Extraction générique de connaissances à partir de données textuelles et mesure de la performance des systèmes d'extraction de relations dans un contexte non supervisé.

Yohann Chasseray, Anne-Marie Barthe-Delanoë, Jean-Marc Le Lann, Stéphane Negny

Yohann CHASSERAY[1], Anne-Marie BARTHE-DELANOË[2], Jean-Marc LE LANN[1], Stéphane NEGNY[1]

[1]Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France
4 allée Émile Monso, 31030 Toulouse, France
yohann.chasseray@ensiacet.fr

[2] Centre Génie Industriel, IMT Mines Albi, Université de Toulouse, Albi, France
1 allée des sciences, 81000 Albi, France
anne-marie.barthe@mines-albi.fr

CIGI 2021
Extraction générique de connaissances à partir de données textuelles et mesure de la performance des systèmes d'extraction de relations dans un contexte non supervisé.

***Résumé* – Parmi les défis à venir dans le monde de l'industrie et dans le pilotage des systèmes industriels, l'agrégation, la synthèse et la gestion des connaissances au travers de structures ontologiques occupent une place primordiale. Beaucoup des systèmes d'extraction de connaissances actuels adoptent une approche supervisée, qui se base sur des données labellisées dont le processus d'annotation est long et fastidieux. Cet article présente une approche par les règles non supervisée, auto-alimentée et indépendante du domaine pour la population d'ontologie à partir de données textuelles. Par ailleurs, l'évaluation de tels systèmes, réalisant de l'extraction de connaissances par les méthodes de traitement automatique du langage, requiert l'utilisation d'indices de performance. Dans le cadre particulier de la population d'ontologie non supervisée, les indices habituellement utilisés pour réaliser ces évaluations présentent des limites dues notamment à l'absence de données annotées. Cet article propose donc également une méthode de mesure des performances dans un contexte où les données de référence et les données extraites ne se recouvrent intrinsèquement pas de manière optimale. Le mode d'évaluation proposé s'appuie sur l'exploitation de données faisant office de références mais qui ne sont pas spécifiquement liées aux données sur lesquelles est réalisée l'extraction, ce qui en fait sa particularité.**

***Abstract* – Among the incoming challenges in the industrial domain and in the monitoring of industrial systems, the aggregation, synthesis and management of knowledge through ontological structures occupy an essential place. Existing knowledge extraction systems often use a supervised approach which rely on labelled data for which the annotation process is fastidious. This paper presents an unsupervised self-feeding rule-based approach for domain-independent ontology population from textual data. Moreover, the evaluation of such systems, performing knowledge extraction using natural language processing methods requires the use of performance indicators. The indicators usually used in such evaluations have limitations in the specific context of knowledge extraction for unsupervised ontology population. Thus, the definition of new evaluation methods becomes a need arising from the singularity of the harvested data, especially when these are unlabelled. Hence, this article also proposes a method for measuring performance in a context where reference data and extracted data do not overlap optimally. The proposed evaluation method is based on the exploitation of data that serve as a reference but are not specifically linked to the data used for extraction, which makes it an original evaluation method.**

***Mots clés* - Traitement automatique du langage, Mesure de performance, Extraction de connaissances, Base de connaissances.**
***Keywords* – Natural language processing, Performance evaluation, Ontologies, Knowledge extraction, Knowledge bases.**

1 Introduction

La construction et l'automatisation de systèmes dans le cadre de la transformation vers l'industrie 4.0 suppose l'appui d'une structure solide basée sur la perception en temps réel du contexte interne et externe dans lequel évolue l'industrie en question. Par exemple, il est primordial pour le pilotage d'une chaîne de production d'avoir une connaissance fine, sinon un aperçu de la demande en aval et des ressources en amont, afin de faire face à tout sur-stock ou arrêt de la chaîne de production. Dans cette optique, des systèmes d'aide à la décision permettent non seulement de représenter l'état des systèmes de production mais également de produire une analyse prédictive quant aux comportements futurs du système (prévision de la demande, maintenance prédictive). La tendance actuelle dans le développement de tels systèmes concerne l'intégration de plus en plus importante de connaissances. Par ce biais, il devient possible pour ces systèmes d'adopter un raisonnement contextualisé de plus en plus proche de celui de l'être humain et ainsi de proposer un meilleur accompagnement dans la prise de décision. L'apport de connaissances aux systèmes d'aide à la décision se matérialise par l'intégration de bases de connaissances qui sont le plus souvent le prolongement d'une ontologie développée pour le domaine au sein duquel interviennent lesdits systèmes. Les ontologies, initialement définies par (Gruber, 1993) comme la *spécification explicite d'une conceptualisation*

constituent un véritable support structuré pour la représentation de la connaissance au sein d'un domaine. Une base de connaissances peut quant à elle être définie comme la version instanciée d'une ontologie dans le cadre d'une application à un problème défini. Une base de connaissance comprend ainsi un ensemble d'instances dérivées des concepts de l'ontologie et un ensemble de relations entre ces instances, dérivées des relations définies au niveau de l'ontologie.

Malheureusement, et malgré l'adoption grandissante de l'approche par les ontologies, la plupart des systèmes développés fonctionnent avec une ontologie et, par extension, avec une base de connaissances propres au cas d'application. En effet, les ontologies et bases de connaissances utilisées sont majoritairement développées en suivant une approche dirigée par les problèmes rendant leur ré-exploitation limitée, voire impossible. Pourtant, des ontologies considérées comme génériques vis-à-vis du domaine de l'industrie 4.0 existent (Cheng et al., 2016; Giustozzi et al., 2018). Cependant, les bases de connaissances qui en découlent, lorsqu'elles existent, ne contiennent pas ou peu de connaissance, rendant leur utilisation également très limitée. Parallèlement, les données générées par l'humain, dans des formats plus ou moins structurés, renferment des éléments de connaissance d'une grande richesse. La multiplication grandissante de ces données et l'avènement de techniques d'analyse des données non-structurées sont autant d'opportunités pour l'enrichissement automatique des ontologies caractéristiques d'un domaine.

Les travaux présentés dans cet article proposent une approche générique pour la population d'ontologie en base de connaissance à partir de données hétérogènes. Cette approche, indépendante du domaine concerné par l'ontologie, se sert des concepts déjà présents dans cette dernière comme d'un support pour la recherche et l'extraction de connaissances au sein de données non structurées.

Cet article aborde, par ailleurs, différents aspects de la construction de bases de connaissances à partir d'ontologies du domaine pré-existantes et est organisé comme suit. La section 2 présente un état de l'art des méthodes utilisées pour la population d'ontologie et plus précisément pour l'extraction de relations. La section 3 introduit un cadre méthodologique générique pour la population d'ontologie indépendamment du domaine et s'intéresse de plus près à la spécification de ce cadre méthodologique pour l'extraction de connaissances à partir de données textuelles. La section 4 présente plus en détail une méthode d'extraction de relations d'hyponymie (Concept-Instance) basée sur l'utilisation de schémas d'extraction. Enfin, la section 5 constitue une prise de recul vis-à-vis de ce cadre en proposant des méthodes de mesure de la performance d'un tel système. Ces méthodes seront discutées dans la section 6.

