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

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

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

Soumia Lilia Berrahou[a,b], Patrice Buche*[a,b], Juliette Dibie[c], Mathieu Roche[a,d]

[a] *LIRMM - 860, rue de Saint Priest, 34095 Montpellier, FRANCE*
[b] *INRA - UMR IATE - 2, place Pierre Viala, 34060 Montpellier, FRANCE*
[c] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, FRANCE*
[d] *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 (Soumia Lilia Berrahou), Patrice.Buche@inra.fr (Patrice Buche*), dibie@agroparistech.fr (Juliette Dibie), 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 μm thickness polypropylene film with a permeability to oxygen of 110 $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ at 23 $°C$ 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

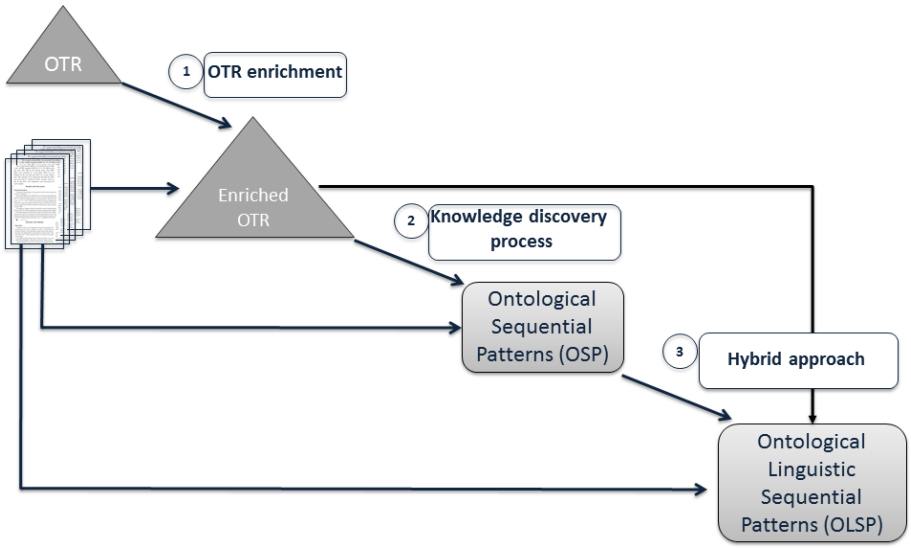

Figure 1: Xart system

regard to their attributes, i.e. the numerical value and measurement unit, between studied objects.

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

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

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

2. Related work

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

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

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

in identifying entities (or arguments) using resources such as ontologies or dictionaries; the second step involves identifying the trigger word of the relation using dictionary-based methods or rule-based approaches to construct patterns from dependency parse results (Le Minh et al., 2011), or using machine learning methods (Buyko et al., 2009; Bui & Sloot, 2011; Björne et al., 2009; Zhou et al., 2014) for predicting which word of the sentence is the trigger word of the relation; and the third step involves constructing a set of binary relations using the trigger word, with a given argument being classified as part or not part of the n-ary relation using machine learning methods. Unfortunately, decomposing the problem of n-ary relation extraction in extracting several binary relations results in lower performance. Our approach relies on the knowledge discovery process using domain knowledge for representing relevant data and for discovering sequential patterns, including several correlated arguments and the trigger word of the relation. The trigger word discovered in the patterns allows all other arguments to be gathered in the sought n-ary relation.

Data mining approaches are used in (Di-Jorio et al., 2008) for enriching ontologies with new concepts, in (Béchet et al., 2012; Cellier et al., 2015) for discovering linguistic patterns without external resources, and in (Qiu, 2007) for adding more semantics and drawing up enhanced association rules. Moreover, in (Jaillet et al., 2006), the authors use association rules and sequential patterns to propose comprehensive and reusable text categorization rules. Those techniques have already been successfully used for processing textual data. In line with these authors, we propose to take advantage of data mining approaches to discover sequential patterns of several correlated arguments in text.

