Un système d'aide à l'extraction de relations sémantiques pour la construction d'ontologies à partir de textes

Rim Bentebibel*, Adeline Nazarenko*
Sylvie Szulman*

* Laboratoire d'Informatique de l'université Paris-Nord (LIPN)
UMR 7030 Université Paris 13 & CNRS
99, avenue Jean-Baptiste Clément
93430 Villetaneuse

prénom.nom@lipn.univ-paris13.fr

Résumé. Cet article présente une méthode d'extraction de relations sémantiques pour la construction d'ontologies à partir de corpus de textes. Notre objectif est de proposer une méthode générique, qui soit indépendante du domaine et de la langue. Elle repose sur une analyse distributionnelle des unités sémantiques du corpus pour faire émerger des relations sémantiques candidates. Cette méthode ne fait aucune hypothèse sur les types de relations recherchées ni sur leur forme linguistique. Il s'agit de regrouper les associations de termes dans des classes qui représentent des relations sémantiques candidates. L'hypothèse sous-jacente est que les occurrences de ces associations réunies sur la base des éléments de contexte qu'elles partagent ont des chances de relever d'une même relation sémantique et que les relations candidates ainsi proposées peuvent aider le travail de conceptualisation de l'ontologue.

1 Introduction

Les textes sont des sources précieuses pour la construction d'ontologies parce qu'ils portent la trace de connaissances stabilisées et partagées et qu'ils sont souvent plus faciles d'accès que les experts. Les méthodes de construction d'ontologies à partir de textes sont aujourd'hui bien connues : pour identifier les concepts du domaine, elles s'appuient sur l'analyse terminologique pour les unes, sur l'analyse distributionnelle et les classes de mots pour les autres.

Au-delà des concepts et de leurs instances, il est important de repérer les relations conceptuelles qui structurent le domaine. Des approches distributionnelles ont été proposées, mais pour l'explicitation des relations hiérarchiques uniquement. Les approches classiques, héritées de la terminologie traditionnelle, permettent d'extraire aussi des relations transversales. Elles explorent les textes à l'aide de patrons mais ceux-ci diffèrent pour chaque relation et varient souvent d'un corpus à l'autre. Nous proposons ici une méthode générique de découverte de relations sémantiques à partir de textes. Il s'agit d'explorer les textes pour identifier les relations conceptuelles qu'ils véhiculent sans idée préconçue sur le type de relations qu'on recherche.

Bien entendu, il ne saurait s'agir d'acquérir automatiquement un ensemble de relations conceptuelles directement intégrables sous la forme de rôles dans une ontologie. Un travail manuel de validation et de conceptualisation des relations candidates proposées est nécessaire. L'originalité de la méthode proposée est double : elle est guidée par le corpus lui-même (les données) plutôt que par les relations à acquérir (le but) et elle est générique par rapport au corpus et à la tâche. Il s'agit, comme pour l'extraction terminologique, de faire « émerger » la sémantique du domaine du corpus, même si les résultats ont besoin d'être retravaillés.

La section 2 situe notre travail par rapport à l'état de l'art en acquisition de relations et souligne l'intérêt d'une approche guidée par le corpus par rapport aux approches guidées par les relations à construire. La section 3 décrit les différentes étapes de notre méthode et l'ensemble est illustré et évalué dans la section 4 sur un corpus particulier. La section 5 présente une discussion et les perspectives de ce travail.

2 Etat de l'art

Beaucoup de travaux sur l'acquisition d'ontologies à partir de textes s'appuient sur des patrons d'extraction pour retrouver des relations dans les textes (Auger et Barriere, 2008). Ces méthodes, qui s'inspirent à la fois des pratiques traditionnelles en terminologie (Pearson, 1998) et des méthodes classiques d'extraction d'informations, bien connues depuis (Hearst, 1992), permettent notamment d'extraire l'information relationnelle qui est mentionnée explicitement dans les textes.

Comme toujours en extraction d'information, il s'agit de « savoir ce qu'on cherche ». Si on recherche des relations d'hyperonymie, par exemple, on applique des patrons caractéristiques de cette relation (de type `des <N1>, <N2>...et autres <N3>`). Toute la difficulté consiste alors à constuire ces patrons caractéristiques, qui varient d'une relation à l'autre et d'un corpus à l'autre pour une même relation (Jacques et Aussenac-Gilles, 2006). Pour réduire le coût de mise au point des patrons d'extraction nécessaires qui, pour la construction d'une ontologie donnée, est vite apparu prohibitif, on a cherché à tirer profit des recherches menées sur l'apprentissage de patrons, notamment les approches semi-supervisées Pour se passer d'un corpus annoté – souvent aussi difficile à produire que les patrons qu'on cherche à en tirer ! – ces approches partent d'un petit nombre d'exemples de couples d'éléments (e_1, e_2) connus pour entretenir la relation R recherchée. La projection de ce couple en corpus permet de trouver des phrases caractéristiques de la relation R dont on peut extraire un patron qui est projeté à son tour en corpus pour trouver de nouveaux couples d'éléments entretenant la relation R. Ces nouveaux couples servent ensuite à trouver d'autres patrons et ainsi de suite. Les couples `<microsoft, redmond> ...<Intel, Santa Clara>` permettent ainsi de construire le patron `the <LOCATION> based in <ORGANIZATION>`.

