Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications

Martin Lentschat, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche

Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications

Martin Lentschat

TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
IATE, Univ. Montpellier, INRAE, INRIA GraphIK
E-mail: martin.lentschat@umontpellier.fr

Patrice Buche

IATE, Univ. Montpellier, INRAE, INRIA GraphIK
Montpellier, France
E-mail: patrice.buche@inrae.fr

Juliette Dibie-Barthelemy

Univ. Paris-Saclay, INRAE, AgroParisTech, JRU MIA-Paris, Paris, France
E-mail: juliette.dibie_barthelemy@agroparistech.fr

Mathieu Roche

CIRAD, TETIS, F-34398 Montpellier, France
TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
E-mail: mathieu.roche@cirad.fr

Abstract: This article presents an ontological and terminological resource (OTR) guided process for targeted extraction of scientific experimental data. Our method relies on the *scientific publication representation SciPuRe* describing the extracted data through ontological, lexical and structural features. Relevance scores may then be computed according to these features to rank the results and sort out some of the numerous false positives. These scores are based on lexical, semantic (depending on the OTR structure) and contextual (using segments in the scientific documents) criteria. Linear and sequential combinations of these scores are presented, implemented and evaluated.

Experiments were carried out on a corpus of 50 English language scientific papers in the food packaging field for the purpose of extracting the most relevant entities related to permeability relations. The findings revealed that article segment

categories are an effective criterion for filtering out a majority of quantitative entity false positives. The experiments showed that the best results for quantitative entity relevance only rely on lexical scores. Otherwise the best symbolic entity extraction results were obtained for sequential combinations of semantic and lexical scores. The results enable ranking of entities by relevance. A threshold can be used to filter false positive results. These scores also provide a useful measure for experts and advanced processes.

Keywords: data extraction; data relevance; data representation; ontological and terminological resource; information retrieval; web scientific documents.

Reference to this paper should be made as follows: Lentschat, M., Buche, P., Dibie-Barthelemy, J., Roche, M., (2021) 'Representation and Relevance Scores of experimental data extracted with an ontological and Terminological Resource', *International Journal of*, Vol. x, No. x, pp.xxx–xxx.

1 Introduction

The increased availability of online scientific publications offers new opportunities to exploit their data. Publications contain substantial information that can be harvested for potential consultation by experts, inclusion in meta-analyses or usage by advanced systems such as decision-support tools [1, 2]. Numerous research studies have been conducted on information extraction in the biomedical domain. This is due to both the high value (public health and commercial applications) of the extracted data and the abundance of textual resources available in this domain [3].

In the fields where there are few available textual resources dealing with specialised information, alternative strategies are required that take knowledge and expertise into account. This concerns the so-called *smart data* concept [4, 5] in comparison to the well-known *big data* concept [6]. In the smart data paradigm, the contextualisation and reliability of extracted data is a challenging issue. In experimental fields related to smart data, the information extraction process is based on smaller corpora, consisting at most of a few hundred documents [7, 8]. In these corpora, only a subset of experimental data must be extracted—particularly those useful for decision support, numerical prediction and meta-analysis. Targeting only data defined by experts as entities of interest constitutes a fundamental difference. Moreover, specific challenges due to smart data or domain complexity must be overcome (e.g. complex units of measures, terminological variations, studied objects designated by compound nouns).

In this paper, we focus on the extraction of experimental smart data from scientific documents in a specialised domain driven by the Ontological and Terminological Resource (OTR) defined in [9]. Our application domain is the study of food packaging. Published papers on this subject present the design process and investigate the characteristics of new packaging for food conservation. In this paper, only the extraction of information related to the packaging composition and permeability characteristics for decision support purposes is targeted [1]. This constitutes a restricted expertise domain requiring an extraction process able to distinguish relevant (smart) information (i.e. packaging composition and permeability and associated experimental parameters) from other packaging characteristics

(e.g. tensile strength, storage conditions) and packaging design process information. The information are represented as textual entities in the documents.

Two types of entities are considered, i.e. symbolic and quantitative entities. Symbolic entities are expressed in texts in the form of lexical expressions. In our research field, this concerns food packaging names (e.g. "low density polyethylene"), packaging components (e.g. "glycerol", "carboxymethyl cellulose") and experimental methods (e.g. "ASTM D95-96"). Quantitative entities consist of a numerical value and a measurement unit including, for instance, permeability values (e.g. "$4.34 * 10^{-3}\ cm^3 \mu m^{-2} d^{-1} kPa$"), experimental control parameters (e.g. temperature, relative humidity) or package thickness.

The challenges specific to our case study are: (1) the choice and adaptation of entity extraction techniques to the requirements and specificities of the domain and, (2) the discovery of textual entities that are actually related to the information of interest (i.e. food packaging composition and permeability). In order to address the first challenge, we use an OTR [9] to drive the entity extraction process.

The OTR defines the targeted entities through concepts and provides a lexicon describing each of these concepts to drive their recognition in the texts. The OTR coverage is the first challenge to address: the lexicon is usually not fully comprehensive or include all the forms present in the documents, including terminology variations, where a term may be present in plural form (e.g. $temperature \rightarrow temperatures$), adjectival form (e.g. $thickness \rightarrow thick$), or the order of terms may vary (e.g. $oxygen\ permeability \rightarrow permeability\ to\ oxygen$). Authors of scientific papers also make extensive use of acronyms to represent packaging names and their components (e.g. "low density polyethylene" $\rightarrow$ "LDPE") or quantity concepts (e.g. "relative humidity" $\rightarrow$ "RH"). As many acronyms may be found in several publications, while others may only occur in one article, they should be recognised on the fly. The recognition of measurement units is also a recurring problem when extracting experimental data. As these units are not harmonised between papers, any new measurement units must be recognised while also associating them with the corresponding quantitative concepts.

The second challenge concerns the high number of false positive results recognised. Numerous entities are present in each document representing, for instance, different packaging names, components and numerical values describing various parameters. The extraction process may encounter a high number of false positives, a priori indistinguishable from relevant entities. For instance, a packaging name present in a document is not necessarily the one whose permeability has been measured. It can be the result of an external work that is referred to for comparison. False positives may also be the result of an indistinction of morphology. For instance, $25°C$ is a temperature, but is it the one used as control parameter in the permeability measure or the room temperature during the packaging process? It is thus important to distinguish relevant from false positive entities. Relevance has long been a crucial concern in information retrieval [10, 11]. In this study, we consider that extracted information is relevant for users if it is correct (notion of *precision*) and representative (notion of *recall*).

We have addressed these challenges by developing a complete process that includes the following components: extension of OTR label coverage, data representation and valid entity selection.

Here we used *scientific publication representation* (*SciPuRe*) to represent the extracted entities. Different methods are applied to represent lexical data extracted from documents, with one of the most popular involving use of a vector space model to represent words in their lexical context [12]. In our case, we used an external knowledge source (i.e. an ontology) to

pinpoint entities of interest and categorise them. In our representation, we opted to include the text extracted as entities along with the disambiguation terms, linking these terms to the ontological corresponding concepts. Moreover, since different sections of scientific articles contain different pieces of information [13], it could be interesting to take different contexts into account in the data representation process. Sentence-level segmentation is commonly used for this purpose. It provides information on the local context and can improve extraction process, for example, through an analysis of the syntactic dependencies [14].