2 État de l'art

2.1 Population d'ontologie

Les problématiques de population d'ontologie ne sont pas nouvelles et plusieurs études se sont déjà intéressées à l'exploitation de données brutes pour la création d'ontologies.

Beaucoup d'entre elles traitent le problème sous un angle spécifique soit relativement à la diversité des données et des formats de données (Martinez-Rodriguez et al., 2018), soit relativement au domaine d'application (Ferrara et al., 2014; Kaushik & Chatterjee, 2018; Oramas et al., 2016).

D'autres systèmes (Louge et al., 2018; Paukkeri et al., 2012) ne se rattachent pas à une ontologie existante, mais proposent d'extraire directement depuis les données textuelles les concepts permettant d'en générer une. Ce genre d'approche permet dans certains cas de s'affranchir de la création antérieure de l'ontologie, mais ne permet pas de répliquer l'opération avec une ontologie qui serait déjà liée au système d'aide à la décision, par exemple.

D'autres outils pour la population d'ontologie font usage de techniques de machine learning et de deep learning (Faria et al., 2014). Ces outils présentent de bonnes performances même lorsqu'ils sont appliqués à des domaines disjoints. Mais, adoptant une approche supervisée, ces derniers peuvent toutefois se retrouver limités pour des domaines où la disponibilité de données annotées, utilisées à des fins d'apprentissage, est limitée. (Ayadi et al., 2019) proposent une méthode de deep learning non supervisée basée sur l'algorithme de deep-learning Word2Vec (Mikolov et al., 2013). Si cette méthode permet de s'affranchir des jeux de données labellisés pour son fonctionnement, son évaluation se fait toutefois sur des jeux de données annotés manuellement.

2.2 Extraction de relations

La détection et l'extraction de relations, au même titre que la détection d'entités, font partie des sujets les plus présents au sein des problématiques de traitement automatique du langage. On distingue dans la littérature, deux types d'approches pour la détection de relations.

La première est une approche par règles, dans laquelle il s'agit de définir un jeu de schémas d'extraction caractéristiques des relations telles qu'elles apparaissent dans les données. Ainsi, des études telles que (Agichtein & Gravano, 2000; Kim et al., 2009; Niladri Chatterjee & Neha Kaushik, 2017; Snow et al., 2005) s'intéressent à la définition et la déduction de ces schémas d'extraction.

Cette approche se révèle souvent efficace car elle cible directement les schémas représentatifs de ce qui doit être extrait et limite ainsi le taux d'erreur du système. En revanche le taux de connaissance extrait par ces schémas d'extraction reste relativement faible du fait de ce ciblage car des relations apparaissant avec un schéma légèrement différent de celui recherché ne peuvent être détectées. Par ailleurs, la mise en place des règles d'extraction peut s'avérer coûteuse et nécessite le plus souvent une expertise tant dans le domaine ciblé que relativement à la source de données étudiée. Par exemple, pour définir des schémas génériques représentatifs de relations au sein de données textuelles, une connaissance de l'expression syntaxique de ce type de relation dans la langue étudiée est nécessaire.

La seconde approche est statistique, et se base sur des indicateurs caractéristiques de la donnée étudiée. Ces mesures peuvent prendre plusieurs formes. L'une d'entre elles est le TF-IDF (Sparck Jones, 2004) qui est une mesure utilisée pour regrouper des termes au sein d'un corpus de documents dans un même champ sémantique (Paukkeri et al., 2012). Certaines études se réfèrent, pour la détection de relations spécifiques, au taux d'apparition des couples de termes au sein d'un texte (de Boer et al., 2007). Sur la base de ces analyses de co-occurrences, des algorithmes de clustering sont appliqués afin d'extraire les relations qui présentent un intérêt vis-à-vis du domaine métier (Rajpathak, 2013). Également, (Nguyen et al., 2017) utilisent une fonction objectif s'appuyant sur les fréquences d'apparition d'un terme avec son contexte pour construire une représentation statistique des relations d'hyperonymie. Enfin, avec l'apparition des réseaux de neurones utilisant les mécanismes d'attention (Vaswani et al., 2017), la détection de relations par des mécanismes

d'apprentissage précis de la dimension sémantique d'un texte est également une méthode qui peut être adoptée (Geng et al., 2020; Luo et al., 2020). Celle-ci se révèle néanmoins gourmande en termes de volume de données, nécessaires à l'entraînement des modèles.

2.3 Apport de la méthode proposée

Dans les approches présentées dans la section précédente, deux verrous ont été retenus, qui concernent (1) la difficulté à coupler généricité et précision lors de la population d'ontologies et (2) l'incapacité à évaluer la performance des systèmes sur des données qui n'ont pas été labellisées manuellement au préalable.

Pour s'attaquer à ces deux verrous de l'extraction de connaissances, cet article propose donc une méthode de mesure de la performance des systèmes d'extraction pour des données non annotées, applicable au sein d'un système d'extraction générique, réalisant la population d'ontologie en s'appuyant sur la connaissance contenue dans cette dernière.

3 Un framework générique pour la population d'ontologie

3.1 Présentation globale du framework.

Pour réaliser l'extraction de données indépendamment de la source de données et du domaine auquel l'ontologie ciblée fait référence, un cadre méthodologique a été défini, reprenant les principes généraux de l'ingénierie dirigée par les modèles.
Ce cadre méthodologique s'articule autour d'un méta-modèle générique pour la représentation des données hétérogènes. Ce méta-modèle, présenté dans (Chasseray et al., 2021), répartit les données extraites au sein de six classes qui sont les classes : *Objet Ontologique*, *Concept*, *Instance*, *Relation*, *Donnée Extraite* et *Contexte*. Dans ce méta-modèle, la classe *Concept* permet de récupérer au sein des données analysées les éléments représentant des classes de l'ontologie à peupler. La classe *Instance* permet de récupérer dans les données les instances relatives à ces concepts, qui deviendront, par la suite, des individus de l'ontologie. Les classes *Concept* et *Instance*, héritées toutes les deux de la classe *Objet Ontologique* sont liées par la classe *Relation* qui qualifie la nature de la relation identifiée entre deux occurrences de la classe *Objet Ontologique*. Les classes *Donnée extraite* et *Contexte*, permettent quant à elles d'embarquer de l'information supplémentaire sur les concepts et instances extraits. Cela permet de capturer d'une part la donnée brute à partir de laquelle l'élément est extrait, et d'autre part des éléments de contexte qui enrichissent sémantiquement l'élément extrait.