Ces approches sont guidées par le but plus que par les données. Etant donné une relation, on fouille le corpus à la recherche d'occurrences de cette relation et/ou de patrons caractéristiques de cette relation. Cette approche a évidemment son intérêt pour la construction d'ontologies.

Une autre approche est proposée par (Hasegawa et al., 2004) pour enrichir des systèmes de question/réponse ou de résumé de textes : elle vise à repérer des relations sémantiques entre entités nommées. Les entités nommées, qui sont généralement des noms propres, sont intéressantes à considérer car elles désignent des entités bien identifiées du domaine. Les relations entre ces entités ont donc des chances d'être pertinentes pour l'ontologie du domaine. La mé-

thode consiste à faire émerger des classes homogènes de couples d'entités nommées, chaque classe pouvant représenter une relation intéressante pour le domaine. Etant donnés par exemple A et B, deux types d'entités nommées fréquemment associées, on cherche à décomposer l'ensemble des cooccurrences de A et B en groupes sémantiquement homogènes sur la base de la similarité des contextes où elles apparaissent. Chaque groupe d'occurrences reflète une même relation sémantique candidate, qui a été ainsi mise au jour même si elle reste à nommer. Nous nous inspirons de cette méthode en cherchant à en généraliser l'application à d'autres types d'unités textuelles, notamment aux termes qui servent eux aussi d'ancres pour la découverte de relations. Nous voulons également étendre la méthode pour permettre d'associer des patrons d'extraction, ou tout au moins des ébauches de patrons, aux relations sémantiques candidates.

En réalité, la méthode que nous présentons s'apparente aux méthodes distributionnelles de construction de classes sémantiques de mots (voir par ex. (Faure et Nédellec, 1999)) qui rapprochent les mots sur la base des éléments des contextes qu'ils partagent et qui proposent les classes obtenues comme ébauches de concepts. Sauf qu'il s'agit ici de construire des classes d'associations de termes et non pas des classes de termes.

3 Méthode

Notre méthode d'extraction de relations est constituée de quatre étapes (voir figure 1).

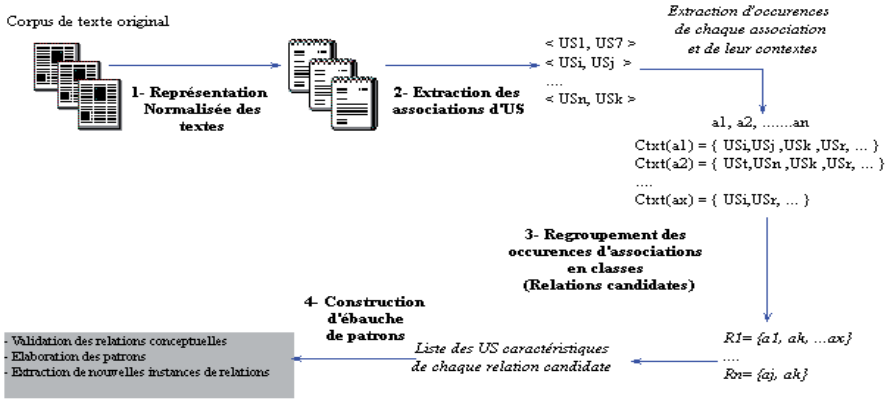

FIG. 1 – *Schéma général de la méthode d'extraction de relations*

3.1 Construction d'une représentation normalisée des textes (étape 1)

La première étape consiste à construire une représentation simplifiée et normalisée du corpus d'acquisition. Comme le but est de construire une ontologie, nous ne nous intéressons pas aux mots du texte mais aux termes qui relèvent du vocabulaire du domaine et qui sont souvent des unités lexicales composées. Nous considérons aussi les entités nommées comme

des éléments sémantiquement pertinents. Une phrase se représente donc comme une séquence d'*unités sémantiques* (US) qui sont des entités nommées ou des termes.

En pratique, nous privilégions les termes les plus longs et, comme l'extracteur de termes que nous utilisons [1] peut laisser passer des termes simples et ne reconnaît pas les termes verbaux, nous conservons aussi les mots « sémantiquement pleins » après élimination des mots grammaticaux et des mots athématiques figurant dans un antidictionnaire (ex. *sorte, faire, être*, etc.). Nous considérons les formes lemmatisées des unités sémantiques [2].