SciPuRe integrates several features for each extracted entity thus enabling an original computation of relevance scores. We designed relevance scores for extracted entities to ensure the selection of valid results. The lexical, ontological and structural features of SciPuRe are used in the computation of lexical and semantic scores. Classical lexical scores used in information retrieval are the *term frequency - inverse document frequency* ($tf - idf$) scores [15]. These lexical scores were extended here by using structural features in order to be able to exploit contextual information provided by the structure of the scientific papers (i.e. sections). The semantic score is based on the ontology structure in order to favour concepts most specific to the target field, as they are considered to carry more informative power. The specificity notion applied here is generally considered to be the opposite of the status notion introduced by [16]. These scores enable, through combinations tailored to each type of entity, selection of the best trade-off between accuracy and coverage when selecting the extracted entity to retain.

A state of the art encompassing the challenge addressed here is presented in section 2. Details on the extraction method proposed to increase the lexical coverage of the OTR are presented in section 3.1. SciPuRe and its lexical, ontological and structural features are described in section 3.2. We then present different lexical and semantic relevance measures based on the *SciPuRe* features in section 4 and describe how to combine them in section 4.3. These scores are designed to rank the results as accurately as possible. We assessed these proposals on a corpus of experimental data related to food packaging permeability measurements in section 5. One characteristic of our approach is its applicability to other domains, this is illustrate is section 6.

This article is a revised and extended version of [17]. Methodological details on the extraction process and SciPuRe have been added. The experiments were expanded to encompass 50 documents and the impact of all relevance scores were calculated. Linear and sequential combinations of semantic and lexical scores based on the SciPuRe features are proposed and evaluated. Finally, a discussion has been introduced on how the proposed approach can be used in other domains.

2 Related Work

In order to extract relevant information from documents while taking into account the requirements of the application domain, the challenge of extracting specific experimental data and sorting out false positive results must be tackled. In the following, those challenges will be considered according to three viewpoints: entity extraction, vocabulary enrichment, relevance of extracted information.

Entity extraction

Conventional information extraction methods are rooted in the field of medicine [18] and bio-medical domains [19]. The methods used for entity recognition in those domains rely

mainly on supervised machine learning (ML) [3] thanks to the high number of documents available on platforms such as PubMed. As the experimental domains targeted in this paper are more related to smart data than big data, the lack of text sources with reliable annotations eliminates the possibility of supervised ML. In similar cases, distant supervision [20] is a possibility to enable use of ML by skipping the building of a learning corpora. However this introduces noise in the data and often delivers lower quality results.

Rule-based methods using conventional state-of-the-art approaches for named entity recognition based on part-of-speech tagging [21], syntactic parsing [22] or statistical methods [23] are not able to take other specificities of the sought entities (e.g. complex units of measures, a lot of terminological variations) into account. However, these techniques can be extended for use in conjunction with external resources to effectively target entities of interest. As specific jargon is used in specialised fields, the resource could be a simple dictionary of terms [24] or a more complex resource such as an ontology [25, 26]. This kind of external resource makes it possible to define entities targeted by the extraction process. Here we use an ontological and terminological resource (OTR) specialised on the smart data of interest to drive the entity recognition process. The OTR is integrated in a complete extraction process and provides information useful to characterize extracted entities and evaluate their relevance via the scientific publication representation (SciPuRe) presented hereafter. SciPuRe integrates a set of ontological, textual and structural features following the common criteria [27] to represent information: discriminate differences, identify similarity, describe accurately and minimize ambiguity.

Vocabulary enrichment

The first concern when using the vocabulary of a resource to drive the entity extraction process is its coverage of the domain of interest. Terminological variations of the vocabulary defining the entities can be extracted from a list of terms present in documents via the analysis of morphological and syntax features [28, 29]. We use a python version of FASTR [28]. Different techniques exist for acronym recognition: use of external resources [a], pattern extraction [30] or syntactic analysis [31]. In this study, we designed our own acronym recognition algorithm based on terms available in the OTR used. This allowed us to adapt to the specificity of the targeted scientific domains. Recognition of new measurement units is also a key concern [32, 33] in order to extract all quantitative entities. It is essential to recognise measurement units that are not present in the OTR, while also linking them to the corresponding quantity concepts of the OTR. The method we propose is a complete extraction process which integrates these extensions as a preprocessing step to expand the OTR vocabulary.

Entity Relevance

Besides entity extraction, we must ensure that the entity recognized corresponds to those expected by the experts. This point is particularly important with regard to smart data extraction when some extracted entities do not belong to information of interest. The relevance of extracted information is a regular concern in information extraction [10, 11]. Information relevance is defined by how it satisfies the user's query [10]. Precision, recall and f-score are the standard criteria upon which information extraction tasks are evaluated. When the system includes a ranking of the results, to ensure that relevant results are retrieved

[a] https://www.acronymfinder.com/

first, other measures are used. Precision@N, or Precision@K, allows to represent changes in the precision value of a ranked series of results depending on the number of the n first results selected. Average Precision and R-Precision are standards used to represent the efficiency of a ranking system through a single number. Average Precision, i.e. a standard in the TREC community [34], provides an overall precision measure across different recall levels. R-Precision is better suited when the proportion of false positive results retrieved is substantial compared to the quantity of expected valid ones [35]. R-Precision is the precision value of the n first ranked results, with n being the number of known valid results. It thus adapts to the proportion of relevant information among all the extracted results. Therefore this is the indicator used to evaluate the entity extraction process proposed in this paper.

We consider that the original contribution of this paper is in proposing a complete pipeline for smart data extraction based on (1) an OTR which specifies the scope of the domain of interest; (2) an entity extraction process able to expand OTR vocabulary that generates a set of extracted entities associated with a SciPuRe representation; (3) a combination of relevance indicators, computed via SciPuRe representation, enabling the ranking of extracted entities by taking their context of appearance in the text into account. Moreover, the extraction process, the SciPuRe representation and the computation of relevance scores associated with extracted entities are designed to be adaptable to any experimental domain.

3 Extraction and Representation of Scientific Data

In this section, we present the overall extraction process as well as the resulting SciPuRe entity representations.

3.1 Entity extraction process

The entity extraction process we developed (see Figure 1) relies on an OTR structured around n-Ary relations [9]. This OTR includes a terminological component for each concept which is used to drive the entity extraction process. The OTR is structured in a core ontology and a domain ontology.

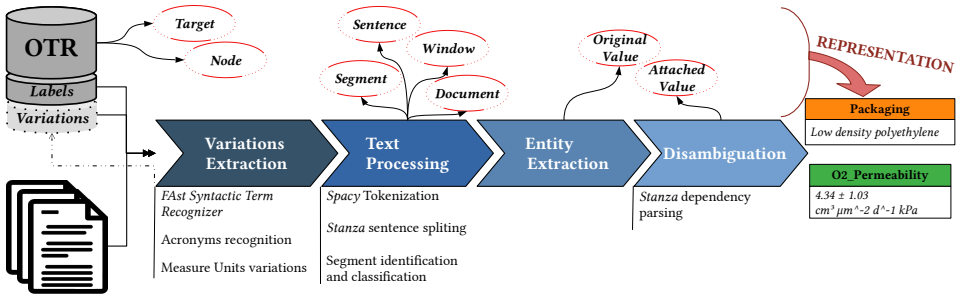


Figure 1: Experimental data extraction in specialised domain driven by an OTR

The up-core ontology includes a representation of the structure of n-Ary relations and their arguments. The down-core ontology contains the main concepts specific to the

experimental fields, such as quantitative or symbolic concepts and measurement units. The domain ontology contains concepts related to our specific field of interest, i.e. food packaging permeability. Figure 2 shows an excerpt of the *TRANSMAT* OTR structure [b].

