Actes de l'atelier Recherche d'Information SEmantique

Jean-Pierre Chevallet, Catherine Roussey, Haïfa Zargayouna

Neuvième Atelier Recherche d'Information SEmantique RISE, Caen 4 juillet 2017

Associé à IC@PFIA 2017

Actes de l'Atelier Recherche d'Information SEmantique RISE 2017

Édité par
Jean-Pierre CHEVALLET, LIG, Grenoble (France)
Catherine ROUSSEY, IRSTEA, Clermont Ferrand (France)
Haïfa ZARGAYOUNA, LIPN, Paris (France)


Atelier Recherche d'Information SEmantique RISE, Caen 04 juillet 2017

Associé à IC@PFIA 2017

1. Introduction

Nous avons le plaisir d'organiser à Caen la neuvième édition de l'atelier Recherche d'Information SEmantique, RISE 2017, associé à la Conférence Ingénierie des Connaissances de la Plateforme d'Intelligence Artificielle. L'atelier est soutenu par l'ARIA (Association Francophone de Recherche d'Information et Applications) et le collège Science de l'Ingénierie des Connaissances de l'AFIA (Association Francophone d'Intelligence Artificielle)

Le but de l'atelier est de proposer un espace d'échange autour de la synergie entre acquisition et gestion de ressources sémantiques (ontologies, terminologies, thesaurii, ...) et la Recherche d'Information. Ces thématiques sont à la croisée du Web Sémantique, de l'Ingénierie des Connaissances, du Traitement Automatique des Langues et de la Recherche d'Information.

Nous avons le plaisir cette année d'accueillir deux conférences invitées :

- Mathieu Lafourcade «10 ans de JeuxDeMots : un gros réseau lexico-sémantique obtenu par crowdsourcing ».
- Kata Gábor «Acquisition automatique de relations entre concepts dans le domaine scientifique»

Mathieu Lafourcade, actuellement maître de conférences en informatique à l'Université de Montpellier a soutenu une thèse en 1994 sur des questions de génie logiciel appliqué à la traduction automatique et au traitement du langage naturel (TALN). Ses activités de recherche, qui s'articulent autour de centres d'intérêt comme l'analyse sémantique du langage, l'intelligence artificielle (IA), l'acquisition automatique de ressources lexicales, l'a conduit à concevoir, puis à mettre en œuvre le projet JeuxDeMots, qui vise à créer un réseau lexico-sémantique du français. Ce réseau de grande taille (actuellement plus d'un million de termes et 90 millions de relations entre eux), en construction depuis 2007, et dont les données sont librement accessibles, est une structure sans équivalent au sein des ressources lexicales, que ce soit en français ou dans d'autres langues. Par ailleurs, la grande originalité de ce réseau est qu'il est en partie alimenté par le public via des GWAP (*Games With A Purpose*) et des contributions directes, et en partie grâce à des procédures automatisées comme l' extraction de relations sémantiques depuis des textes et les inférences. Mathieu Lafourcade enseigne le TALN et l'IA, mais également la compilation, l'algorithmique et la programmation.

Kata Gábor est chercheuse contractuelle au LIPN-CNRS & Université Paris 13 et chercheuse associée au LATTICE-CNRS. Ses sujets de recherche portent sur la sémantique distributionnelle et l'acquisition non supervisée de propriétés lexicales, avec une application à l'extraction d'informations. Ses activités actuelles visent l'induction automatique de patrons lexico-grammaticaux pour l'extraction des relations sémantiques afin d'améliorer l'accès au contenu textuel. Elle réalise ses travaux dans le cadre du projet LABEX-*Empirical Foundations of Linguistics.*
Précédemment, elle a été post-doctorante à l'équipe Alpage de l'INRIA, et boursière 'jeune chercheur' de l'Académie des Sciences Hongroise. Elle est co-créatrice du corpus annoté ACL-RelAcS, et co-organisatrice de la tâche d'évaluation *'Semantic Relation Extraction and Classification in Scientific Papers*' à SemEval 2018.

Pour la deuxième année consécutive, nous organisons une session entreprise. Deux entreprises sont invitées à présenter leurs avancées en acquisition de connaissance pour l'accès à l'information.

- Nicolas Delaforge (Co-fondateur & Gérant, Société Coopérative Mnémotix) «La terminologie structurée, élément structurant de l'activité de l'entreprise ? Ses atouts, ses inconvénients : exemple d'application dans une fondation d'art contemporain».
- Marc Dutoo (responsable de projet R&D, SmileLab, Smile) «Constitution d'un thésaurus pour la recherche de produits».

2. Comité de programme

- BERTIN Marc, STIH Paris (France), CIRST Montreal (Canada)
- BUSCALDI Davide, LIPN, Paris (France)
- CALABRETTO Sylvie, LIRIS Lyon (France)
- CHEVALLET Jean-Pierre, LIG, Grenoble (France)
- GRAU Brigitte, ENSIIE (France)
- KAMEL Mouna, IRIT Toulouse (France)
- ROUSSEY Catherine, IRSTEA, Clermont Ferrand (France)
- SALLABERRY Christian, LIUPPA, Pau (France)
- SCHWAB Didier , LIG-GETALP, Grenoble (France)
- TAMINE LECHANI Lynda, IRIT, Toulouse (France)
- ZARGAYOUNA Haïfa , LIPN, Paris (France)

3. Remerciements

Nous tenons à remercier les auteurs, les membres du comité de programme, les organisateurs de la plateforme PFIA ainsi que les participants qui assurent chaque année la bonne tenue de l'atelier.

4. Table des matières

10 ans de JeuxDeMots : un gros réseau lexico-sémantique obtenu par crowdsourcing

Mathieu Lafourcade

LIRMM, Université de Montpellier & CNRS
mathieu.afourcade@lirmm.fr

Résumé : Le projet JeuxDeMots a pour objet de construire un réseau lexical de sens commun (et de spécialité) en français à l'aide de jeux (gwaps - games with a purpose), d'approches contributives mais également de mécanismes d'inférences. Une dizaine de jeux ont été conçus dans le cadre du projet, chacun permettant de collecter des informations spécifiques ou encore de vérifier la qualité de données acquise via un autre jeu. Cet exposé s'attachera à décrire la nature des données que nous avons collectées et construites depuis le lancement du projet durant l'été 2007.

Nous décrirons en particulier les aspects suivant : la structure de réseau lexical obtenu, les types de relations sémantiques représentées (ontologiques, subjectives, rôles sémantiques, associations d'idées), les questions liées à l'activation et l'inhibition de termes et relations, l'annotation de relations (méta-informations), les raffinements sémantiques (gestion de la polysémie), la création d'agglomérations permettant la représentation de connaissances plus riches.

Ce réseau lexical, distribué sous licence libre, est exploité dans de nombreux laboratoires de recherche et entreprises. Les applications en cours utilisant le réseau JeuxDeMots concernent principalement l'interprétation sémantique de textes, la compréhension de l'écrit, la recherche d'information, l'inférence de faits, l'analyse d'opinions et de sentiments - et ce dans des domaines comme la radiologie, le tourisme, la nutrition. Construit à partir d'une liste de 150 000 termes sans aucune relation entre eux, le réseau lexical de JeuxDeMots contient maintenant plus de 1000 000 termes et plus de 80 millions de relations.

Acquisition automatique de relations entre concepts dans le domaine scientifique

Kata Gábor
LIPN, UMR 7030, Université Paris 13
Kata.gabor@lipn.univ-paris13.fr

Résumé : De nos jours, la production d'articles scientifiques croît à un rythme accéléré. Cette explosion d'information rend le travail des chercheurs, des experts et des relecteurs de plus en plus difficile et nécessite de nouvelles méthodes pour la compréhension, l'extraction et la structuration automatique de l'information dans les textes de spécialité. Comme la disponibilité et la couverture des bases de connaissances existantes est souvent insuffisante, nous proposons de prendre comme point de départ l'analyse sémantique du contenu afin de faire émerger un modèle de connaissances. Nous présentons deux approches non supervisées pour l'acquisition des relations sémantiques dans un corpus de spécialité. L'identification des relations ne nécessite pas des données d'apprentissage annotées et bien qu'elle soit spécifiquement dédiée à la littérature scientifique, elle reste applicable sur n'importe quel domaine pour lequel une telle littérature existe.

La présentation explorera les problématiques spécifiques à la tâche non supervisée. Deux approches complémentaires seront distinguées et explorées. La première se concentre principalement sur les relations lexicales, qui se caractérisent par une sélection sémantique des arguments, et qui ne dépendent pas du contexte. Cette approche est basée sur la représentation du sens des mots individuels par des vecteurs distributionnels (word embeddings). Les vecteurs sont créés à partir de corpus et combinés pour représenter le sens et la relation sémantique du couple d'entités. Nous proposons une nouvelle méthode de combinaison de vecteurs distributionnels qui permet de mieux estimer la similarité relationnelle entre deux couples d'entités. L'avantage de cette méthode est de pouvoir s'appliquer à des couples d'entités qui ont peu de co-occurrences dans le corpus. La deuxième approche, à son tour, s'applique aux relations contextuelles et s'appuie sur les contextes de co-occurrence des entités. Les couples d'entités sont caractérisés par leurs co-occurrences avec des motifs spécifiques à la relation, qui sont extraits automatiquement à partir du corpus. Nous montrons que cette approche peut bénéficier de la fouille de motifs séquentiels, qui crée un espace vectoriel plus adapté (moins creux) pour un clustering non supervisé.

Enhancing Translation Language Models with Word Embedding for Information Retrieval

Jibril Frej, Jean-Pierre Chevallet, Didier Schwab

LABORATOIRE D'INFORMATIQUE DE GRENOBLE (LIG)
UNIV. GRENOBLE ALPES (UGA)
jibril.frej@etu.univ-grenoble-alpes.fr
jean-pierre.chevallet@imag.fr
didier.schwab@imag.fr

Abstract : In this paper, we explore the usage of Word Embedding semantic resources for Information Retrieval (IR) task. This embedding, produced by a shallow neural network, have been shown to catch semantic similarities between words (Mikolov *et al.*, 2013). Hence, our goal is to enhance IR Language Models by addressing the term mismatch problem. To do so, we applied the model presented in the paper *Integrating and Evaluating Neural Word Embedding in Information Retrieval* by Zuccon *et al.* (2015) that proposes to estimate the translation probability of a Translation Language Model using the cosine similarity between Word Embedding. The results we obtained so far did not show a statistically significant improvement compared to classical Language Model.

Keywords— Information Retrieval, Language Model , Word Embedding

1 Introduction

Information Retrieval Systems (IRS) are computer assistants that help to retrieve digital documents, in which user is supposed to found relevant information for his task. Hence an IRS *is* about semantics, because user information needs is topically related and serve to help to accomplish user's task.

Curiously, most commercial and experimental search engine do not handle any sort of semantics nor knowledge to solve queries. This is because matching computations are mostly based on statistical word distributions, and intersection between query and documents. Also, most IR models see query and document as simple bag of word. Though, these systems provide satisfaction to their user, as long as statistics are possible, i.e. documents are long enough and queries are expressed using the same vocabulary as documents.

When we are not in this situation, that is, if documents to be retrieved are very short and/or there is a strong discrepancy between document vocabulary and user's one, IRS are facing the problem of *term mismatch*. For example, the collection Europeanna[1] gives access to millions of digital objects from cultural heritage, by a very small textual meta-data description. Even, descriptions are made by specialists that are prone to use technical terms.

Hence, in this paper we propose to study the effect of exploiting semantic resources to reduce term mismatch negative effect on collection with specialized vocabularies. We focus on automatically constructed resources, namely word Embedding resources, first because they cover a very large vocabulary, and second, because they seem to capture interesting word semantics (Mikolov *et al.*, 2013).