Ce choix de représentation efface toute information syntaxique – une fois les unités sémantiques identifiées, seul l'ordre des unités est conservé – mais simplifie les phrases et facilite donc leur rapprochement. Cette normalisation est importante puisque notre approche repose sur la récurrence des unités et de leurs associations dans le corpus, à la différence des méthodes à base de patrons qui s'appuient davantage sur la structure des phrases. La figure 2 montre que le corpus est représenté récursivement à l'issue de cette étape de normalisation comme une séquence de documents, de phrases et d'unités sémantiques.

```
- <Corpus>
  - <Doc1>
    - <Sentence Id="1">
        <Unité_Sémantique type="NamedEnt" value="nuit avant l'accident" Pos="NN-Prep-Art-NN" lemme="nuit-avant-l'accident"/>
        <Unité_Sémantique type="Terme" value="conditions météorologiques" Pos="NN_NN" lemme="condition-météorologique" />
        <Unité_Sémantique type="word" value="avaient été variées" Pos="Vpas" lemme="varier" />
      </Sentence> .....
    </Doc1> .....
  - <Docn>
      <Sentence Id="1" /> .....
      <Sentence Id="2" /> .....
    </Docn>
  </Corpus>
```

FIG. 2 – *Représentation xml du texte comme séquence d'unités sémantiques*

3.2 Extraction des associations d'unités sémantiques (étape 2)

Notre méthode repose sur le repérage dans le corpus de couples d'unités sémantiques fortement associées, ces associations étant potentiellement des indices de relations sémantiques du domaine. Pour extraire les associations d'unités sémantiques, nous nous appuyons sur un calcul de cooccurrence : plus les unités apparaissent ensemble (dans les mêmes phrases), plus elles sont considérées comme fortement associées.

Il existe plusieurs manières de mesurer la force de cette association. Dans nos premières expériences, nous avons considéré la simple fréquence des associations, cette mesure devant nous permettre ultérieurement de tester et de comparer des mesures alternatives pour apprécier leurs effets respectifs puis ajuster notre méthode en fonction. Cette force d'association est calculée pour tous les couples d'unités sémantiques présentes dans le corpus. Nous conservons au final comme « associations » tous les couples d'unités sémantiques dont la force est supérieure à un seuil donné [3].

1. YaTeA (Aubin et Hamon, 2006).
2. La lemmatisation du corpus est effectuée par TreeTager (Schmid, 1994) et l'extracteur de Gate est utilisé pour la reconnaissance des entités nommées.
3. Pour la mise au point de la méthode, ce seuil a volontairement été fixé bas et déterminé expérimentalement, voir la section 4.

3.3 Regroupement des occurrences d'association en classes (étape 3)

On applique ensuite un processus de classification ascendante (ou regroupement, *clustering* en anglais) sur toutes les occurrences a_k des associations $< US_i, US_j >$ identifiées à l'étape précédente [4]. Il s'agit de regrouper en classes les occurrences d'associations les plus similaires.

Pour calculer la similarité entre les occurrences d'association, on représente chaque occurrence a_k par son contexte $ctxt(a_k)$, *i.e.* par l'ensemble des unités sémantiques figurant dans la même phrase que a_k. Considérons par exemple l'association $< US_1, US_{12} >$. Si on a $ctxt(a_1) = \{US_{20}, US_{33}, US_{45}\}$, cela signifie que ces trois unités sémantiques apparaissent aux côtés de US_1 et US_{12} dans la phrase où cette occurrence particulière a_1 de l'association apparaît. Plus formellement, $ctxt(a_k)$ est un vecteur dans l'espace du vocabulaire des unités sémantiques du corpus. La coordonnée de $ctxt(a_k)$ sur l'axe de l'unité US_i est 1 si US_i figure dans le contexte de a_k et 0 sinon. L'ensemble des vecteurs de contexte se représente sous la forme d'une matrice (M_{ij}) comme dans l'exemple de la figure 3 qui décrit deux associations : <US_1, US_{12}> et <US_{19}, US_{24}>, avec leurs occurrences respectives : a_1, a_2, a_3 et a_4, a_x.

