Classification Supervisée de Questions: Rôle de l'Expansion Sémantique

Ali Harb, Michel Beigbeder, Jean-Jacques Girardot

EMSE-G2I, 158, Cours Fauriel 42023 Saint-Étienne Cedex 2
{harb, mbeig, girardot}@emse.fr

RÉSUMÉ. *Fournir de bonnes réponses à une question donnée en cherchant au sein d'un grand corpus de documents est une tâche difficile. Il est nécessaire de percevoir et de reconnaître la question à un niveau qui permet d'imposer des contraintes sur l'ensemble des réponses possibles. Une contrainte fréquemment utilisée est la catégorie des questions qui permet de déduire le type de réponse attendue. L'objectif est de fournir des informations supplémentaires afin de réduire l'écart entre la question et sa réponse. Nous proposons ici une approche pour améliorer l'efficacité des classifieurs, basée sur l'analyse linguistique et des approches statistiques. Ce travail propose également deux méthodes d'expansion des questions. Pour cela, différentes caractéristiques pour la représentation des questions, différentes pondérations et plusieurs algorithmes d'apprentissage automatique ont été étudiés. Les expériences menées sur des jeux de données réels montrent une amélioration de la précision dans la classification des questions.*

ABSTRACT. *Responding correctly to a question given a large collection of textual data is not an easy task. There is a need to perceive and recognize the question at a level that permits to detect some constraints that the question imposes on possible answers. The question classification task is used in Question Answering systems. This deduces the type of the expected answer, to perform a semantic classification to the target answer. The purpose is to provide additional information to reduce the gap between answer and question. An approach to improve the effectiveness of classifiers focusing on linguistic analysis and statistical approaches. This work also proposes two methods of questions expansion. Various questions representation, term weighting and diverse machine learning algorithms are studied. Experiments conducted on actual data are presented. Of interest is the improvement in the precision on the classification of questions.*

MOTS-CLÉS : *Classification, Sélections des descripteurs, Expansion sémantique, Apprentissage, Fouille de texte.*

KEYWORDS: *Classification, Feature selection, Semantic expansion, Machine learning, Text mining.*

1. Introduction

Le nombre de documents disponibles avec le développement du Web devient de plus en plus important. Il est devenu difficile pour les utilisateurs de trouver des informations spécifiques. Ils sont confrontés à de nombreux résultats de recherche, et beaucoup d'entre eux ne sont pas pertinents. Les systèmes sont vus comme un moyen de fournir rapidement des informations, plus particulièrement en réponse à des questions précises. Pour fournir des réponses correctes à une question donnée en cherchant au sein d'un grand corpus de documents, on a besoin d'imposer des contraintes sur l'ensemble des réponses possibles. Une étape fréquemment utilisée est la classification des questions. Cette étape assigne à chaque question une catégorie qui définit le type de réponse attendue. Cette tâche réduit significativement l'espace de recherche pour identifier la bonne réponse et améliorer la qualité du système.

Une approche pour la classification des questions est d'utiliser des règles construites manuellement (Kangavari *et al.*, 2008 ; Saxena *et al.*, 2007). Il est cependant difficile de les maintenir pour être à jour. Récemment, avec la croissance de la popularité des approches statistiques, l'apprentissage automatique a été appliquée pour détecter les catégories de questions (Xin et Dan, 2006 ; Zhang et Lee, 2003 ; Krishnan *et al.*, 2005). L'avantage est que les algorithmes d'apprentissage peuvent détecter les descripteurs discriminants. La représentation par *sac des mots* (*Bag-of-Words*) est fréquemment utilisée dans la tâche de classification en utilisant l'apprentissage automatique. Toutefois, comme les questions sont courtes, nous devons combiner plusieurs descripteurs pour la représentation des questions et ajouter des informations pertinentes pour permettre au classifieur d'obtenir une grande précision.

Nous avons utilisé plusieurs sources d'informations sémantiques de différents niveaux de granularité. Ces ressources sémantiques permettent d'expliciter et de préciser le sens des termes des questions. Par exemple, on peut obtenir la catégorie grammaticale des termes (e.g. nom, verbe, adjectif), les catégories sémantiques des noms (e.g. personne, lieu), ou des synonymes, des hyperonymes, des hyponymes des noms. Dans ces opérations le véritable sens des termes doit être conservé et pour cela, nous utilisons le contexte.