[1] http://www.europeana.eu

A very simple way to exploit a resource to solve term mismatch, is to expand queries or documents with words that are semantically similar according to the resource. This approach has been heavily studied (Carpineto & Romano, 2012) and suffers from a definitive problem: how to control query meaning shift when a-priory the query is statically modified, i.e. for all documents ?

In this paper, we propose not to change the query nor the document but to adapt the matching function. This adaptation requires to change the formula that computes the Relevant Status Value (RSV). This formula depends on the Information Retrieval (IR) model.

There are several large categories of IR model: Vector Space, Logical Based, Probabilist, Graph Based, and Language Models (LM). Among these models, only the Probabilist models like the famous BM25 formula of Robertson (Robertson *et al.*, 1995), or Language Models provide state of the art results. Some recent proposals of IR Graph Modeling (Bannour *et al.*, 2016) have the advantage to fuse one single model document index and knowledge base, and exploit only one matching function: activation propagation from query term to document through indexing terms and knowledge concept nodes. Beside the nice property of this model to exploit a Knowledge Base at query time, without query or document expansion, experiments of this model are still bellow the state of the art of Probabilistic or Language models. For this reason, we have decided to work on the transformation of a Language Model formula. We did not chose a Probabilistic model because formulas of the LM model are much simpler for similar results (Manning *et al.*, 2008), hence they are easier to transform.

A simple way to include a Knowledge Resource in a LM matching function, is Translation Language Models (Berger & Lafferty, 1999) where the translation probability between query and document terms is taken into account, in addition to the exact term matching. Probability are estimated using the mutual information between the two terms and is the highest if the considered terms have the same distribution over the collection. This computation can be considered as an automatic basic knowledge resource extracted from the corpus itself. Recently, it has been proposed to estimate the translation probability using Neural Word Embedding (Zuccon *et al.*, 2015).

Word Embedding denotes a set of methods to produce Knowledge resources with vector representation of words[2] that express some semantics learned from word usage in very large text collection. Usually these methods are based on the distributional hypothesis (Harris, 1954): *words that occur in the same contexts tend to have similar meaning*. The vector representation of words are computed using their context so that words with similar meaning will have similar vector representation. Word Embedding includes dimension reduction techniques on the word co-occurrence matrix (Latent Semantic Analysis (Deerwester *et al.*, 1990)), probabilistic models (Latent Dirichlet Allocation (Blei *et al.*, 2003)) and more recently shallow neural network-based methods such as the skip-gram model that can also learn phrases vector representations and is very effective on word similarity and word analogy tasks (Mikolov *et al.*, 2013). As we said before, these vector representations can be used to capture semantic relationships between words by measuring the similarity between the vectors, with the cosine similarity measure for example.

Such vector representation could help us with the vocabulary mismatch problem since they

[2]Most of the time the vectors are real-valued.

provide us a new way to estimate the translation probability between query and document terms by considering their semantic similarity. The empirical results show that improving Dirichlet Language Model using Word Embedding is possible.

In the rest of this paper, we first recall LM formulas in section 2, present the extension with the translation model in section 3, then the usage of Word Embedding within this model in section 4. We present the implementation of this model in section 5 and results on the Cultural Heritage in CLEF (CHiC) Dataset[3], a sub part of the Europeanna, in section 6.

2 Language models

In this part, we recall basics on the Language Model in order to detail our modifications of the formula in the next section. The query likelihood Language Model aims to rank documents d by computing the probability of a document interpreted as the likelihood that it is relevant to the query q. Hence, the Relevance Status Value (RSV) of a Language Model based IR is expressed by:

$$RSV(q,d) = p(d|q) \tag{1}$$

Using Bayes rule, we have:

$$RSV(q,d) = \frac{p(q|d)p(d)}{p(q)} \tag{2}$$

We can ignore the constant $p(q)$ to rank documents and we also consider that $p(d)$ is uniform over all the documents of the collection and can also be ignored. Therefore, to rank documents with respect to a given query, we only have to compute the probability of having q, knowing d. In LM model, document d is replaced by its language model, i.e. the probability distribution of terms of the vocabulary in d denoted θ_d. So the RSV is:

$$RSV(q,d) \simeq_{rank} p(q|\theta_d) \tag{3}$$

In this work, we chose the multinomial event model for θ_d, and the unigram language model[4], so the probability of a given word does not depend on its context:

$$RSV(q,d) \simeq_{rank} log\left(p(q|\theta_d)\right) = \sum_{i=1}^{|q|} log\left(p(q_i|\theta_d)\right) \tag{4}$$

With q_i the i^{th} term of the query and $|q|$ the query size. To estimate $p(q_i|\theta_d)$, Dirichlet Language Model is used since it has been shown to work better with Translation Language Models (Karimzadehgan & Zhai, 2010) :

$$p(q_i|\theta_d) = \frac{|d|}{\mu + |d|} p_{ml}(q_i|\theta_d) + \frac{\mu}{\mu + |d|} p(q_i|C) \tag{5}$$

[3]http://ims.dei.unipd.it/data/chic/

[4]Usually in IR the unigram language model give the best results with the lowest computational cost (Manning *et al.*, 2008)

$\mu \in \mathbb{R}^+$ is the smoothing parameter and $p_{ml}(q_i|\theta_d)$ is estimated using the maximum likelihood [5] and is equal to:

$$p_{ml}(q_i|\theta_d) = \frac{c(q_i, d)}{|d|} \tag{6}$$

With $c(q_i, d)$ the frequency of q_i in d and $|d|$ the document size. The same goes for $p(q_i|C)$ the smoothing term which is also estimated with the maximum likelihood : $p(q_i|C) = c(q_i, C)/|C|$. This estimation of $p(q_i|\theta_d)$ leads to the following ranking formula (see *the appendix* for more details) :

$$RSV(q, d) \simeq_{rank} \sum_{i:c(q_i,d)>0} \left[log\left(1 + \frac{c(q_i, d)}{\mu p(q_i|C)}\right)\right] + |q| log\left(\frac{\mu}{\mu + |d|}\right) \tag{7}$$

We decided to present equations (5) and (7) even though they are equivalent for ranking documents because both of them can be found to describe Dirichlet Language Model. In the frame of our work, and for reasons that are developed in *the appendix*, we will introduce the next models by giving the expression of $p(q_i|\theta_d)$ as in equation (5).

As we said previously, one issue of these models is the term mismatch problem: as we can see on equation (7) ranking takes into account only the terms that appear in both the considered document and query. Consequently, relevant documents that do not contain the exact query terms will not be considered. One approach to solve this problem is to adapt the language model to take into account the semantic similarities between terms.

3 Translation Language Models

Translation Language Models (TLM) try to estimate the semantic similarity between two terms by using tools from statistical translation. The main idea is to estimate the likelihood of translating a document to a query using the translation probability between terms (Karimzadehgan & Zhai, 2010). To do so, the maximum likelihood estimator in the Dirichlet language model $p_{ml}(q_i|\theta_d)$ is replaced with the likelihood that the query has been produced by a translation of the document $p_t(q_i|\theta_d)$:

$$p(q_i|\theta_d) = \frac{|d|}{\mu + |d|} p_t(q_i|\theta_d) + \frac{\mu}{\mu + |d|} p(q_i|C) \tag{8}$$

$p_t(q_i|\theta_d)$ is calculated the following way :

$$p_t(q_i|\theta_d) = \sum_{u \in d} p_t(q_i|u) p_{ml}(u|\theta_d) \tag{9}$$

With $p_t(w|u)$ the probability to translate term u into term w which is estimated using mutual information between u and w (Karimzadehgan & Zhai, 2010) :

$$p_t(w|u) = \frac{I(w, u)}{\sum\limits_{w' \in V} I(w', u)} \tag{10}$$

[5] We consider that terms in a document follow a multinomial distribution

$I(w,u)$ is the mutual information score between word u and w, defined as follow :

$$I(w,u) = \sum_{X_w=0,1} \sum_{X_u=0,1} p(X_w, X_u) log \left(\frac{p(X_w, X_u)}{p(X_w)p(X_u)} \right) \tag{11}$$

With X_w and X_u binary random variables indicating if a word is absent or present (refer to (Karimzadehgan & Zhai, 2010) for more details).

4 Word Embedding-based Translation Language Model

Within the frame of this work, word Embedding are used instead of mutual information in order to estimate the translation probability $p_t(q_i|u)$. The model is named Word Embedding-based Translation Language Model (WETLM). We consider the new estimation of the translation probability, denoted $p_{cos}(q_i|u)$, to be proportional to the similarity between q_i and u that is measured with the cosine between the vectors of the two terms :

$$p_{cos}(q_i|u) = \frac{cos(q_i, u)}{\sum\limits_{u' \in V} cos(u', u)} \tag{12}$$

Consequently the ranking formula becomes :

$$p(q_i|\theta_d) = \frac{|d|}{\mu + |d|} p_{cos}(q_i|\theta_d) + \frac{\mu}{\mu + |d|} p(q_i|C) \tag{13}$$

With $p_{cos}(q_i|\theta_d) = \sum_{u \in d} p_{cos}(q_i|u) p_{ml}(u|\theta_d)$. Both the estimation of the translation probability p_t and p_{cos} underestimate the self-translation probability: we can have $p_{cos}(u|w) > p_{cos}(u|u)$ for $u \neq w$ which is not desirable for a translation language model (Karimzadehgan & Zhai, 2012).

One way to make sure that the self-translation probability is the highest for a given term is to redefine it by introducing a hyper-parameter $\alpha \in [0, 1]$ that "controls" the self-translation probability (Karimzadehgan & Zhai, 2010):

$$p_{cos-\alpha}(w|u) = \begin{cases} \alpha + (1-\alpha) p_{cos}(w|u) & \text{if } u = w \\ (1-\alpha) p_{cos}(w|u) & \text{if } u \neq w \end{cases} \tag{14}$$

This formula ensures the fact that we have $p_{cos-\alpha}(u|u) > p_{cos-\alpha}(u|w) \forall u, w \in V$ for $\alpha > 0.5$. The model that uses $p_{cos-\alpha}$ to estimate the translation probability will be referred as WETLM-α. We set the value of α to 0.45 since it is the one that produced the best results for the Threshold T = 0.7. This approach was not developed in the paper *Integrating and Evaluating Neural Word Embedding in Information Retrieval* by Zuccon *et al.* (2015) since they reported that the word Embedding they used did not underestimate the self-translation probability.

5 Implementation and data

For this work, instead of using an already existing Information Retrieval System (IRS) such as *Terrier*, we developed our own IRS in C++ to easily add word Embedding to the classical mod-

els and also because having a low retrieval time[6] is not an objective: before trying to compute a fast IRS, we should make sure that the Word Embedding-based language models outperforms state of the art Language models. In order for our results to be comparable with other work that used Terrier, we made sure that with the same pre-processing on the corpus, we obtain the same results as Terrier (see *the appendix* for more details about our IRS). When doing so, we noticed that Terrier does not implement Dirichlet language model using equation (7) (more details in *the appendix*).

During pre-processing, we used a Stop List and replaced capital letters with lower case letters on both the collection and the queries. In order to have the same results as Terrier with our IRS, we also removed characters that were not digits or letters and deleted words that contained more that 4 digits or more than 3 consecutive identical characters. We did not use any stemmer on our collection since the best results were obtained without any stemming.