$Ctxt(a_1) = \{US_{20}, US_{33}, US_{45}\}$
$Ctxt(a_2) = \{US_4, US_{20}, US_{33}\}$
$Ctxt(a_3) = \{US_{20}, US_{31}\}$
$Ctxt(a_4) = \{US_1, US_4, US_{95}, US_{120}\}$
...
$Ctxt(a_x) = \{US_{20}, US_{31}, US_{33}\}$

	US 1..	US 4 ..	US 20 ..	US 33 ..	US 45 ..	US n
a1	0 ..	0 ..	1 ..	1 ..	1 ..	0
a2	0 ..	1 ..	1 ..	1 ..	0 ..	0
a3	...	..	1 ..	0 ..	0 ..	0
a4	1 ..	1 ..	0 ..	0 ..	0 ..	
a5	0 ..	0 ..	0 ..	0 ..	..	0
...		0 ..	..	..	1 ..	1
ax	0	0 ..	1 ..	1 ..	0 ..	0

FIG. 3 – *Exemple de vecteurs de contexte et matrice associée*

On regroupe ensuite les contextes dont les unités sémantiques sont les plus similaires entre elles par la méthode de l'analyse relationnelle (Benhadda et Marcotorchino, 1998), qui permet de classifier de manière non supervisée des données textuelles. Cette méthode ne présupose pas la connaissance *a priori* de la structure du corpus et elle n'impose pas de fixer au départ le nombre de classes, ce qui permet une détermination automatique non biaisée des partitions. D'un point de vue technique, cette méthode est linéaire et permet de traiter en un temps très raisonnable de grands corpus (Ah-Pine et Jacquet, 2009). De manière classique, cette méthode repose sur la définition d'une mesure de similarité (ou dissimilarité) entre deux contextes, et un critère à optimiser, en l'occurrence le critère de Condorcet. Pour déterminer à quel point deux contextes $ctxt(a_x)$ et $ctxt(a_y)$ (noté i et i' dans la formule) sont similaires, nous utilisons la similarité de Condorcet $C_{ii'}$ (formule 1) où i et i' sont deux contextes et j parcourt l'ensemble des US. A partir de cette matrice, on construit une matrice de dissimilarité $\overline{C_{ii'}}$ (formule 2)

$$C_{ii'} = \sum_j (M_{ij}.M_{i'j}) \quad (1) \qquad C_{ii'} = \frac{\sum_j M_{ij} + \sum_j M_{i'j}}{2} - C_{ii'} \quad (2)$$

Le critère de Condorcet à maximiser est défini par la formule 3, la matrice X représentant la partition à obtenir.

$$C(X) = \sum_i \sum_{i'} (C_{ii'} X_{ii'}) + \sum_i \sum_{i'} (\overline{C_{ii'} X_{ii'}}) \quad (3)$$

4. On trouve en général plusieurs occurrences de d'une même association dans le corpus.

$$\text{où } X_{ii'} = \begin{cases} 1 & \text{si les contextes i et i' appartiennent à une même classe} \\ 0 & \text{sinon} \end{cases}$$

$$\text{et } \overline{X_{ii'}} = 1 - X_{ii'} = \begin{cases} 1 & \text{si les contextes i et i' n'appartiennent pas à une même classe} \\ 0 & \text{sinon} \end{cases}$$

En reprenant l'exemple précédent, l'analyse relationnelle met en évidence la similarité des occurrences de contextes a_1, a_2 et a_x, et les regroupe dans une même classe qui peut être présentée comme une relation sémantique candidate.

L'intuition sous-jacente est que les occurrences d'association regroupées partagent un même « sens », une même relation sémantique. Comme les classes obtenues sont néanmoins bruitées, il faut les analyser, en supprimer ou en redécouper certaines, y supprimer des intrus et, au final, nommer la relation sous-jacente si la classe apparaît suffisamment cohérente. L'intérêt de ce type d'approche, aussi bruitée soit-elle, est de faire émerger du sens en rapprochant des éléments diffus dans le corpus, ce que ne permettent pas de faire les approches par patrons.

3.4 Construction d'ébauches de patrons (étape 4)

Pour faciliter l'analyse des classes obtenues, il est intéressant de considérer les éléments qui ont permis le rapprochement des occurrences d'association qui la composent et, une fois la relation sémantique identifiée et sélectionnée, il est précieux de pouvoir lui associer un patron d'extraction pour en repérer de nouvelles occurrences.

Les éléments de contexte qui caractérisent le mieux les classes obtenues sont en réalité donnés par le processus de classification précédent : ce sont les unités sémantiques que les contextes des occurrences d'association regroupées partagent et celles qui entrent dans l'association. Dans la classe $\{a_1, a_2, a_x\}$ de l'exemple précédent, on a donc deux unités sémantiques caractéristiques contextuelles : $\{US_{20}, US_{33}\}$ et deux couples d'unités sémantiques associées : $< US_1, US_{12} >$ et $< US_{19}, US_{24} >$. En réordonnant ces unités sémantiques en fonction de leur ordre d'apparition en corpus et en distinguant le rôle des unités sémantiques associées et contextuelles, on peut construire des ébauches de patrons. Pour la classe d'occurrences d'association $\{a_1, a_2, a_x\}$, on obtient ainsi deux ébauches de patrons [5] : [$\underline{US_1}$ ...US_{20} ...US_{33} ...$\underline{US_{12}}$] et [$\underline{US_{19}}$ ...US_{20} ...US_{33} ...$\underline{US_{24}}$].

