetermined recognition patterns and wrapper based ap-  
206 proaches. Indeed, related work, e.g. in (Jessop et al., 2011a), has revealed that  
207 most quantitative data extraction failures are due to typographic variations of  
208 units in text. In chemistry, an efficient tool for text-mining, i.e. ChemicalTagger  
209 (Hawizy et al., 2011) is proposed not only for the identification and annotation  
210 of chemical entities (Jessop et al., 2011b) but also of relationships linking these

211 entities. The tool relies on the use of a Regex-tagger based on regular expres-  
212 sions in order to identify sentences where quantitative data, chemical entities  
213 and units appear. However, in (Jessop et al., 2011a), the authors note that  
214 ChemicalTagger fails in the process of recognizing chemical names as reagents  
215 because of typographic variations of units in text.

216 Several domain ontologies have been modeled for units and measurements, such  
217 as EngMath (Gruber & Olsen, 1994), Measurement units in clinical information  
218 systems, UCUM (Schadow et al., 1999), Quantities, Units, Dimensions and Data  
219 Types Ontologies, QUDT (Hodgson et al., 2013), units.obo (Gkoutos, 2011) or  
220 Ontology of Measurement units and Related Concepts, OM (Rijgersberg et al.,  
221 2013) in order to exchange and process quantitative information. However, do-  
222 main authors can freely use typographic variations to write measurement units  
223 in scientific documents. Moreover, domain ontologies often do not entirely over-  
224 lap and several units do not exist in those ontologies, especially when considering  
225 documents of a specific scientific area (e.g. food packaging, biorefinery). Thus,  
226 enriching the ontology is a key step in the proposed process. Since units do not  
227 follow syntactic rules of common words, using specific patterns to identify units  
228 in text is not a trivial task. Our approach aims at addressing this issue using  
229 supervised learning methods and proposing a new edit measure.

#### 230 *4.1. Locating units*

231 In this subsection, the aim of the Xart system is to reduce the search space  
232 of units having typographic variations using a text mining approach. The pro-  
233 posed method is intended to predict whether a part of a text contains a unit  
234 (typographic variations) or not by applying binary classification.

235 ***Data preparation.*** Data preparation involves text processing and text trans-  
236 formation tasks. Text processing consists of:

- 237 • text segmentation in order to generate a set of sentences;
- 238 • text cleanup, which removes punctuation and special characters from text,  
239 except those involved in units;

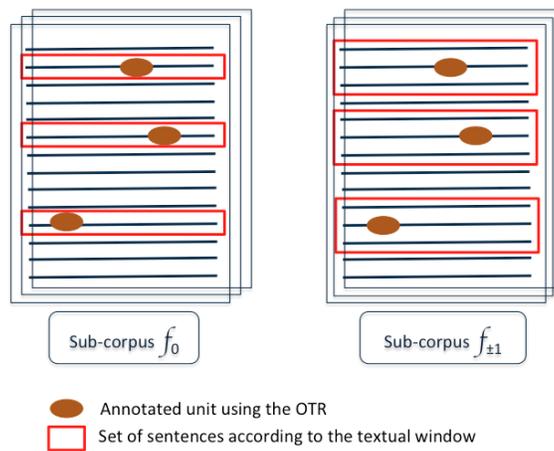


Figure 4: Subcorpus preparation

- 240 • text tokenization, which splits a string of characters into a set of tokens;
- 241 • text reduction, which prunes away tokens containing junk words according
- 242 to a list of stop words;
- 243 • text tagging, which automatically annotates the units in text using all
- 244 unit terms referenced in the OTR.

245 ***Subcorpus preparation.*** After text tagging, the corpus is divided into several  
 246 subcorpora according to several textual windows, as shown in Figure 4. Those  
 247 subcorpora are used as training data.

248 ***Data transformation.*** This process aims at transforming each sentence in  
 249 a vector to constitute the training matrix for the learning step. It involves  
 250 representing a text according to the words it contains and their occurrences  
 251 in the sentences. Selected words (features) make the *bag-of-words* and their  
 252 occurrence in each sentence is computed according to the three following word  
 253 weighting measures:

- 254 • Term Frequency (TF), which considers that the word is more important
- 255 if it appears several times in sentences;

- 256 • Term Frequency-Inverse Document Frequency (TF.IDF (Hiemstra, 2000)),  
257 which considers that the word is more important if it appears in fewer  
258 sentences;
- 259 • Okapi BM25 (Jones et al., 2000), which also takes into consideration the  
260 length of the sentence in which the word appears to define its relevance.

261 In this work, positive examples, i.e. sentences containing measurement units,  
262 and negative examples, i.e. sentences randomly selected in the corpus and that  
263 do not contain any measurement units, are used in order to create the training  
264 matrix. The learning step of the training matrix proposes a model able to  
265 predict whether a part of text contains a unit or not (i.e class "unit" and class  
266 "non-unit").

267 ***Model learning.*** Each evaluated training matrix is run under several learning  
268 algorithms:

- 269 • Naive Bayes classifier and the Discriminative Multinomial Naive Bayes  
270 (DMNB) classifier;
- 271 • J48 decision tree classifier ;
- 272 • Sequential Minimal Optimization (SMO), which is a Support Vector Ma-  
273 chine classifier (SVM).

274 The aim of the assessment is to carry out exhaustive experiments in order to con-  
275 clude on the best classification model. Those widely known learning algorithms  
276 are chosen by comparing their behavior on corpora containing many quantita-  
277 tive data. Naive Bayes (John & Langley, 1995) is competitive for computational  
278 efficiency. Decision tree (Kohavi & Quinlan, 2002) classifiers are known to ob-  
279 tain good classification results but are less competitive in execution speed. SMO  
280 (Platt, 1999) is a discriminative classifier known to efficiently behave on text  
281 classification and, DMNB is an original text classification algorithm (Su et al.,  
282 2008) which performs competitively with discriminative classifiers in terms of

283 accuracy, without losing the computational efficiency advantages of Naive Bayes  
284 classifiers.

285 **Results assessment.** The obtained results are compared in terms of the pre-  
286 cision, recall, and F-measure. The recall value is an important measure to assess  
287 relevant sentences that are retrieved without too much precision loss. The con-  
288 fusion matrix is interesting to compare the results of the tested classifier with  
289 trusted external judgements. As we want to estimate how accurately the model  
290 of each classifier will perform in practice, a 10-fold cross-validation is used: The  
291 original sample is randomly partitioned into 10 equal sized subsamples. One  
292 subsample is used as validation data for testing the model while the other sub-  
293 samples are used as training data. This process is repeated 10 times with each  
294 subsample used once as validation data. The average result produces the model  
295 estimation. Using cross-validation is crucial to avoid "overfitting" effects of the  
296 model. According to the compared results, the best model is then reused to  
297 predict whether or not a new sentence from any text contains new units to be  
298 identified.