Available annotated Corpora. As cited in (Zhou et al., 2014), several corpora have been designed for binary relation extraction, e.g. GENIA[1], LLL05[2], AIMed[3]. Those corpora essentially contain sentences with interactions

[1] http://www.nactem.ac.uk/genia/genia-corpus

[2] http://genome.jouy.inra.fr/texte/LLLchallenge/

[3] ftp://ftp.cs.utexas.edu/pub/mooney/bio-data

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

3. Xart system key elements

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

3.1. An ontology for n-ary relation representation

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

[4]https://catalog.ldc.upenn.edu/LDC2014T27

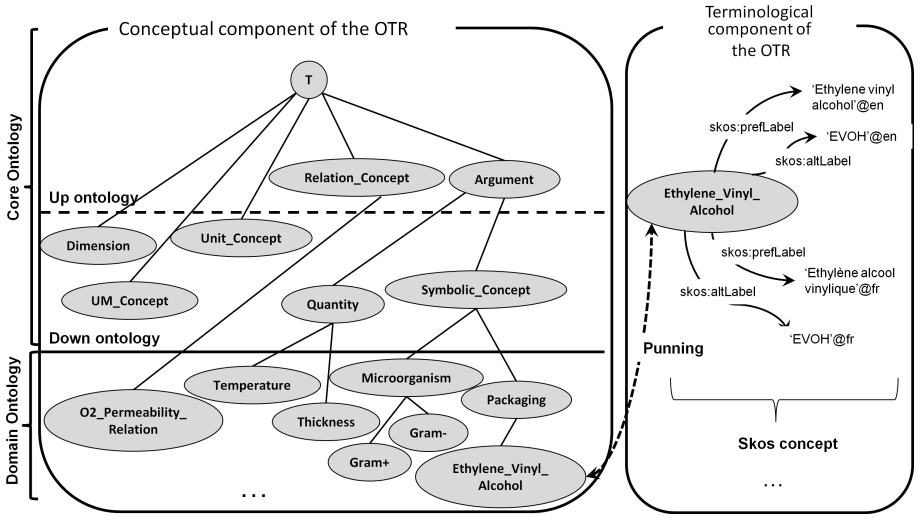

Figure 2: An excerpt of the naRyQ concept hierarchy in the food packaging domain

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

Figure 2 illustrates an application of naRyQ in the food packaging domain. In the up core ontology, generic concepts Relation_Concept and Argument respectively represent n-ary relations and their arguments. In the down core ontology, generic concepts Dimension, UM_Concept, Unit_Concept and Quantity allow the management of quantities and their associated measurement units. Note that the measurement units are represented by instances of the generic concept Unit_Concept. The subconcepts of the generic concept Symbolic_Concept represent the non-numerical arguments of n-ary relations. The domain ontology contains specific concepts of a given application domain. They appear in naRyQ as subconcepts of the generic concepts of the core ontology. The terminological component of naRyQ contains the set of terms describing the studied domain. naRyQ presented in (Touhami et al., 2011; Buche et al., 2013b) may be formally define as follows.

Definition 1.

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

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

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

Definition 2.

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

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

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

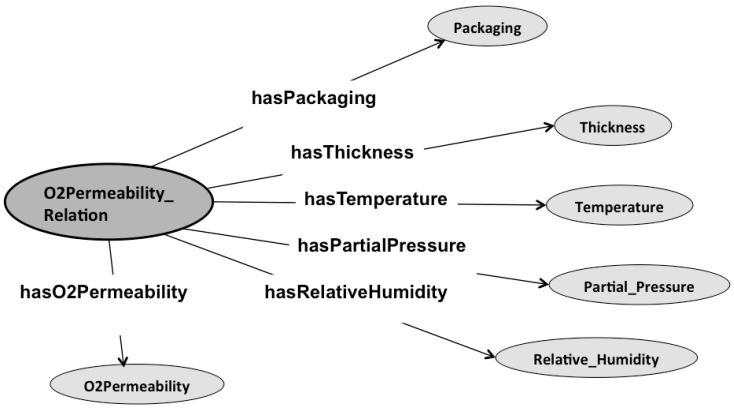

Figure 3: The n-ary relation O2Permeability_Relation

Definition 3.

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

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

An example of a n-ary relation concept is given in Figure 3 and represents the relation O2Permeability_relation in the naRyQ_pack OTR (food packaging domain OTR). The signature of the n-ary relation O2Permeability_Relation is: signatureR(O2Permeability_Relation)= {Packaging, Thickness, Temperature, Partial_Pressure, Relative_Humidity, O2Permeability}.