4 Expérience

L'approche proposée ici est testée sur un corpus TSB issu de la *Transportation Safety Board* du Canada [6] qui comporte 162 rapports d'incidents de l'air publiés entre 2002 et 2005 en français. Le nombre de mots varie pour chaque document entre 1500 et 5000 mots.

4.1 Exemple détaillé

Nous avons sélectionné quelques exemples pour illustrer notre méthode. Pour la phrase

5. Sur cet exemple artificiel, l'ordre n'est pas significatif. Les unités sémantiques associées sont soulignées.
6. (http ://www.tsb.gc.ca/fra/rapports-reports/aviation/2003/index.asp).

Même si l'équipage (a coupé) les moteurs de l'avion après (avoir commandé) la sortie du (train d'atterrissage), ce dernier s'(est affaissé) sur la piste. [7]

on obtient la représentation normalisée suivante :

équipage couper moteur avion commander sortie
train_d'atterrissage affaisser piste

où *train d'atterrissage* est identifié comme terme composé et *dernier* et a été éliminé comme mot vide de même que tous les mots grammaticaux.

Considérons deux couples d'unités sémantiques retenus lors de la phase d'extraction des associations parce qu'elles apparaissent fréquemment ensemble : <équipage, train_d'atterrissage> et <instrument_de_vol, avion>. Considérons également des occurrences de ces associations : a_1, a_2, a_3, a_4, a_5 sont cinq occurrences de la première association et a_6 est une occurrence de la deuxième association [8].

a_1 Même si l'équipage a coupé les moteurs de l'avion après avoir *commandé* la *sortie* du train d'atterrissage, ce dernier s'est affaissé sur la piste.

a_2 Les rapports de l'équipage et des témoins indiquent que l'équipage avait *commandé* la *sortie* du train pendant l'approche, et que le train d'atterrissage s'était déployé en position *sortie* bien avant que l'avion se pose...

a_3 L'équipage a *commandé* la *sortie* du train d'atterrissage pendant l'approche, et le train était *sorti* avant que l'avion ne se pose.

a_4 L'équipage possédait les qualifications nécessaires pour *commander* la *sortie* du train d'atterrissage.

a_5 Lors de la formation, l'équipage a été impressionné par l'*avancée technologique* des trains d'atterrissage.

a_6 Notons l'*avancée technologique* des instruments de vol des avions.

On obtient les vecteurs de contexte suivants :

$Ctxt(a_1)$ {couper, moteur, avion, commander, sortie, affaisser, piste}

$Ctxt(a_2)$ {rapport, témoin, indiquer, commander, sortie, train, approche}

$Ctxt(a_3)$ {commander, sortie, approche, train, avion, poser}

$Ctxt(a_4)$ {posséder, qualification, commander, sortie}

$Ctxt(a_5)$ {formation, impressionner, avancée_technologique}

$Ctxt(a_6)$ {avancée_technologique}

La comparaison des contextes met en évidence les éléments partagés par les occurrences d'association a_1, a_2, a_3, a_4 d'une part et par les occurrences a_5, a_6 d'autre part. Le processus de regroupement propose donc deux relations sémantiques représentées par deux classes d'occurrences d'association : $R1 = \{a_1, a_2, a_3, a_4\}$ et $R2 = \{a_5, a_6\}$ avec resp. {commande, sortie} et {avancée_technologique} comme unités sémantiques caractéristiques contextuelles.

7. Les parenthèses montrent les expressions composées identifiées lors l'analyse du corpus.

8. Toutes les occurrences sont extraites de notre corpus. Nous donnons les phrases non transformées par souci de lisibilité. Le texte souligné correspond aux unités sémantiques associées. Le texte en italique correspond aux autres unités sémantiques caractéristiques des classes construites.

Ces résultats permettent d'associer des ébauches de patrons aux relations proposées. La classe de R1 est homogène du point de vue de l'association puisque toutes les occurrences ont les mêmes unités sémantiques associées. On a donc une seule ébauche de patron (voit patron $R1$ ci-dessous) et la classe peut effectivement s'interpréter comme une relation sémantique spécialisée « commander la sortie de ». La deuxième classe est plus difficile à analyser parce qu'elle réunit des occurrences de deux associations différentes. On a donc deux ébauches de patrons (voir patrons $R2$ ci-dessous) mais cette classe paraît difficile à interpréter en relation sémantique : il faut sans doute la supprimer ou la décomposer.