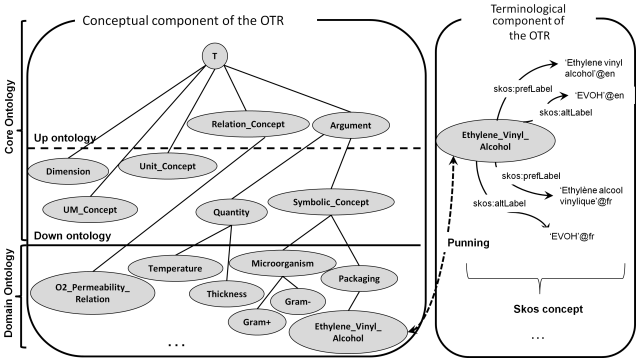


Figure 2: An excerpt of the structure of TRANSMAT Ontological and Terminological Resource

Each symbolic or quantitative concept is associated with a terminological component in the form of *labels* (*preferred* or *alternative*). Quantitative concepts are also associated with measurement unit concepts. Each measurement unit is associated with a set of labels. All of these labels are used for text entity recognition. The TRANSMAT ontology describes 62 relation concepts based on the use of 2,432 symbolic concepts, 82 quantity concepts and 62 unit concepts.

Note that the entity extraction process proposed in this paper is independent of the specific domain (here food packaging). It may be applied to another scientific domain by simply replacing the domain ontology part of the OTR.

The OTR drives the entity extraction process based on its concepts and associated vocabularies. The measurement units and *labels* of the concepts involved in the n-Ary relations of interest define the tokens forming the entities in the documents (see Example 1).

Example 1: *Recognised entities:*
The permeability of (*low density* (*polyethylene*)) *films* ((*LDPE*)) *was measured with the* (*ASTM D95-96*) *method at* $\underline{25 \pm 1}$ $\boxed{^{\circ}C}$. *The film had a* (*thickness*) *of* $\underline{15}$ $\boxed{\mu m}$ *and showed optimal barrier properties with a* (*permeability to oxygen*) *of* $\underline{4.34 * 10^{-3}}$ $\boxed{cm^3 \mu m m^{-2} d^{-1} kPa}$. *This measurement was obtained at a constant* (*RH*) *of* $\underline{85.0}$ $\boxed{\%}$.
Legend: (*OTR label*) | *measurement unit* | *numerical value*

To improve entity extraction, the OTR vocabulary was expanded with terminological variations using *FASTR* [29]. This tool extracts terminological variations of a list of terms in a document via the analysis of morphological and syntactic characteristics. It can recognise

[b] https://ico.iate.inra.fr/atWeb/ - 15/11/2020

OTR labels present in plural form (e.g. $temperature \rightarrow temperatures$) or adjectival form (e.g. $thickness \rightarrow thick$). It can deal with single or multi-word terms based on rules such as the insertion of modifiers, determiners and prepositions (e.g. $linear\ polyethylene \rightarrow linear\ low\ density\ polyethylene$). We modified to be able to capture more multi-word terms by lifting the word-order conservation restrictions (e.g. $oxygen\ permeability \rightarrow permeability\ to\ oxygen$).

We used the method described in [32] to recognise measurement unit variations. This method was rewritten in Python language in order to be compatible with other extraction codes. It begins by finding a candidate for a new measurement unit: a sequence of tokens located between two terms of a standard dictionary and containing at least one token present in an existing unit. The candidates are then filtered using a Jaccard index score and an extended Damerau-Levenshtein measure. These scores validate candidates close to units of measure already existing in the OTR. It also enables the association of each candidate with an existing measurement unit in order to link the new measure units to quantity concepts.

On the other hand, the acronym recognition task we developed begins with the identification of OTR labels present in the texts. Candidate acronyms for each label are selected via straightforward heuristics (e.g. proximity, mostly composed of upper case letters, parenthesis). Similarity scores are then computed using *Dice coefficient* [36] and results above a determined threshold are added to the OTR alternative labels. This allows us to consider the similarity between the first letters of an OTR label and a candidate acronym without being overly restrictive regarding the order. Indeed, specialised terms are often broken down into several characters in an acronym (e.g. $Dice(\text{"}lowdensitypolyethylene\text{"}, \text{"}LDPE\text{"}) = .86$).

After extension of the OTR coverage, the texts are split into tokens using *spacy 2* [c] while *stanza* is used for sentences [37, 38]. The names of document sections are also automatically gathered in classes based on low *Levenshtein proximity* [39] and inclusion (e.g. $Classe_{\{Result\ and\ Discussions\}} = \{Results\ and\ Discussions\} \cup \{Result\ and\ Discussion\} \cup \{Results\} \cup \{Discusion\}$).

When these pre-processing steps are completed, the extended OTR vocabulary is used to recognise terms and measurement units (along with numerical values) that will constitute the entities of interest. Our extraction process separates recognized terms into symbolic entities and quantitative entities. Symbolic entities correspond to terms of symbolic concepts in the OTR (e.g. `Packaging` or `Method`). Quantitative entities are composed of an OTR measure unit and the numerical value associated. The association between a measure unit and its numerical value is found through dependency parsing. Token proximity is used if no direct dependency can be found. Recognised tokens related to unambiguous entities (e.g. symbolic entities and quantitative entities with fundamental measurement units like $"^{\circ}C"$) are matched with their corresponding concepts in the OTR. Other quantitative entities are disambiguated by associating the measurement unit with an identified term denoting the concept. Such associations are also discovered via syntactic analysis, dependency parsing using stanza [37, 38] and token proximity. In example 1, $"4.34 * 10^{-3}"$, $"cm^3 \mu m m^{-2} d^{-1} kPa"$ can be disambiguated into an `O2_permeability` entity through *"permeability to oxygen"*.

This extraction process does not require a training corpus and relies only on straight forward recognition techniques driven by the OTR vocabulary. Section 5 presents the recall, precision and f-score for each of the entity of interest.

[c] https://spacy.io/

3.2 Scientific publication representation (SciPuRe)

SciPuRe, first proposed in [17], is associated with each extracted entity to gather useful information for entity relevance evaluation. *SciPuRe* involves three categories of features:

- **Ontological Features**: The *Target* feature indicates the OTR top concept to which the entity is associated. A top concept is an argument of a n-Ary relation defined in the OTR, e.g. the `O2Permeability_Relation` relation links these arguments together, i.e. the symbolic concept `Packaging` and quantitative concepts such as `Temperature`, `O2_Permeability` or `Relative_Humidity`. The *Node* feature specifies the sub-concept the entity represents (i.e. the sub-concept containing the label used for entity recognition).

  The extracted entity *LDPE* of Example 1 corresponds to an alternative label of the `Low_Density_Polyethylene`concept which in turn is a sub-concept of `Packaging`.

- **Lexical Features**: The *Original Value* feature contains the text corresponding to the extracted entity. The *Attached Value* feature corresponds to the terms in the sentence associated with quantities in the OTR which are used to disambiguate the measurement unit when necessary. Otherwise, the *Attached Value* feature is the same as the *Original Value* for symbolic entities, or the PrefLabel of the *Node* concept for quantitative entities.

  In example 1 *RH* is an alternative label of the `Relative_Humidity` concept and allows disambiguation of *50 %*.

- **Structural Features**: The *Sentence* and *Window* (i.e. previous, current and next sentences) features indicate the textual context in which the entity appears. The *Segment* feature (e.g. sections like Introduction, Materials and Methods) allows the structure of a scientific article to be taken into account. The *Document* feature provides references to the article (e.g. title, authors, year).