Le document est organisé comme suit : la section 2 décrit les principales techniques de classification de questions. Dans la section 3, nous présentons une étude sur les différents descripteurs de représentation des questions. La section 4 décrit en détail les expériences avec différents descripteurs en utilisant trois algorithmes d'apprentissage. Notre approche est décrite dans la section 5, où nous décrivons la méthode d'expansion sémantique des questions. La section 6 présente les résultats des expériences réalisées sur des données réelles.

2. Travaux antérieurs

Les travaux réalisés dans le domaine de la classification des questions peuvent être classés en deux groupes principaux : le premier où les méthodes sont basées sur

des règles prédéfinies et le deuxième où sont utilisés des algorithmes d'apprentissage automatique.

2.1. *Règles construites manuellement*

Dans les approches du premier groupe, des règles construites manuellement sont employées pour analyser et reconnaître des patrons (*patterns*) discriminants (Kosseim et Yousefi, 2008). Le système QUANTUM (Plamondon *et al.*, 2003) définit un ensemble de 40 règles qui lui permet de classifier correctement 88% des 492 questions provenant de TREC-10 [1]. Dans les travaux de (Kangavari *et al.*, 2008 ; Saxena *et al.*, 2007), une liste des mots déterminants «*who*, *where*, ... » est associée à des règles prédéfinies. Les auteurs s'intéressent à la détection des marqueurs. Ainsi, pour chaque marqueur ils créent une catégorie de type de questions. Par exemple, dans la question «*who was the first president of France ?*», le type attendu après la détection de «*who*» est *Person*. Mais si le mot déterminant «*what*» est trouvé, le type de réponse peut être indéfini. Réciproquement, il existe des questions dont le type attendu est *Person* mais où « who » n'apparaît pas.

Ces règles manuelles sont difficiles et coûteuses en temps à construire. Leur couverture est limitée, car il est peu probable qu'elles couvrent toutes les catégories de questions et qu'elles permettent de classifier finement des questions plus détaillées. Lors de chaque évolution de la taxonomie, de nombreuses règles doivent être modifiées ou complètement réécrites. Du fait de ces limitations, la majorité des systèmes qui font appel à des règles construites manuellement utilisent un nombre réduit de catégories de question. Ce faible nombre de catégories va donc influer sur les performances du système de Q/R tout entier.

2.2. *Algorithme d'apprentissage*

Dans le second groupe de méthodes de classification où des algorithmes d'apprentissage sont utilisés, la connaissance des experts est remplacée par un corpus d'apprentissage contenant des questions annotées. En utilisant ce corpus, le classifieur est entraîné en mode supervisé. Plusieurs choix de classifieurs sont possibles, en voici une liste non limitative : réseaux de neurones (*Neural Network, NN*), classifieur Bayésien (*Naive Bayes, NB*), K plus proches voisins (*K nearest neighbours, K-NN*), les séparateurs à vaste marges (*Support Vector Machine, SVM*) et les arbres de décision (*Decision Trees, DT*). Ces approches utilisant les algorithmes d'apprentissage résolvent quelques limitations des approches basées sur les règles construites manuellement. La reconstruction d'un classifieur appris est plus souple que la réécriture manuelle de règles. De plus, la couverture est plus large et peut être améliorée en fournissant de nouvelles questions étiquetées d'apprentissage.

1. http ://trec.nist.gov/data/qa.html

Dans les travaux de (Zhang et Lee, 2003), les auteurs comparent un système de classification basé sur SVM avec d'autres algorithmes d'apprentissage (KNN, NB, DT). La représentation sacs de mots est utilisée par tous ces classifieurs et ils sont entrainés sur le même corpus d'apprentissage. Récemment, une nouvelle architecture d'apprentissage SNoW (*Sparse Network of Winnows*) (Khardon *et al.*, 1999) a été utilisée pour la classification des questions dans le travail de (Xin et Dan, 2002). Ils ont construit le corpus de classification de question UIUC. Dans ce travail, ils ont utilisé les catégories grammaticales *part-of-speech*, les résultats d'une analyse syntaxique, le premier nom dans la question et les entités nommées. Ils ont obtenu une précision de 78,8%.