Ces six classes permettent ainsi de structurer assez d'informations pour réaliser par la suite des transformations vers une ontologie cible. Le framework général dans lequel s'intègre ce méta-modèle a été présenté dans (Chasseray, 2020) dont est issue la figure 1. Ce dernier est constitué de deux chaînes de traitement qui, si elles peuvent fonctionner indépendamment l'une de l'autre, sont complémentaires car elles permettent de combiner à la fois l'approche sémantique et l'approche par règles. Ce framework fonctionne ainsi de manière itérative, le processus étant initié par l'approche par règles que rejoint et complète par la suite l'approche sémantique. Les deux boucles de rétroaction se nourrissent ainsi des instances extraites lors de l'étape initiale pour, d'un côté, déduire de manière automatique de nouvelles règles d'extraction, et de l'autre, dériver des matchings sémantiques.

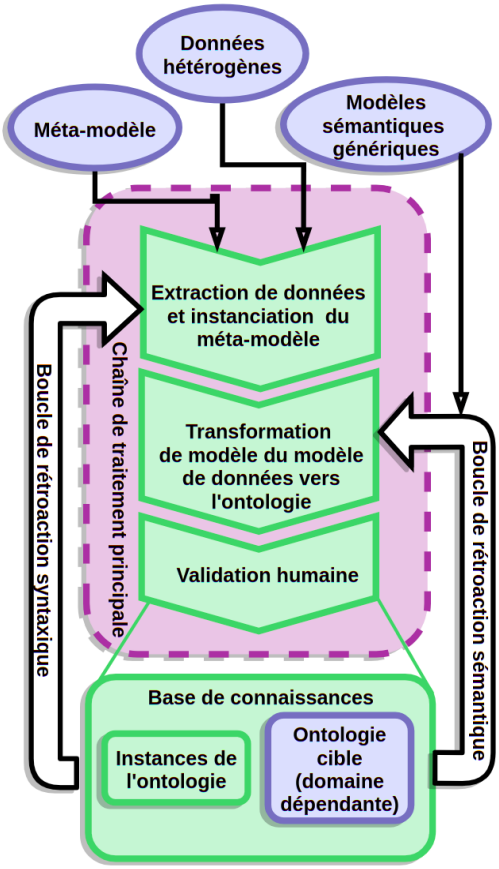

Figure 1. Framework générique pour la population d'ontologie (issu de (Chasseray, 2020)).

La section 4 de cet article se focalise sur la boucle de rétroaction qui s'appuie sur les règles. Ainsi, le processus initial d'extraction et le processus de déduction de nouvelles règles d'extraction, utilisant respectivement les schémas d'extraction proposés par (Hearst, 1992) et le principe de bootstrapping introduit par (Herbelot & Copestake, 2006) et (Pennacchiotti & Pantel, 2006) sont présentés en section 4.

3.2 Spécification du framework pour le traitement de données textuelles.

Le framework générique présenté dans la section 3.1 peut être spécifié pour le traitement de données textuelles grâce à l'utilisation des techniques de traitement automatique du langage. Dans cette section, on présente succinctement les méthodes de traitement automatique du langage qui sont utilisées afin de déduire un modèle de données et nourrir les boucles de rétroaction du framework lors de cette spécification. Ainsi sur la figure 2, on peut distinguer une chaîne de traitement commune qui donne lieu par la suite à trois chaînes de traitement spécifiques. La chaîne de traitement commune contient des opérations de traitement classique de tokenisation et de tagging. Sur la base de ce pré-traitement, les chaînes de traitement différenciées (1), (2), et (3) sont appliquées dans un but qui leur est propre. La chaîne de traitement notée (1) sur la figure 2 a pour objectif d'extraire des couples d'entités liées par une relation. Cette relation peut être de différentes natures (hyponymie, hyperonymie, relation définie par l'ontologie). Ces couples seront par la suite utilisés pour instancier le modèle de données et enfin être rattachés à l'ontologie. Cette chaîne de traitement contient trois étapes. La première permet d'identifier et de labelliser des concepts qui apparaissent dans le texte sur la base des concepts contenus dans l'ontologie. La deuxième

étape consiste à dresser l'arbre des dépendances syntaxiques qui existent entre les tokens du texte. Sur la base des concepts identifiés et de cet arbre des dépendances syntaxiques, des schémas d'extraction sont appliqués au cours de la troisième étape du traitement. Cette chaîne de traitement intervient dans l'étape d'initialisation du système évoquée dans les sections précédentes.

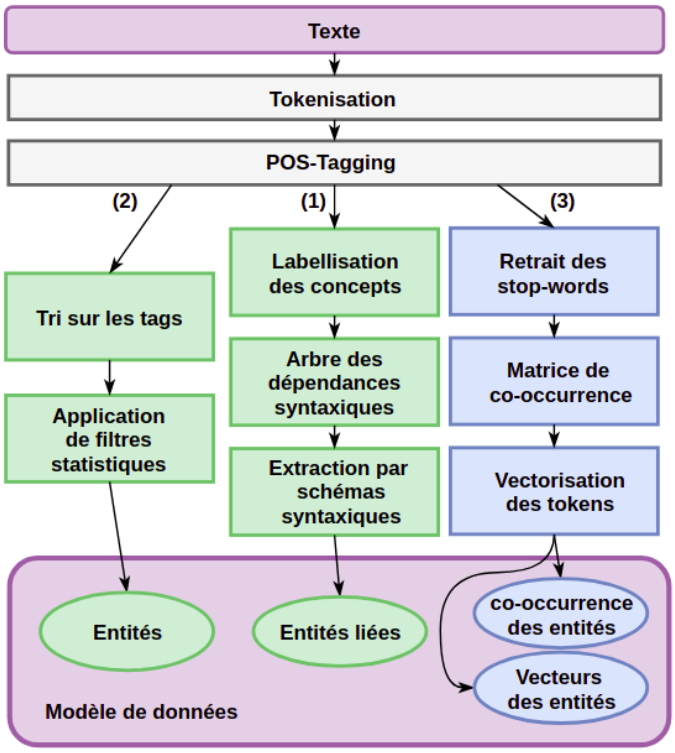

Figure 2. Chaînes de traitement NLP pour la création d'un modèle de données à partir de données textuelles

La chaîne d'extraction notée (2) sur la figure 2 permet d'extraire d'autres entités, cette fois non liées entre elles. L'extraction se fait selon deux filtres. Le premier filtre porte sur les POS-tags représentatifs des termes pouvant s'apparenter à des instances. Un deuxième filtre, statistique, sélectionne les entités les plus susceptibles d'être des instances liées à des concepts du domaine, définis par l'ontologie. Bien que la recherche concerne des instances du domaine, cette chaîne d'extraction n'est pas guidée par l'ontologie. Ainsi, les entités qui en sont issues ne seront pas directement reliées à l'ontologie mais devront subir une étape de matching sémantique et viendront donc alimenter la boucle de rétroaction sémantique.