Two examples of *SciPuRe* representations extracted from the sentence of example 1 are presented in Tables 1 and 2.

	Feature	Example
ONT.	Target	`Permeability`
	Node	`O2_Permeability`
LEX.	Original Value	['$4.34 * 10^{-3}$', '$cm^3 \mu m m^{-2} d^{-1} kPa$']
	Attached Value	*'permeability', 'to', 'oxygen'*
STRUCT.	Sentence	*'The film had ...* $d^{-1}kPa$*'*
	Window	[*'The permeability ...* $25 \pm 1\ °C$*'*, *'The film ...* $d^{-1}kPa$*'*, $\emptyset$]
	Segment	*'Results and Discussion'*
	Document	*A. Farro and al. - Development of films based on quinoa starch*

Table 1 *SciPuRe* representation of a quantitative entity

	Feature	Example
ONT.	Target	`Packaging`
	Node	`Low_Density_Polyethylene`
LEX.	Original Value	*'LDPE'*
	Attached Value	*'LDPE'*
STRUCT.	Sentence	*'The permeability of ... at* $25 \pm 1\ °C$*'*
	Window	[$\emptyset$, *'The permeability ...* $25 \pm 1\ °C$*'*, *'The film ...* $d^{-1}kPa$*'*]
	Segment	*'Results and Discussion'*
	Document	*A. Farro and al. - Development of films based on quinoa starch*

Table 2 *SciPuRe* representation of a symbolic entity

4 Relevance scores

Our entity extraction method must be able to distinguish between relevant and irrelevant terms, e.g. between terms related to the studied packaging and those quoted for comparison purposes, or between the controlled temperature value during the experiment and those involved in the packaging preparation. We decided to address this issue through relevance scores computed from SciPuRe features. A relevance score is associated with each extracted entity. The aim is to associate valid extracted entities with high relevance scores in order to choose a threshold to filter the valid results. An evaluation by Precision@N [40] (also known as Precision@K [35]) is presented in Section 5 to assess the relevance scores proposed below.

4.1 Lexical relevance scores

A relevance score based on the notion of term discrimination by computing the *term frequency - inverse document frequency* (*tf-idf*) indicator [15] is proposed. *Tf* is based on the hypothesis that the most frequent terms in a document are the most important. *idf* aims to reflect the discriminating nature of the terms, while giving greater importance to those just specific to few documents.

Table 3 lists the different lexical relevance scores proposed. The SciPuRe features provide the elements to compute the lexical scores of an extracted entity at different levels. The `Attached Value` feature is the element indicating the manifestation of the entity in the text. Its frequency (tf) or presence (idf) is considered in relation to its context. This context is usually the `Document` but the text segment may also be considered. `Segment` informs on the section in which an entity is present. Since *Segments* are grouped into segment classes in order to consider sections named differently but with seemingly similar contents, the more generic $tf - icf$ indicator (*term frequency - inverse category frequency*) proposed in [41] is used for the computation. Note that `Target` can be used instead of `Attached Value` to consider generic concepts pooling entities.

4.2 Semantic relevance scores

As the entities extracted from scientific documents are intended to be used by experts or advanced systems, their relevance measurement must also reflect their informative power. This informative power could be considered through the concept specificity. For example,

Nom	feature 1	feature 2	Equation
$TF^{term}_{document}$	`Attached Value` t	`Document` d	$\frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$
$TF^{term}_{segment}$	`Attached Value` t	`Segment` s	$\frac{f_{t,s}}{\sum_{t' \in s} f_{t',s}}$
$TF^{target}_{segment}$	`Target` a	`Segment` s	$\frac{f_{a,s}}{\sum_{a' \in s} f_{a',s}}$
$IDF^{term}_{document}$	`Attached Value` t	`Document` d	$log \frac{\|D\|}{\|d \in D : t \in d\|}$
$ICF^{term}_{segment}$	`Attached Value` t	`Segment` s	$log \frac{\|S\|}{\|s \in S : t \in s\|}$
$ICF^{target}_{segment}$	`Target` a	`Segment` s	$log \frac{\|S\|}{\|s \in S : a \in s\|}$

Table 3 Definition of the lexical relevance scores based on SciPure features

if *multilayer film* is actually a kind of `packaging`, the more specific *PE films coated with chitosan* would be preferred.

SciPuRe includes the OTR concept associated with the (`Node`) entity and its generic concept (`Target`). The distance (i.e. number of edges) from `Node` to `Target` is computed using the OTR concept hierarchy. It expresses the entity specificity measurement, inspired by [16], in the *Conceptual Distance* CD^{node}_{target} relevance score (see equation 1). The relevance of each entity corresponds to the distance between `Node` n and `Target` a denoted $dist(n, a)$. This is compared to the maximum distance between the generic concept considered a and all of its sub-concepts n', denoted $max(dist(n', a) : n' \sqsubseteq a)$, where $\sqsubseteq$ denotes the specialisation relationship (i.e. subsumption) in the OTR. The relevance measurement of CD^{node}_{target} is assumed to be more useful for symbolic entities, as these are described at more specialisation levels in the OTR.

$$Conceptual\ Distance \quad CD^{node}_{target} = \frac{1 + dist(n, a)}{1 + max(dist(n', a) : n' \sqsubseteq a)} \tag{1}$$

4.3 Combination of scores

The relevance scores presented above can be used alone or in combination. Conventionally, *tf* and *icf* scores are combined by multiplication in order to jointly consider the relative frequency and the discriminating character of a term in a document. Concerning relevance scores associated with extracted entities, it would seem more appropriate to consider other ways of combining the different scores so as to fine-tune the effects. Both linear and sequential score combinations are proposed in order to benefit from properties associated with each type of relevance score. For example, combining a lexical score of type *tf* with the semantic score CD^{node}_{target} enables us to take both the frequency of the extracted entity in the texts (i.e. lexical criterion) and the specificity of its associated concept (i.e. semantic criterion) into account. Note that before combining the relevance scores they must first be normalised with a *min-max* function on a $[0, 1]$ scale.

Linear combination

The linear combination (see equation 2) sums the different scores after assigning a weight to each of them. The total sum of all α_i weights is always equal to 1.

$$Linear(Score_i) = \sum_{i=1} \alpha_i.Score_i \quad : \quad \sum_{i=1} \alpha_i = 1 \tag{2}$$

Sequential combination

The purpose of sequential combination is to assign a relevance score to a set of extracted entities that have been pre-filtered by another score. For example, the CD^{node}_{target} score can be used to first eliminate the less specific according to their semantic (i.e. ontological) aspect. Then a lexical score such as $TF^{term}_{segment}$ will select the most frequent extracted entities in the subset of the remaining results.

Sequential combination thus involves ranking the extracted entities according to a first $Score_1$. A subset consisting of a proportion θ (%) of the first results is then re-ranked according to a $Score_2$. This process can be replicated i times until the last score to be considered $Score_i$ in order to benefit from the specific effects of each score sequentially. Naturally, the choice of the combination order is important.

5 Experiments

The experiments conducted here aimed to measure how relevance scores (and their combinations) could be employed to improve the reliability of the extracted entities. Starting from the extraction results of the method described in section 3, we show how the different relevance scores affect the different entities. Finally, we tried linear and sequential combinations of lexical and semantic scores in an attempt to improve the ranking accuracy.