#### 299 4.2. Identifying units

300 From the previous step, the studied corpora were reduced to the significant  
301 sentences, i.e. those classified as potentially containing a typographic variation.  
302 Typographic variations of units are then extracted and identified in order to  
303 enrich the OTR. Units with typographic variations are extracted from the sen-  
304 tences using a dictionary of common words. All common words or numerical  
305 values identified in the sentence are eliminated, so that we only keep the unit  
306 with typographic variations to identify.

307 The identification process relies on a similarity value obtained when the unit  
308 is compared to a set of reference units in the OTR: the higher the value, the  
309 closer the two units. Let us consider a simple example of a unit using a ty-  
310 pographic variation *amol/m.sec.Pa* compared to the reference unit in the OTR  
311 *amol/(m.s.Pa)*. In the identification process, we consider that units are com-  
312 posed of blocks, which represent subunits. In our example, *amol/(m.s.Pa)* is

313 composed of four blocks, *amol*, *m*, *s*, and *Pa* whereas *amol/m.sec.Pa* is composed  
 314 of *amol*, *m*, *sec*, and *Pa*. The identification process consists of:

- (1) Pre-selecting a set of relevant candidate units to be compared (i.e. a unit having typographic variations and a unit from the OTR) using a Jaccard measure that allows the common blocks to be evaluated in the two units ( $u_1, u_2$ ) without the block order constraint using *bl* a function associating a unit with its set of blocks:

$$Jaccard(u_1, u_2) = \frac{|bl(u_1) \cap bl(u_2)|}{|bl(u_1) \cup bl(u_2)|}$$

- (2) Pre-selected candidate units are then compared using our new edit measure,  $SM_{D_b}$ , we adapted from the Damerau-Levenshtein distance ( $D_c$ ) (Damerau, 1964) used to compare characters. The distance  $D_c$  between two strings is defined as the minimum number of edits needed to transform one string into another, with the edit operations being insertion, deletion, or substitution of a single character. The distance  $D_c$  can then be normalized by using the approach detailed in (Maedche & Staab, 2002):

$$SM_{D_c}(u1, u2) = \max\left[0; \frac{\min(|u1|, |u2|) - D_c(u1, u2)}{\min(|u1|, |u2|)}\right] \\ \in [0; 1]$$

315 The similarity measure is computed and the higher this measure is, the  
 316 closer the unit  $u1$  is to the unit  $u2$ .

317  $SM_{D_b}$  considers the same edit operations as being an insertion, deletion,  
 318 or substitution of blocks, not of a single character. Example 1 shows  
 319 the relevance of  $SM_{D_b}$  to identify units with typographic variations as  
 320 compared to the classical measure  $SM_{D_c}$ .

321 **Example 1.** *Let us consider the similarity between  $kg\ m\ Pa^{-1}\ s^{-1}\ m^{-2}$  and*  
 322 *its OTR referent  $lb.m.m^{-2}.s^{-1}.Pa^{-1}$ . Those two units cannot be directly com-*  
 323 *pared to the classical distance  $D_c$ , which can only compare strings of characters.*  
 324 *Actually, the first unit is composed of several blank spaces that do not allow*

325 comparison. If we try to replace those blank spaces with another character to  
 326 make the comparison possible, we need to choose a non-specific unit character  
 327 (e.g. the underscore '\_') because other characters such as '×', '.', ',', '/' symbolize  
 328 specific operations in units.  $kg\ m\ Pa^{-1}\ s^{-1}\ m^{-2}$  becomes  $kg\_m\_Pa^{-1}\_s^{-1}\_m^{-2}$ .  
 329 The classical distance  $D_c$  computes the similarity by considering all differences  
 330 between the two units: 12 different characters, 3 new characters are inserted, 4  
 331 substitutions of characters. The  $D_c$  (the distance between those units) is there-  
 332 fore 19 and the similarity distance normalized according to  $D_c$  is:

$$SM_{D_c}(kg\ m\ Pa^{-1}\ s^{-1}\ m^{-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}) = \max[0; \frac{|17-19|}{|17|}]$$

$$SM_{D_c} = 0.12$$

334 Our new approach allows those two units to be directly compared:

335

336 (1) They are first pre-selected with the Jaccard measure as relevant for com-  
 337 parison with our  $SM_{D_b}$  measure.

338

339 (2) The new measure  $SM_{D_b}$  then allows us to more accurately identify those  
 340 units:

$$341\ SM_{D_b}(kg\ m\ Pa^{-1}\ s^{-1}\ m^{-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}) = \max[0; \frac{5-1}{5}] = 0.8.$$

342

343 The unit  $kg\ m\ Pa^{-1}\ s^{-1}\ m^{-2}$  is finally associated with its OTR referent and  
 344 validated to enrich the OTR.

345 In this first step, the ontology is enriched with new units and terminological  
 346 variations of existing units. These units will be used to define more relevant  
 347 textual contexts in the second step of the Xart system.

## 348 5. The second Xart step: knowledge discovery process

349 This section presents the second step, called the *Knowledge discovery pro-*  
 350 *cess*, (see Figure 1) of the Xart system. This step aims at discovering frequent

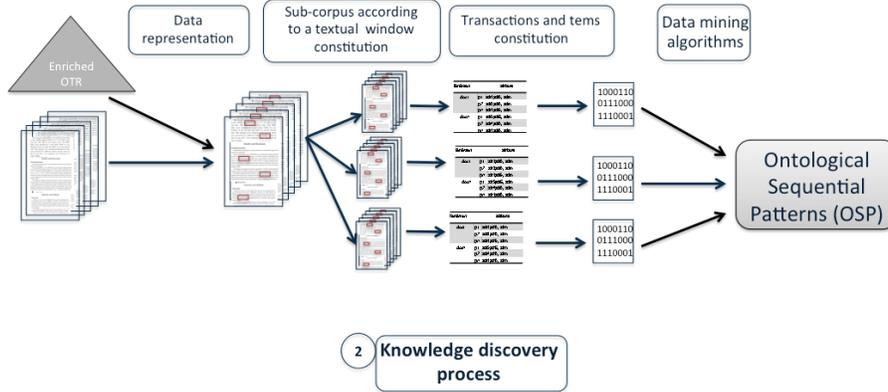


Figure 5: Second step of the Xart system: Knowledge discovery process driven by a domain OTR

351 patterns involving arguments of n-ary relations using data mining approaches.  
 352 In this second step of the Xart system, the discovery of Ontological Sequential  
 353 Patterns (OSP) is driven by the OTR and is composed of the four substeps pre-  
 354 sented in Figure 5: (1) a new data representation; (2) subcorpus constitution;  
 355 (3) transactions and items; (4) data mining.