Récemment, plusieurs taxonomies de questions ont été présentées, mais il n'y a pas de standard utilisé par tous les systèmes. La plupart des systèmes de Q/A participant à TREC utilisent leur propre taxonomie. Ces taxonomies doivent évoluer pour durer dans le temps. Elles sont typiquement composées d'une vingtaine de catégories, ce qui est peu, car, comme montré par plusieurs systèmes Q/A, l'utilisation d'une taxonomie détaillée contenant des catégories fines favorise la détection du type de réponse attendu (Zhang et Lee, 2003 ; Xin et Dan, 2002). Plus récemment, la taxonomie et le corpus décrits dans (Xin et Dan, 2006) sont devenus les plus fréquemment utilisés dans la recherche actuelle.

Un système utilisant un classifieur SVM et une représentation par bigrammes de mots qui a obtenu une précision de 80,2% a été présenté dans (Hacioglu et Ward, 2003). Plus récemment, dans les travaux de (Krishnan *et al.*, 2005), les auteurs ont utilisé des représentations par N-grammes avec N égal à 1 ou 2, et intégré tous les hyperonymes des mots. Ils ont obtenu une précision de 86,2% en utilisant la même taxonomie et le même corpus d'apprentissage utilisé par (Krishnan *et al.*, 2005). Plus tard, dans les travaux de (Xin et Dan, 2006), plus d'information sémantique est utilisée, y compris les entités nommées, WordNet et des classes spécifiques pour des mots proches. Du fait de leur utilisation, ils ont pu atteindre la précision de 86,3%.

Dans cet article, nous étudions également différentes méthodes de transformation et de combinaison des descripteurs en particulier nous allons enrichir les questions par des termes qui préservent le sens original des mots. Pour cette tâche, nous proposons et testons plusieurs types d'expansions sémantiques et leur combinaison.

3. Descripteurs

Dans toute classification automatique de texte, les choix de représentation des instances à traiter (dans notre cas les questions) et des opérations à leur appliquer sont cruciaux. Une approche fréquente consiste à faire appel à la représentation dite en « sac de mots », où la seule information utilisée est la présence et/ou la fréquence de certains mots. Un grand nombre de chercheurs du domaine ont choisi d'utiliser une représentation vectorielle selon le modèle de Salton (Salton et Buckley, 1988), appliquée pour la première fois dans (Salton, 1971). Cette représentation transforme

chaque texte en un vecteur de n mots pondérés. A la base, les n descripteurs du texte sont tout simplement les n différents mots apparaissant dans les documents. Il est possible d'utiliser d'autres types d'attributs pour caractériser les vecteurs, dont certains seront présentés ci-dessous.

Notre analyse a montré que certaines informations sémantiques et syntaxiques qui existent fréquemment dans les questions appartenant à la même catégorie n'existent pas dans les autres catégories. Ainsi, l'exploitation de ces informations fournit des indices précieux pour les classifieurs par rapport à la simple représentation en sacs de mots.

3.1. *Descripteurs syntaxiques*

Dans une question, outre le terme lui-même on peut considérer pour chaque mot d'autres caractéristiques syntaxiques :les lemmes, les catégories grammaticales *Part-of-Speech* (e.g. verbe, nom, adjectif), les résultats de l'analyse syntaxique et en particulier la dépendance grammaticale. Ces dernières permettent de détecter les mots particulièrement significatifs (e.g. sujet, objet).

3.2. *Relation sémantique entre les mots*

Les mots sont sémantiquement liés de plusieurs façons différentes. Ces relations sont décrites dans des vocabulaires contrôlés. Il y a deux catégories principales pour ces relations : équivalence et hiérarchie. La principale relation dans la classe *équivalence* est celle de synonymie. Plus précisément, la synonymie décrit la relation entre deux mots qui ont la même signification (e.g. *city* et *town*). Du côté des relations hiérarchiques, l'hyperonymie décrit la relation sémantique de généralisation (e.g. *flower* et *plant*).

3.3. *Entité nommée*

Cette caractéristique attribue une catégorie sémantique pour certains noms dans les questions. La présence de ces annotations d'entité nommée dans les questions met en évidence la sémantique commune et discriminante appartenant au même type de question. Par exemple dans la question *Who is the first president of France ?*, l'annotation des entités nommées fournit : *Who is the [Num first] [Vocation president] of [Country France]*. Comme nous pouvons le voir, nous obtenons plus d'informations sémantiques exprimées dans les catégories de *first* (Number), *president* (Vocation) et *France* (Country).