5.1 Gold Standard to assess entity extraction results

The ontology-driven method for extracting entities described in section 3 was applied on a corpus of 50 documents (≈ 258000 words and 9400 sentences after cleaning). These documents were manually collected from *ScienceDirect* in `html` format and then processed to obtain a version stripped of unnecessary elements and retaining only text and structural information. These 50 documents are considered to be representative of the packaging permeability domain. Indeed, @Web platform [d] which stores this kind of data in PackPermXXX folders currently hosts around 200 annotated documents in this domain which contain the required information.

Note that other experiments on automatic extraction of scientific concepts have been conducted on small corpora. This reflects the limited amount of textual resources available in specialized domains beyond the medical scope. A recent study [7] uses a corpus of 110 documents encompassing 110 abstracts from diverse domains (i.e. from agriculture to computer sciences). The authors compare their results with other works using larger corpora and ML [42, 43]. They obtained similar results and concluded that a corpus of 110 abstracts is sufficient such tasks. An other study [8] used 300 radiology reports for the extraction of quantitative and symbolic entities related to research on kidneys.

[d] https://ico.iate.inra.fr/atWeb/

The *Gold Standard* of the 50 documents was established by three annotators on a *WebAnno* [44] server (see figure 3). All of them were familiar with annotation while two were well acquainted with the subject area. The instructions given to the annotators were to identify only entities related to packaging permeability relations. For example, a packaging name quoted as a bibliographical reference or a temperature other than a permeability control parameter were not to be annotated. Symbolic entity identification was straightforward: a word or sequence of words were annotated (e.g. the `method` entity in figure 3). Quantitative entities required linkage of the identified numerical value and measurement unit, and sometimes a term was used to disambiguate the unit. Annotations were thus linked together: the numerical value to the measurement unit (e.g. the `temperature` entity in figure 3), and then the measurement unit to a term if necessary (e.g. the `relative_humidity` entity in figure 3).

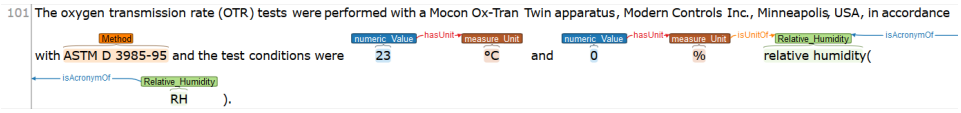


Figure 3: Entity annotation in WebAnno

The annotations were then automatically recovered to constitute the Gold Standard. Its content include entities annotated in articles with different character sequences, and the position of entities in the documents was not taken into account. This choice was made because many duplicates were present in the documents, thus the annotation of all occurrences would not generate any necessary information for the task at hand.

A *Gwet's Kappa* score [45] was computed, with the average score being $\kappa_{average} = 0.62$, indicating a moderate level of agreement, reflecting the difficulty for annotators to determine the relevance of some entities. This is a regular concern in the annotation of scientific documents. For instance, [7] obtained $\kappa_{Cohen's}$ values ranging from .94 (medicine) to .57 (astronomy). For instance, *multilayer films* is generic yet it does reflect a `packaging` entity, so one annotator may decide to annotate it in addition to the name of the specific packaging (e.g. *PE films coated with chitosan*) whereas another may decide to only annotate the latter. The Gold Standard is the result of merging the annotations provided by the three annotators. The annotations of the main annotator prevailed in the event of a conflict over the category assigned to a term. The pre-processing and extraction algorithms [e], the version of the OTR used [f] and the Gold Standard [46] are available online.

5.2 Extraction results

Table 4 presents, by entity type, the number of distinct entities annotated in the Gold Standard, the number of recognised entities and the extraction results according to the recall, precision and F-score (micro). The general recall value was .85, with some variations depending on the categories of the considered entities. The general precision value was .41, and was subject to more variations, with an average of .47 for symbolic entities and .14 for quantitative ones. This was due to the larger number of false positives in the extraction of entities that included numerical values. For example, many temperatures were identified

[e] https://github.com/Eskode/ARTEXT4LOD - 15/11/2020
[f] http://pfl.grignon.inra.fr/atWeb/ - 15/11/2020

(1925) compared to the number of annotated temperatures (54), namely those associated with permeability measures. This disparity between the number of annotated entities in the Gold Standard and the number of extracted entities was also noted for symbolic entities. This had a lower impact on precision because more duplicates were found. The precision also depended on the type of considered symbolic concept: the precision of `method` (.16) was much lower than that of `component` (.56). This was due to the high number of occurrences of the generic term *method*, a false positive, compared to specific designations such as *ASMT D95-96*). The recall values obtained enabled extraction of a large number of valid entities. As the precision was more uneven, the extracted entities had to be filtered to obtain relevant information. This was the aim of the experiments described hereafter based on the relevance scores described in section 4.

`Target`	#distinct	#recognised	recall (%)	precision (%)	F-score
SYMBOLIC	988	16665	85	47	61
packaging	431	6940	86	37	51
component	514	9506	84	56	67
method	43	219	77	16	26
QUANTITATIVE	303	3994	86	14	24
permeability	150	832	83	16	27
relative_humidity	55	696	88	28	43
thickness	44	541	100	14	24
temperature	54	1925	83	08	15
GENERAL	1291	20659	85	41	55

#distinct : number of distinct entities present in the Gold Standard.
#recognised : number of entities recognised by the extraction process.

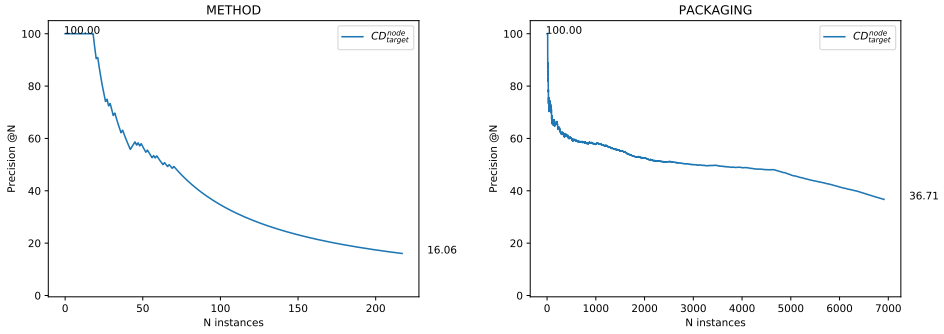
Table 4 Entity extraction results using the Gold Standard.

5.3 Ranking score evaluation

We used *Precision@N* [40] (also known as Precision@K [35]) to assess the usefulness of relevance scores for ranking the results. For a set of entities ordered according to a given score, this involved computing the precision value of the first *N* results. Variations in N, from $N = 1$ to $N = all$, represented the precision variation pattern according to N on a curve. This evaluation procedure highlights the precision obtained with a relevance score according to the number of considered entities. The Precision@N plots facilitated selection of a threshold to filter the results according to a relevance score or helped decide on its use in combination with other scores (see section 5.4).

Figure 4 displays the Precision@N of entities associated with the generic concept `method` and `packaging` ranked accordingly to their CD^{node}_{target} scores. The x-axis indicates the number of N *best* selected entities according to CD^{node}_{target}, while the y-axis indicates the associated Precision@N values. In figure 4a, the Precision@N is 100% up to $N = 24$ entities, and it then gradually decreases and reaches the average precision of `method` at $N = all$. Above $N \approx 70$ ($\approx 35\%$ of the population), the curve is monotonic and decreasing (i.e. entities selected beyond that threshold are only false positives). Figure 4b shows a similar, though less pronounced, behaviour for `packaging`. We observed the same behaviour for entities associated with `component`. These results indicate that sorting

out entities of symbolic concepts with the CD^{node}_{target} score and retaining only those with top values is an efficient way to filter valid results.