356 *5.1. First substep: a new data representation*

357 In this section, we propose a new data representation using the OTR concep-  
 358 tual level in order to increase relevant data expressiveness in text. Our aim  
 359 is to extract argument instances, whose forms of expression frequently change  
 360 in text and whose numerical values frequently change according to the measure-  
 361 ments obtained on the studied object. Mining frequent patterns directly on text  
 362 without increasing the expressiveness of n-ary relation arguments substantially  
 363 decreases the knowledge discovery process efficiency. We propose to tackle this  
 364 issue by taking the data expressiveness into consideration using a new repre-  
 365 sentation. This new representation relies on the signature of the sought n-ary  
 366 relation in terms of symbolic and quantitative arguments. We propose, with

367 Definition 6, to increase the expressiveness of the symbolic arguments by rep-  
 368 resenting them with their corresponding concepts, subconcepts of the generic  
 369 concept *Symbolic\_Concept*, which belong to the signature of the sought n-ary  
 370 relation. For example, in our experiments on the packaging domain corpus,  
 371 we choose the subconcept *Packaging* which belongs to the signature of the re-  
 372 lation *O2Permeability\_Relation* in order to represent the studied packaging in  
 373 text (e.g. gluten).

374 **Definition 6.** (*Symbolic concept representation for a sought relation rel*)  
 375 Let us consider  $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$  of Definition 1.  $\forall t$ , a term of  
 376 the text,  $t$  is annotated by  $c_j \in C_{Symb}$ , denoted by  $\langle C_{Symb} \rangle$ , in the new data  
 377 representation if:

- 378 (i)  $c_j \in SignatureR(rel)$  and  $rel \in C_{Rel}$ ,
- 379 (ii)  $\exists c_i \in C_{Symb}, c_i \leq_o c_j$ ,
- 380 (iii)  $\exists w_i \in W_{oi}$  which denotes  $c_i$  such that  $sim(w_i, t)=1$ , *sim* being a similarity  
 381 measure.

382 We propose, with Definition 7, to increase the expressiveness of the quanti-  
 383 tative arguments of the sought n-ary relation by increasing the expressiveness  
 384 of their numerical values using their associated measurement units. Numeri-  
 385 cal values indeed represent the relevant information we want to discover and  
 386 which often vary, depending on the measurements obtained on the studied  
 387 object. Measurement units related to numerical values are associated in the  
 388 OTR with specific subconcepts of the generic concept *Quantity* by the relation  
 389 *hasUnit* ( $hasUnit \in R$ ) (e.g. the measurement unit  $^{\circ}C$  is associated in the  
 390 naRyQ OTR with the quantity *Temperature*). We use those *Quantity* subcon-  
 391 cepts, which belong to the signature of the sought n-ary relation, to represent  
 392 numerical values.

393 **Definition 7.** (*Quantity concept representation for a sought relation rel*)  
 394 Let us consider  $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$  of Definition 1. Let us also

395 consider  $v_i$ , a numerical value in the text associated, in the text, with  $t_i$ , a term,  
 396 and  $t_u$ , a unit term. Then  $v_i$  is annotated by  $C_i^u$ , denoted by  $\langle \text{numval}C_i^u \rangle$ ,  
 397  $t_i$  by  $\langle \text{Quantity} \rangle$  and  $t_u$  by  $\langle \text{um} \rangle$  in the new data representation if:

398 (i)  $\exists w_i \in W_{oi}$  which denotes  $C_i^u \in C_{Qty}$  with  $C_i^u \in \text{Signature}R(\text{rel})$ ,  $\text{rel} \in$   
 399  $C_{Rel}$  and  $\text{sim}(w_i, t_i) = 1$ ,

400 (ii)  $\exists w_j \in W_{oi}$  such that  $w_j$  denotes  $i \in I_{um}$  and  $\text{sim}(w_j, t_u) = 1$  where  
 401  $i \in \text{hasUnit}(C_i^u)$ .

402 The two previous definitions are illustrated in Example 2. In sentence (1),  
 403 the expressiveness of the underlined data is improved using Definitions 6 and 7.  
 404 Sentence (2) corresponds to the new data representation of sentence (1). This  
 405 sentence contains an instance of `O2Permeability_Relation`, as described in Fig-  
 406 ure 3, which represents the oxygen permeability of a packaging under given  
 407 experimental conditions. The experimental conditions are defined by the pack-  
 408 aging thickness ( $64 \mu\text{m}$ ), temperature ( $23^\circ\text{C}$ ) and relative humidity (0%). More  
 409 precisely, note that the numerical value  $64$  is followed by the unit  $\mu\text{m}$  that is  
 410 associated with the *Thickness* concept. Thus,  $\langle \text{numvalthick} \rangle$  is used to anno-  
 411 tate the value  $64$ , which is the relevant instance to be identified in text, and we  
 412 represent the term "thickness" by  $\langle \text{Quantity} \rangle$  and the term " $\mu\text{m}$ " by  $\langle \text{um} \rangle$ .

413 **Example 2.**

414

415 (1) *Eight apple wedges were packaged in polypropylene trays and wrap-sealed*  
 416 *using a  $64 \mu\text{m}$  thick polypropylene film with an oxygen permeability of*  
 417  *$110 \text{ cm}^3 \text{ m}^{-2} \text{ bar}^{-1} \text{ day}^{-1}$  at  $23^\circ\text{C}$  and  $0\%$  RH.*

418

419 (2) *Eight apple wedges were packaged in polypropylene  $\langle \text{Packaging} \rangle$  trays*  
 420 *and wrap-sealed using a  $64 \langle \text{numvalthick} \rangle \mu\text{m} \langle \text{um} \rangle$  thick  $\langle \text{Quantity} \rangle$*   
 421 *polypropylene  $\langle \text{Packaging} \rangle$  film with an oxygen permeability  $\langle \text{Quantity} \rangle$*   
 422 *of  $110 \langle \text{numvalperm} \rangle \text{ cm}^3 \text{ m}^{-2} \text{ bar}^{-1} \text{ day}^{-1} \langle \text{um} \rangle$  at  $23 \langle \text{numvaltemp} \rangle$*   
 423  *$^\circ\text{C} \langle \text{um} \rangle$  and  $0 \langle \text{numvalrh} \rangle \% \langle \text{um} \rangle$  RH  $\langle \text{Quantity} \rangle$ .*

424 In the second substep, we propose to define relevant textual contexts in order  
425 to encompass the involved arguments of the sought n-ary relation.

### 426 5.2. Second substep: subcorpus constitution

427 We obtained our subcorpus in the same way as we did in the first step of  
428 the Xart system (see Figure 4) by applying Definitions 4 and 5. The process  
429 allows, in the data mining step, several subcorpora obtained in different textual  
430 windows to be assessed.

### 431 5.3. Third substep: transactions and items

432 This subsection presents the data preparation in the knowledge discovery  
433 process. The data must be organised in two sets in order to be efficiently mined  
434 by the algorithms. A set of transactions, according to Definition 8, and an  
435 itemset, according to Definition 9, are proposed and are associated with each  
436 studied subcorpus, i.e. with each relevant textual window.