We used the Cultural Heritage in CLEF (CHiC) 2012 English collection for ad hoc retrieval. This collection is composed of 1 107 176 documents containing "metadata records describing digital representations of cultural heritage objects" (Petras *et al.*, 2012) and 50 queries for *ad hoc* retrieval tasks. Below is a table summing up some statistics of the collection:

#d	Avdl	Vocabulary Size	#q	Avql
1 107 176	30.92	290 265	50	1.84

Table 1: CHiC 2012 statistics

With $\#d$ and $\#q$ being respectively the number of documents in the collection and the number of queries. Avdl is the average document length and Avql is the average query length. To evaluate the models, the top 1000 documents were returned for each query, the MAP and P@10 were computed using the standard tool for evaluating an *ad hoc* retrieval: *trec_eval*.

We used the word2vec-GoogleNews-vectors word Embedding of dimension 300, pre-trained on the google news Corpus (3 billion words) that are available *here* [7]. As we said earlier, the CHiC collection we used have a very specific vocabulary and even if the word Embedding we used were trained on a 3 billion words corpus, a lot of word of the vocabulary were missing :

- only 42.68% of the words of the vocabulary have an Embedding;
- but 91.92% of word occurrences in the collection have an Embedding.

On the other hand, most of the queries's terms had an Embedding :

- 94.95% of the queries terms have an Embedding;
- 2% of the queries have none of their term that posses an Embedding.

Finally, the translation probability described in equation (12) is not the one that was implemented: we computed the cosine similarity between two words if it was above a given threshold

[6] Which is the case with Terrier

[7] `https://github.com/mmihaltz/word2vec-GoogleNews-vectors`

T. This allowed us to reduce the number of cosine similarities to compute and also it acts like a noise reducer since we did not take into account the similarity between non similar terms. We found that **T** = 0.7 produces the highest MAP on the CHiC collection using word2vec-GoogleNews-vectors.

6 Results

Instead of giving the value of the parameter μ that produces the optimal results, we decided to display the MAP for a range of values of μ to see if the WETLM outperforms (or not) the Dirichlet LM consistently or if for some values of μ one model performs better than the other.

At first we evaluated the Dirichlet LM, the optimal value we found for μ was 44. Retrieved documents are evaluated using the Mean Average Precision and P@10:

μ	12	16	20	24	28	32	36	40	44	48
MAP (%)	35.61	36.09	36.16	36.24	36.30	36.36	36.39	36.23	36.43	36.09
μ	52	56	60	64	68	72	76	80	84	88
MAP (%)	36.05	36.06	35.82	35.86	35.92	35.92	36.04	35.77	35.68	35.68

Table 2: Values of the MAP on the CHIC2012 collection using the Dirichlet Language Model

μ	12	16	20	24	28	32	36	40	44	48
P@10 (%)	33.54	33.75	33.75	33.96	34.17	34.38	34.17	34.17	34.38	34.38
μ	52	56	60	64	68	72	76	80	84	88
P@10 (%)	34.38	34.17	34.38	34.38	34.38	34.38	34.38	34.38	34.38	34.58

Table 3: Values of the P@10 on the CHIC2012 collection using the Dirichlet Language Model

We checked that the results obtained are identical to the ones produced by Terrier with the same pre-processing on the collection. Also we decided to explore values of μ that are close to the Average Document Length (Avdl) since the optimal value of μ in the Dirichlet Language Model is usually around the Avdl. Table 4 below represents the results obtained with the WE-based Translation Language Model :

μ	12	16	20	24	28	32	36	40	44	48
MAP (%)	36.89	37.71	37.76	37.86	37.78	37.79	37.81	37.65	37.67	37.29
μ	52	56	60	64	68	72	76	80	84	88
MAP (%)	37.17	36.87	36.66	36.69	36.57	36.56	36.55	36.50	36.37	36.35

Table 4: Values of the MAP on the CHIC2012 collection using WETLM

μ	12	16	20	24	28	32	36	40	44	48
P@10 (%)	34.38	34.38	35.21	35.42	35.63	35.63	35.42	35.42	35.42	35.63
μ	52	56	60	64	68	72	76	80	84	88
P@10 (%)	35.42	35.21	35.21	35.21	35.21	35.21	35.21	35.21	35.21	35.21

Table 5: Values of the P@10 on the CHIC2012 collection using WETLM

As we can see the WE-based Translation Language Model seems to sightly outperform the Dirichlet Language model for every μ. The optimal value of μ we found is different for the two models : $\mu_{opt} = 44$ for the Dirichlet LM and $\mu_{opt} = 24$ for the WETLM. We performed a paired t-test over the average precision of each query for μ = 36 for both models, the measured p-value with R is $0.1733 > 0.01$. Unfortunately the improvement is not statistically significant.

In the table below we show some of the results collections presented in the work of (Zuccon *et al.*, 2015) over the 3 collections AP88-89 , WSJ87-92 and DOTGOV :

Method	AP88-89		WSJ87-92		DOTGOV	
	MAP	P@10	MAP	P@10	MAP	P@10
Dirichlet LM	22.69	39.60	21.71	40.80	18.73	24.60
WETLM	24.27*	41.00	22.66*	42.40*	19.32	25.00

Table 6: Values of the MAP and P@10 reported by (Zuccon *et al.*, 2015) on the collections AP88-89 , WSJ87-92 and DOTGOV using Dirichlet LM and WETLM. The statistically significant differences are indicated by *.

As we can see, according to (Zuccon *et al.*, 2015), depending on the collection, the WETLM can produce results that exhibit a statistically significant improvement of the MAP compared to Dirichlet LM.

Table 7 below represents the results obtained with the WE-based Translation Language Model that "controls" the self translation probability with the parameter α:

μ	12	16	20	24	28	32	36	40	44	48
MAP (%)	37.35	37.92	38.07	38.15	38.28	38.31	38.35	38.19	38.17	37.81
μ	52	56	60	64	68	72	76	80	84	88
MAP (%)	37.72	37.73	37.53	37.58	37.59	37.47	37.40	37.10	37.06	37.05

Table 7: Values of the MAP on the CHIC2012 collection using WETLM-α

μ	12	16	20	24	28	32	36	40	44	48
P@10 (%)	34.79	34.79	35.21	36.04	36.25	36.46	36.46	36.46	36.25	36.67
μ	52	56	60	64	68	72	76	80	84	88
P@10 (%)	36.25	36.25	36.25	36.25	36.25	36.25	36.25	36.25	36.25	36.25

Table 8: Values of the P@10 on the CHIC2012 collection using WETLM-α

We performed a paired t-test over the average precision of each query for $\mu = 36$ to compare LM and WETLM-α models : the measured p-value with R is $0.01219 > 0.01$: the improvement is still not statistically significant.

7 Conclusion and future work

The results we obtained are consistent with the ones in (Zuccon *et al.*, 2015) since they observed the same improvement as we did in the MAP. They also checked that their improvements were independent of the corpus and the training set for the Word Embedding: they do not need to be trained on the same corpus used in retrieval. For now our results are limited to one corpus and one set of word embedding, one of our objective in the near future is to perform the experiments on different corpora and also to improve our model by considering the context of the terms of the query by using phrase vectors (Mikolov *et al.*, 2013) to replace the query or to perform query expansion and by modifying the translation probability so that it satisfies a set of constraints (Karimzadehgan & Zhai, 2012).

References

BANNOUR I., ZARGAYOUNA H. & NAZARENKO A. (2016). Modèle unifié pour la recherche d'information sémantique. In N. PERNELLE, Ed., *IC 2016 : 27es Journées francophones d'Ingénierie des Connaissances (Proceedings of the 27th French Knowledge Engineering Conference), Montpellier, France, June 6-10, 2016.*, p. 155–160. 2

BERGER A. & LAFFERTY J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, p. 222–229, New York, NY, USA: ACM. 2

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022. 2

CARPINETO C. & ROMANO G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, **44**(1), 1:1–1:50. 2

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, **41**(6), 391–407. 2

HARRIS Z. (1954). Distributional structure. *Word*, **10**(23), 146–162. 2

KARIMZADEHGAN M. & ZHAI C. (2010). Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, p. 323–330, New York, NY, USA: ACM. 3, 4, 5

KARIMZADEHGAN M. & ZHAI C. (2012). Axiomatic analysis of translation language model for information retrieval. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, p. 268–280, Berlin, Heidelberg: Springer-Verlag. 5, 9

MANNING C. D., RAGHAVAN P. & SCHÜTZE H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press. 2, 3

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv e-prints*. 2, 9

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, p. 3111–3119, USA: Curran Associates Inc. 1

PETRAS V., FERRO N., GÄDE M., ISAAC A., KLEINEBERG M., MASIERO I., NICCHIO M. & STILLER J. (2012). Cultural heritage in clef (chic) overview 2012. 6

ROBERTSON S. E., WALKER S., JONES S., HANCOCK-BEAULIEU M. M., GATFORD M. *et al.* (1995). Okapi at trec-3. *Nist Special Publication Sp*, **109**, 109. 2

ZUCCON G., KOOPMAN B., BRUZA P. & AZZOPARDI L. (2015). Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, p. 12:1–12:8, New York, NY, USA: ACM. 1, 2, 5, 8, 9

8 Appendix

8.1 From equation (5) to equation (7) :

$$
\begin{aligned}
& log\left(p(q|\theta_d)\right) = \sum_{i=1}^{|q|} log\left(p(q_i|\theta_d)\right) \\
&= \sum_{i=1}^{|q|} log\left(\frac{|d|}{\mu+|d|}p_{ml}(q_i|\theta_d) + \frac{\mu}{\mu+|d|}p(q_i|C)\right) \\
&= \sum_{i=1}^{|q|} log\left(\frac{c(q_i,d)+\mu p(q_i|C)}{|d|+\mu}\right) \\
&= \sum_{i:c(q_i,d)>0} log\left(\frac{c(q_i,\theta_d)+\mu p(q_i|C)}{|d|+\mu}\right) + \sum_{i:c(q_i,d)=0} log\left(\frac{\mu p(q_i|C)}{|d|+\mu}\right) \\
&= \sum_{i:c(q_i,d)>0} log\left(\frac{c(q_i,d)+\mu p(q_i|C)}{|d|+\mu}\right) - \sum_{i:c(q_i,d)>0} log\left(\frac{\mu p(q_i|C)}{|d|+\mu}\right) + \sum_{i=1}^{|q|} log\left(\frac{\mu p(q_i|C)}{|d|+\mu}\right) \\
&= \sum_{i:c(q_i,d)>0} log\left(\frac{c(q_i,d)+\mu p(q_i|C)}{|d|+\mu} \times \frac{|d|+\mu}{\mu p(q_i|C)}\right) + \sum_{i=1}^{|q|} log\left(\frac{\mu p(q_i|C)}{|d|+\mu}\right) \\
&= \sum_{i:c(q_i,d)>0} log\left(1+\frac{c(q_i,d)}{\mu p(q_i|C)}\right) + |q| log\left(\frac{\mu}{|d|+\mu}\right) + \sum_{i=1}^{|q|} log\left(p(q_i|C)\right)
\end{aligned}
$$

The last term of the equation above depends only on the query and collection, therefore it can be ignored to rank documents, which leads to :

$$
log\left(p(q|\theta_d)\right) = \sum_{i:c(q_i,d)>0} \left[log\left(1+\frac{c(q_i,d)}{\mu p(q_i|C)}\right)\right] + |q| log\left(\frac{\mu}{\mu+|d|}\right)
$$

8.2 Why equation (5) instead of equation (7) :

As we said previously, usually language model RSV functions are presented using a sum that goes through terms that appear in both the document and the query, as in equation (7). The reason is that this formulation is more compatible with the usage of an inverted index. In fact an inverted index just associates to a term, all documents with non null term frequency: this is called the posting list. When using inverted index, the matching computation have to compute partial RSV sum for all documents found in posting lists, in parallel. If the formula only requires terms to have non null frequency in documents as in formula (7) the matching algorithm is trivial. If the formula need all terms of the query, as in formula (5) then some partial matching sums has to be corrected during matching computation and the matching algorithm is much more complex and less efficient.