3.4. *N-grammes*

Un N-gramme de mots est une séquence de N mots consécutifs. Ce modèle est fondé sur l'hypothèse que la présence d'un mot n'est pertinente que dans le contexte des mots qui le précédent et ceux qui le suivent. La représentation par N-grammes prend en compte l'ordre des mots, et par conséquent peut refléter le thème de la phrase d'une manière plus précise que la représentation par mots isolés. Par exemple la phrase "*On ne change pas une équipe qui gagne*" se décompose en N-grammes de mots de la façon suivante pour N=1, 2, 3 :

– **1-grammes** : on ne change pas une équipe qui gagne ;

– **2-grammes** : on-ne ne-change change-pas pas-une une-équipe équipe-qui qui-gagne

– **3-grammes** : on-ne-change ne-change-pas change-pas-une pas-une-équipe une-équipe-qui équipe-qui-gagne

3.5. *Pondération*

Soit n le nombre total de descripteurs distincts (e.g. mots, N-grammes, etc.) dans le corpus. Chaque question est représentée par un vecteur de n éléments. Chaque composante de ce vecteur de document peut être tout simplement binaire mais aussi le nombre d'occurrences du descripteur dans le document. Cependant en procédant avec la représentation fréquentielle on donne une importance trop grande aux descripteurs qui apparaissent dans beaucoup de classes (voire toutes) et qui sont peu représentatifs d'une classe en particulier. Pour pallier à ce défaut bien connu en recherche d'information, plutôt qu'utiliser la fréquence brute, on utilise d'autres mesure connue sous le nom $tf \cdot idf$ (Salton et Buckley, 1988) (*Term Frequency Inverse Document Frequency*). Elles permettent de mesurer l'importance $tf(t,d)$) du mot t en fonction de sa fréquence dans le document d, pondérée par la fréquence d'apparition du terme dans tout le corpus, $idf(t) = log\frac{|S|}{df(t)}$.

$$tf \cdot idf(t,d) = tf(t,d) \cdot idf(t) \quad [1]$$

Où $|S|$ est le nombre de documents dans le corpus et $df(t)$ est le nombre de documents contenant t. Cette mesure permet de donner un poids plus important aux mots discriminants d'une collection. Inversement, un terme apparaissant dans tous les documents du corpus aura un poids faible.

4. Classification des questions

La tâche de classification peut être définie comme l'association à chaque question une classe parmi N, dont le but est de déterminer le type de réponse attendu. Chaque classe comporte sa propre sémantique, qui va nous aider dans la suite de la tâche de Q/R à trouver la réponse pertinente.

Nous avons présenté plusieurs méthodes de représentation des données pour la classification automatique des questions. Nous avons vu en quoi consistait le principe de chaque descripteur, quels bénéfices ils apportent et de quelle façon les utiliser. L'objectif est de choisir parmi les différentes représentations issues de ces différents descripteurs celles qui permettent la meilleure performance de classification.

Pour évaluer la pertinence des différents descripteurs, nous les avons utilisé dans un système de classification complet. Nous avons choisi d'utiliser trois algorithmes d'apprentissage supervisés : KNN, SVM et NB. Nous comparons les résultats obtenus avec ces trois algorithmes. Les résultats de la classification sont présentés en termes de précision des algorithmes (capacité à bien classer la question dans la catégorie adaptée). Nous avons utilisé le processus de la validation croisée en dix sous-ensembles (90% du corpus est utilisé pour l'apprentissage, 10% pour le test).

4.1. *Corpus*

Nous avons utilisé un corpus de questions composé de 10 343 questions, qui est la réunion de plusieurs collections : 5 500 de *UIUC DATA* (Xin et Dan, 2006), 1343 de *TREC10*, *TREC9*, et *TREC8*, 200 de *QA@CLEF2006* [2] et 2 249 de *CRL-QA* [3] et 1 011 de *NTCIR-QAC1* [4].