#### 437 **Definition 8.** (*Transaction*)

438 *A transaction is defined as a set of sentences according to a textual window  $f_{sn}$ .*

#### 439 **Example 3.**

440 *In a textual window  $f_{\pm 1}$ , each transaction corresponds to a set of sentences  
441 composed of the pivot sentence, the previous and subsequent sentences.*

#### 442 **Definition 9.** (*Itemset*)

443 *An itemset  $IS^n$  is the set of  $n$  nearest terms or annotations associated with a  
444 given argument of a sought relation  $rel$  in the data representation detailed in  
445 Definitions 6 and 7.*

#### 446 **Example 4.**

447 *Let us consider the sentence (2) of Example 2, if we choose to select the 1-  
448 term nearest neighbors of the annotation <Packaging>, we obtain an item-  
449 set composed of <Packaging>, polypropylene, trays, films. For the annota-  
450 tion <Quantity>, we obtain an itemset composed of <Quantity>, thickness,  
451 polypropylene, oxygen, <numvalperm>, RH.*

452 *5.4. Fourth substep: Data mining*

453 The fourth substep of the knowledge discovery process is based on data  
 454 mining. Each studied subcorpus associated with its sets of transactions and  
 455 itemsets according to a relevant textual window is mined by the algorithms.  
 456 This substep is intended to extract the Ontological Sequential Patterns (OSP)  
 457 that allow the correlations of arguments expressed in text to be discovered.

458 Based on data mining definitions of (Agrawal & Srikant, 1995), we propose  
 459 Definitions 10 and 11 tailored from previous definitions to our context of the  
 460 knowledge discovery process driven by the OTR and based on our new data  
 461 representation.

462 **Definition 10.** (*OS - Ontological Sequence*)

463 *Let us consider  $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$  of Definition 1. An ontological*  
 464 *sequence  $OS_{f_{sn}}$  is a non-empty ordered list of itemsets  $IS_j^n$  extracted in a textual*  
 465 *window  $f_{sn}$ , denoted  $\langle IS_1^n IS_2^n \dots IS_p^n \rangle$ .*

466 **An ontological sequential pattern** is a frequent ontological subsequence  
 467 characterized by a support, which represents the number of occurrences of a  
 468 pattern in a set  $\mathcal{OS}$  of ontological sequences. Extracting frequent ontological  
 469 sequential patterns involves extracting patterns with a support value greater  
 470 than a minimum support parameter  $\theta$ . Let  $\mathcal{M}$  be a set of extracted ontological  
 471 sequential patterns, then  $\forall M \in \mathcal{M}, Support(M) \geq \theta$ . Thus extracting onto-  
 472 logical sequential patterns involves searching frequent ontological subsequences  
 473 from  $\mathcal{OS}$ .

475 **Definition 11.** (*OSP - Ontological Sequential Pattern*)

476 *Let  $(OS_{f_{sn}})_A = \langle IS_1^n IS_2^n \dots IS_p^n \rangle$  be an ontological subsequence of another*  
 477 *ontological sequence  $(OS_{f_{sn}})_B = \langle IS_1^m IS_2^m \dots IS_m^m \rangle$ , then  $((OS_{f_{sn}})_A \preceq$   
 478  $OS_{f_{sn}})_B$  if  $p \leq m$  and  $IS_1^n \subseteq IS_{j_1}^m, IS_2^n \subseteq IS_{j_2}^m, \dots, IS_p^n \subseteq IS_{j_p}^m$  with  $1 <$   
 479  $j_1 < j_2 < \dots < j_k < \dots < j_p < m$ . Let  $\theta$  be a minimum support, then the  
 480 *Ontological Sequential Pattern OSP is defined as a set of frequent subsequences*  
 481 *from  $OS_{f_{sn}}$  such that  $Support(OSP) \geq \theta$ .**

482 **Example 5.**

483 *Let us consider the OSP  $\langle(\text{Packaging})(\text{numvalthick } um)\rangle$  supported by  $OS_{f_{\pm 1}}$*   
484 *obtained with a large dataset. This OSP is extracted from the set of sequences*  
485 *in the textual window  $f_{\pm 1}$ . It allows us to obtain a correlation between the pack-*  
486 *aging concept, defined in the OTR of the food packaging domain, and the repre-*  
487 *sentation of its thickness given by  $\text{numvalthick}$ . The pattern given in the OSP*  
488 *shows that the expression of the studied object (i.e. the packaging) frequently*  
489 *occurs with its thickness in text and this cooccurrence is frequently discovered in*  
490 *a textual window  $f_{\pm 1}$  (i.e. a context extended to three sentences).*

491 The third step of the Xart system is the hybrid approach detailed in (Berra-  
492 hou et al., 2016), which proposes to combine OSP with syntactic analysis in  
493 order to construct Ontological Linguistic Sequential Patterns (OLSP) for iden-  
494 tifying correlated arguments directly in text.

495 **6. Experiments and results**

496 *6.1. OTR enrichment*

497 **Subcorpus constitution.** From the food packaging corpus, we organised  
498 several subcorpora according to textual windows (e.g. a corpus  $f_0, f_{-2}$ ). The  
499 number of sentences changes according to the chosen subcorpus from 5 000 to  
500 more than 35 000 sentences. During the experiments, we can set the number  
501 of instances that will constitute our training data. The results are based on  
502 a training set of 2 000 instances randomly chosen and size balanced between  
503 positive (i.e. containing units) and negative instances. The *bag-of-words* used  
504 to construct the model changes from 3 000 to 4 800 features depending on  
505 the chosen subcorpus. We used a list of 211 unit terms referenced in the food  
506 packaging domain OTR.

507 **Learning results.** Table 1 pools the results according to the textual win-  
508 dows tested. This first table helps us to determine which textual window is the  
509 most relevant context to locate units in text. We are particularly interested in  
510 recall, since our aim is to obtain the most relevant instances that are retrieved

511 considering the "unit" class, but without losing too much precision in the re-  
512 sults, which is described by the F-measure. First, we can say that Naive Bayes  
513 returns F-measure rates ranging from 0.85 to 0.88. Decision tree (i.e. J48)  
514 returns better rates from 0.93 to 0.96. DMNB and SMO<sup>5</sup> return better values  
515 (0.95 to 0.99). Second, we can note that a larger context (i.e. composed of two  
516 sentences –  $f_{+2}$  and  $f_{-2}$ ) does not improve the results. We can conclude that  
517 considering the smallest context based on one sentence (i.e.  $f_0$ ) is enough for  
518 unit location. This allows us to significantly reduce the search space while being  
519 in an optimal discovery context.

520 Table 2 pools the results on the  $f_0$  textual window, previously underlined, ac-  
521 cording to the three weight-based measures and the Boolean matrix. This second  
522 table shows us algorithm behaviors according to several weight-based measures.  
523 Note that, with all weight-based measures included, Naive Bayes returns rates  
524 that decrease from 0.88 (Boolean matrix) to 0.76 (other weightings). SMO  
525 loses around 17%, with a rate decreasing from 0.99 (Boolean) to 0.82 (okapi).  
526 DMNB (F-measure at 0.95) and Decision Tree J48 (F-measure at 0.92-0.93) stay  
527 constant regardless of weight-based measures.