In our experimental system, we do not use any inverted index: we store the all index in a direct structure in main memory. Hence, matching is done by simply browsing all documents

and computing RSV strait using the formula.

The models we presented that used statistical translation and word Embedding cannot be written with only a sum over $q \cap d$ as in equation (7) : this is why we used equation (5) to give the expression of the rank. Hence the implementation using a inverted fil is more problematic, and for example, cannot be done in a simple way in Terrier. That is the reason why we could not use Terrier for these experiments.

8.3 Terrier's Dirichlet Language Model :

Moreover during the experiment, when we tried to obtain the same results as Terrier, we noticed that they did not implement exactly equation (5) to compute the score of each document. Below is the formula they used to compute the score of documents :

$$log\left(p(q|\theta_d)\right) = \sum_{i:c(q_i,d)>0} \left[log\left(1 + \frac{c(q_i,d)}{\mu p(q_i|C)}\right) + log\left(\frac{\mu}{\mu + |d|}\right)\right] \tag{15}$$

Compared to equation (5) where the quantity $log\left(\frac{\mu}{\mu+|d|}\right)$ is computed |q| times, this formula computes it only $|q \cap d|$ times. Since $log\left(\frac{\mu}{\mu+|d|}\right) < 0$, equation(15) gives a little bit more importance to query terms that do not appear in documents. For our experiment, the Dirichlet language model was implemented using equation (5) but we also checked that we obtained the same results as Terrier when computing equation (15).

8.4 Some details about the implemented Information Retrieval System

In order to have the same results as Terrier, our IRS performed the following pre-processing on the collection :

- We replaced with a space all the characters that were not integers or letters, for example the term *pre-processing* is broken down into the two terms *pre* and *processing*.
- We also replaced majuscule letters with their minuscule equivalent
- We deleted terms that contained more than 4 digits
- We deleted terms that had more than 3 consecutive identical characters

Also, when we implemented equation (13) (the ranking formula of the WETLM), we decided to remove the normalization terms when $p_{cos}(q_i, \theta_d)$ or $p(q_i|C)$ were equal to 0 :

$$p(q_i|\theta_d) = \begin{cases} \frac{|d|}{\mu+|d|} p_{cos}(q_i|\theta_d) + \frac{\mu}{\mu+|d|} p(q_i|C) & \text{if } p_{cos}(q_i|\theta_d) \neq 0 \text{ and } p(q_i|C) \neq 0 \\ p_{cos}(q_i|\theta_d) & \text{if } p_{cos}(q_i|\theta_d) \neq 0 \text{ and } p(q_i|C) = 0 \\ p(q_i|C) & \text{if } p_{cos}(q_i|\theta_d) = 0 \text{ and } p(q_i|C) \neq 0 \end{cases} \tag{16}$$

We decided to do so in order to avoid underweighting the translation probability of a query term that is not in the collection and conversely to avoid underweighting the smoothing term of associated to a word that is in the collection but does not have any similar terms in the considered document.

Améliorer la qualité d'un thésaurus à l'aide de requêtes SPARQL

Catherine Roussey1, Stephan Bernard1

[1] TSCF, Irstea centre de Clermont-Ferrand,
{prénom.nom}@irstea.fr

Résumé : SKOS est un schéma RDF utilisé pour stocker et publier sur le web de données liées des thésaurus ou des taxonomies. Lors du développement d'un thésaurus, il est utile de détecter automatiquement les concepts qui violent les contraintes d'intégrité. Cet article présente le cas d'usage d'un développement manuel d'un thésaurus monolingue. Ce thésaurus porte sur l'usage des cultures en France. Dans cet article, nous avons spécifié diverses contraintes d'intégrité. Pour vérifier ces contraintes un ensemble de requêtes SPARQL a été défini. Ces requêtes permettent à la fois de vérifier des contraintes d'intégrité propres à tous les thésaurus stockés au format SKOS, mais aussi des contraintes propres à notre thésaurus.

Mots-clés : SKOS, requête SPARQL, contrainte d'intégrité, développement de thésaurus, thésaurus agricole.

1 Introduction

SKOS ou Simple Knowledge Organization System (Miles & Bechhofer , 2009) est un schéma RDF utilisé pour publier sur le web de données liées des thésaurus, des taxonomies, des vocabulaires contrôlés. Ainsi les modèles SKOS peuvent être réutilisés pour indexer tout type de sources d'information. Ces modèles deviennent une ressource centrale pour les systèmes de recherche d'information dédiés. Cependant la qualité de ces modèles va impacter directement les résultats des systèmes de recherche d'information qui les utilisent. Comme le montre le Tableau 1, les spécifications SKOS définissent des contraintes d'intégrité que tout modèle SKOS doit vérifier pour être considéré comme cohérent (Miles & Bechhofer , 2009). Les spécifications SKOS définissent les contraintes minimales que doivent remplir les modèles SKOS. En fonction du domaine et de l'usage du modèle SKOS, de nouvelles contraintes peuvent être définies. Par exemple le tableau 2 présente les conventions d'usage proposées dans les spécifications SKOS. Les développeurs de thésaurus peuvent décider ou non de suivre ces conventions.

TABLE 1 – *Les contraintes d'intégrité des modèles SKOS issues de (Miles & Bechhofer , 2009).*

S9	la classe skos:ConceptScheme est disjointe de la classe skos:Concept.
S13	skos:prefLabel, skos:altLabel et skos:hiddenLabel sont deux à deux des propriétés disjointes.
S14	Une ressource ne peut pas avoir plus d'un skos:prefLabel par langue.
S27	la propriété skos:related est disjointe de la propriété skos:broaderTransitive.
S37	la classe skos:Collection est disjointe des classes skos:Concept et skos:ConceptScheme.
S46	la propriété skos:exactMatch est disjointe des propriétés skos:broadMatch et skos:relatedMatch.
	La propriété skos:memberList est une propriété fonctionnelle, c'est-à-dire

	qu'elle ne peut pas avoir plus d'une valeur. Cette contrainte essaye d'exprimer qu'il n'y a aucun sens pour une skos:OrderedCollection d'avoir deux fois le même élément. Cette contrainte n'est pas spécifiée totalement par la propriété fonctionnelle.

TABLE 2 – *Les conventions d'usage des modèles SKOS issues de (Miles & Bechhofer , 2009).*

Aucune contrainte n'est spécifiée sur l'existence de top concept appartenant à un modèle. Une convention d'usage veut que les top concepts définis dans un modèle donné n'aient pas de concept parent appartenant à ce modèle.
Une convention d'usage veut que l'existence des cycles et de chemins alternatifs dans le graphe hiérarchique des concepts soit évitée.

Lors du développement d'un modèle SKOS, il est utile de détecter automatiquement les concepts SKOS qui violent les contraintes d'intégrité. Cet article présente le cas d'usage d'un développement manuel d'un thésaurus monolingue. Ce thésaurus porte sur l'usage des cultures en France. Dans ce cas d'usage, nous avons défini un ensemble de contraintes d'intégrité que doivent vérifier les concepts de ce thésaurus. Nous avons formalisé ces contraintes sous forme de requêtes SPARQL pour détecter les concepts qui violent ces contraintes.

2 Contexte

Les thésaurus agricoles sont utilisés principalement pour indexer des ressources produites ou utilisées par de grandes institutions travaillant dans le domaine de l'agriculture (INRA, Irstea, FAO). La FAO en particulier développe son thésaurus multilingue Agrovoc. Ce thésaurus, mondialement connu, est publié sur le web de données liées au format SKOS. Ce thésaurus couvre un domaine large et répond à des contraintes internationales. Les spécificités nationales n'ont pas vocation à être décrites dans ce thésaurus.

Pour les besoins d'indexation d'un corpus d'alerte agricole, nous recherchons un thésaurus capable de lister les cultures françaises mais aussi d'indiquer une parentalité d'usage dans ces cultures. À notre connaissance, il n'existait pas de ressource structurée française permettant de décrire les cultures par leurs usages ou leur destination. Les grandes classes d'usage de l'agriculture sont l'alimentation humaine ou l'alimentation animale. Certaines productions sont destinées à être transformées pour faciliter leur consommation. Par exemple, la production de houblon est destinée à la fabrication de la bière. Très peu de productions agricoles sont destinées à l'industrie sans avoir un but alimentaire. Nous pouvons citer par exemple le chanvre, qui est utilisé pour la fabrication de textile.

Notre but était de construire une hiérarchie des cultures en fonction de leur usage. Les liens hiérarchiques représentaient des relations de généralisation/spécialisation entre cultures (céréale/blé). Pour construire notre thésaurus intitulé FrenchCropUsage, nous avons étudié les termes contenus dans des documents disponibles librement. Les documents étudiés sont :

- Les statistiques agricoles annuelles publiées sur le site de l'Agreste. Le document intitulé "la statistique agricole annuelle : présentation générale" décrit la hiérarchie des cultures pour répertorier l'ensemble de la production agricole (Agreste).
- Les métadonnées du registre parcellaire graphique présentent une nomenclature des cultures ou groupes de culture (Registre Parcellaire).
- Les listes des rubriques utilisées pour organiser les bulletins d'alerte sur chacun des sites web des chambres d'agriculture (une liste contient les rubriques "Arboriculture", "Grandes cultures", …).

- Le classement des cultures par groupe d'usage proposé par Wikipédia (Wikipédia France Culture).
- Pour compléter avec des définitions chacun des concepts de la hiérarchie, nous avons recherché les définitions dans le Larousse Agricole (Larousse Agricole).
- En cas d'absence d'information dans le Larousse Agricole, nous avons utilisé le portail français de l'agriculture de Wikipédia (Portail Agricole). Des absences de définition sont à noter surtout pour tous les fruits tropicaux.

3 Modélisation du thésaurus

L'ensemble de la hiérarchie a été modélisée à l'aide du vocabulaire SKOS proposé par le W3C (Miles & Bechhofer, 2009). Notre vocabulaire de types de cultures est disponible sur le web de données liées[1]. Il contient 272 concepts. La profondeur maximale de la hiérarchie est de 6 niveaux. Comme le montre la figure 1, chaque concept est défini par les propriétés suivantes:

- skos:prefLabel contient le terme préféré utilisé comme étiquette du concept en français. En général, le terme est le nom usuel de la plante cultivée suivi de son usage.
- skos:altLabel contient les autres termes qui peuvent être utilisés comme étiquettes du concept.
- skos:definition contient la définition en français du concept justifiant sa position dans la hiérarchie.
- skos:inScheme exprime l'appartenance du concept au thésaurus.
- rdfs:seeAlso contient un lien web vers une définition retenue lors de la construction du thésaurus, comme par exemple les définitions du Larousse Agricole
- skos:note contient au moins une définition trouvée dans une autre source comme l'agreste ou Wikipédia.
- skos:editorialNote contient la définition du Larousse Agricole. Pour des raisons de propriété intellectuelle cette propriété est supprimée dans les versions en ligne.
- skos:broader indique le lien vers le concept plus générique.
- skos:narrower indique le lien vers le concept plus spécifique

Les grandes classes de cultures représentées dans notre thésaurus sont : culture légumière (l'alimentation humaine), culture fruitière (l'alimentation humaine), grande culture (industrie, alimentation humaine et animale), culture fourragère (l'alimentation animale), horticulture ornementale, zone non agricole. Toute culture doit être un descendant d'un de ces concepts (top concepts dans le modèle skos).