$R1$ [*equipage* ... *commander* ... *sortie* ... *train_d'atterrissage*]

$R2$ [*equipage* ... *avancée_technologique*... *avions* ... *train_d'atterrissage*]

$R2$ [*instruments_de_vol* ... *avancée_technologique* ...*avions*]

4.2 Evaluation

L'approche proposée a été évaluée sur un sous-corpus d'une centaine de phrases [9].

Résultats de la construction d'une représentation normalisée des textes Notre corpus d'évaluation a été segmenté par l'analyseur en 93 phrases. Chaque phrase a été représentée par ses unités sémantiques (figure 4). 524 unités sémantiques ont été recensées au total (tableau 2).

Phrase N°	Phrase
1	[Rapport_d'_enquête, événement_aéronautique, Collision, andain, Beech, exploité, Labrador_Airways, aéroport, international, St_John's, Terre-Neuve-et-Labrador, 11_janvier_2003, Rapport, numéro, Sommaire, Beech_1900D, immatriculé, C-GLHO, portant, numéro_de_série, UE-266, identifié, étant, vol_8333, Lab_Air, roule, piste, heurte, andain, haut, pieds, trouvant, travers, piste, juste, nord, voie_de_circulation, Charlie]
2	[n', a, blessé, passagers, membres_d'_équipage]
3	[avion, subit, importants, dommages]
4	[accident, survient, h, heure, normale, Terre-Neuve]
5	[Autres, renseignements, base, aéroport_de_St_John's, est, aéroport, certifié, contrôlé, comporte, pistes, utilisées, année]
6	[priorités, matière, déneigement, maintenance_hivernale, pistes, établies, plan, maintenance_hivernale, aéroport_de_St_John's]
7	[cause, possibilités, approche, instruments, catégorie, piste, moins, ne, soit, évident, conditions_de_vent, favorisent, utilisation, piste, concentre, d'abord, efforts, déneigement, préparation, piste, pistes, voies_de_circulation, connexes, nécessaires, à, accès]
8	[différentes, surfaces, aéroport, sont, désignées, fonction, priorités, déneigement, voir, figure]
9	[surfaces, priorité, sont, déneigées, dernier]
10	[partie, piste, située, nord, voie_de_circulation, Charlie, entrée, voie_de_circulation, Bravo, voit, octroyer, priorité, pistes, sont, en_service]
11	[plan_de_maintenance_hivernale, aéroport_de_St_John's, comporte, critères, relatifs, fermeture, piste, en_service, cas_de_contamination]
12	[vertu, critères, piste, en_service, doit, être, fermée, a, dernière, andains, hauts, plus, pouces]
13	[plan, ne, renferme, pas, directives, concernant, fermeture, pistes, non, en_service, partie, piste, située, nord, voie_de_circulation, Charlie, autres, surfaces, faible, priorité]
14	[zones, priorité, trouvant, nord, voie_de_circulation, Charlie, aéronefs, circulant, sol, ne, rencontrent, normalement, pas, andains, derniers, ne, sont, habituellement, pas, signalés, contrôleur_au_sol, mentionnés, rapports, état, surface_des_pistes]

FIG. 4 – *Extrait de la représentation normalisée des textes*

Résultats de l'extraction des associations d'unités sémantiques Les associations de termes fournies par le système sont classées par ordre de fréquence décroissante (tableau 2). La fréquence maximale pour notre corpus test est de 9. La courbe de fréquence montrant un point de cassure au niveau de la fréquence 5, c'est à cette valeur que nous avons fixé le seuil. Nous avons considéré les 26 associations apparaissant 5 fois ou plus. Ces 26 associations ont été analysées manuellement et 77% d'entres elles ont été jugées recevables.

9. Le rapport d'aviation du 11 janvier 2003 http ://www.tsb.gc.ca/fra/rapports-reports/aviation/2003/index.asp

US	US	US
rapport_de_enquête	*Beech*	*collision*
événement_aéronautique	*exploiter*	*aéroport*
rapport_de_enquête	*andain*	*St_John's*

TAB. 1 – *Extrait de la liste des 524 US*

Fréq.	Terme 1	Terme 2	Fréq.	Terme 1	Terme 2
9	*voie_de_circulation*	*piste*	7	*trouver*	*andain*
9	*Charlie*	*piste*	7	*piste*	*andain*
9	*Charlie*	*nord*	7	*signaler*	*andain*
9	*Charlie*	*voie_de_circulation*	7	*vol*	*piste*
8	*nord*	*piste*	7	*chel_de_équipe*	*piste*
8	*en_service*	*piste*	7	*appareil*	*andain*
8	*équipage*	*andain*	6	*trouver*	*piste*
7	*voie_de_circulation*	*nord*	6	*heure*	*piste*

TAB. 2 – *Extrait de la liste des associations classées par fréquence décroissante*

Notre étude montre des résultats qui sont dans l'ensemble cohérents : ils mettent en évidence des relations évoquées explicitement ou implicitement dans le texte. Nous pouvons, par exemple, interpréter assez naturellement l'association <voie_de_circulation, piste> puisque nous savons qu'une piste est située sur une voie de circulation.