(a) (b)

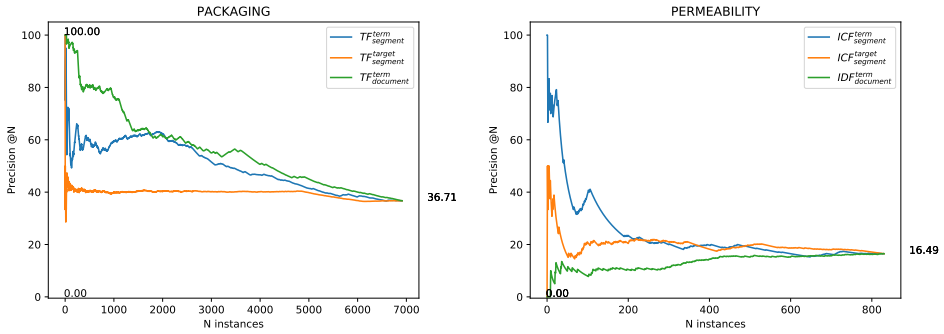
Figure 4: *Precision@N* computed for semantic relevance scores associated with for `method` and `packaging` concepts of the OTR

The impact of type tf lexical relevance scores for a ranking symbolic entities is presented in figure 5a with `packaging` entities. The reason for this is that the names of the packaging upon which the study was focused were repeated in each document. We observed similar results for `component` entities. $TF^{term}_{segment}$ also yielded exploitable results and proved to be better for `method` entities (see Tables 5 and 6). Therefore the names of the sought methods seemed to be more frequently present in specific sections (e.g. *"Material and Method"*). It would be possible to decide to filter part of the results with lexical relevance scores of type tf by removing, for example, the last 25% while accepting the fact that some valid entities would be lost (risk reduced by the presence of duplicates).

Scores such as idf and icf performed well in measuring the relevance of quantitative entities (see figure 5b). The use of `Segments` in the $ICF^{term}_{segment}$ score produced the strongest results. Relevant quantitative experimental data were present in specific sections (e.g. *"Material and Method"*), as reflected in the lexical relevance scores. The Precision@N curve rapidly declined for quantitative entities with low overall precision. Relevance scores of the icf type could thus be used to roughly filter the results (by removing $\approx 75\%$ of the population) without the risk of excluding too much relevant information.

Tables 5 and 6 show precision values of entities ranked using the presented relevance scores. The average precision [35] in Table 5 was computed using all the Precision@N values, from $N = 1$ to $N = all$. Table 6 presents the R-precision [35] indicating the precision value at an N equal to the number of valid entities for the considered `Target`.

Overall, the semantic scores revealed improvements in the precision values for semantic entities but not for the quantitative entities. R-Precision shows a clear improvement in the precision of symbolic entities. The symbolic entity precision was equal to .47 and improved up to .55 with CD^{node}_{target} and up to .66 with $TF^{term}_{document}$ This varies depending of the entity considered and is most visible with `method` entities: CD^{node}_{target} improved the precision of .16 up to .64. This confirmed our intuition that CD^{node}_{target} was highly dependent on the ontology structure and therefore not applicable to quantitative entities. We observed that



(a) (b)

Figure 5: Lexical relevance measures for entities of `packaging` and `permeability` entity

Target	p*	CD_a^n	TF_d^t	TF_s^t	TF_s^a	IDF_d^t	ICF_s^t	ICF_s^a
SYMBOLIC	47	55	**64**	51	52	53	49	47
packaging	37	49	**56**	50	40	31	30	36
component	56	60	**71**	52	61	70	63	56
method	16	41	18	**28**	25	20	25	18
QUANTITATIVE	14	13	13	13	14	12	13	14
permeability	16	16	13	15	17	14	**21**	15
relative_humidity	28	27	28	27	33	22	33	**35**
thickness	14	14	14	14	13	11	19	**20**
temperature	08	07	08	08	08	06	05	05

p*: baseline precision

Table 5 Average Precision values using relevance scores

lexical scores involving frequency (i.e. tf) were suitable for measuring the relevance score of symbolic entities. $TF_{document}^{term}$ was well adapted for `packaging` and `component`, the reason being that the terms related to the symbolic entities are the main subject of the papers and therefore highly frequent. `method` entities were more specific to certain sections of the documents, therefore $TF_{segment}^{term}$ was more suitable.

The quantitative entities sought were specific to certain sections of the articles yet they were not present in large numbers. The improvement achieved when comparing R-Precision to the previous precision of quantitative entities depends on the entity considered. Except for `temperature` entities, the R-Precision showed significant improvement: .16 to .30 for `permeability` and .14 to .24 for `thickness` with the $ICF_{segment}^{term}$ score. For the `relative_humidity` entities, $ICF_{segment}^{target}$ shwed better improvement with a precision of .28 and a R-Precision of .49 after ranking. The lexical relevance score using the icf model is thus appropriate for sorting out the valid quantitative entities. `temperature` entities were a remarkable exception, with no relevance scores presenting interesting results. This was due to the lack of explicit terms upon which to base a score computation as the *"temperature"* term was uniformly found throughout the documents.

Target	p*	CD_a^n	TF_d^t	TF_s^t	TF_s^a	IDF_d^t	ICF_s^t	ICF_s^a
SYMBOLIC	47	55	**66**	55	52	45	46	47
packaging	37	50	**59**	57	41	22	24	36
component	56	58	**73**	53	61	63	63	56
method	16	64	17	**44**	28	25	31	19
QUANTITATIVE	14	13	13	12	17	12	12	15
permeability	16	17	9	17	29	7	**30**	10
relative_humidity	28	27	28	26	31	19	35	**49**
thickness	14	14	13	12	12	11	**24**	22
temperature	08	06	08	04	06	02	04	03

p*: baseline precision

Table 6 R-Precision values using relevance scores

5.4 Evaluation of scores combinations

As lexical relevance scores are well suited for symbolic entities, we conducted experiments on combining them with the semantic score to improve their effects. Linear and sequential combinations of $TF_{document}^{term}$ with the semantic score CD_{target}^{node} were then evaluated for `packaging` entities in order to compare their respective effects. This allowed us to combine scores using different types of information (lexical and semantic) in order to fine-tune the entity relevance assessments.

Linear combination

The linear combination in figure 6a is a combination of the CD_{target}^{node} and $TF_{document}^{term}$ scores of the `packaging` symbolic concept. This combination thus enhanced the semantic specificity of the entities with respect to their frequencies in the documents after giving them different weights. We did not observe significant gains using linear combinations, as illustrated in Table 7, at any α_i value (see Equation 2). This suggests that the linear combination did not really take advantage of the specific criteria associated with the combined scores, but instead balanced them.