Enfin, la troisième chaîne de traitement spécifique, notée (3) sur la figure 2, a un objectif légèrement différent des deux premières puisqu'elle est destinée à extraire l'ensemble des éléments de contexte (co-occurrences, vecteurs) qui permettent de caractériser les entités précédemment extraites. Pour ce faire, le traitement s'effectue en trois étapes. La première étape consiste à éliminer les mots vides (stop-words) afin de se concentrer sur les termes porteurs de sens. La seconde étape permet de construire la matrice de co-occurrence qui peut être utilisée pour lier les termes qui co-occurrent avec une entité extraite ou un concept de l'ontologie par exemple. Enfin, la troisième étape utilise un modèle du langage pour obtenir une représentation vectorielle des entités extraites. Ces éléments de contexte ont pour objectif d'être réutilisés par la suite pour identifier des instances parmi les entités non liées dans la boucle de rétroaction sémantique.

Dans cet article, il a été choisi de détailler les méthodes d'extraction et de déduction utilisées pour le fonctionnement de la boucle de rétroaction basée sur les règles. Cette boucle s'appuie sur la chaîne de traitement (1). Une application de l'extraction par schémas syntaxiques sur des données textuelles est ainsi présentée en section 4.1. Cette dernière permet notamment de détailler la manière de procéder à l'extraction à partir de concepts extraits d'une ontologie.

4 Approche par schémas d'extraction (approche par les règles)

Dans de nombreux cas d'application, il n'existe pas ou très peu de données déjà disponibles qui permettraient d'adopter une approche supervisée pour l'extraction de connaissances. Ainsi, un système de population d'ontologie générique doit être en mesure de fonctionner dans un contexte non supervisé, c'est-à-dire sans connaissance à priori sur les entités qui doivent être extraites. De plus, ce système doit se révéler adaptable à différents domaines. L'approche par règles avec l'utilisation de schémas d'extraction génériques prend alors tout son sens.

Dans cette section, la définition de schémas d'extraction générique pour l'extraction à partir de données textuelles est explicitée et le fonctionnement de la boucle de rétroaction basée sur les règles, incluant la déduction de nouveaux schémas est illustré.

4.1 Initialisation par schémas d'extraction génériques

Comme évoqué dans la section 3, la boucle de rétroaction basée sur les règles, et par extension le système global, contient une étape d'initialisation pour l'extraction des premières instances. Cette étape est primordiale car elle conditionne l'apprentissage futur de nouveaux schémas d'extraction et doit également garantir la performance de la boucle de rétroaction sémantique.

La figure 3 illustre un schéma d'extraction syntaxique se basant sur les schémas de (Hearst, 1992). Ce schéma représente la relation d'hyponymie qui peut apparaître dans le texte entre un concept et son instance. Le schéma ainsi construit est en réalité la superposition de trois séquences d'extraction.

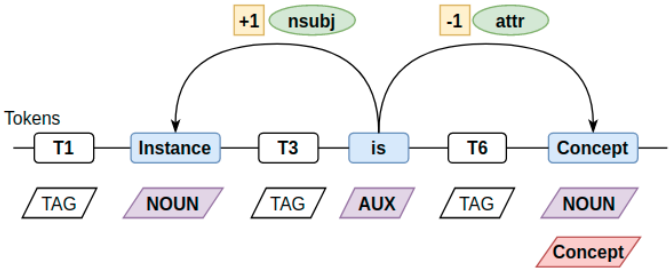

Figure 3. Exemple de schéma générique d'extraction pour les relations d'hyponymie.

La première séquence est relative aux POS-tags et décrit l'enchaînement de tags recherché. Dans l'exemple illustré par la figure 3, la séquence recherchée est la suivante : *(Concept)->(AUX)-> (NOUN)OR(PROPN)OR(PROP)*.

La deuxième séquence est relative aux tags affectés aux dépendances dans l'arbre des dépendances syntaxiques et décrit donc l'enchaînement des dépendances syntaxiques recherché. Dans cet exemple, la séquence recherchée est la suivante : *(attr)->(nsubj)*.

Une troisième séquence permet de préciser le sens de navigation dans l'arbre des dépendances syntaxiques, c'est-à-dire d'indiquer si le schéma remonte (-1) ou descend (+1) l'arbre des dépendances syntaxiques. Il est important de préciser que l'ordre de parcours du schéma d'extraction ne respecte pas, à priori, l'ordre de lecture du texte étudié. Dans l'exemple proposé, la séquence utilisée pour préciser la navigation est la suivante : *(-1)->(+1)*.
Au-delà de l'utilisation des dépendances syntaxiques pour plus de généricité, cette manière de définir des schémas d'extraction contient une particularité. En effet, si la séquence de tags utilisée est majoritairement constituée de POS-tags classiques, un tag supplémentaire est utilisé pour indiquer que le token représente un concept de l'ontologie. C'est ce tag qui déclenche la recherche d'une relation à partir de ce concept en suivant le schéma générique prédéfini. Le procédé de labellisation des concepts sur la base des classes de l'ontologie permettant d'assurer l'identification de ces concepts est effectué en amont, dans la chaîne de traitement automatique du langage (voir figure 2).
Si le schéma peut être respecté par les données du texte, alors une ou plusieurs instances peuvent être détectées. Dans l'exemple donné, un concept et une instance seront systématiquement détectés aux deux extrémités du schéma d'extraction, donnant lieu à la création de deux objets ontologiques liées par une relation d'hyponymie dans le modèle de données.

4.2 Mécanisme de déduction de schémas syntaxiques.

L'extraction par schémas génériques ou spécifiques est une méthode qui existe déjà depuis plusieurs années. La figure 4 donne une illustration globale du fonctionnement de la boucle de rétroaction basée sur les règles au sein du framework d'extraction. Cette figure présente le principe de bootstrapping mis en avant par (Pennacchiotti & Pantel, 2006).