3.2. Xart textual context

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

arguments. From this hypothesis, we propose two relevant textual search contexts: pivot sentence and textual window defined as follows:

Definition 4. *(Pivot sentence)*
A pivot sentence is defined as the sentence where at least one unit referenced in the OTR is identified

Definition 5. *(Textual window)*
A textual window denoted f_{sn} is defined as the set of sentences composed of the pivot sentence and the n previous sentences, and/or the n subsequent sentences, where n is the window dimension. The search direction in sentences, denoted s, is represented with - considering previous sentences, + considering subsequent sentences or $\pm$ considering previous and subsequent sentences

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

4. The Xart first step: Enrichment of the OTR with measurement units

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

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

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

4.1. Locating units

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

Data preparation. Data preparation involves text processing and text transformation tasks. Text processing consists of:

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

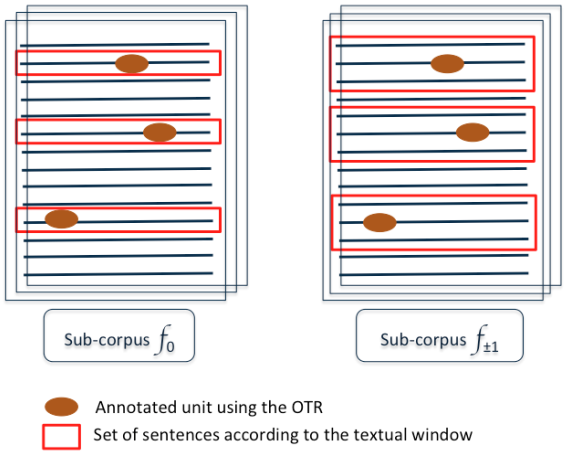

Figure 4: Subcorpus preparation

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

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

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

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

- Term Frequency-Inverse Document Frequency (TF.IDF (Hiemstra, 2000)), which considers that the word is more important if it appears in fewer sentences;

- Okapi BM25 (Jones et al., 2000), which also takes into consideration the length of the sentence in which the word appears to define its relevance.

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

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

- Naive Bayes classifier and the Discriminative Multinominal Naive Bayes (DMNB) classifier;

- J48 decision tree classifier ;

- Sequential Minimal Optimization (SMO), which is a Support Vector Machine classifier (SVM).

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

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

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

4.2. Identifying units

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

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

composed of four blocks, *amol*, *m*, *s*, and *Pa* whereas *amol/m.sec.Pa* is composed 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):

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

$$\in [0; 1]$$

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

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

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

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

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

$$SM_{D_c} = 0.12$$

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

(1) They are first pre-selected with the Jaccard measure as relevant for comparison with our SM_{D_b} *measure.*

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

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

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

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

5. The second Xart step: knowledge discovery process

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

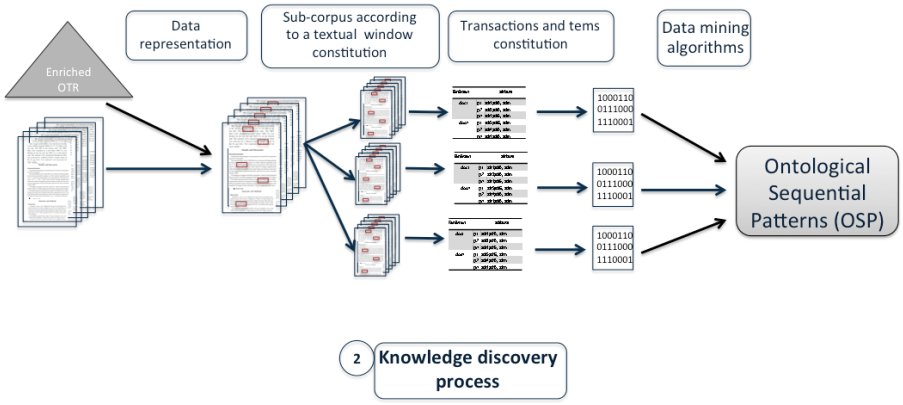

Figure 5: Second step of the Xart system: Knowledge discovery process driven by a domain OTR

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

5.1. First substep: a new data representation

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

Definition 6, to increase the expressiveness of the symbolic arguments by representing them with their corresponding concepts, subconcepts of the generic concept *Symbolic_Concept*, which belong to the signature of the sought n-ary relation. For example, in our experiments on the packaging domain corpus, we choose the subconcept *Packaging* which belongs to the signature of the relation O2Permeability_Relation in order to represent the studied packaging in text (e.g. gluten).