$\alpha_i =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Average precision	56	57	57	57	56	56	55	54	54
R-Precision	58	59	59	58	58	58	58	58	58

Table 7 $Linear(CD_{target}^{node}, TF_{document}^{term})$ precision for `packaging` entities

Sequential combination

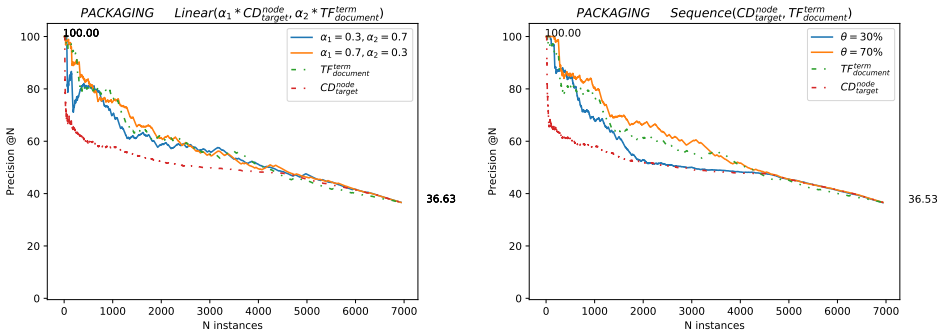
Sequential combination improved the symbolic entity relevance measurements. Figure 6b shows the effects of sequential combination for `packaging` entities: $Sequence$-$(CD_{target}^{node}, TF_{document}^{term})$, where CD_{target}^{node} was used to filter out a θ proportion of the results before $TF_{document}^{term}$. Since non-specific entities can be very frequent in documents (e.g. like the word *"packaging"*), this sequential combination resulted in a better final ranking than either of the two scores alone. Table 8 displays the impact of different re-ranking

proportions, different θ values (see section 4.3), on sequential combination of semantic and lexical scores for ranking `packaging` entities. Filtering out a small portion ($\approx 30\%$ to 20%) of entities using the semantic score before the lexical score was found to enhance the relevance measure. The R-Precision of `packaging` with this sequential combination is .63, while R-Precision of the semantic and lexical scores previously used were .50 and .59, respectively. This represents an improvement of .04, which is still noteworthy.

$\theta(\%) =$	10	20	30	40	50	60	70	80	90
Average precision	51	53	53	54	55	56	57	57	57
R-Precision	50	50	49	51	57	61	**63**	**63**	60

Table 8 $Sequence(CD^{node}_{target}, TF^{term}_{document})$ precision values for `packaging` entities

A similar behaviour was observed with `component` entities. The $Sequence$-$(CD^{node}_{target}, TF^{term}_{segment})$ sequential combination had an even better behaviour for `method` entities, indicating that the use of text segments could be more efficient for some entities of symbolic concepts.



(a) (b)

Figure 6: Linear and sequential combinations

Sequential combination of scores therefore proved to be more suited than linear combination for entities of symbolic concepts. This makes it possible to combine criteria of different types (semantic and lexical) to measure the relevance of these entities.

Our experiments did not reveal any score combinations adapted to the ranking of entities of quantitative concepts. Moreover, semantic scores such as CD^{node}_{target} are not adapted to such entities because they are generally described on a small number of levels in the ontology. We therefore recommend filtering entities of quantitative concepts using icf type scores for text segments identified in the scientific publications. This supports the intuition that the sections of the articles have an important discriminating power that should be included in the process of extracting targeted experimental data.

These relevance scores and their combinations might present similarly behaviour on an experimental data extraction task on other domain. Indeed, the OTR-driven extraction method presented in section 3 is domain independent. This assumption requires an OTR of

an other experimental domain and a new Gold Standard. This could lead to an more generic application the relevance score presented.

6 Discussion: Applicability to other domains

Our entity extraction method could be applied to other experimental domains. Firstly, it is suited for domains in which entities are complex because they include both studied object names with terminological variations and complex units of measures. Secondly, only a subset of experimental data present in the articles is useful and must be extracted.

Changing the domain ontology for another experimental field is the main requirement to apply our method to an other experimental domain. As presented in section 3.1, the entity extraction process we developed relies on an OTR structured with symbolic, quantity and unit concepts [9]. This OTR includes a terminological component for each concept, which is used to drive the entity extraction process. The entire extraction process is highly dependant of the completeness of the OTR description, both for the detection of new terms or measure units and the disambiguation of quantity concepts. The use of a domain OTR also enables representation of entities in SciPuRe, which itself enables the computation of relevance scores.

Examples of application to other domains are provided hereafter. The vocabulary in these domains is highly specialised and contains both symbolic and quantitative entities that may be of interest. Biorefinery and food spoilage are application domains that manage experimental information for dedicated tasks. Preliminary studies have already been conducted on food spoilage assessment [47] and biorefinery [2]. Dedicated OTR and annotated data-sets have been created [48]. Below is an example of an extraction using the VALORCARN OTR [g] [49] in the domain of meat food spoilage by pathogens. This OTR is aimed at relations regarding microorganism growth conditions and defines symbolic entities of interest such as `microorganism` or `matrix` and quantitative entities (i.e. experimental conditions) such as `temperature` or `time`. The microbial growth relation described by this OTR covers 49 symbolic concepts, 10 quantity concepts and 14 unit concept.

Example 2: *Recognised entities:*
For macro-morphological observations, the isolates were three-point inoculated on (*MEA*) *medium and grown for* 7 |*days*| *at* 25 |°*C*| *in the dark. The isolates from the genera* (*Aspergillus*) *and* (*Penicillium*) *were additionally three-point inoculated ...*
Legend: (*OTR label*) |*measurement unit*| *numerical value*

Example 2 is an excerpt of [50], available on ScienceDirect. Two `microorganism` can be recognized: *Aspergillus* and *Penicillium*. The `matrix` is *MEA*. As it is the acronym of a three word term, it would probably require term variation extraction to be recognized. Control parameters, `time` and `temperature` are extracted according to the process detailed in section 3.1.

These entities can then be represented using SciPuRe based on features depending on the ontology, structure and lexicon (see section 3.2). Exactly as in the food packaging domain, considerable information is present in the articles. The entities extracted do not

[g] available on https://ico.iate.inra.fr/atWeb/ under Ontology thumbnail- 15/03/2021

systematically concern the microorganism growth conditions studied in the article. They may involve in entities from an external source, quoted for comparison, or entities specific to the article but not related to the microorganism growth conditions (e.g. the time and temperature at which an organism was stored before the experimentation). Relevance scores may then be computed using SciPuRe features and used to sort out false positive results (see section 4).

7 Conclusion

The findings of the experiments presented in this paper show that lexical and relevance scores can be employed to rank experimental entities.

The extraction method applied to entities related to food packaging permeability led to a high proportion of false positives. SciPuRe of the extracted entities was used to compute lexical and semantic relevance scores to sort out the results. The relevance of semantic entities was better measured with type tf lexical scores using the frequency in the documents or, in some cases, in textual segments (i.e. sections) of the articles. Quantitative entity relevance was efficiently assessed with the icf type score which makes extensive use of textual segments of articles.

Score combinations were also considered to boost the effects of lexical scores using a semantic score measuring the concept specificity. Linear combination led to no improvements due to its inability to take advantage of the specific criteria that support the different scores. However, sequential combination had interesting effects when the semantic score was used to filter a small proportion of less specific results prior to using a lexical score to rank symbolic entities.

Relevance scores and their combination could thus be used to sort out some of the extracted entities, thus making it possible to find trade-offs between the completeness and validity of the results. The selected entities could then be proposed to experts or be integrated in subsequent processes.

On a larger scale, the SciPuRe features used for score computations provided essential material for *word embedding*. Contexts enhanced by *word embedding* (e.g. *word2vec*, or *BERT*) will help generate new information to incorporate into the representation and associated relevance scores. SciPuRe and the relevance scores constitute information that can be used in advanced processing, such as the reconstitution of the n-Ary relations involving our instances. This latter point will be addressed in future work. Moreover, future studies are needed on other domains to assess the adaptability of the method. This would require OTR and annotated corpora in specialised experimental domains. Preliminary studies have already been conducted [47],[2].