528 **Identification step.** At the end of the learning step, we get a set of sen-  
529 tences that potentially contain units having typographic variations. The ex-  
530 tracted units are first pre-selected to be compared to relevant units referenced  
531 in the OTR according to the Jaccard measure. The candidate units are then  
532 compared according to the new  $SM_{D_b}$  measure. The first experiments were  
533 conducted on 11 articles in which 25 manually annotated unit terms had to be  
534 extracted and identified. Those first results obtained on a sample allowed us to  
535 assess the precision and recall of the proposed method since we did not have  
536 a complete annotated corpus. Then we applied our method with the  $SM_{D_b}$   
537 measure on the complete corpus. The results are given in Table 3 for each  
538 identification step (i.e. Jaccard and  $SM_{D_b}$  measures) and according to several  
539 similarity thresholds. We focused specifically on precision in order to facilitate

---

<sup>5</sup>with a polynomial kernel

|          | Dec. Tree J48 |          |          | Naive Bayes |          |          | DMNB     |          |             | SMO      |          |             |
|----------|---------------|----------|----------|-------------|----------|----------|----------|----------|-------------|----------|----------|-------------|
|          | <i>P</i>      | <i>R</i> | <i>F</i> | <i>P</i>    | <i>R</i> | <i>F</i> | <i>P</i> | <i>R</i> | <i>F</i>    | <i>P</i> | <i>R</i> | <i>F</i>    |
| $f_0$    | 0.99          | 0.87     | 0.93     | 0.83        | 0.93     | 0.88     | 0.95     | 0.96     | <b>0.95</b> | 0.99     | 0.99     | <b>0.99</b> |
| $f_{+2}$ | 0.99          | 0.92     | 0.96     | 0.95        | 0.77     | 0.85     | 0.93     | 0.96     | <b>0.95</b> | 0.99     | 0.97     | <b>0.99</b> |
| $f_{-2}$ | 0.99          | 0.92     | 0.95     | 0.77        | 0.98     | 0.86     | 0.94     | 0.96     | <b>0.95</b> | 0.99     | 0.97     | <b>0.98</b> |

Table 1: Results of "Unit" instances: Precision (P), Recall (R), F-measure (F) are given for each textual window. Best results are in bold considering F.

|         | Dec. Tree J48 |          |          | Naive Bayes |          |          | DMNB        |             |             | SMO      |          |          |
|---------|---------------|----------|----------|-------------|----------|----------|-------------|-------------|-------------|----------|----------|----------|
|         | <i>P</i>      | <i>R</i> | <i>F</i> | <i>P</i>    | <i>R</i> | <i>F</i> | <i>P</i>    | <i>R</i>    | <i>F</i>    | <i>P</i> | <i>R</i> | <i>F</i> |
| Boolean | 0.99          | 0.87     | 0.93     | 0.83        | 0.93     | 0.88     | <b>0.95</b> | <b>0.96</b> | <b>0.95</b> | 0.99     | 0.99     | 0.99     |
| TF      | 0.99          | 0.86     | 0.92     | 0.69        | 0.85     | 0.76     | <b>0.95</b> | <b>0.96</b> | <b>0.95</b> | 0.84     | 0.90     | 0.87     |
| TF.IDF  | 0.99          | 0.86     | 0.92     | 0.69        | 0.85     | 0.76     | <b>0.95</b> | <b>0.96</b> | <b>0.95</b> | 0.84     | 0.90     | 0.87     |
| Okapi   | 0.99          | 0.86     | 0.92     | 0.69        | 0.86     | 0.76     | <b>0.95</b> | <b>0.96</b> | <b>0.95</b> | 0.77     | 0.88     | 0.82     |

Table 2: Results of "Unit" instances: Precision (P), Recall (R), F-measure (F) are given for each weight-based measure and Boolean matrix. The best results are in bold.

540 the expert validation step without too noisy results. The results showed that the  
 541 complete process, including the Jaccard and  $SM_{D_b}$  measure, was more accurate  
 542 and relevant. First applying the Jaccard measure to get pre-selected candidate  
 543 units substantially decreased the extent of noisy results in the second validation  
 544 step, with  $SM_{D_b}$  (F-measure  $>0.7$  for thresholds under 0.6). Then the process  
 545 was applied on the complete food packaging corpus. 121 new unit terms were  
 546 identified and enriched the food packaging OTR (originally composed of 211  
 547 terms).

| Similarity threshold | <b>Jaccard pre-selection</b> |               |                  | $SM_{D_b}$ selection |               |                  |
|----------------------|------------------------------|---------------|------------------|----------------------|---------------|------------------|
|                      | <i>Precision</i>             | <i>Recall</i> | <i>F-measure</i> | <i>Precision</i>     | <i>Recall</i> | <i>F-measure</i> |
| [0.9-1]              | 0.7                          | 0.4           | 0.5              | 0.8                  | 0.4           | 0.5              |
| [0.8-1]              | 0.8                          | 0.5           | 0.6              | 0.8                  | 0.6           | 0.7              |
| [0.7-1]              | 0.8                          | 0.7           | 0.7              | 0.8                  | 0.6           | 0.7              |
| [0.6-1]              | 0.7                          | 0.7           | 0.7              | <b>0.7</b>           | <b>0.8</b>    | <b>0.7</b>       |
| [0.5-1]              | 0.7                          | 0.8           | 0.7              | <b>0.7</b>           | <b>0.8</b>    | <b>0.7</b>       |
| [0.4-1]              | 0.5                          | 0.8           | 0.6              | <b>0.6</b>           | <b>1</b>      | <b>0.8</b>       |

Table 3: Identification step: Jaccard pre-selection and  $SM_{D_b}$  selection.

548 At the end of the first step, the Xart system identified several new unit terms  
 549 to enrich the domain OTR. Those important features are then used to define  
 550 several relevant textual contexts. Those relevant textual windows are mined  
 551 during the knowledge discovery process in the second step of the Xart system,  
 552 as described in the following section.

### 553 6.2. *Ontological sequential patterns*

554 **subcorpus constitution.** From the food packaging corpus, we organised  
 555 several subcorpora according to the textual windows represented (e.g. a cor-

556 pus  $f_0, f_{\pm 2}$ ). We applied our knowledge discovery process and obtained several  
557 matrices for each subcorpus tested. The number of transactions tested changed  
558 according to the textual window represented, i.e. from 5 000 to 35 000. The  
559 number of items also changed according to the textual window represented, i.e.  
560 from 2 000 to more than 10 000.

561 **Algorithms used in the experiments.** A substantial number of data min-  
562 ing algorithms currently exist, such as Apriori (Agrawal & Srikant, 1994), Spade  
563 (Zaki, 2001), and PrefixSpan (Pei et al., 2001). The experiments were conducted  
564 using Clospan (Yan et al., 2003) to extract sequential patterns. Clospan imple-  
565 ments one of the most efficient algorithms to date, PrefixSpan, and allows the  
566 discovery of a set of sequential patterns without redundancy and without loss  
567 of informativeness.