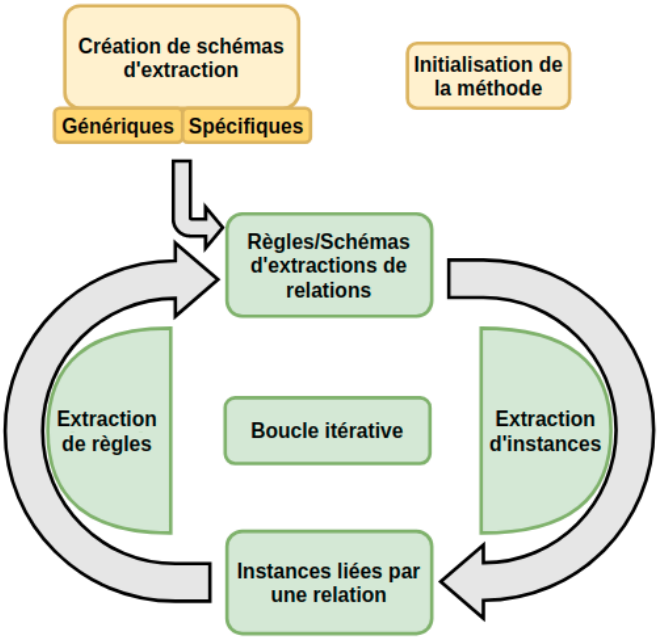

Figure 4. Illustration de l'approche bootstrapping pour la déduction de schémas d'extraction de relations.

Cette méthode consiste en l'auto-alimentation des étapes de déduction des schémas et d'application des schémas d'extraction. La déduction de schémas se fait en inversant le processus d'extraction. Ainsi, lorsque des instances connues apparaissent liées à leur concept dans les données, le principe de bootstrapping consiste à déduire un schéma qui lie ces deux instances. Si ce schéma est représentatif de la relation exprimée entre deux instances ou entre une instance et son concept associé alors celui-ci peut être ajouté à l'ensemble des schémas. Il est ensuite utilisé pour détecter de nouvelles relations dans les données. Cette méthode suppose donc bel et bien l'existence préalable de couples concept-instance identifiés, d'où l'importance de l'étape d'initialisation et de l'utilisation de schémas génériques prédéfinis par l'humain.
Dans le cadre du traitement de données textuelles, le schéma présenté dans cet article est un schéma syntaxique, appliqué sur l'arbre des dépendances syntaxiques préalablement construit. Toutefois, cette approche peut être étendue à d'autres types de schémas, applicables à d'autres types de données (structure, enchaînement de balises).

5 Méthode de mesure de la performance du système d'extraction

Le manque de données annotées comme frein à l'adoption d'une approche supervisée a été évoqué en introduction de la section 4. Cette problématique affecte également les démarches qui touchent à l'évaluation d'un système d'extraction. En effet, dans des tâches de traitement automatique de données, la performance d'une méthode est de manière générale évaluée par application de cette dernière sur un jeu de données de référence (communément identifié sous le terme de jeu de test). C'est le cas par exemple dans la plupart des tâches de machine learning qui font appel à des méthodes d'évaluation telles que le calcul du F1-Score et le tracé des courbes ROC (Fawcett, 2006). Malheureusement, dans un contexte où les données annotées ne sont pas répandues, et avec des méthodes non supervisées telles que celle présentée dans ce papier, ces méthodes sont limitées car :

- Elles ne permettent pas d'évaluer la performance d'un système lorsque celui-ci est appliqué à de nouveaux jeux de données.
- Elles ne tiennent pas compte des connaissances nouvelles, éventuellement extraites par le système mais ignorées au cours de l'annotation.

L'apport de la proposition faite dans cet article, est de traiter ces deux aspects en définissant une mesure de performance ne s'appliquant pas strictement aux extractions faites sur un jeu de test donné et défini à l'avance.

On supposera donc que les données que l'on traite ne sont pas, ou très rarement, annotées. Deux solutions peuvent alors être envisagées. La première consiste à labelliser manuellement les données une fois l'extraction réalisée. Cette méthode demande un investissement fort d'experts métiers spécialisés dans le domaine de l'ontologie, mais permet en contrepartie, de déterminer la précision du système de manière rigoureuse. La deuxième s'appuie sur d'éventuelles données de référence existantes indépendamment du processus de population. Cette deuxième méthode présente l'avantage de ne pas faire intervenir l'expert du domaine pour effectuer une validation manuelle. Cependant, si elle est rapide à mettre en place, elle ne permet pas de déterminer avec rigueur la précision du modèle. Cette section se concentre sur la deuxième méthode, appliquée dans le cadre de relations concept-instance et qui demande une redéfinition des mesures de performance. Pour aborder ce problème, il convient de redéfinir les notions de *concept*, d'*instance* et de *couple*. Un concept est ici la représentation dans les données (référence ou extraites) d'une des classes initialement contenues dans l'ontologie.

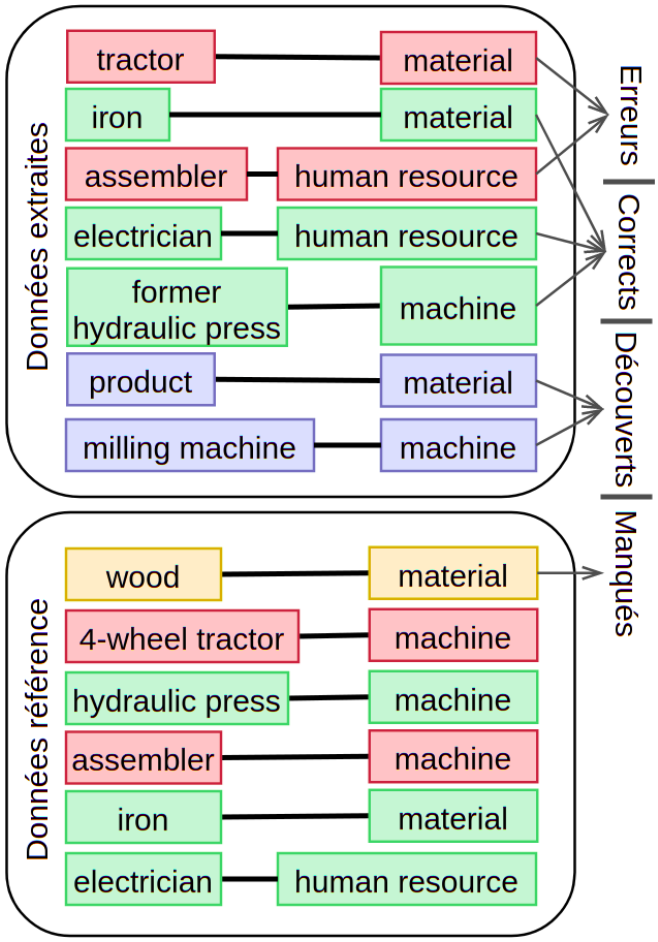

Figure 5. Représentation des jeux de données extrait et référence exemples.

Une instance (reliée à un concept) est la représentation dans les données de l'instance d'une classe de l'ontologie, qui a été extraite ou qui fait partie des données de référence. Enfin un couple est la réunion d'un concept et d'une instance (qu'il soit issu de l'extraction ou des données de référence) reliés par une relation. Pour illustrer à titre d'exemple la mesure, celle-ci sera appliquée aux jeux de couples extraites et référence simulés présentés dans la figure 5. Il est important de préciser que ces jeux de couples sont utilisés à titre d'exemple, et ne sont pas liés à un jeu de données en particulier. En revanche, des mesures de performance utilisant la méthode ont été réalisées sur un jeu de données réel. Ces résultats sont présentés en section 5.4.