Finally, there are prospects for increasing the relevance score application. Several textual segment levels (e.g. sections, subsections, table captions) could also be used or combined to extend the relevance scores. Inclusion of the frequency at which an entity related to a concept appears in a semantic score is also a standard addition to measure the specificity of an entity. More complex combinations including more than two scores and some ML techniques to adjust the parameters, also seem promising.

Acknowledgement

This project received funding from the IDEX/I-SITE MUSE [h] Univ. Montpellier (France).

We warmly thank all of our colleagues for their involvement and useful criticism. We are especially grateful to Julien Cufi (INRAE, IATE, Montpellier, France) for his technical assistance with the annotation server.

References

[1] Valérie Guillard, Patrice Buche, Sébastien Destercke, Nouredine Tamani, Madalina Croitoru, Luc Menut, Carole Guillaume, and Nathalie Gontard. A Decision Support System to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. *Computers and Electronics in Agriculture*, 111:131–139, February 2015.

[2] Charlotte Lousteau-Cazalet, Abdellatif Barakat, Jean-Pierre Belaud, Patrice Buche, Guillaume Busset, Brigitte Charnomordic, Stéphane Dervaux, Sébastien Destercke, Juliette Dibie, Caroline Sablayrolles, et al. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture*, 127:351–367, 2016.

[3] Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, 4(1):78, 2015.

[4] Marcia Lei Zeng. Smart data for digital humanities. *Journal of data and information science*, 2(1):1–12, 2017.

[5] Trong H Duong, Hong Q Nguyen, and Geun S Jo. Smart data: where the big data meets the semantics, 2017.

[6] Valentina Janev, Dea Pujić, Marko Jelić, and Maria-Esther Vidal. *Chapter 9 Survey on Big Data Applications*, pages 149–164. Springer International Publishing, Cham, 2020.

[7] Arthur Brack, Jennifer D'Souza, Anett Hoppe, Sören Auer, and Ralph Ewerth. Domain-independent extraction of scientific concepts from research articles. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 251–266, Cham, 2020. Springer International Publishing.

[8] Anne-Lyse Minard, Anne-Laure Ligozat, and Brigitte Grau. Extraction de résultats expérimentaux d'articles scientifiques pour le peuplement d'une base de données. In *10th International Conference on statistical analysis of textual data (JADT)*, volume 73, Roma, Italy, June 2010.

[9] Valérie Guillard, Patrice Buche, Luc Menut, and Stéphane Dervaux. Matter transfer ontology. `https://doi.org/10.15454/NK24ID`, 2018. Accessed: 2019-09-11.

[h] https://muse.edu.umontpellier.fr/en/muse-i-site/

[10] William S Cooper. A definition of relevance for information retrieval. *Information storage and retrieval*, 7(1):19–37, 1971.

[11] Stefano Mizzaro. How many relevances in information retrieval? *Interacting with computers*, 10(3):303–320, 1998.

[12] Jun Yan. *Text Representation*, pages 3069–3072. Springer US, Boston, MA, 2009.

[13] K Bretonnel Cohen, Helen L Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E Hunter. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC bioinformatics*, 11(1):492, 2010.

[14] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):55, 2015.

[15] M.J. McGill Gerard Salton. *Introduction to modern information retrieval*. McGraw-Hill, 1983.

[16] Robert Z Norman et al. Structural models: An introduction to the theory of directed graphs. 1965.

[17] Martin Lentschat, Patrice Buche, Juliette Dibie-Barthelemy, and Mathieu Roche. Scipure: a new representation of textual data for entity identification from scientific publications. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, pages 220–226, 2020.

[18] Erwin Marsi and Pinar Öztürk. Extraction and generalisation of variables from scientific publications. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 505–511, 2015.

[19] Miguel A Andrade and Peer Bork. Automated extraction of information in molecular biology. *FEBS letters*, 476(1-2):12–17, 2000.

[20] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[21] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[22] Behrang Mohit and Rebecca Hwa. Syntax-based semi-supervised named entity tagging. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 57–60, 2005.

[23] Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, 2005.

[24] Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444, 2017.

[25] Luke K. McDowell and Michael Cafarella. Ontology-driven information extraction with ontosyphon. In *The Semantic Web - ISWC 2006*, pages 428–444, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[26] Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie, and Mathieu Roche. Xart: Discovery of correlated arguments of n-ary relations in text. *Expert Systems with Applications*, 73:115–124, 2017.

[27] Bert R Boyce, Bert R Boyce, Charles T Meadow, Donald H Kraft, Donald H Kraft, and Charles T Meadow. *Text information retrieval systems*. Elsevier, 2017.

[28] Christian Jacquemin and Evelyne Tzoukermann. Nlp for term variant extraction: synergy between morphology, lexicon, and syntax. In *Natural language information retrieval*, pages 25–74. Springer, 1999.

[29] Didier Bourigault and Christian Jacquemin. Term extraction-i-term clustering: An integrated platform for computer-aided terminology. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, 1999.

[30] Jonathan D Wren and Harold R Garner. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of information in medicine*, 41(05):426–434, 2002.

[31] Jun Xu and Ya-Lou Huang. A machine learning approach to recognizing acronyms and their expansion. In *2005 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2313–2319. IEEE, 2005.

[32] Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, and Mathieu Roche. Identification des unités de mesure dans les textes scientifiques. In *Actes de la 22e conf\'erence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 88–94, 2015.

[33] Luca Foppiano, Laurent Romary, Masashi Ishii, and Mikiko Tanifuji. Automatic identification and normalisation of physical measurements in scientific literature. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–4, 2019.

[34] Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, MA, 2005.

[35] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Evaluation in information retrieval. *Introduction to information retrieval*, 1:188–210, 2008.

[36] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[37] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages, 2020.

[38] Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. Biomedical and clinical english model packages in the stanza python nlp library, 2020.

[39] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[40] Nick Craswell. *Precision at n*, pages 2127–2128. Springer US, Boston, MA, 2009.

[41] Deqing Wang and Hui Zhang. Inverse-category-frequency based supervised term weighting scheme for text categorization. *arXiv preprint arXiv:1012.2609*, 2010.

[42] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[43] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*, 2018.

[44] Richard Eckart de Castilho and all. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[45] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[46] Martin Lentschat, Patrice Buche, and Luc Menut. Transmat gold standard. 2020. CIRAD Dataverse. 10.18167/DVN1/U7HK8J.

[47] Valérie Guillard, Olivier Couvert, Valérie Stahl, Patrice Buche, Aurélie Hanin, Catherine Denis, Juliette Dibie-Barthelemy, Stéphane Dervaux, Catherine Loriot, Thierry Vincelot, et al. Map-opt: a software for supporting decision-making in the field of modified atmosphere packaging of fresh non respiring foods. *Packaging Research*, 2(1):28, 2017.

[48] Charlène Fabre, Patrice Buche, Xavier Rouau, and Claire Mayer-Laigle. Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. *Data in brief*, 33:106430, 2020.

[49] Mathieu Roche, Maguelonne Teisseire, and Gaurav Shrivastava. Valorcarn-tetis: Terms extracted with fastr (free extraction). 2020. CIRAD Dataverse. 10.18167/DVN1/FC2YXC.

[50] Silva Sonjak, Mia Ličen, Jens Christian Frisvad, and Nina Gunde-Cimerman. The mycobiota of three dry-cured meat products from slovenia. *Food Microbiology*, 28(3):373–376, 2011.