Résultats du regroupement des occurrences d'association en classes Le tableau 3 présente les résultats de la classification des contextes où chaque ligne représente une classe. Par exemple, les contextes 4, 18 et 90 ont été jugés similaires et sont regroupés dans la même classe 2. Sur notre corpus d'évaluation de 93 phrases, l'algorithme de classification a trouvé 72 classes avec une cardinalité maximale de 4 contextes par classe. On peut vite réduire ce nombre de classes à 14 car 58 classes sont des singletons (voir tableau 4).

Classe (clusteur)	Ctxt			
1	1	0	0	0
2	4	18	90	0
3	7	0	0	0
4	8	9	0	0
...	...	...	...	...
9	10	13	69	81
10	11	12	0	0
...	...	...	...	

TAB. 3 – *Résultats de la classification des contextes*

Cardinalité	NbClasses
Card = 1	58
Card = 2	9
Card = 3	3
Card = 4	2

TAB. 4 – *Résultats chiffrés de la classification.*

A chaque association correspond une liste de contextes. Par exemple pour l'association <voie_de_circulation, piste> le système nous fournit les 9 contextes : {Ctxt(1), Ctxt(7), Ctxt(10), Ctxt(13), Ctxt(19), Ctxt(28), Ctxt(54), Ctxt(55), Ctxt(81)}. En faisant le rapprochement avec les classes de contextes du système (tableau 3) on note que les contextes Ctxt(10), Ctxt(13), et Ctxt(81) apparaissent ensemble dans la classe 9. Le retour au texte montre que ces trois contextes évoquent bien la même relation [10] : 'être_situé_au_nord' entre *voie_de_circulation* et *piste*.

Résultats de la construction d'ébauches de patrons L'étape précédente permet de rapprocher dans une même classe les instances de relations similaires. Le rapprochement se fait en repérant les US partagées par leurs contextes. Ces US partagées, qui caractérisent la classe, sont réutilisées pour construire une ébauche de patron. Ainsi, pour l'association étudiée <voie_de_circulation, piste>, nous réutilisons les US partagées : *situé* et *nord* dans leur ordre d'apparition. Il en résulte le patron : $[\underline{voie\_de\_circulation}\ldots\text{situé}\ldots\text{nord}\ldots\underline{piste}]$

5 Discussion et travaux futurs

Notre méthode doit être testée de manière plus approfondie et sur d'autres corpus mais se faire une idée de la qualité des résultats est toujours difficile quand les résultats sont bruités et « à retravailler ». Il faut apprécier à la fois la quantité de travail supplémentaire à fournir pour aboutir à un résultat acceptable et l'aide apportée par les suggestions du système. Pour notre méthode de mise en lumière de relations sémantiques et sur notre corpus TSB, nous prévoyons de construire en parallèle deux ontologies des incidents aériens [11]. La première, l'« ontologie manuelle » (O_m), est en cours de construction, à l'aide de l'outil TERMINAE (Aussenac-Gilles et al., 2008) mais manuellement pour ce qui concerne les relations sémantiques. La seconde, l'« ontologie assistée » (O_a), sera construite à partir de l'analyse des classes d'occurrences d'association proposées par notre méthode, donc avec l'aide de notre système. L'évaluation consistera alors à

10. Les contextes correspondent aux phrases suivantes :
Ctxt(10) : *La partie de la piste 02 située au nord de la voie de circulation Charlie jusqu'à l'entrée de la voie de circulation Bravo, se voit octroyer la priorité 3 lorsque les pistes 11-29 ou 16-34 sont en service.*
Ctxt(13) : *Mais ce plan ne renferme pas de directives concernant la fermeture de pistes non en service, comme la partie de la piste 02-20 située au nord de la voie de circulation Charlie ou d'autres surfaces à faible priorité.*
Ctxt(81) : *La piste 02, située au nord de la voie de circulation Charlie, est demeurée ouverte pendant toute la durée du déneigement même si elle n'était ni utilisable, ni nécessaire.*

11. Les ontologies produites doivent permettre de comparer les incidents rapportés.

1. comparer la sortie brute du système (les classes d'occurrences d'association) avec les relations conceptuelles de O_a pour mesurer l'effort fourni par l'auteur de l'ontologie pour construire O_a à partir des résultats du système ;
2. comparer la sortie brute du système avec O_m pour mesurer la part du travail qui aurait réellement pu être assisté ;
3. comparer O_a et O_m pour comprendre comment le système influence le travail de conceptualisation.