Il existe des plantes qui ont plusieurs usages : c'est le cas par exemple des céréales fourragères qui sont à la fois destinées à l'alimentation humaine (céréale, contenue dans l'épi) et à l'alimentation animale (paille, parfois utilisée comme fourrage). Quand un nom de plante est suffisamment ambigu pour qu'aucun usage ne puisse être déduit de ce nom, il faut pouvoir modéliser cette ambiguïté. Pour ce faire, nous avons défini la notion de concepts agricoles ambigus.

Concept agricole ambigus: correspond à un nom de plante cultivée pour des usages différents amenant une indécision dans le choix de l'usage de la production de la culture. Par exemple les céréales fourragères, par définition, sont destinées à l'alimentation humaine pour leur grain et à l'alimentation animale pour leur tige. Le concept représentant les céréales fourragères n'est pas ambigu. Par compte, le terme « pomme de terre » est ambigu car la

[1] FrenchCropUsage est disponible sur la plateforme AgroPortal du LIRMM et aussi sur le site http://ontology.irstea.fr

pomme de terre cultivée en maraîchage est destinée à l'alimentation humaine directe et la pomme de terre cultivée en grande culture est destinée à l'industrie pour la consommation humaine indirecte ou pour la production de matériaux. Le terme « pomme de terre » ne permet pas de déterminer de manière certaine quelle est la destination de cette production.

Ces concepts ambigus ne peuvent pas être utilisés pour représenter une culture. Un de ses fils doit être choisi en remplacement. Pour la pomme de terre par exemple, on utilisera donc «pomme de terre féculière» ou «pomme de terre potagère» pour lever l'ambiguïté.

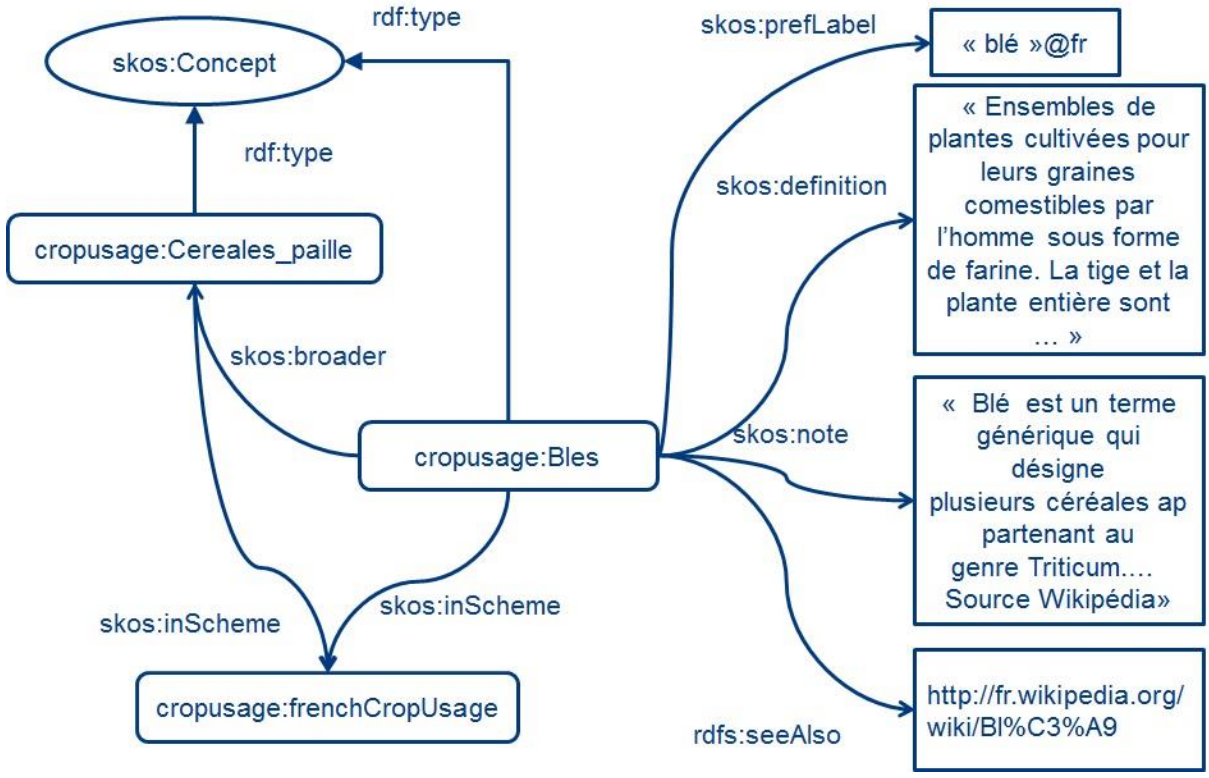

FIGURE 1 : extrait du thésaurus FrenchCropUsage

4 Liste des contraintes

Cette section liste les contraintes que nous avons utilisées pour le développement de notre thésaurus. Une contrainte est décrite par son identifiant, son niveau de priorité et une description de la requête associée. Le niveau de priorité varie entre contrainte forte > contrainte faible > indication. Pour des raisons de lisibilité nous ne présentons que le code des requêtes les plus courtes. L'ensemble des requêtes SPARQL associées est disponible sur le site ontology.irstea.fr.

4.1 Contraintes sur les labels

LabLang

Contrainte forte : tous les labels doivent être associés à une langue, le français dans notre cas.

Requête : Afficher les labels (préféré, alternatif, caché) qui n'ont pas la langue française associée.

ConcLabPref

Contrainte forte : dans notre thesaurus, un concept skos doit avoir un label préféré.
Requête : afficher les concepts qui n'ont pas de label préféré en français.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?concept WHERE {
 ?concept rdf:type skos:Concept.
FILTER NOT EXISTS {
         ?concept skos:prefLabel ?label.
        BIND ( LANG(?label) AS ?LANG)
        VALUES ?LANG {"fr"}
        }
}
```

UniLabPref

Contrainte forte : un label préféré doit être associé à un unique concept.
Requête : afficher les labels préférés qui sont associés à plusieurs concepts.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?label (COUNT(?concept) AS ?nb) WHERE {
                ?concept rdf:type skos:Concept.
                ?concept skos:prefLabel ?label.
        }
        GROUP BY ?label
        HAVING (?nb > 1)
```

LabDiff

Contrainte faible : un label préféré doit être distinct d'un label alternatif pour un concept donné.
Requête : trouver les concepts dont un label alternatif est équivalent à un label préféré.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?label1 ?concept
WHERE{
        ?concept rdf:type skos:Concept.
        ?concept skos:prefLabel ?label.
        ?concept skos:altLabel ?label.
}
```

Polysem

Indication : éviter autant que possible les répétitions entre les différents types de labels, sauf en cas de polysémie où un label est associé à deux concepts.
Requête : trouver les concepts qui partagent des labels (préférés ou alternatifs) en commun.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?label1 (COUNT (DISTINCT ?concept1) AS ?nbc)
```

```
WHERE{
        ?concept1 rdf:type skos:Concept.
        ?concept1 ?p ?label1.
        {
        SELECT ?label ?concept
        WHERE {
                {
                        ?concept skos:prefLabel ?label.
                }
                UNION
                {
                        ?concept skos:altLabel ?label.
                }
                 }
        }
        FILTER (?label1=?label && ?concept1=?concept).
}
GROUP BY (?label1)
HAVING (?nbc > 1)
```

4.2 Contraintes sur les notes, définitions

NoteLang

Contrainte forte : une définition, une note, une note éditoriale doit être associée à une langue, le français dans notre cas.

Requête : trouver les concepts dont la définition, la note ou la note éditoriale ne sont pas associées à la langue française.

ConcDef

Contrainte forte : un concept doit avoir une définition.

Requête : trouver les concepts qui ne sont pas associés à une définition.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?concept WHERE {
        ?concept rdf:type skos:Concept.
        FILTER NOT EXISTS {
                ?concept skos:definition ?def.
        }
}
```

SupNoteEdition

Contrainte forte propre à notre thésaurus : pour la publication sur le web, les notes éditoriales doivent être supprimées.

Requête : trouver les concepts qui ont des notes éditoriales.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?concept WHERE {
        ?concept rdf:type skos:Concept.
        FILTER EXISTS {
```

```
                ?concept skos:editorialNote ?note.
        }
}
```

4.3 Contraintes sur les schemas

HasSchema

Contrainte forte : un concept doit être associé à un schéma.
Requête : trouver les concepts qui ne sont pas associés à un schéma.

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT ?concept WHERE {
        ?concept a skos:Concept.
        FILTER NOT EXISTS
        {?concept skos:inScheme ?aScheme.}
}
```

HasTopConc

Contrainte forte : un schéma doit avoir des top concepts.
Requête : trouver les schémas qui n'ont pas de top concepts.

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT ?schema  WHERE {
        ?schema a skos:ConceptScheme.
        FILTER NOT EXISTS
        {?schema skos:hasTopConcept ?concept.}
}
```

4.4 Contraintes sur la hiérarchie

CheckRoot

Contrainte forte : tout concept doit être rattaché à la racine par un chemin de propriétés skos:broader.

Requête : recherche les concepts qui ne sont pas rattachés à la racine par un chemin de propriétés skos:broader. Le respect de la règle hasTopConc est une précondition à cette requête.

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT ?concept WHERE {
?concept rdf:type skos:Concept.
FILTER NOT EXISTS {

  ?concept skos:broader* ?parent.
 ?parent a skos:Concept.
 ?parent  skos:topConceptOf ?schema.
 ?schema a skos:ConceptScheme .
}
}
```

TopOrphelin

Contrainte forte : les top concepts du schéma ne doivent pas avoir de père.

Requête : trouver les top concepts qui ont un père.

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT ?concept WHERE {
        ?concept rdf:type skos:Concept.
        ?schema skos:hasTopConcept ?concept.
        FILTER EXISTS {
                ?concept skos:broader ?parent.
        }
}
```

CheckInverse

Contrainte forte : si un lien skos:broader existe le lien inverse skos:narrower doit aussi exister (et inversement).
Requête : trouver un lien skos:narrower qui n'a pas son lien inverse skos:broader.
Requête : trouver un lien skos:broader qui n'a pas son lien inverse skos:narrower.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?parent ?enfant WHERE {
  ?parent a skos:Concept.
  ?enfant a skos:Concept.
  ?parent skos:narrower ?enfant.
FILTER NOT EXISTS{  ?enfant skos:broader ?parent.}
}
```

CheckBi

Contrainte forte : les liens skos:related doivent être bidirectionnels.
Requête : trouver un lien skos:related qui n'a pas son inverse.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?source ?cible WHERE {
  ?source a skos:Concept.
  ?cible a skos:Concept.
  ?source skos:related ?cible.
FILTER NOT EXISTS{  ?cible skos:related ?source.}
}
```

CheckTrans

Indication : rechercher des liens transitifs dans la hiérarchie des skos:broader ou skos:narrower. Ces propriétés ne doivent indiquer que des liens directs.
Requête : trouver un lien skos:broader (ou skos:narrower) qui s'exprime aussi par un chemin de liens skos:broader (ou skos:narrower).

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?grandparent ?parent ?enfant WHERE {
  ?grandparent a skos:Concept.
  ?parent a skos:Concept.
  ?enfant a skos:Concept.
  ?grandparent skos:narrower+ ?parent.
  ?parent skos:narrower+ ?enfant.
FILTER EXISTS{  ?grandparent skos:narrower ?enfant.}
}
```

CheckCycle

Contrainte forte : il ne doit pas y avoir de cycle entre les liens skos:narrower (idem pour les liens skos:broader).

Requête : recherche des cycles.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?a ?b WHERE {
  ?a a skos:Concept.
  ?b a skos:Concept.
  ?a skos:broader+ ?b.
FILTER EXISTS{  ?b skos:broader+ ?a.}
}
```

4.5 Contraintes de domaine propre à notre thésaurus

CheckCoherence

Indication, contrainte propre à notre thésaurus. Il faut vérifier que le parent est compatible avec son enfant. Nous avons 6 grandes catégories de cultures. Un concept parent qui est un descendant d'une catégorie doit avoir généralement sa descendance sous cette catégorie uniquement.