568 **Selection criteria.** A well-known data mining issue concerns managing the  
569 number of sequential patterns generated from algorithms. Thus, the support is  
570 an important measure used to eliminate uninteresting sequential patterns and  
571 can be exploited for the efficient discovery of sequential patterns.

572 Beyond those classical support measure, we propose to use two new selection  
573 criteria based on both statistical and semantic criteria. The first one will se-  
574 lect only the OSP where at least one argument of n-ary relations represented  
575 in the domain OTR is identified. The second one will select the OSP from the  
576 intersection of several studied textual windows.

577 **Quantitative results.** The number of ontological sequential patterns varies  
578 according to the selection criteria applied. For example, we obtained more than  
579 52 000 patterns in the subcorpus  $f_{\pm 2}$  according to a minimum support of 0.5 and  
580 the criteria for selecting patterns containing at least one argument referenced  
581 in the OTR. When we added the intersection selection criteria, we reduced this  
582 number to around 1 000 OSP.

583 **Qualitative results.** We applied the knowledge discovery process without  
584 increasing the expressiveness of arguments with data representation in text. We  
585 obtained a small set of patterns as compared to other results, i.e. around 500,  
586 and the extracted patterns were meaningless, e.g. none of the patterns retrieved

| Textual window | Ontological sequential pattern        | Support |
|----------------|---------------------------------------|---------|
| $f_{\pm 1}$    | <(Packaging)(numvalthick um)>         | 0.5     |
|                | <(numvalthick)(films)>                | 0.5     |
|                | <(film)(mm)(thickness)>               | 0.1     |
|                | <(film thickness)(rh)>                | 0.1     |
|                | <(Packaging)(Quantity)(permeability)> | 0.5     |
|                | <(Packaging)(permeability)>           | 0.6     |
| $f_0$          | <(pressure)(water permeability)>      | 0.05    |
|                | <(oxygen permeability)(pressure)>     | 0.05    |
| $\cap f_n$     | <(numvaltemp)(numvalrh%)>             |         |
|                | <(Packaging)(numvalthick)>            |         |
|                | <(Packaging)(numvaltemp °C)>          |         |

Table 4: Excerpt of OSP -  $\cap$  window intersection criteria

587 numerical values, whereas they are important for discovering new instances in  
588 text.

589 Table 4 gives an excerpt of OSP obtained with the knowledge discovery process  
590 using our data representation. First, the results show the advantages of the  
591 new data representation to extract more meaningful patterns. Second, they  
592 show that extracted patterns allow us to discover implicit argument expressions  
593 in text. We came up with the three following patterns.

- 594 1. OSP <(Packaging)(numvalthick um)> highlights that *packaging* and *thick-*  
595 *ness* arguments frequently appear to be correlated in text and that corre-  
596 lations frequently occur in a maximal textual window of  $f_{\pm 1}$ ;
- 597 2. <(pressure)(water permeability)> shows that the *partial pressure* and *per-*  
598 *meability* arguments frequently occur in the same sentence;
- 599 3. Several OSP suggest that the terms denoting the *packaging* concept could  
600 be the trigger of the relation since they frequently occur in OSP of previ-

601       ous correlations, e.g.  $\langle (Packaging)(permeability) \rangle$ ,  $\langle (Packaging) (num-$   
602        $valtemp \text{ } ^\circ C) \rangle$ .

603       Interested readers will find in (Berrahou et al., 2016) additional experimen-  
604       tal results associated with the third step of the Xart system, which is the hybrid  
605       approach combining OSP with syntactic analysis in order to construct Ontolog-  
606       ical Linguistic Sequential Patterns (OLSP) for identifying correlated arguments  
607       directly in text.

## 608 **7. Conclusion**

609       We presented the Xart system based on a hybrid approach driven by an  
610       OTR that takes advantage of data mining techniques and syntactic analysis for  
611       complex data extraction from plain text. Thanks to the generic structure of the  
612       OTR, the Xart system may be used for different domains by only redefining the  
613       domain part of the OTR.

614       The first step of the Xart system proposes to enrich an Ontological and Ter-  
615       minological Resource (OTR) with new unit terms that are specific attributes  
616       of the sought n-ary relations. The proposed method enabled the identification  
617       of more than 57% of new units and units with typographic variations. In the  
618       second step, we propose a knowledge discovery process that takes the data ex-  
619       pressiveness into consideration using the conceptual level given by the OTR,  
620       and defined the new notion of Ontological Sequential Patterns (OSP).

621  
622       The different steps of the Xart system were tested on a specific domain  
623       (i.e. packaging). Note that our approach was also tested on another domain  
624       (i.e. biorefinery) in order to assess the relevance and genericity of the proposed  
625       methods. For instance, the first step of the Xart system allowed us to identify 38  
626       new units for enriching the biorefinery OTR (originally composed of 36 terms).

627  
628       To sum up, the Xart approach applies a complete process *Data Informa-*  
629       *tion Knowledge*. This "information chain" is a key feature of the Xart system.

630 In order to implement this "information chain", our system has to integrate a  
631 lot of different techniques (e.g. NLP tools, data-mining approaches, statistic  
632 weightings, etc.), and semantic resources. Although each tool of our system is  
633 efficient, the combination of the approaches, the pretreatment of textual data,  
634 and the analysis of the obtained results can be time consuming.

635

636 We used machine learning methods associated with a *bag-of-words* repre-  
637 sentation of documents to locate units in text. As future work, we plan to  
638 implement a feature selection approach in order to select relevant features for  
639 the *bag-of-words* representation. This method is close to the wrapper approach,  
640 as explained in (Kohavi, 1998). More precisely, we plan to select two types of fea-  
641 tures: experimental verbs and relevant domain terms (i.e. words and multi-word  
642 terms) extracted using the weight-based measures presented in (Lossio-Ventura  
643 et al., 2016).

644

645 There are two further prospects. The first one is a potential application.  
646 The OLSP of the Xart system will be integrated in a tool, @web<sup>6</sup> software,  
647 that allows researchers to manually annotate data tables extracted from doc-  
648 uments (Buche et al., 2013a). Indeed, during the annotation process of n-ary  
649 relation instances in tables, it often turns out that several argument instances  
650 (e.g. thickness) are missing in the table and are expressed in the text. Specific  
651 OLSP (e.g. packaging and thickness correlated arguments) can help to retrieve  
652 the sentences in which the relevant information appears and help researchers to  
653 complete the annotation of data given in the tables.

654 The second prospect is methodological. In future work, we intend to propose  
655 a formal definition of n-ary relation instantiation in an ontological sequential  
656 pattern context. Here we have shown that OSP enables detection of correlated  
657 arguments and the trigger word of the n-ary relation. This trigger word helps  
658 to gather all correlated arguments, expressed in several sentences, in the same

---

<sup>6</sup><http://www6.inra.fr/cati-icat-atweb/Web-platform>

659 n-ary relation. Another methodological prospect is then to propose a formal  
660 definition of the extraction of the complete n-ary relation.