5.1 Réconciliation des données extraites et référence par le ROUGE Score.

L'évaluation du système doit donc pouvoir se faire à partir de données de référence qui ne sont pas nécessairement liées au texte étudié, mais qui constituent un bon exemple de la base de connaissance qui est visée à travers la population de l'ontologie du domaine ciblé.

La problématique induite par ce contexte, est la non correspondance exacte des données extraites aux données de référence, ce qui ne signifie pas pour autant une erreur de la part du système d'extraction. En effet, dans la mesure où sont comparées des chaînes de caractères (représentatives des instances extraites), il peut arriver que certaines d'entre elles diffèrent sans pour autant que leur sens ne s'en voit modifié. Afin d'éviter les défauts d'appréciation dus à cette spécificité, il est possible de définir une distance entre les chaînes de caractères et d'y appliquer un seuil.

De nombreuses métriques existent pour faire le calcul de la distance entre deux chaînes de caractères. Parmi elles, la distance de Levenshtein (Levenshtein, 1966) ou encore la distance de Hamming (Hamming, 1950) font référence. Cependant, ces distances sont généralement utilisées sur des chaînes de caractères soit déjà fortement similaires, soit très longues, ce qui n'est pas nécessairement le cas ici. De plus, ces mesures peuvent se révéler très inefficaces pour la comparaison de chaînes de caractères proches d'un point de vue purement basé sur les caractères mais ayant une sémantique totalement différente. Par exemple, les chaînes de caractères *iron* et *wood* ont une distance de Levenshtein relativement faible (6) mais représentent cependant des instances différentes.

L'évaluation du ROUGE Score (Lin, 2004) entre deux chaînes de caractères a donc été retenue pour caractériser l'égalité des instances qu'elles représentent. Le ROUGE Score est un indice de similarité simple qui adopte une méthode de mesure semblable à la distance de Levenshtein à la différence près que ce dernier est adapté à l'échelle des termes (et non pas à l'échelle du caractère). De cette manière, les chaînes de caractères *iron* et *wood* apparaissent bien comme plus éloignées (ROUGE Score = 0) que les chaînes de caractères *hydraulic press* et *former hydraulic press* par exemple (ROUGE Score = 0.79), ce qui n'est pas le cas si les calculs de distance sont effectués à l'échelle du caractère comme dans le cadre de mesures de distance classiques (Levenshtein(*wood*, *iron*) = 6, Levenshtein(*hydraulic press*, *former hydraulic press*) = 7) . Couplée à un seuil d'acceptation, la valeur du ROUGE Score fait donc office de validateur de l'égalité entre deux instances représentées par deux chaînes de caractères.

5.2 Construction des ensembles Corrects, Manqués, Découverts, et Erreurs

La méthode de comparaison entre les données extraites et les données de référence établie, il est possible de catégoriser les instances extraites selon que celles-ci ont été affectées ou non au concept leur correspondant dans les données de référence.

Comme les données de référence ne sont pas nécessairement liées au texte utilisé pour l'extraction, il peut cependant arriver que certaines instances des données de référence ne soient pas retrouvées dans les données extraites et inversement, que certaines données extraites ne soient pas représentées dans les données de référence. Ainsi quatre ensembles sont définis dans cette section afin de traduire cette spécificité :

- *Corrects* (C) : L'ensemble *Corrects* contient les couples dont les instances sont considérées identiques dans le jeu de données référence et dans le jeu de données extrait et pour lesquelles le concept associé est conforme au concept du jeu de référence.
- *Erreurs* (E) : L'ensemble *Erreurs* contient les couples extraits dont l'instance apparaît bien dans le jeu annoté mais est associée à un concept différent de celui du jeu annoté, soit par erreur, soit par sur-classification.
- *Manqués* (M) : L'ensemble *Manqués* contient les couples du jeu de données de référence qui n'apparaissent pas dans le jeu de données extrait, parce que l'instance n'a pas été remontée du tout dans les données extraites.
- *Découverts* (D) : L'ensemble *Découverts* contient les couples identifiés par le système mais dont l'instance n'apparaît pas dans le jeu de données annoté.

Le critère d'égalité entre deux instances est déclaré lorsque la distance entre ces deux instances passe un seuil défini par l'évaluateur. La distance utilisée s'appuie quant à elle sur le ROUGE Score. Pour l'exemple, la valeur seuil utilisée est fixée à 0.6. Les ensembles *E* et *C* contiennent des couples extraits dont les instances sont égales au sens de la distance définie. L'ensemble *D* contient des couples extraits dont l'instance ne présente de similarité suffisante avec aucune des instances de référence pour être classée dans l'un des ensembles *C* ou *E*.

L'ensemble *M,* à l'inverse, contient des couples de référence qui n'apparaissent pas dans les données extraites et ne peuvent être classées ni dans *C*, ni dans *E*, ni dans *D*. Le fait de ne pas inclure dans *M* les couples dont l'instance apparaît, dans les couples extraits, associée à un concept différent est un choix fait pour distinguer les couples manqués des couples mal définis (erreur sur le concept). Cette décision induit un biais car elle ne prend pas en compte les cas particuliers où une instance a été reliée par le système à seulement une partie des concepts présents dans les couples impliquant l'instance dans les données de référence. Par souci de clarté, ce cas de figure n'est pas présenté dans l'exemple.

5.3 Estimation de la précision et du rappel

Les ensembles *M* et *D* présentés dans la section 5.2 sont singuliers car ils concernent des couples dont les instances ne sont présentes que dans un seul des deux jeux de données.

Des couples peuvent ainsi apparaître dans l'ensemble *M* soit parce que l'instance associée au concept n'apparaît pas dans les données et n'a donc pas pu être extraite, soit parce que celle-ci n'a pas été extraite, malgré sa présence dans le texte. Dans le premier cas, le couple ne peut pas être estimé comme un réel manquement du système. Dans le deuxième, il s'agit bel et bien d'une instance qui n'a pas été extraite.

De la même manière, il est compliqué d'estimer la valeur des couples contenus dans l'ensemble *D* qui peuvent être soit de nouveaux couples pertinents, qui n'étaient pas listés dans les données de référence, soit un ensemble d'erreurs de la part du système (faux positifs).