4.2. *Taxonomie*

Notre avons choisi la taxonomie proposée dans (Xin et Dan, 2006) parce qu'elle a une bonne couverture sur les différentes types de questions et qu'elle a bénéficié de mises à jour régulières depuis sa constitution initiale. Elle représente une classification sémantique naturelle pour les catégories des questions. Elle comporte 6 grandes classes et 50 classes fines présentées dans le tableau 1. Nous avons annoté manuellement les questions de notre corpus selon cette taxonomie à la fois avec l'une des six classes générales et l'une des 50 classes détaillées.

4.3. *Résultat d'expérimentation*

Lors des premières expérimentations, la représentation des questions est simplement réalisée en filtrant les mots vides appartenant à une liste. Dans notre approche, les mots vides qui sont fréquents et inutiles sont filtrés en utilisant une liste de *Stop Words* [5], modifiée pour ne pas filtrer ce qui est utile pour notre application, par exemple

2. http ://clef-qa.itc.it/2006bis/downloads.html
3. http ://www.cs.nyu.edu/ sekine/PROJECT/CRLQA
4. http ://www.nlp.cs.ritsumei.ac.jp/qac/
5. http ://www.lsi.upc.es/padro/lists.html

Classes générales	Classes détaillées
ABBREVIATION	abbreviation, expression
DESCRIPTION	definition, description, manner, reason
ENTITY	animal, body, color, creative, currency, disease, event, food instrument, lang, letter, other, plant, product, religion, sport substance, symbol, technique, term, vehicle, word
HUMAN	description, group, individual, title
LOCATION	city, country, mountain, other, state
NUMERIC	code, count, date, distance, money, order, other, period, percentage, speed, temp, volumesize, weight

Tableau 1. *Taxonomie de Catégories de Questions*

Who, *Where* sont des mots discriminants dans des questions, et ne sont donc pas supprimés. Nous avons appliqué les trois algorithmes d'apprentissage. les précisions obtenues sont présentées dans le tableau 2, ligne **filtrage**.

Algorithmes	K-NN		SVM		NB	
Pondération	Fréq.	$tf \cdot idf$	Fréq.	$tf \cdot idf$	Fréq.	$tf \cdot idf$
Filtrage	52,5	55,7	57	58,6	56,7	57,2
Lemmatisation	60,9	62,1	62,8	63,9	62,2	62,4
1-grammes	52,5	55,7	57	58,6	56,7	57,2
2-grammes	65,8	66,9	67,7	68,2	67,9	68,1
3-grammes	62,1	63,8	64,3	66,2	65	64,2
4-grammes	54,4	56,7	55,8	56	53,1	54,9

Tableau 2. *Résultats de la classification.*

Dans une deuxième série d'expériences, nous avons utilisé TreeTagger (Schmid, 1994) [6] un outil développé par l'*Institute for Computational Linguistics* de l'Université de Stuttgart. TreeTagger peut faire de la lemmatisation mais permet également d'attribuer une catégorie grammaticale à chaque mot. Il est fondé sur un algorithme d'arbre de décision (Quinlan, 1986) pour effectuer l'analyse grammaticale. Ensuite nous procédons à une étape de filtrage utilisant la liste de mots vides. Les résultats sont présentés dans le tableau 2 ligne **lemmatisation** et ils montrent une amélioration par rapport aux précédents.

Le deuxième descripteur expérimenté est le N-gramme, les résultats sont présentés dans le tableau 2. Les résultats obtenues avec une taille de N-grammes supérieure à 4 n'étant pas bons, nous ne reportons dans ce document que ceux pour N variant de 1 à 4. Les meilleurs résultats sont obtenus avec N=2, puis l'efficacité décroît quand N croît pour toutes les variantes de classifieurs et de pondération. Nous constatons que c'est avec l'algo SVM que la représentation avec les N-grammes donne les meilleurs scores.

6. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Avec ces expérimentations nous avons démontré que l'utilisation de l'une ou l'autre des représentations par N-grammes ou la lemmatisation ou par rapport au filtrage des mots vides améliore les résultats. Nous avons ensuite essayé de combiner ces descripteurs ensemble. Nous avons donc appliqué TreeTagger, suivi par le filtrage des mots vides et une représentation par N-grammes avec N variant de 2 à 4.