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

(i) $c_j \in SignatureR(rel)$ *and* $rel \in C_{Rel}$*,*

(ii) $\exists\ c_i \in C_{Symb}$*,* $c_i \leq_o c_j$*,*

(iii) $\exists\ w_i \in W_{oi}$ *which denotes* c_i *such that sim(*w_i*, t)=1, sim being a similarity measure.*

We propose, with Definition 7, to increase the expressiveness of the quantitative arguments of the sought n-ary relation by increasing the expressiveness of their numerical values using their associated measurement units. Numerical values indeed represent the relevant information we want to discover and which often vary, depending on the measurements obtained on the studied object. Measurement units related to numerical values are associated in the OTR with specific subconcepts of the generic concept *Quantity* by the relation $hasUnit(hasUnit \in R)$ (e.g. the measurement unit $^\circ C$ is associated in the naRyQ OTR with the quantity $Temperature$). We use those *Quantity* subconcepts, which belong to the signature of the sought n-ary relation, to represent numerical values.

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

consider v_i, a numerical value in the text associated, in the text, with t_i, a term, and t_u, a unit term. Then v_i is annotated by C_i^u, denoted by $< numvalC_i^u >$, t_i by <Quantity> and t_u by <um> in the new data representation if:

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

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

The two previous definitions are illustrated in Example 2. In sentence (1), the expressiveness of the underlined data is improved using Definitions 6 and 7. Sentence (2) corresponds to the new data representation of sentence (1). This sentence contains an instance of O2Permeability_Relation, as described in Figure 3, which represents the oxygen permeability of a packaging under given experimental conditions. The experimental conditions are defined by the packaging thickness (64 μm), temperature (23 $°C$) and relative humidity (0%). More precisely, note that the numerical value *64* is followed by the unit μm that is associated with the *Thickness* concept. Thus, <numvalthick> is used to annotate the value 64, which is the relevant instance to be identified in text, and we represent the term "thickness" by <Quantity> and the term "μm" by <um>.

Example 2.

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

(2) Eight apple wedges were packaged in polypropylene <Packaging> trays and wrap-sealed using a 64 <numvalthick> μm <um> thick <Quantity> polypropylene <Packaging> film with an oxygen permeability <Quantity> of 110 <numvalperm> $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ <um> at 23 <numvaltemp> $°C$ <um> and 0 <numvalrh> % <um> RH <Quantity>.

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

5.2. Second substep: subcorpus constitution

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

5.3. Third substep: transactions and items

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

Definition 8. *(Transaction)*
A transaction is defined as a set of sentences according to a textual window f_{sn}.

Example 3.
In a textual window $f_{\pm 1}$*, each transaction corresponds to a set of sentences composed of the pivot sentence, the previous and subsequent sentences.*

Definition 9. *(Itemset)*
An itemset IS^n *is the set of n nearest terms or annotations associated with a given argument of a sought relation rel in the data representation detailed in Definitions 6 and 7.*

Example 4.
Let us consider the sentence (2) of Example 2, if we choose to select the 1-term nearest neighbors of the annotation <Packaging>, we obtain an itemset composed of <Packaging>, polypropylene, trays, films. For the annotation <Quantity>, we obtain an itemset composed of <Quantity>, thickness, polypropylene, oxygen, <numvalperm>, RH.

5.4. Fourth substep: Data mining

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

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

Definition 10. *(OS - Ontological Sequence)*
Let us consider $OTR = \langle C_{OTR}; R; I; V; \leq_o; W_{oi} \rangle$ *of Definition 1. An ontological sequence* OS_{fsn} *is a non-empty ordered list of itemsets* IS_j^n *extracted in a textual window* f_{sn}*, denoted* $< IS_1^n IS_2^n ... IS_p^n >$.

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

Definition 11. *(OSP - Ontological Sequential Pattern)*
Let $(OS_{fsn})_A = < IS_1^n IS_2^n ... IS_p^n >$ *be an ontological subsequence of another ontological sequence* $(OS_{fsn})_B = < IS_1'^n IS_2'^n ... IS_m'^n >$*, then* $((OS_{fsn})_A \preceq OS_{fsn})_B)$ *if* $p \leq m$ *and* $IS_1^n \subseteq IS_{j_1}'^n, IS_2^n \subseteq IS_{j_2}'^n, ..., IS_p^n \subseteq IS_{j_p}'^n$ *whith* $1 < j_1 < j_2 < ... < j_k < ... < j_p < m$*. Let* θ *be a minimum support, then the Ontological Sequential Pattern OSP is defined as a set of frequent subsequences from* OS_{fsn} *such that* $Support(OSP) \geq \theta$.