661

662 **Acknowledgements:** This work was partially funded by the **Labex NUMEV**  
663 (ANR-10-LABX-20), **INRA, 3BCAR IC2ACV** project, and **Valorcarn project**  
664 (**Glofood**). We also thank Valérie Guillard, a food packaging expert, who helped us  
665 in the experimental validation of the method.

## 666 References

667 Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules  
668 in large databases. In *Proceedings of the 20th International Conference on*  
669 *Very Large Data Bases VLDB '94* (pp. 487–499). San Francisco, CA, USA:  
670 Morgan Kaufmann Publishers Inc.

671 Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings*  
672 *of the Eleventh International Conference on Data Engineering ICDE '95* (pp.  
673 3–14). Washington, DC, USA: IEEE Computer Society.

674 B chet, N., Cellier, P., Charnois, T., & Cr milleux, B. (2012). Discovering  
675 linguistic patterns using sequence mining. In *Proceedings of the 13th Interna-*  
676 *tional Conference on Computational Linguistics and Intelligent Text Process-*  
677 *ing - Volume Part I CICLing'12* (pp. 154–165). Berlin, Heidelberg: Springer-  
678 Verlag.

679 Berrahou, S. L., Buche, P., Dibie, J., & Roche, M. (2016). Xart system: dis-  
680 covering and extracting correlated arguments of n-ary relations from text. In  
681 *Proceedings of the 6th International Conference on Web Intelligence, Mining*  
682 *and Semantics, WIMS 2016, N mes, France, June 13-15, 2016* (pp. 8:1–8:12).

683 Bj rne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., & Salakoski, T.  
684 (2009). Extracting complex biological events with rich graph-based feature  
685 sets. In *Proceedings of the Workshop on Current Trends in Biomedical Natural*

- 686 *Language Processing: Shared Task BioNLP '09* (pp. 10–18). Stroudsburg, PA,  
687 USA: Association for Computational Linguistics.
- 688 Buche, P., Dervaux, S., Dibia-Barthélemy, J., Soler, L., Ibanescu, L., &  
689 Touhami, R. (2013a). Intégration de données hétérogènes et imprécises guidée  
690 par une ressource termino-ontologique. *Revue d'Intelligence Artificielle*, 27,  
691 539–568.
- 692 Buche, P., Dibia-Barthélemy, J., Ibanescu, L., & Soler, L. (2013b). Fuzzy Web  
693 Data Tables Integration Guided by an Ontological and Terminological Re-  
694 source. *IEEE Trans. Knowl. Data Eng.*, 25, 805–819.
- 695 Bui, Q.-C., & Sloot, P. M. A. (2011). Extracting biological events from text  
696 using simple syntactic patterns. In *Proceedings of the BioNLP Shared Task*  
697 *2011 Workshop BioNLP Shared Task '11* (pp. 143–146). Stroudsburg, PA,  
698 USA: Association for Computational Linguistics.
- 699 Buyko, E., Faessler, E., Wermter, J., & Hahn, U. (2009). Event extraction from  
700 trimmed dependency graphs. In *Proceedings of the Workshop on Current*  
701 *Trends in Biomedical Natural Language Processing: Shared Task BioNLP '09*  
702 (pp. 19–27). Stroudsburg, PA, USA: Association for Computational Linguis-  
703 tics.
- 704 Cellier, P., Charnois, T., Plantevit, M., Rigotti, C., Crémilleux, B., Gandrillon,  
705 O., Kléma, J., & Manguin, J. (2015). Sequential pattern mining for discover-  
706 ing gene interactions and their contextual information from biomedical texts.  
707 *J. Biomedical Semantics*, 6, 27.
- 708 Damerau, F. J. (1964). A technique for computer detection and correction of  
709 spelling errors. *Commun. ACM*, 7, 171–176.
- 710 Di-Jorio, L., Bringay, S., Fiot, C., Laurent, A., & Teisseire, M. (2008). Se-  
711 quential patterns for maintaining ontologies over time. In *On the Move to*  
712 *Meaningful Internet Systems: OTM 2008, OTM 2008 Confederated Interna-*

- 713 *tional Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monter-*  
714 *rey, Mexico, November 9-14, 2008, Proceedings, Part II* (pp. 1385–1403).
- 715 Gkoutos, G. (2011). Units ontology. URL: [http://www.obofoundry.org/](http://www.obofoundry.org/ontology/uo.html)  
716 [ontology/uo.html](http://www.obofoundry.org/ontology/uo.html).
- 717 Gruber, T. R., & Olsen, G. (1994). An ontology for engineering mathemat-  
718 ics. In J. Doyle, P. Torasso, & E. Sandewall (Eds.), *Principles of Knowledge*  
719 *Representation and Reasoning: Proceedings of the 4th International Confer-*  
720 *ence (KR '94): Bonn, Germany: 1994, May, 24 - 27* The Morgan Kaufmann  
721 Series in Representation and Reasoning (pp. 258–269). Morgan Kaufmann  
722 Publishers.
- 723 Guillard, V., Buche, P., Destercke, S., Tamani, N., Croitoru, M., Menut, L.,  
724 Guillaume, C., & Gontard, N. (2015). A Decision Support System to design  
725 modified atmosphere packaging for fresh produce based on a bipolar flexible  
726 querying approach. *CEA*, (pp. 131–139).
- 727 Hao, Y., Zhu, X., Huang, M., & Li, M. (2005). Discovering patterns to extract  
728 proteinprotein interactions from the literature: part ii. *Bioinformatics*, *21*,  
729 32943300.
- 730 Hawizy, L., Jessop, D., Adams, N., & Murray-Rust, P. (2011). ChemicalTagger:  
731 a tool for semantic text-mining in chemistry. *Journal of cheminformatics*, *3*,  
732 17.
- 733 Hiemstra, D. (2000). A probabilistic justification for using tf x idf term weighting  
734 in information retrieval. *Int. J. on Digital Libraries*, *3*, 131–139.
- 735 Hodgson, R., Paul, J., Jack, H., & Jack, S. (2013). Qudt-quantities, units,  
736 dimensions and data types ontologies. URL: <http://www.qudt.org>.
- 737 Huang, M., Zhu, X., Payan, D. G., Qu, K., & Li, M. (2004). Discovering patterns  
738 to extract protein-protein interactions from full biomedical texts. In *Proceed-*  
739 *ings of the International Joint Workshop on Natural Language Processing in*