Algorithmes	K-NN		SVM		NB	
Pondération	Fréq.	$tf \cdot idf$	Fréq.	$tf \cdot idf$	Fréq.	$tf \cdot idf$
1-grammes	60,9	62,1	62,8	63,9	62,2	62,4
2-grammes	71,3	71,2	72,5	73,5	71,9	72,3
3-grammes	64,7	66,3	65,9	66,7	65,3	65,7
4-grammes	55,9	57,1	56,7	57,2	56,3	56,6

Tableau 3. *Résultat de la classification en utilisant lemmatisation, filtrage et N-grams.*

Le tableau 3 montre la précision obtenue pour la classification avec les trois algorithmes d'apprentissage sur le même corpus des questions. En général, nous remarquons que les résultats se détériorent quand N égale 3 ou 4. Les meilleurs résultats sont obtenus avec SVM par rapport à KNN et NB. Le pourcentage des questions bien classées, grâce à cette combinaison, passe de 68.2% à **73.5**% avec SVM et l'application de $tf \cdot idf$.

4.4. *Discussion*

Dans les étapes d'apprentissage de l'algorithme et dans l'étape de classification des nouvelles questions, aucune information sémantique n'est utilisée. Cependant des questions qui devraient être classées dans une même catégorie peuvent utiliser un vocabulaire différent, soit par synonymie (p.ex. *birth* et *cradle*), soit par hyperonymie (p.ex. *city* et *town*). Dans ce qui suit, nous allons présenter notre approche fondée sur la combinaison des différents descripteurs traités dans cette section, et l'expansion des questions par hyperonymie, synonymie, et entité nommées.

5. L'Approche SACSEQ

L'objectif de cette section est de présenter l'approche SACSEQ. Le processus général, composé de quatre phases, est décrit dans la figure 5. Dans les sous sections suivantes, nous présentons en détail ces différentes phases.

5.1. *Phase 1 : Pré-traitement du corpus*

Nous annotons toutes les questions du corpus selon les classes générales et détaillées de la taxonomie. Ensuite, nous utilisons TreeTagger pour la lemmatisation, et pour identifier la catégorie grammaticale des mots.

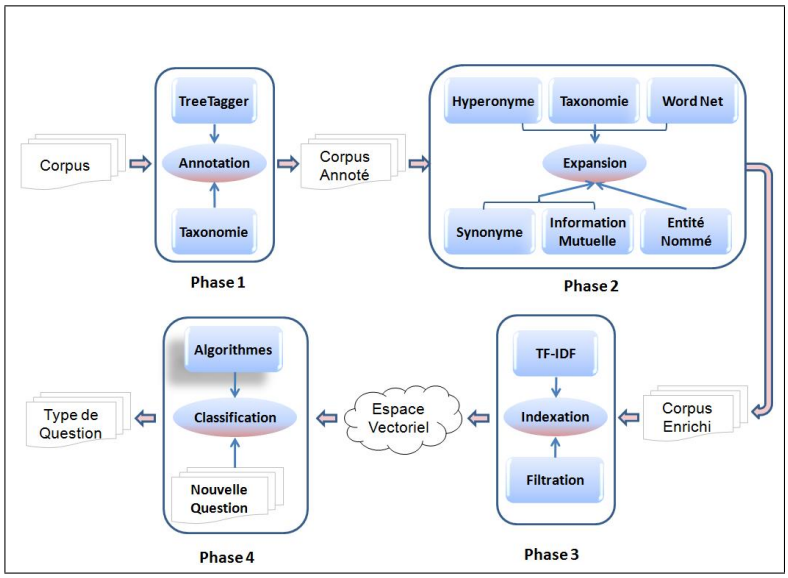

Figure 1. *Le processus général de l'approche* SACSEQ

5.2. *Phase 2 : Expansion des Questions*

En ce qui concerne la diversité du vocabulaire et la taille des questions, plusieurs mots sémantiquement corrélés sont traités comme différents (e.g. *City* and *Town*). Nous essayons donc d'enrichir les questions avec des termes plus généraux (les *hyperonymes*), des synonymes et avec des entités nommées.

1- Projection et expansion avec les hyperonymes : La projection désigne un appariement exact entre un mot donné et un terme représentant un concept de la taxonomie. L'idée est d'enrichir les questions avec les hyperonymes des noms appartenant aux questions. Pour cette tâche nous avons utilisé WordNet. Pour un nom donné, la fonctionnalité d'hyperonymie de WordNet fournit plusieurs mots génériques à différents niveaux, en commençant par le plus spécifique et en allant vers le plus général. En cherchant ces représentations plus générales d'un nom donné, la procédure de découverte des hyperonymes doit aussi s'assurer qu'ils préservent bien la sémantique originale du mot.