Example 5.

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

The third step of the Xart system is the hybrid approach detailed in (Berrahou et al., 2016), which proposes to combine OSP with syntactic analysis in order to construct Ontological Linguistic Sequential Patterns (OLSP) for identifying correlated arguments directly in text.

6. Experiments and results

6.1. OTR enrichment

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

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

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

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

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

[5] with a polynomial kernel

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

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**		
	P	R	F	P	R	F	P	R	F	P	R	F
Boolean	0.99	0.87	0.93	0.83	0.93	0.88	**0.95**	**0.96**	**0.95**	0.99	0.99	0.99
TF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
TF.IDF	0.99	0.86	0.92	0.69	0.85	0.76	**0.95**	**0.96**	**0.95**	0.84	0.90	0.87
Okapi	0.99	0.86	0.92	0.69	0.86	0.76	**0.95**	**0.96**	**0.95**	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.

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

Similarity threshold	**Jaccard pre-selection**			SM_{Db} **selection**		
	Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
[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	**0.7**	**0.8**	**0.7**
[0.5-1]	0.7	0.8	0.7	**0.7**	**0.8**	**0.7**
[0.4-1]	0.5	0.8	0.6	**0.6**	**1**	**0.8**

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

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

6.2. Ontological sequential patterns

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

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

Algorithms used in the experiments. A substantial number of data mining algorithms currently exist, such as Apriori (Agrawal & Srikant, 1994), Spade (Zaki, 2001), and PrefixSpan (Pei et al., 2001). The experiments were conducted using Clospan (Yan et al., 2003) to extract sequential patterns. Clospan implements one of the most efficient algorithms to date, PrefixSpan, and allows the discovery of a set of sequential patterns without redundancy and without loss of informativeness.

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

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

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

Qualitative results. We applied the knowledge discovery process without increasing the expressiveness of arguments with data representation in text. We obtained a small set of patterns as compared to other results, i.e. around 500, 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
$\bigcap f_n$	<**(numvaltemp)(numvalrh%)**>	
	<(Packaging)(numvalthick)>	
	<**(Packaging)(numvaltemp** $^\circ C$**)**>	

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

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

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

1. OSP <*(Packaging)(numvalthick um)*> highlights that *packaging* and *thickness* arguments frequently appear to be correlated in text and that correlations frequently occur in a maximal textual window of $f_{\pm 1}$;
2. <(pressure)(water permeability)> shows that the *partial pressure* and *permeability* arguments frequently occur in the same sentence;
3. Several OSP suggest that the terms denoting the *packaging* concept could be the trigger of the relation since they frequently occur in OSP of previ-

ous correlations, e.g. *<(Packaging)(permeability)>*, *<(Packaging) (numvaltemp °C)>*.

Interested readers will find in (Berrahou et al., 2016) additional experimental results associated with the third step of the Xart system, which is the hybrid approach combining OSP with syntactic analysis in order to construct Ontological Linguistic Sequential Patterns (OLSP) for identifying correlated arguments directly in text.

7. Conclusion

We presented the Xart system based on a hybrid approach driven by an OTR that takes advantage of data mining techniques and syntactic analysis for complex data extraction from plain text. Thanks to the generic structure of the OTR, the Xart system may be used for different domains by only redefining the domain part of the OTR.
The first step of the Xart system proposes to enrich an Ontological and Terminological Resource (OTR) with new unit terms that are specific attributes of the sought n-ary relations. The proposed method enabled the identification of more than 57% of new units and units with typographic variations. In the second step, we propose a knowledge discovery process that takes the data expressiveness into consideration using the conceptual level given by the OTR, and defined the new notion of Ontological Sequential Patterns (OSP).

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

To sum up, the Xart approach applies a complete process *Data Information Knowledge*. This "information chain" is a key feature of the Xart system.

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

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

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

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

[6]http://www6.inra.fr/cati-icat-atweb/Web-platform

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

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

References

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

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

Béchet, N., Cellier, P., Charnois, T., & Crémilleux, B. (2012). Discovering linguistic patterns using sequence mining. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I* CICLing'12 (pp. 154–165). Berlin, Heidelberg: Springer-Verlag.