En l'absence d'annotation manuelle, il devient alors difficile d'établir exactement les performances du système sans formuler d'hypothèses sur la nature des couples contenus dans les ensembles *M* et *D*. Il est toutefois possible d'en donner une estimation en supposant que les couples de l'ensemble *M* sont bel et bien des couples manqués et que les couples de l'ensemble *D*, comme ils ne font pas partie du jeu de référence, ne constituent pas une connaissance suffisante pour être considérés comme des vrais positifs. En suivant ces hypothèses, on peut définir génériquement les sous-ensembles Ens_{ci} et $\overline{Ens_{ci}}$:

$$Ens_{c_i} = \{cpl = (con, ins) \in Ens \mid con = c_i\}$$

$$\overline{Ens_{c_i}} = \{cpl = (con, ins) \notin Ens \mid con = c_i\}$$

et ainsi redéfinir la matrice de confusion. Les vrais positifs sont alors estimés à partir de l'ensemble C_{ci}, les faux positifs sont estimés à partir des ensembles E_{ci} et D_{ci}, les vrais négatifs à partir des ensembles $\overline{E_{ci}}$, $\overline{D_{ci}}$, et $\overline{C_{ci}}$, et les faux négatifs à partir de l'ensemble M_{ci}:

$$VP_{c_i} = \sum_{cpl \in C_{c_i}} sim_{cpl}$$

$$FP_{c_i} = \sum_{cpl \in E_{c_i} \cup D_{c_i}} sim_{cpl}$$

$$VN_{c_i} = \sum_{cpl \in \overline{E_{c_i}} \cup \overline{C_{c_i}} \cup \overline{M_{c_i}} \cup \overline{D_{c_i}}} sim_{cpl}$$

$$FN_{c_i} = \sum_{cpl \in M_{c_i}} sim_{cpl}$$

Où sim_{cpl} vaut 1 pour les couples des ensembles *D* et *M* et correspond à l'indice de similarité calculé à partir du ROUGE Score entre un couple extrait et son homologue au sein des données de référence pour les couples des ensembles *C* et *E*.

En effet, pour les couples appartenant aux ensembles C et *E*, le critère d'égalité a été déclaré à partir d'un calcul de similarité. Pour prendre en compte cette similarité, plutôt que d'ajouter une unité aux groupes de faux positifs ou vrais positifs, c'est la valeur du ROUGE Score qui est utilisée. Cette mesure étant comprise entre 0.6 (valeur seuil) et 1, elle constitue un marqueur du degré de similarité entre les deux couples. Ainsi, un couple considéré comme appartenant à l'ensemble *C* car il a juste dépassé le seuil d'admission fixé aura moins de poids qu'un couple pour lequel la valeur du ROUGE Score vaut 1 car ce dernier est parfaitement identique au couple de référence. Pour l'exemple présenté, on obtient par cette méthode les valeurs détaillées dans le tableau 1 pour chacun des concepts de la figure 5.

Tableau 1. Calcul des matrices de confusion et des performances après définition de *E*, *C*, *M* et *D*.

	human resource	**machine**	**material**
VP	1	0.79	1
FP	1	1	1.66
VN	5.45	5.66	3.79
FN	0	0	1
Précision	0.50	0.56	0.38
Rappel	1	1.00	0.74
F1-Score	0.67	0.72	0.50

Le F1-Score global peut quant à lui être calculé en pondérant les F1-Scores de chaque concept par la participation (en termes de similarité), de chacun des couples dans lesquels il intervient. Dans cet exemple simple, on obtient un F1-Score global égal à 0.6.

5.4 Application de la méthode à un jeu de données traitant de la biochimie.

La méthode illustrée dans les sections précédentes a été appliquée en utilisant un jeu de données référence lié au domaine de la biochimie (abstracts annotés) et en réalisant une extraction sur des articles de recherche associés à ce jeu de données référence (Shardlow et al., 2018). Il est important de noter ici que les données de référence et les données sur lesquelles est appliquée l'extraction sont distinctes, mettant en avant l'intérêt de la méthode proposée.

Le tableau 2 indique le volume (somme des similarités) des

ensembles C, E, M et D pour un seuil d'égalité fixé à 0.5. On obtient pour ce jeu de données une valeur de précision de 0.65. Il s'agit néanmoins, du fait de la méthode, et pour les raisons évoquées dans la section 5.3, d'une évaluation pessimiste de la précision. La valeur du rappel est quant à elle très faible (environ 5%) dans la mesure où de nombreux couples sont considérés comme manqués par l'extraction. Encore une fois, cette estimation stricte est discutée dans la section 5.3.

Tableau 2 : Volume (somme des similarités) des différents ensembles, après extraction sur des données liées au domaine de la biochimie (Shardlow et al., 2018).

Corrects (C)	142.27
Erreurs (E)	10.83
Manqués (M)	2503
Découverts (D)	78

5.5 Discussion sur la méthode d'évaluation des performances.

Il est important de souligner que la définition des vrais positifs, faux positifs, vrais négatifs et faux négatifs adopte une vision pessimiste de l'extraction pour les raisons évoquées au début de la section 5.3. Le pendant optimiste consisterait à considérer que :

- Les couples contenus dans M ne peuvent pas être assimilés à des faux négatifs dans la mesure où l'on suppose que ceux-ci n'apparaissent pas dans les données étudiées.
- Les couples contenus dans D ne sont plus assimilés à des faux positifs mais à des vrais positifs.

Néanmoins, cette vision conduit à un rappel faussement maximal (égal à 1), et à une précision dopée, qui ne seront que très rarement représentatifs des performances réelles de l'extraction. Une meilleure estimation de la précision peut toutefois être envisagée en ne traitant pas le cas des couples de l'ensemble D et en se limitant aux ensembles C et E pour définir vrais positifs et faux positifs.

Par ailleurs, on omet dans la définition de M les couples non détectés mais dont l'instance a été détectée dans un autre couple ce qui éloigne la mesure du rappel de la réalité. Une autre définition de M, incluant ces couples, reviendrait à considérer dans certains cas deux fois un même résultat du système, d'abord comme une erreur du point de vue des données extraites, puis comme un manquement du point de vue des données de référence.

Enfin, une vérification automatique de la présence des instances des couples de M dans le texte sur lequel l'extraction a été réalisée est une solution qui permet de mieux estimer la part de réels faux négatifs dans l'ensemble M. Ainsi seuls les couples dont les instances sont présentes dans le texte traité peuvent être considérées dans la définition des faux négatifs.

6 Conclusion

Nous avons présenté dans cet article des méthodes d'évaluation de la performance pour le cas particulier de l'extraction de relations d'hyponymie entre concept et instance au sein de données textuelles non annotées. La méthode proposée permet d'estimer la performance d'un tel modèle d'extraction à partir de données de référence qui ne sont pas nécessairement liées ou construites à partir de la source de données dont sont extraites les relations à évaluer.

Cette définition de nouvelles méthodes d'évaluation des performances s'inscrit dans une démarche plus globale de population d'ontologie dont a été présentée la stratégie générale, et détaillée l'une des méthodes permettant de mener à bien l'extraction de connaissances à partir de données textuelles.