Les différentes étapes du processus de « Projection et Hyperonymes » sont les suivantes :1) extraire les noms de la question ; 2) pour chaque nom extrait, le projeter sur les 50 concepts de la taxonomie et sur leurs instances ; si ce mot appartient à l'ensemble des instances d'un concept, le question sera enrichie par ce concept ; 3) si il est identique à un concept, pas de changement ; 4) sinon, trouver l'ensemble des hyperonymes du mot ; 5) en préservant l'ordre des hyperonymes reflétant le niveau

d'abstraction, nous cherchons le premier parmi ces hyperonymes qui se projette sur les instances des concepts de la taxonomie ou les concepts eux-mêmes.

2- Expansion avec les synonymes : Cette étape utilise une nouvelle méthodologie pour sélectionner les synomymes. De nouveau nous avons utilisé WordNet pour lister les synonymes d'un terme. Pour maintenir un enrichissement sémantique, nous étudions la corrélation sémantique entre un mot et chacun de ses synonymes. Une mesure couramment utilisée pour calculer la force de corrélation entre deux mots est l'information mutuelle au cube (MI^3)(Downey *et al.*, 2007).

Cette mesure dépend d'un contexte c. Elle est basée sur trois fréquences : $nb(x, c)$, le nombre de co-occurrences de x et c, $nb(y, c)$, le nombre de co-occurrences de y et c, et $nb(x, y, c)$, le nombre de co-occurrences de x, y et c. La mesure est calculée avec la formule $AcroDef_{MI^3}$ (Roche et Prince, 2007) :

$$AcroDef_{MI^3}(x, y, c) = \frac{nb(x, y, c)^3}{nb(x, c) \cdot nb(y, c)} \quad [2]$$

Nous ne gardons que le synonyme jugé sémantiquement proche, c'est-à-dire dont la mesure $AcroDef$ avec le terme original a la plus grande similarité parmi le groupe de synonymes. Nous enrichissons les questions avec ces synonymes.

Pour mettre en œuvre cette mesure, nous avons besoin d'un contexte c. Pour l'obtenir nous avons utilisé *Stanford parser* (Marneffe et Manning, 2008) pour extraire les dépendances grammaticales. Nous utilisons cet analyseur pour extraire les relations syntaxiques des questions. Ces relations définissent les *dépendances grammaticales* entre les mots de questions qui seront utilisés ultérieurement comme le contexte des mots adressés.

L'exemple suivant illustre l'enrichissement par synonymes. **Question** : *What is the capital of the French Republic ?* **Dépendances grammaticales** : attr(is, Who), det(capital, the), nsubj(is, capital), det(republic, the), amod(republic, French) and prep_of(capital, republic.)

Après l'analyse, nous avons l'information que *capital* est le sujet de la phrase. En se basant sur les dépendances grammaticales *amod* et *prep_of*, nous trouvons que *French Republic* est le contexte de *capital*. Lors de la recherche de l'ensemble des synonymes de *capital* sur WordNet nous trouvons : *Seat of Government*, *City*, *Principal*, *Assets*, *Wealth*.

Pour évaluer les trois fréquences nous envoyons des requêtes à Google, puis nous calculons les mesures d'information mutuelle au cube. La liste ci-dessous montre les valeurs de ces mesures pour tous les synonymes du terme original *capital* où le contexte est *French Republic* :

1) $AcroDef_{MI^3}(capital, seat\ of\ government, French\ Republic) = 3,57 \times 10^{-1}$
2) $AcroDef_{MI^3}(capital, city, French\ Republic) = 2,24 \times 10^{-2}$
3) $AcroDef_{MI^3}(capital, assets, French\ Republic) = 1,16 \times 10^{-4}$
4) $AcroDef_{MI^3}(capital, wealth, French\ Republic) = 1,0078 \times 10^{-4}$

5) $AcroDef_{MI^3}(capital, principal, French\ Republic) = 3,097 \times 10^{-6}$

Les deux premiers synonymes *seat of governement* et *city* sont les plus appropriés et ils obtiennent effectivement les plus grandes valeurs avec la mesure $AcroDef$. Ainsi, nous ne conservons que ces deux synonymes pour enrichir la question.