Requête : rechercher les descendants d'un concept qui ne sont pas descendants de la même catégorie de culture.Les résultats de cette requête doivent être validés par un expert.Les céréales sont par exemple classées dans « grande culture » et dans « culture fourragère » donc l'association « grande culture » et « culture fourragère » est acceptée pour toutes les céréales.

CheckAmbiguity

Indication : contrainte propre à notre thésaurus. Il faut vérifier qu'un concept ambigu a bien des enfants qui sont catégorisés dans des cultures différentes (parmi les 6 grandes catégories de culture).

Requête : trouver parmi les concepts ambigus (les fils directs de la racine qui ne sont pas dans les 6 grandes catégories) ceux dont les enfants sont tous classés dans la même catégorie.

CheckPatternCropMultipleUsage

Indication : il faut valider auprès d'experts les cultures qui ont plusieurs usages (cas des céréales fourragères).

Requête : trouver les concepts de culture qui ont plusieurs parents directs.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?cropLabel (COUNT(DISTINCT ?parentLabel) AS ?nb)  WHERE {
  ?crop a skos:Concept.
  ?crop skos:prefLabel ?cropLabel.
  ?crop skos:broader ?parent.
  ?parent skos:prefLabel ?parentLabel.
}
GROUP BY ?cropLabel
HAVING (?nb > 2)
```

5 Travaux connexes

Le W3C met à disposition une liste d'outils pour valider les modèles SKOS. Les plus connus sont Poolparty Thesaurus Consistency Checker et Skosify (Suominen & Hyvönen, 2012). Ces outils vérifient non seulement les contraintes d'intégrité spécifiées dans (Miles & Bechhofer, 2009) mais aussi d'autres contraintes structurelles propres au thésaurus multilingue, comme la couverture homogène des langues dans le thésaurus, la détection de concepts isolés etc… Comme le montre le tableau 3, certaines de nos requêtes sont une spécialisation de ces contraintes d'intégrité génériques au thésaurus.

L'outil qSKOS (Mader et al, 2012) propose 15 contraintes que les modèles SKOS doivent vérifier. qSKOS met à disposition les fonctions pour vérifier ces contraintes. Parmi ces contraintes on retrouve celles testées dans les deux outils précédents, plus de nouvelles contraintes liées à la publication sur le web de données liées (test des liens entrants, des liens sortants, de la validité des URI etc…). Les travaux de (Lacasta et al, 2016) vérifient que les thésaurus suivent la norme ISO 25964. Ils vérifient en plus que les liens entre concepts soient informatifs. Par exemple un lien skos:related ne doit pas relier deux concepts liés par des liens hiérarchiques (skos:broder ou skos:narrower). Pour valider les liens hiérarchiques, ces travaux projettent le thésaurus sur une ressource générique comme Wordnet ou dolce. Ainsi la hiérarchie du thésaurus est comparée à la hiérarchie extraite de ces ressources.

TABLE 3 *: Comparaison entre 4 outils de validation de modèle SKOS*

Nom de contraintes	Poolparty checker	Skosify (Suominen & Hyvönen, 2012)	qSKOS (Mader et al, 2012)	(Lacasta et al. 2016)
LabLang	X	X	X	X
ConcLabPref	X	X	X	X
UniLabPref	X	X	X	X
LabDiff	X	X	X	X
Polysem	X	X	X	X
NoteLang	X	X	X	X
ConcDef			X	X
hasSchema				
hasTopConc			X	
CheckRoot		X		
topOrphelin			X	
CheckInverse				
CheckBi				
CheckTrans				
CheckCycle		X	X	X

Nos travaux sont une implémentation en SPARQL de la vérification de ces contraintes mais en utilisant uniquement l'interface d'interrogation d'un Sparql-endpoint. Nous pouvons ainsi tester de nouvelles contraintes propres à ce thésaurus. Certains des travaux précédents ont eu pour but d'évaluer les thésaurus en proposant des métriques. Le but est de choisir le meilleur thésaurus pour une tache donnée. Notre proposition a uniquement pour but de détecter les erreurs pour aider le développeur à modifier son thésaurus. Nos requêtes sont des tests unitaires faciles à mettre en œuvre qui s'appliqueront à chaque modification du thésaurus. Nous espérons que cette base de requêtes mise à disposition pourra être réutilisée et adaptée au cours du développement de notre thésaurus et d'autres. Les outils précédents peuvent nous permettre de proposer de nouvelles requêtes. En effet notre thésaurus ne contient pas actuellement de liens skos:related ou de liens de correspondance (skos:exactMatch ou skos:broadMatch).

6 Conclusion

SKOS est un schéma RDF utilisé pour stocker et publier sur le web de données liées les thésaurus ou taxonomies. Lors du développement d'un thésaurus, il est utile de détecter automatiquement les concepts qui violent les contraintes d'intégrité. Cet article présente le cas d'usage d'un développement manuel d'un thésaurus monolingue. Ce thésaurus porte sur l'usage des cultures en France.

Dans cet article, nous avons spécifié des contraintes propres à un thésaurus. Pour vérifier ces contraintes un ensemble de requêtes SPARQL a été défini. Ces requêtes permettent à la fois de vérifier des contraintes d'intégrité propres à tous les thésaurus stockés au format SKOS mais aussi des contraintes propres à ce thésaurus agricole.

Références

la statistique agricole annuelle présentation générale. Disponible à l'url
http://www.agreste.agriculture.gouv.fr/IMG/pdf_methosaa.pdf

LACASTA J., FALQUET G., , F. ZARAZAGA-SORIA J., NOGUERAS-ISO J (2016). An automatic method for reporting the quality of thesauri. Data & Knowledge Engineering Vol 104, Pages 1–14
https://doi.org/10.1016/j.datak.2016.05.002

Larousse agricole Édition 2002. Disponible à l'url http://www.larousse.fr/archives/agricole/

MADER C., HASLHOFER B., ISAAC A. (2012) Finding quality issues in SKOS vocabularies. Theory and Practice of Digital Libraries, 2012, p. 222-233.

MILES A. & BECHHOFER S. (2009) SKOS simple knowledge organization system reference.

Portail: Agriculture et agronomie. Disponible à l'url
https://fr.wikipedia.org/wiki/Portail:Agriculture_et_agronomie

Description de la couche Registre parcellaire graphique 2012 (îlots PAC) Métadonnée du 24/09/2013. Disponible à l'url
http://piece-jointe-carto.developpement-durable.gouv.fr/DEPT063A/METADONNEES/N_RPG_2012_S_063_metadonnees.pdf

SUOMINEN O. & HYVÖNEN E.(2012) Improving the quality of SKOS vocabularies with Skosify.In : International Conference on Knowledge Engineering and Knowledge Management. Springer Berlin Heidelberg, p. 383-397.

page de Wikipédia sur les classements des cultures disponible à l'url:
https://fr.wikipedia.org/wiki/Classement_en_France_des_cultures_par_groupes_d'usage

Annotation sémantique à partir de textes : Cas des observations dans les Bulletins de Santé du végétal

Haïfa Zargayouna[**], Catherine Roussey[*], Synda Ouardani[**]

[*]Irstea UR TSCF, 9 avenue Blaise Pascal, CS 20085, Aubière, France
catherine.roussey@irstea.fr
[**]Laboratoire d'Informatique de Paris Nord (LIPN, UMR 7030), Université Paris 13
prenom.nom@lipn.univ-paris13.fr

Résumé : Dans cet article nous décrivons un schéma d'annotation pour annoter les observations extraites des bulletins agricoles disponibles sur le Web. Le but est de proposer un processus d'annotation automatique qui permet de générer des annotations complexes accessibles via un sparql endpoint. Nous partons d'annotations manuelles des portions de textes, ces annotations permettent d'identifier les entités sémantiques à repérer dans le texte. Nous nous appuyons, dans la mesure du possible, sur des schémas de données et des ontologies disponibles.
Mots-clés : annotation sémantique, ontologie, domaine agricole, bulletins de santé du végétal, schéma d'annotation

1 Introduction

Dans cet article, nous présentons la mise en place d'un schéma d'annotation pour guider le processus d'annotation à partir de plusieurs ontologies. Les annotations produites doivent être structurées pour être exploitables par un moteur d'interrogation tel que SPARQL (*SPARQL Protocol and RDF Query Language* (Segaran *et al.*, 2009). L'objectif de l'interrogation est d'effectuer des recherches au sein d'un corpus d'alerte agricole pour rendre visible une dynamique temporelle et spatiale de phénomènes agricoles. Ceci permettra aux agronomes de visualiser ces dynamiques et facilitera leurs interprétations et leurs prévisions de l'état sanitaire des cultures dans les régions françaises.

Nous nous inscrivons dans une optique d'ouverture de données de l'agriculture. Le domaine agricole est à l'intersection de plusieurs autres domaines : environnement, santé, alimentation etc. En effet les opérations agricoles sont impactées et impactent plusieurs facteurs tels que l'eau, le climat, la biodiversité, etc. L'ouverture des données agricoles pose débat. Le ministère a prévu la création d'une plateforme AgGate pour l'ouverture des données agricoles en France (Bournigal, 2016). Des plateformes existantes tels que API-AGRO (Plateforme de données et de services pour l'écosystème agricole) participent au partage de données avec néanmoins un contrôle en droits de lecture. Même si il y a actuellement de grandes bases de données d'observations, leur ouverture peut être problématique pour leurs auteurs. Les observations scientifiques faites sur les parcelles agricoles ont besoin d'être anonymisées et protégées pour que les agriculteurs puissent faire leur travail correctement.

2 Corpus et cas d'usage

Nous présentons dans ce qui suit les bulletins de santé du végétal et les besoins que nous prenons en compte dans ce travail.

2.1 Les Bulletins de Santé du Végétal

Le Grenelle de l'Environnement [1] et le plan Ecophyto [2] ont renforcé les réseaux de surveillance sur les cultures et les pratiques agricoles. Les Bulletins de Santé du Végétal sont une des modalités mises en place par ces réseaux de surveillance. Le Bulletin de Santé du Végétal (BSV) est un document d'information technique et réglementaire, rédigé sous la responsabilité d'un représentant régional du ministère de l'agriculture. Les BSV diffusent des informations relatives à la situation sanitaire des principales productions végétales de la région et proposent une évaluation des risques encourus pour les cultures. Des données générales concernant les stratégies de lutte (notes nationales, etc.) ou sur la réglementation peuvent figurer également dans les BSV. Les BSV sont une synthèse des observations effectuées sur les cultures. Il existe des bases de données d'observations mais la rédaction des BSV oblige leurs auteurs à décider si une observation est un phénomène unique non représentatif ou un phénomène important représentatif d'une réalité. Les BSV ne sont pas une agrégation automatique de données mesurées mais bien une synthèse humaine des jugements sur des observations.

Une archive pérenne de ces bulletins agricoles a été constituée, afin d'en extraire un ensemble d'information sur les cultures. Cette archive est disponible comme jeux de données sur le Web de données (Roussey *et al.*, 2016). Une première série d'annotations de ces bulletins a été réalisé pour retrouver un bulletin au sein du corpus. Les annotations portent sur la grande catégorie de culture indiquées dans le titre du bulletin, la date de publication du bulletin et sa région de publication. Ces méta-données ont été extraites des sites Web de publication des BSV et non des contenus des BSV (Roussey & Bernard, 2015). La figure 1 présente un exemple de BSV. La figure 2 présente les annotations actuelles d'un fichier BSV.


BULLETIN DE SANTÉ DU VÉGÉTAL
BOURGOGNE
Grandes cultures n° 20 du 5 avril 2011
A RETENIR CETTE SEMAINE
- Grande majorité de parcelle avec des fleurs
- Avec le développement rapide des colzas le risque méligèthes s'estompe
- Signalement de quelques dégâts de charançons de la tige du colza
- Premiers charançons des siliques piégés mais sans incidence car stade de sensibilité des colzas pas atteint (à partir du stade G2)
Réseau 2010-2011
Cette semaine, 53 parcelles ont fait l'objet d'au moins une observation.
Stade des colzas
Rappel : un stade est atteint lorsque 50% des plantes sont à ce stade.
Les conditions climatiques estivales de la semaine dernière ont permis une accélération des stades et a fortiori dans les secteurs qui ont eu la chance de bénéficier des pluies.