- 740 *Biomedicine and Its Applications JNLPBA '04* (pp. 22–28). Stroudsburg, PA,  
741 USA: Association for Computational Linguistics.
- 742 Jaillet, S., Laurent, A., & Teisseire, M. (2006). Sequential patterns for text  
743 categorization. *Intell. Data Anal.*, *10*, 199–214.
- 744 Jessop, D. M., Adams, S. E., & Murray-Rust, P. (2011a). Mining chemical  
745 information from open patents. *Journal of cheminformatics*, *3*, 40.
- 746 Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., & Murray-Rust, P.  
747 (2011b). OSCAR4: a flexible architecture for chemical text-mining. *Journal*  
748 *of cheminformatics*, *3*, 1–12.
- 749 John, G. H., & Langley, P. (1995). Estimating continuous distributions in  
750 bayesian classifiers. In *Proc. of the conf. on Uncertainty in artificial intel-*  
751 *ligence* (pp. 338–345).
- 752 Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model  
753 of information retrieval: development and comparative experiments - part 1.  
754 *Inf. Process. Manage.*, *36*, 779–808.
- 755 Kohavi, G. H., Ronand John (1998). The wrapper approach. In H. Liu, Hua-  
756 nand Motoda (Ed.), *Feature Extraction, Construction and Selection: A Data*  
757 *Mining Perspective* (pp. 33–50). Boston, MA: Springer US.
- 758 Kohavi, R., & Quinlan, J. R. (2002). Data mining tasks and methods: Classi-  
759 fication: decision-tree discovery. In *Handbook of data mining and knowledge*  
760 *discovery* (pp. 267–276). Oxford University Press, Inc.
- 761 Le Minh, Q., Truong, S. N., & Bao, Q. H. (2011). A pattern approach for  
762 biomedical event annotation. In *Proceedings of the BioNLP Shared Task 2011*  
763 *Workshop BioNLP Shared Task '11* (pp. 149–150). Stroudsburg, PA, USA:  
764 Association for Computational Linguistics.
- 765 Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016). Biomed-  
766 ical term extraction: overview and a new methodology. *Information Retrieval*  
767 *Journal*, *19*, 59–99.

- 768 Maedche, A., & Staab, S. (2002). Measuring similarity between ontologies.  
769 In *Knowledge Engineering and Knowledge Management: Ontologies and the*  
770 *Semantic Web* (pp. 251–263). Springer volume 2473 of *LNCS*.
- 771 Minard, A.-L., Ligozat, A.-L., & Grau, B. (2011). Multi-class svm for relation  
772 extraction from clinical reports. In G. Angelova, K. Bontcheva, R. Mitkov,  
773 & N. Nicolov (Eds.), *RANLP* (pp. 604–609). RANLP 2011 Organising Com-  
774 mittee.
- 775 Miwa, M., Sætre, R., Miyao, Y., & Tsujii, J. (2009). A rich feature vector for  
776 protein-protein interaction extraction from multiple corpora. In *Proceedings of*  
777 *the 2009 Conference on Empirical Methods in Natural Language Processing:*  
778 *Volume 1 - Volume 1 EMNLP '09* (pp. 121–130). Stroudsburg, PA, USA:  
779 Association for Computational Linguistics.
- 780 Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M.  
781 (2001). Prefixspan: Mining sequential patterns by prefix-projected growth.  
782 In *Proceedings of the 17th International Conference on Data Engineering* (pp.  
783 215–224). Washington, DC, USA: IEEE Computer Society.
- 784 Platt, J. C. (1999). Fast training of support vector machines using sequential  
785 minimal optimization. In *Advances in kernel methods* (pp. 185–208). MIT  
786 Press.
- 787 Proux, D., Rechenmann, F., & Julliard, L. (2000). A pragmatic information  
788 extraction strategy for gathering data on genetic interactions. In P. E. Bourne,  
789 M. Gribskov, R. B. Altman, N. Jensen, D. A. Hope, T. Lengauer, J. C.  
790 Mitchell, E. D. Scheeff, C. Smith, S. Strande, & H. Weissig (Eds.), *ISMB*  
791 (pp. 279–285). AAAI.
- 792 Qiu, C. Z. T. Q. S. Z. J. C. P. Z. J., Jiangtaoand Tang (2007). A novel text  
793 classification approach based on enhanced association rule. In H. L. J. L. X.  
794 Z. O. R. Alhajj, Redaand Gao (Ed.), *Advanced Data Mining and Applica-*  
795 *tions: Third International Conference, ADMA 2007 Harbin, China, August*

- 796 6-8, 2007. *Proceedings* (pp. 252–263). Berlin, Heidelberg: Springer Berlin  
797 Heidelberg.
- 798 Raja, K., Subramani, S., & Natarajan, J. (2013). Ppinterfinder - a mining tool  
799 for extracting causal relations on human proteins from literature. *Database*,  
800 2013.
- 801 Rijgersberg, H., van Assem, M., & Top, J. L. (2013). Ontology of units of  
802 measure and related concepts. *Semantic Web*, 4, 3–13.
- 803 Rosario, B., & Hearst, M. A. (2005). Multi-way relation classification: Ap-  
804 plication to protein-protein interactions. In *Proceedings of the Conference*  
805 *on Human Language Technology and Empirical Methods in Natural Language*  
806 *Processing HLT '05* (pp. 732–739). Stroudsburg, PA, USA: Association for  
807 Computational Linguistics.
- 808 Schadow, G., McDonald, C. J., Suico, J. G., Föhring, U., & Tolxdorff, T. (1999).  
809 Model formulation: Units of measure in clinical information systems. *JAMIA*,  
810 6, 151–162.
- 811 Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008). Discriminative parameter  
812 learning for bayesian networks. In *Proc. of the int. conf. on Machine learning*  
813 (pp. 1016–1023).
- 814 Touhami, R., Buche, P., Dibia-Barthélemy, J., & Ibanescu, L. (2011). An On-  
815 tological and Terminological Resource for n-ary Relation Annotation in Web  
816 Data Tables. In *ODBASE Conferences (2)* (pp. 662–679).
- 817 Van Landeghem, S., Saeys, Y., De Baets, B., & Van de Peer, Y. (2009). Analyz-  
818 ing text in search of bio-molecular events: A high-precision machine learning  
819 framework. In *Proceedings of the Workshop on Current Trends in Biomed-*  
820 *ical Natural Language Processing: Shared Task BioNLP '09* (pp. 128–136).  
821 Stroudsburg, PA, USA: Association for Computational Linguistics.

- 822 Yan, X., Han, J., & Afshar, R. (2003). Clospan: Mining closed sequential  
823 patterns in large databases. In D. Barbará, & C. Kamath (Eds.), *SDM*.  
824 SIAM.
- 825 Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences.  
826 *Mach. Learn.*, 42, 31–60.
- 827 Zhang, H., Huang, M., & Zhu, X. (2011). Protein-protein interaction extraction  
828 from bio-literature with compact features and data sampling strategy. In *4th*  
829 *International Conference on Biomedical Engineering and Informatics, BMEI*  
830 *2011, Shanghai, China, October 15-17, 2011* (pp. 1767–1771).
- 831 Zhou, D., Zhong, D., & He, Y. (2014). Event trigger identification for biomedical  
832 events extraction using domain knowledge. *Bioinformatics*, 30, 1587–1594.