Néanmoins, la méthode d'évaluation définie repose en grande partie sur la mesure d'égalité qui a également ses propres limites. Le ROUGE Score ne prend pas par exemple en compte la dimension sémantique entre deux instances. Cette méthode de mesure des performances doit ainsi faire l'objet de tests plus poussés, sur des jeux de données représentatifs afin d'éprouver sa pertinence. L'étude d'autres mesures de distances, couplées au ROUGE-Score, pour définir l'égalité entre deux couples est également envisagée.

7 Références

Agichtein, E., & Gravano, L. (2000). *Snowball* : Extracting relations from large plain-text collections. *Proceedings of the fifth ACM conference on Digital libraries*, 85-94. https://doi.org/10.1145/336597.336644

Ayadi, A., Samet, A., de Beuvron, F. de B., & Zanni-Merk, C. (2019). Ontology population with deep learning-based NLP : A case study on the Biomolecular Network Ontology. *Procedia Computer Science*, *159*, 572-581. https://doi.org/10.1016/j.procs.2019.09.212

Chasseray, Y. (2020). Un meta-modèle pour la population d'ontologie indépendamment du domaine. *Forum Jeunes Chercheuses Jeunes Chercheurs*, 17-20.

Chasseray, Y., Barthe-Delanoë, A.-M., Négny, S., & Le Lann, J.-M. (2021). A generic metamodel for data extraction and generic ontology population. *Journal of Information Science*, 0165551521989641. https://doi.org/10.1177/0165551521989641

Cheng, H., Zeng, P., Xue, L., Shi, Z., Wang, P., & Yu, H. (2016). Manufacturing Ontology Development Based on Industry 4.0 Demonstration Production Line. *2016 Third International Conference on Trustworthy Systems and Their Applications (TSA)*, 42-47. https://doi.org/10.1109/TSA.2016.17

de Boer, V., van Someren, M., & Wielinga, B. J. (2007). A redundancy-based method for the extraction of relation instances from the Web. *International Journal of Human-Computer Studies*, *65*(9), 816-831. https://doi.org/10.1016/j.ijhcs.2007.05.002

Faria, C., Serra, I., & Girardi, R. (2014). A domain-independent process for automatic ontology population from text. *Science of Computer Programming*, *95*, 26-43. https://doi.org/10.1016/j.scico.2013.12.005

Fawcett, T. (2006). Introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques : A survey. *Knowledge-Based Systems*, *70*, 301-323. https://doi.org/10.1016/j.knosys.2014.07.007

Geng, Z., Chen, G., Han, Y., Lu, G., & Li, F. (2020). Semantic relation extraction using sequential and tree-structured LSTM with attention. *Information*

Sciences, *509*, 183-192. https://doi.org/10.1016/j.ins.2019.09.006
Giustozzi, F., Saunier, J., & Zanni-Merk, C. (2018). Context Modeling for Industry 4.0 : An Ontology-Based Proposal. *Procedia Computer Science*, *126*, 675-684. https://doi.org/10.1016/j.procs.2018.08.001
Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199-220. https://doi.org/10.1006/knac.1993.1008
Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, *29*(2), 147-160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x
Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *COLING*.
Herbelot, A., & Copestake, A. (2006). *Acquiring ontological relationships from Wikipedia using RMRS*.
Kaushik, N., & Chatterjee, N. (2018). Automatic relationship extraction from agricultural text for ontology construction. *Information Processing in Agriculture*, *5*(1), 60-73. https://doi.org/10.1016/j.inpa.2017.11.003
Kim, S.-J., Lee, Y.-H., & Lee, J.-H. (2009). *Method of Extracting Is-A and Part-Of Relations Using Pattern Pairs in Mass Corpus ■*. 9.
Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707. http://adsabs.harvard.edu/abs/1966SPhD...10..707L
Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*. 8.
Louge, T., Karray, M.-H., & Archimède, B. (2018). Investigating a Method for Automatic Construction and Population of Ontologies for Services : Performances and Limitations. *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, 1-6. https://doi.org/10.1109/AICCSA.2018.8612844
Luo, L., Yang, Z., Cao, M., Wang, L., Zhang, Y., & Lin, H. (2020). A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of Biomedical Informatics*, *103*, 103384. https://doi.org/10.1016/j.jbi.2020.103384
Martinez-Rodriguez, J. L., Lopez-Arevalo, I., & Rios-Alvarado, A. B. (2018). OpenIE-based approach for Knowledge Graph construction from text. *Expert Systems with Applications*, *113*, 339-355. https://doi.org/10.1016/j.eswa.2018.07.017
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. 9.
Nguyen, K. A., Köper, M., Walde, S. S. im, & Vu, N. T. (2017). Hierarchical Embeddings for Hypernymy Detection and Directionality. *arXiv:1707.07273 [cs]*. http://arxiv.org/abs/1707.07273
Niladri Chatterjee & Neha Kaushik. (2017). RENT : Regular Expression and NLP-Based Term Extraction Scheme for Agricultural Domain. In S. C. Satapathy, V. Bhateja, & A. Joshi (Éds.), *Proceedings of the International Conference on Data Engineering and Communication Technology* (p. 511-522). Springer Singapore.
Oramas, S., Espinosa-Anke, L., Sordo, M., Saggion, H., & Serra, X. (2016). Information extraction for knowledge base construction in the music domain. *Data & Knowledge Engineering*, *106*, 70-83. https://doi.org/10.1016/j.datak.2016.06.001
Paukkeri, M.-S., García-Plaza, A. P., Fresno, V., Unanue, R. M., & Honkela, T. (2012). Learning a taxonomy from a set of text documents. *Applied Soft Computing*, *12*(3), 1138-1148. https://doi.org/10.1016/j.asoc.2011.11.009
Pennacchiotti, M., & Pantel, P. (2006). *A Bootstrapping Algorithm for Automatically Harvesting Semantic Relations*.
Rajpathak, D. G. (2013). An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain. *Computers in Industry*, *64*(5), 565-580. https://doi.org/10.1016/j.compind.2013.03.001
Shardlow, M. J., Nguyen, N., Owen, G., O'Donovan, C., Leach, A., McNaught, J., Turner, S., & Ananiadou, S. (2018). A New Corpus to Support Text Mining for the Curation of Metabolites in the ChEBI Database. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 280-285. http://www.lrec-conf.org/proceedings/lrec2018/summaries/229.html
Snow, R., Jurafsky, D., & Ng, A. Y. (2005). *Learning Syntactic Patterns for Automatic Hypernym Discovery*. 8.
Sparck Jones, K. (2004). A Statistical Interpretation of Term Specificity in Retrieval. *Journal of Documentation*, *60*, 493-502. https://doi.org/10.1108/00220410410560573
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*. http://arxiv.org/abs/1706.03762