- D1 boutons accolés encore cachés par les feuilles terminales : 2%
- D2 inflorescence principale dégagée, boutons accolés, inflorescences secondaires visibles : 8%
- E boutons séparés, les pédoncules floraux s'allongent en commençant par ceux de la périphérie : 66%
- F1 premières fleurs ouvertes : 20%
- F2 allongement de la hampe florale, nombreuses fleurs ouvertes : 4%
Méligèthes
Météo
Prévisions météorologiques du mercredi 06 au mardi 12 avril :
Temps très ensoleillé et sec pendant toute cette période. Les températures sont en hausse avec des minimales ne descendant pas en dessous de 9°C et des maximales atteignant 20 à 22°C (source météociel).

FIGURE 1 – Un exemple de BSV de la région bourgogne catégorie grande culture à la date du 5 avril 2011

1. Le Grenelle de l'Environnement a eu lieu en 2007 et avait pour but de proposer un ensemble d'actions afin de permettre la mise en place d'une nouvelle politique environnementale.

2. Le plan Ecophyto fut lancé en 2008 dans le but de réduire progressivement l'utilisation des pesticides en France tout en maintenant une agriculture économiquement performante.

2.2 Cas d'usage

Notre objectif est d'extraire du contenu des bulletins agricoles des informations précises. Les informations extraites du contenu textuel sont de plusieurs types : des observation sur les stades de développement des cultures, des observations de la présence des ravageurs et de maladies sur les cultures, à partir de ces deux types d'observations des estimations de risques sur les cultures sont présentées. Dans cet article, nous prenons comme exemple les observations sur les stades de développement des cultures. Le but à terme est de généraliser la proposition à tout type d'observation. Les acteurs sont de deux types :

— agriculteurs et conseillers techniques : ils recherchent l'état général des cultures par type de culture dans leur région ou les régions limitrophes. Leur objectif est de caractériser l'état de leurs cultures par rapport aux autres (retard dans le développement etc...).

— agronomes : ils sont intéressés pour faire le bilan d'une année culturale c'est à dire : savoir quelles cultures sont produites en France par région et leur rendement, et de suivre l'évolution temporelle des cultures pour identifier les causes des variations dans le rendement. Ils ont besoin de connaître précisément l'échantillon sur lequel portent les observations c'est-à-dire combien de parcelles culturales sont observées.

3 État de l'art

L'annotation sémantique est le processus qui consiste à établir des liens entre une ressource (le texte) et une autre ressource (ressource sémantique). L'annotation sémantique fait référence aussi au résultat de ce processus. Zargayouna *et al.* (2015) présentent les différentes étapes du processus d'annotation ainsi que les plates-formes de l'état de l'art. Nous nous intéressons spécifiquement aux schémas d'annotation. Un schéma d'annotation permet de guider le processus d'extraction en définissant les entités sémantiques pertinentes et les liens entre elles. Le schéma d'annotation peut être assez simple et générique tel que modèle appelé OADM (*Open Annotation Data Model*) proposé par le groupe de travail du W3C (2014) (*Web Annotation Working Group*). D'autres schémas correspondent au schéma de l'ontologie projeté comme dans (Abacha & Zweigenbaum, 2010). Nous proposons un schéma d'annotation complexe adossé à plusieurs ontologies et commençons donc l'élaboration de ce schéma par une analyse manuelle des différentes entités et relations pertinentes à annoter. Cette phase d'analyse est suivie par une phase de recherche de ressources et ontologies existantes. Le schéma d'annotation défini manuellement permet de lier ensemble les différentes entités repérés et de les intégrer dans une représentation commune.

4 Exemple d'annotation

Nous présentons dans ce qui suit un extrait de texte présentant une observation sur des colzas. Ce texte est issu du bulletin présenté dans la figure 1. Les entités sémantiques intéressantes sont mises en exergue.

Grandes cultures n° 20 du **5 avril 2011**

..

Cette semaine, **53 parcelles** ont fait l'objet d'au moins **une observation**. ..

Stade des colzas

Rappel : un stade est atteint lorsque 50% des plantes sont à ce stade. Les conditions climatiques estivales de la semaine dernière ont permis une accélération des stades et a fortiori dans les secteurs qui ont eu la chance de bénéficier des pluies.

- **D1** boutons accolés encore cachés par les feuilles terminales : **2%**
- ...

Plusieurs informations sont intéressantes dans ce texte. Tout d'abord nous savons que ce bulletin concerne uniquement des cultures catégorisées sous "Grande culture" et qu'il a été publié le 5 avril 2011. Ainsi toutes les observations détectées dans ce texte seront datées du 5 avril 2011. Nous savons ensuite que des parcelles de colza ont été observées pour déterminer le stade de développement de cette culture dans la région Bourgogne. À noter que le nom de la région (Bourgogne) n'apparaît pas dans le texte car le titre est une image intégrée dans le fichier pdf. En revanche, cette information peut être trouvée dans les annotations déjà disponibles sur les BSV extraites au moment de la constitution du jeu de données [3]. L'échantillon des parcelles de colza observées en Bourgogne se monte à 53 parcelles. Parmi ces 53 parcelles seules 2% d'entre elles ont atteint le stade D1, soit une parcelle. En conclusion dans ce texte nous pouvons déduire qu'il y a une observation du stade de développement des cultures de colza. Le résultat de cette observation est D1. L'échantillon est une parcelle de colza. Cette échantillon est lié à un échantillon globale de 53 parcelles de colza.

5 Schéma d'annotation pour les observations

Le but est d'enrichir le schéma d'annotation des BSV (voir figure 2) avec des annotations sur les observations du stade de développement des cultures. Le schéma d'annotation proposé par Roussey & Bernard (2015) s'appuie sur des concepts et instances publiés par l'IGN [4] ainsi que les concepts FOAF(Brickley & Miller, 2007), SKOS(Miles *et al.*, 2005) et les propriétés définies par Dublin Core.

3. Informations trouvées dans les url des pages aspirées.
4. L'Institut Géographique National publie ses données liées sur `http://data.ign.fr/`

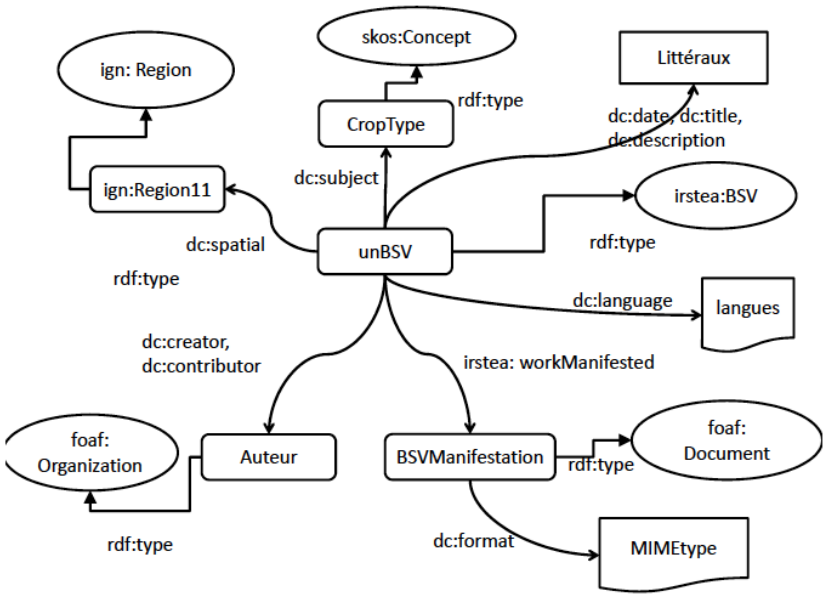

FIGURE 2 – Schéma d'annotation des BSV (Roussey & Bernard, 2015)

Peu de ressources sémantiques existent dans le domaine agricole, notamment pour le français. Il existe, néanmoins des ontologies et schémas de données qui peuvent être réutilisées. Nous nous appuyons également sur le thésaurus qui a été construit manuellement à partir du Larousse Agricole pour référencer les types de culture. Nous avons effectué une recherche sur le Web de données liées ainsi que sur des plates-formes dédiées telle AgroPortal [5]. Nous présentons dans ce qui suit l'ensemble des ressources que nous exploitons pour définir le schéma d'annotation sur les observations.

Thésaurus FrenchCropUsage est un thésaurus construit à l'Irstea. il référence les types de cultures organisés en fonction de leurs usages, intitulé FrenchCropUsage. Les classes d'usage de la production agricole sont destinées, à l'alimentation humaine, ou à l'alimentation animale ou à l'industrie. Les besoins de l'industrie sont la conception de textile (lin, chanvre) ou d'autres utilisations industrielles (huiles à usage particulier, biocarburants). Dans ce thésaurus, les liens hiérarchiques représentent des relations de généralisation/spécialisation entre cultures (céréale/blé). Ce vocabulaire de type de culture est disponible sur le Web de données liées sous format SKOS [6]. Il contient 272 concepts, la profondeur maximale de la hiérarchie est de 6 niveaux.

Ontologie Sosa ou Sensor Observation Sample Actuator (Haller *et al.*, 2017) est une ontologie du domaine des capteurs en cours de validation au W3C. Elle est une évolution de l'ontologie Semantic Sensor Network (SSN). Elle est liée à SSN grâce à une architecture de modularisation horizontale et verticale. Cette ontologie suit le patron de conception de l'ontologie SSN en ajoutant de nouvelles classes et propriétés pour les

5. AgroPortal accessible à `http://agroportal.lirmm.fr/`.
6. Accessible à partir d'agroportal `http://agroportal.lirmm.fr/ontologies/CROPUSAGE`

actionneurs et l'échantillonnage. Cette ontologie est en cours de validation et continue à être modifiée. Par exemple, la définition de la propriété sosa:hasFeatureOfInterest n'est pas encore stabilisée, car cette propriété peut avoir comme co-domaine une instance de sosa:FeatureOfInterest ou une instance de sosa:Sample.

QU QU (Lefort, 2011) repose partiellement sur l'ontologie des systèmes des quantités, des unités, des dimensions et des valeurs (QUDV). Elle est créée pour définir les unités et les quantités.

Prov L'ontologie PROV (PROV-O) (Timothy *et al.*, 2013) fournit un ensemble de classes et de propriétés pour représenter des relations de provenance entre différentes entités. PROV a la caractéristique d'être légère et facilement réutilisable pour à la fois être intégrée dans de nouvelles ontologies ou pour caractériser des jeux de données.

À l'aide de ces 4 ressources nous allons définir un nouveau schéma d'annotation des BSV. La figure 3 présente le schéma d'annotation que nous allons utiliser.

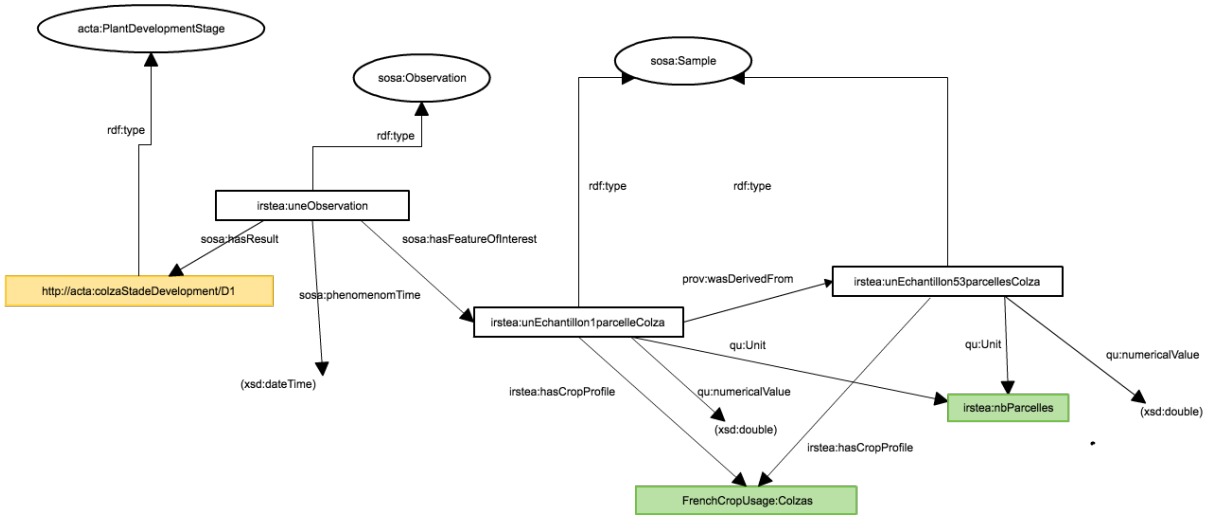

FIGURE 3 – Schéma d'annotation sur les observations. Les concepts sont représentés par des ovales, les instances par des rectangles. Les couleurs montrent les instances déjà définies dans des ressources (en vert dans la ressource locale irstea et en jaune les ressources extérieures).

Nous avons utilisé la classe sosa:Observation pour définir chaque observation. Une observation porte sur l'étude d'un phénomène. Dans notre cas ce phénomène est une parcelle de colza. Cette information est représenté par le lien sosa:hasFeatureOfInterest entre une instance de sosa:Observation et une instance de sosa:Sample. L'instance de Sample représente l'échantillon sur lequel porte cette observation. Cet échantillon est relié à un échantillon plus large représentant les 53 parcelles de colza observées dans la région Bourgogne. Nous utilisons la propriété prov:wasDerivedFrom pour relier l'instance de sosa:Sample représentant une parcelle de colza à l'instance de sosa:Sample représentant les 53 parcelles de colza. Le résultat de l'observation est le stade de développement D1 représenté par la ressource `http://acta:colzaStadeDevelopment/D1`. Ce code est issu d'une nomenclature des stades de déve-

loppement des cultures défini par l'acta [7].

Nous avons utilisé qu:unit pour représenter les unités associées aux échantillons. Cette propriété pointe sur une unité "nombre de parcelles". La propriété qu:numericalValue relie l'échantillon à une valeur xsd :double. Pour préciser la date des observations, nous avons employé la propriété sosa:phenomenonTime pour lier une observation à une date au format xsd:DateTime.

6 Annotations attendues

Nous présentons dans ce qui suit les annotations attendues à partir du texte présenté en exemple plus haut. Nous présentons ces annotations en précisant les calculs nécessaires pour l'automatisation de leur génération à partir de texte.

Repérage des instances d'entités sémantiques

Le repérage des instances d'entités revient à un problème classique de reconnaissance de termes et d'entités nommées. Ainsi par exemple, il faut reconnaître que « 53 parcelles » fait référence à une instance de sosa:Sample et donnera lieu à la création de <irstea:sample_ParcellesColza_Bourgogne_53>. Le nommage des URI nécessite une prise en compte du contexte ainsi qu'une structuration et typage de l'information (par exemple Région/Culture/unité/sample/nombre pour l'exemple précédent).

Voici les annotations qui devraient être obtenues :