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

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

Language Processing: Shared Task BioNLP '09 (pp. 10–18). Stroudsburg, PA, USA: Association for Computational Linguistics.

Buche, P., Dervaux, S., Dibie-Barthélemy, J., Soler, L., Ibanescu, L., & Touhami, R. (2013a). Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique. *Revue d'Intelligence Artificielle*, *27*, 539–568.

Buche, P., Dibie-Barthélemy, J., Ibanescu, L., & Soler, L. (2013b). Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource. *IEEE Trans. Knowl. Data Eng.*, *25*, 805–819.

Bui, Q.-C., & Sloot, P. M. A. (2011). Extracting biological events from text using simple syntactic patterns. In *Proceedings of the BioNLP Shared Task 2011 Workshop* BioNLP Shared Task '11 (pp. 143–146). Stroudsburg, PA, USA: Association for Computational Linguistics.

Buyko, E., Faessler, E., Wermter, J., & Hahn, U. (2009). Event extraction from trimmed dependency graphs. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* BioNLP '09 (pp. 19–27). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cellier, P., Charnois, T., Plantevit, M., Rigotti, C., Crémilleux, B., Gandrillon, O., Kléma, J., & Manguin, J. (2015). Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. *J. Biomedical Semantics*, *6*, 27.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, *7*, 171–176.

Di-Jorio, L., Bringay, S., Fiot, C., Laurent, A., & Teisseire, M. (2008). Sequential patterns for maintaining ontologies over time. In *On the Move to Meaningful Internet Systems: OTM 2008, OTM 2008 Confederated Interna-*

tional Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part II (pp. 1385–1403).

Gkoutos, G. (2011). Units ontology. URL: `http://www.obofoundry.org/ontology/uo.html`.

Gruber, T. R., & Olsen, G. (1994). An ontology for engineering mathematics. In J. Doyle, P. Torasso, & E. Sandewall (Eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the 4th International Conference (KR '94): Bonn, Germany: 1994, May, 24 - 27* The Morgan Kaufmann Series in Representation and Reasoning (pp. 258–269). Morgan Kaufmann Publishers.

Guillard, V., Buche, P., Destercke, S., Tamani, N., Croitoru, M., Menut, L., Guillaume, C., & Gontard, N. (2015). A Decision Support System to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. *CEA*, (pp. 131–139).

Hao, Y., Zhu, X., Huang, M., & Li, M. (2005). Discovering patterns to extract proteinprotein interactions from the literature: part ii. *Bioinformatics*, *21*, 32943300.

Hawizy, L., Jessop, D., Adams, N., & Murray-Rust, P. (2011). ChemicalTagger: a tool for semantic text-mining in chemistry. *Journal of cheminformatics*, *3*, 17.

Hiemstra, D. (2000). A probabilistic justification for using tf x idf term weighting in information retrieval. *Int. J. on Digital Libraries*, *3*, 131–139.

Hodgson, R., Paul, J., Jack, H., & Jack, S. (2013). Qudt-quantities, units, dimensions and data types ontologies. URL: `http://www.qudt.org`.

Huang, M., Zhu, X., Payan, D. G., Qu, K., & Li, M. (2004). Discovering patterns to extract protein-protein interactions from full biomedical texts. In *Proceedings of the International Joint Workshop on Natural Language Processing in*

Biomedicine and Its Applications JNLPBA '04 (pp. 22–28). Stroudsburg, PA, USA: Association for Computational Linguistics.

Jaillet, S., Laurent, A., & Teisseire, M. (2006). Sequential patterns for text categorization. *Intell. Data Anal.*, *10*, 199–214.

Jessop, D. M., Adams, S. E., & Murray-Rust, P. (2011a). Mining chemical information from open patents. *Journal of cheminformatics*, *3*, 40.

Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., & Murray-Rust, P. (2011b). OSCAR4: a flexible architecture for chemical text-mining. *Journal of cheminformatics*, *3*, 1–12.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proc. of the conf. on Uncertainty in artificial intelligence* (pp. 338–345).

Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments - part 1. *Inf. Process. Manage.*, *36*, 779–808.

Kohavi, G. H., Ronand John (1998). The wrapper approach. In H. Liu, Huanand Motoda (Ed.), *Feature Extraction, Construction and Selection: A Data Mining Perspective* (pp. 33–50). Boston, MA: Springer US.