Les retours de cette évaluation permettront de finaliser notre méthode avant son intrégration dans la plateforme Dafoe. Deux aspects importants restent à travailler. Nous avons présenté une version de base de notre méthode distributionnelle mais différents paramètres peuvent être ajustés (méthode de seuillage, choix des mesures de coocurrences et de similarité, pondération des unités sémantiques dans les vecteurs de contexte, etc.). L'analyse des premiers résultats devrait permettre d'avancer dans cette direction. Il faut aussi concevoir un module de présentation et d'analyse des résultats et un retour d'expérience est nécessaire pour comprendre comment ces résultats peuvent être utilisés en pratique : on peut par exemple ajouter des étiquettes sémantiques sur les unités sémantiques caractéristiques des classes obtenues [12] mais il faut veiller à ce que l'ensemble reste simple à analyser.

Soulevons un autre point important, les matrices de contextes sont réparties de façon non uniforme car certaines unités sémantiques sont plus utilisées que d'autres. Nous prévoyons alors de remplacer la similarité logique de condorcet par la similarité statistique qualifiée de "présence-rareté" (Benhadda et Marcotorchino, 1998). Cette similarité tient compte non seulement de la modalité partagée par les deux individus, mais aussi du nombre d'individus possédant cette modalité. En d'autres termes, plus ces deux individus sont rares à partager la même modalité, plus ils sont semblables et *vice-versa*.

6 Conclusion

Nous proposons dans cet article une méthode générique pour mettre en lumière les relations sémantiques d'un domaine à partir de textes de ce domaine. En tant que telle, cette méthode est indépendante du domaine et de la langue. Elle doit à terme s'intégrer dans la plateforme Dafoe de construction d'ontologies à partir de textes. Il s'agit d'extraire tout type de relation sans connaitre *a priori* les relations à extraire. La méthode est fondée sur une représentation normalisée des textes comme séquences d'unités sémantiques du domaine (essentiellement des termes et des entités nommées) et l'idée maîtresse consiste à construire des classes d'associations d'unités sémantiques de manière distributionnelle. L'hypothèse sous-jacente est que les occurrences d'association réunies sur la base des éléments de contexte qu'elles partagent ont des chances de relever d'une même relation sémantique et que les relations candidates ainsi proposées peuvent aider le travail de conceptualisation de l'ontologie de domaine.

Remerciements

Ce travail a été partiellement financé par le projet ANR Dafoe4App.

12. Les outils d'étiquetage d'entités nommées leur associent généralement un type sémantique et on peut utiliser le fait que les termes sont associés à des concepts de l'ontologie lorsqu'ils le sont.

Références

Ah-Pine, J. et G. Jacquet (2009). Clique-based clustering for improving named entity recognition systems. In *EACL*, pp. 51–59.

Aubin, S. et T. Hamon (2006). Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo, et T. Pahikkala (Eds.), *Advances in Natural Language Processing*, 5th International Conference on NLP (finTAL), pp. 380–387. LNAI 4139.

Auger, A. et C. Barriere (2008). Pattern-based approaches to semantic relation extraction : A state-of-the-art. *Terminology 14*(1), 1–19.

Aussenac-Gilles, N., S. Despres, et S. Szulman (2008). The terminae method and platform for ontology engineering from texts. In P. Buitelaar et P. Cimiano (Eds.), *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text*. IOS Press.

Benhadda, H. et J. F. Marcotorchino (1998). Introduction à la similarité régularisée en analyse relationnelle. *Revue de statistique appliquée vol. 46*(no1), pp. 45–69.

Faure, D. et C. Nédellec (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : the system asium. In D. Fensel et R. Stude (Eds.), *Proceedings of the 11th International Conference on Knowledge Engineering and Knowledge Management*, pp. 329–334. Springer-Verlag.

Hasegawa, T., S. Sekine, et R. Grishman (2004). Discovering Relations among Named Entities from Large Corpora. *Proc. of ACL-2004*, 415–422.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International conference on Computational Linguistics*, Volume 2, Nantes, pp. 539–545.

Jacques, M.-P. et N. Aussenac-Gilles (2006). Variabilité des performances des outils de tal et genre textuel. cas des patrons lexico-syntaxiques. *Traitement Automatique des Langues (TAL) 47*(1), 11–32.

Pearson, J. (1998). *Terms in Context*, Volume 10. John Benjamins Publishing Company.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Summary

This paper presents a method for extracting semantic relations from text corpora in the perspective of ontology design. Our goal is to propose a generic method that is both domain and language independent. It is based on the distributional analysis of semantic units to bring out semantic relations that may be relevant to model at the ontology level. This method makes no hypothesis on the types or linguistic forms of the relations. It clusters term associations in classes that represent candidate semantic relations. The underlying assumption is that the occurrences of these associations, which are clustered according to the contextual elements they share, belong to the same semantic relation.