```
 <irstea:sample_ParcellesColza_Bourgogne_53>
  rdf:type sosa:Sample.

 <irstea:obs_parcelles_colza_bourgogne_stadeDeveloppement_20110405>
     rdf:type sosa:Observation ;
     sosa:phenomenomTime "05/04/2011".

<irstea:sample_ParcellesColza_Bourgogne_1> rdf:type sosa:Sample.

<acta:colzaStadeDevelopment/D1> rdf:type <acta:PlantDevelopmentStage>.
```

Structuration des informations extraites

Certaines annotations nécessitent de structurer l'information et de reconnaître les types des entités extraites. Il est, par exemple, nécessaire de repérer les valeurs numériques et l'unité.

```
<irstea:sample_ParcellesColza_Bourgogne_53>
  hasCropProfile <FrenchCropUsage:Colzas> ;
  qu:Unit <irstea:nbParcelles> ;
  qu:numericalValue "53".
```

Le typage permet d'effectuer des calculs. Cela est le cas par exemple pour retrouver que 2% de 53 parcelles fait référence à une seule parcelle. Le repérage du pourcentage permet de déduire

7. Les instituts techniques agricoles `http://www.acta.asso.fr/`

le nombre de parcelles concernés par l'observation sans que cela soit mentionné explicitement dans le texte.

```
<irstea:sample_ParcellesColza_Bourgogne_1>
  hasCropProfile <FrenchCropUsage:Colzas> ;
  qu:Unit <irstea:nbParcelles> ;
  qu:numericalValue "1" ;
  prov:wasDerivedFrom <irstea:sample_ParcellesColza_Bourgogne_53>.
```

Détection de relations

Il s'agit de mettre en lien les entités annotées. Comme par exemple lier le stade de développement à l'échantillon concerné.

```
<irstea:obs_parcelles_colza_bourgogne_stadeDevelopement_20110405>
  sosa:hasResult <acta:colzaStadeDevelopment/D1> ;
  sosa:hasFeatureOfInterest <irstea:sample_ParcellesColza_Bourgogne_1>.
```

7 Conclusion

Nous avons présenté la méthodologie de construction d'un schéma d'annotation des observations dans les bulletins de santé du végétal et les annotations RDF qui en découlent. Nous allons mettre en place l'automatisation des annotations à partir de textes en définissant des heuristiques utiles pour l'extraction. L'automatisation des annotations nécessitent une segmentation préalable des BSV selon la culture et une définition de contextes pour l'extraction afin de réduire le bruit. Nous avons commencé à utiliser l'outil OMTAT, qui permet une annotation automatique ou manuelle de textes, et comptons l'enrichir par les heuristiques propres à notre domaine. Notre objectif est de généraliser l'annotation à partir de textes à d'autres types d'observation telles que l'observation des ravageurs et des maladies sur les cultures.

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083.

Références

ABACHA A. B. & ZWEIGENBAUM P. (2010). Annotation et interrogation sémantiques de textes médicaux. In *Atelier Web Sémantique Médical, IC*.

BOURNIGAL J.-M. (2016). Portail de données pour l'innovation en agriculture. http://agriculture.gouv.fr/un-portail-de-donnees-pour-linnovation-en-agriculture-la-synthese-du-rapport.

BRICKLEY D. & MILLER L. (2007). Foaf vocabulary specification 0.91.

HALLER A., JANOWICZ K., COX S., PHUOC D. L., TAYLOR K. & LEFRANÇOIS M. (2017). Semantic sensor network ontology. https ://www.w3.org/TR/vocab-ssn/.

LEFORT L. (2011). Ontology for quantity kinds and units : units and quantities definitions. https ://www.w3.org/2005/Incubator/ssn/ssnx/qu/qu-rec20.html.

MILES A., MATTHEWS B., WILSON M. & BRICKLEY D. (2005). Skos core : simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, p. 3–10.

ROUSSEY C. & BERNARD S. (2015). Annotation des bulletins de santé du végétal. In *Actes de la 7ème édition de l'Atelier Recherche d'Information SEmantique (RISE)*, p. 12 pages.

ROUSSEY C., BERNARD S., PINET F., REBOUD X. & CELLIER V. (2016). Gestion sémantique des bulletins de santé du végétal dans le projet vespa. In *Actes de l'atelier IN-OVIVE @IC 2016*, p. 12 pages.

SEGARAN T., EVANS C., TAYLOR J., TOBY S., COLIN E. & JAMIE T. (2009). *Programming the semantic web*. O'Reilly Media, Inc.

TIMOTHY L., SATYA S. & DEBORAH M. (2013). Prov-o : The prov ontology. https ://www.w3.org/TR/prov-o/.

W3C (2014). Web annotation data model. http ://www.w3.org/TR/annotation-model/.

ZARGAYOUNA H., ROUSSEY C. & CHEVALLET J.-P. (2015). Recherche d'information sémantique : état des lieux. *Traitement Automatique des Langues*, **56**(3), 49–73.

La terminologie structurée, élément structurant de l'activité de l'entreprises ? Ses atouts, ses inconvénients : exemple d'application dans une fondation d'art contemporain

Nicolas Delaforge
Société Coopérative Mnémotix
nicolas.delaforge@mnemotix.com

Résumé : La présentation concernera Mnémotix ainsi que l'intégration du standard SKOS qui est au coeur de la démarche de structuration de l'activité métier chez les clients. Cette première étape de formalisation et d'explicitation est un préalable à tout développement de service à forte valeur ajoutée. La présentation décrira, ensuite, l'outil développé pour rendre la gestion de terminologie SKOS plus accessible à des utilisateurs non-experts. La présentation finira par la présentation des développements en cours pour la mise en oeuvre d'un workflow collaboratif d'édition de terminologie métier en cours de réalisation pour la fondation d'art contemporain Lafayette Anticipation.

Constitution d'un thésaurus pour la recherche de produits

Marc Dutoo
SmileLab
marc.dutoo@smile.fr

Résumé : La présentation introduira l'important de la recherche de produit dans les solutions e-commerce [60% du CA vient de là], puis la solution e-commerce Smile Magento Elastic Suite qui l'adresse. Elle décrira ensuite les problématiques d'expansion de recherche et comment son thésaurus y répond. Elle se focalisera enfin sur l'usage dans la pratique de ce thésaurus et des problématiques concrètes auxquelles il a permis de répondre chez les clients Smile.

Bio : Marc Dutoo dirige les projets R&D chez Smile, le premier intégrateur de solutions Open Source en Europe, où il conduit l'innovation en Big & Linked Data, Cloud, SOA et BPM. Il coordonne actuellement le projet PCU et ses 5 partenaires dans un effort de 3 ans vers une plateforme de *Machine Learning* unifiée pour les applications métier telles le *ecommerce*. Fervent partisan de l'open source, il appartient aux comités techniques d'Eclipse SOA et OW2, intervient régulièrement en conférences (Cloud Expo, Eclipse Con, Linux Solutions) et publie en ce moment le *playground* d'*API cloud OCCInterface*.

Smile est le premier intégrateur de solutions Open Source en Europe, avec plus de 1000 collaborateurs, une offre complète en 5 axes : e-business, Systèmes d'Information, embedded & IoT, digital, infrastructure, et le pôle Smile Lab qui assure sa position de leader dans l'innovation.