Kohavi, R., & Quinlan, J. R. (2002). Data mining tasks and methods: Classification: decision-tree discovery. In *Handbook of data mining and knowledge discovery* (pp. 267–276). Oxford University Press, Inc.

Le Minh, Q., Truong, S. N., & Bao, Q. H. (2011). A pattern approach for biomedical event annotation. In *Proceedings of the BioNLP Shared Task 2011 Workshop* BioNLP Shared Task '11 (pp. 149–150). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, *19*, 59–99.

Maedche, A., & Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web* (pp. 251–263). Springer volume 2473 of *LNCS*.

Minard, A.-L., Ligozat, A.-L., & Grau, B. (2011). Multi-class svm for relation extraction from clinical reports. In G. Angelova, K. Bontcheva, R. Mitkov, & N. Nicolov (Eds.), *RANLP* (pp. 604–609). RANLP 2011 Organising Committee.

Miwa, M., Sætre, R., Miyao, Y., & Tsujii, J. (2009). A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1* EMNLP '09 (pp. 121–130). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. (2001). Prefixspan: Mining sequential patterns by prefix-projected growth. In *Proceedings of the 17th International Conference on Data Engineering* (pp. 215–224). Washington, DC, USA: IEEE Computer Society.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods* (pp. 185–208). MIT Press.

Proux, D., Rechenmann, F., & Julliard, L. (2000). A pragmatic information extraction strategy for gathering data on genetic interactions. In P. E. Bourne, M. Gribskov, R. B. Altman, N. Jensen, D. A. Hope, T. Lengauer, J. C. Mitchell, E. D. Scheeff, C. Smith, S. Strande, & H. Weissig (Eds.), *ISMB* (pp. 279–285). AAAI.

Qiu, C. Z. T. Q. S. Z. J. C. P. Z. J., Jiangtaoand Tang (2007). A novel text classification approach based on enhanced association rule. In H. L. J. L. X. Z. O. R. Alhajj, Redaand Gao (Ed.), *Advanced Data Mining and Applications: Third International Conference, ADMA 2007 Harbin, China, August*

6-8, 2007. Proceedings (pp. 252–263). Berlin, Heidelberg: Springer Berlin Heidelberg.

Raja, K., Subramani, S., & Natarajan, J. (2013). Ppinterfinder - a mining tool for extracting causal relations on human proteins from literature. *Database*, *2013*.

Rijgersberg, H., van Assem, M., & Top, J. L. (2013). Ontology of units of measure and related concepts. *Semantic Web*, *4*, 3–13.

Rosario, B., & Hearst, M. A. (2005). Multi-way relation classification: Application to protein-protein interactions. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* HLT '05 (pp. 732–739). Stroudsburg, PA, USA: Association for Computational Linguistics.

Schadow, G., McDonald, C. J., Suico, J. G., Föhring, U., & Tolxdorff, T. (1999). Model formulation: Units of measure in clinical information systems. *JAMIA*, *6*, 151–162.

Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008). Discriminative parameter learning for bayesian networks. In *Proc. of the int. conf. on Machine learning* (pp. 1016–1023).

Touhami, R., Buche, P., Dibie-Barthélemy, J., & Ibanescu, L. (2011). An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. In *ODBASE Conferences (2)* (pp. 662–679).

Van Landeghem, S., Saeys, Y., De Baets, B., & Van de Peer, Y. (2009). Analyzing text in search of bio-molecular events: A high-precision machine learning framework. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task* BioNLP '09 (pp. 128–136). Stroudsburg, PA, USA: Association for Computational Linguistics.

Yan, X., Han, J., & Afshar, R. (2003). Clospan: Mining closed sequential patterns in large databases. In D. Barbará, & C. Kamath (Eds.), *SDM*. SIAM.

Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Mach. Learn.*, *42*, 31–60.

Zhang, H., Huang, M., & Zhu, X. (2011). Protein-protein interaction extraction from bio-literature with compact features and data sampling strategy. In *4th International Conference on Biomedical Engineering and Informatics, BMEI 2011, Shanghai, China, October 15-17, 2011* (pp. 1767–1771).

Zhou, D., Zhong, D., & He, Y. (2014). Event trigger identification for biomedical events extraction using domain knowledge. *Bioinformatics*, *30*, 1587–1594.