3- Entités Nommées : Après les deux étapes d'expansion des questions par *Projection & Hyperonymes* et *Synonymes*, nous utilisons IdentFinder (Bikel *et al.*, 1999) pour affecter une catégorie sémantique pour certains noms dans les questions. IdentFinder est capable d'annoter 7 types d'entités nommées *Person*, *Description*, *Location*, *Profession*, *Money*, *Number* and *Date*.

5.3. *Phase 3 : Espace Vectoriel*

Dans cette phase, d'abord les mots vides sont éliminés grâce à une liste. Ensuite, tous les N-grammes sont extraits. Chaque N-gramme est considéré comme une dimension d'un espace vectoriel. Chaque question est alors convertie en un vecteur où le nombre d'occurrences est pondéré avec la mesure $tf \cdot idf$.

5.4. *Phase 4 : Classification*

Dans un premier temps, le modèle du classifieur est appris en appliquant la validation croisée. Le classifieur est appris en utilisant le corpus d'apprentissage constitué de 90% du corpus initial. De nouveau, les trois algorithmes d'apprentissages sont SVM, KNN et NB. Le modèle de classification est construit en combinant une séquence de deux classifieurs. Le premier classifie les questions dans les 6 classes générales, et le deuxième dans les 50 classes détaillées. Puis la paire de classifieurs est utilisée pour assigner une classe générale et une classe détaillée à chaque nouvelle question des 10% du corpus qui constitue notre corpus de test.

Le principe de l'algorithme 1 est le suivant. Pour chaque question nous appliquons *TreeTagger* pour lemmatiser les mots et pour avoir les catégories grammaticales, et nous appliquons *Stanford parser* pour extraire les dépendances grammaticales. Pour tous les noms de ce corpus, nous appliquons *Project & Hyperonym* Nous essayons de projeter ces noms sur les instances de notre taxonomie. Si la projection n'aboutit pas, nous avons recours à $Wordnet_{Hyp}$ pour chercher les hyperonymes. Ensuite, nous essayons de projeter ces hyperonymes sur les instances de la taxonomie. Dans une autre étape, nous extrayons tous les synonymes des noms en utilisant WordNet. Ensuite nous calculons la force de correlation entre le nom et ses synonymes $Noun$ en utilisant $AcroDef_{MI^3}(s, Noun, S_{GD})$, suivi par $Filtre(S_s, \beta)$ pour ne garder que ceux qui sont les plus pertinents. À cette étape, nous appliquons la fonction *Filter* pour supprimer les mots vides, nous obtenons un corpus lemmatisé, filtré et enrichi Q_{L_F}. L'application de *N-grammes* suivi de $tf \cdot idf$ va nous donner l'espace vectoriel *V*, qui va être utilisé ensuite par l'algorithme d'apprentissage choisi.

Algorithm 1: SACSEQ

Input: The Learning Corpus of questions Q, Taxonomy T, WordNet, AcroDef$_{MI^3}$, Threshold β

Output: Space Vector V ;

```
1  begin
2      foreach q in Q do
3          q_L = TreeTagger(q);
4          Q_L = Q_L ∪ q_L;
5      foreach q_L in Q_L do
6          S_GD = Dependencies(q_L) ;
7          foreach Noun in q_L do
8              S = φ ;
9              if Project(Noun, T) == True then
10                 Expand(Noun, q_L);
11             else
12                 S = Wordnet_Hyp(Noun);
13                 foreach s in S do
14                     if Project(s, T) == True then
15                         Expand(s, q_L);
16                         break ;
17             S_s = φ ;
18             S_s = Wordnet_Syn(Noun);
19             foreach s in S_s do
20                 AcroDef_MI3(s, Noun, S_GD) ;
21             Sort(S_s) ;
22             Filtre(S_s, β) ;
23             Expand(S_s, q_L);
24     Q_LF = Filter_St−W(Q_L);
25     Q_A = Part(90%, Q_LF);
26     foreach q in Q_A do
27         N − grammes(Q_A);
28     tf · idf(Q_A);
29     Vectorise(Q_A);
30     return V ;
31 end
```

6. Expérimentations

Dans cette section, nous présentons les résultats des différentes expérimentations que nous avons réalisées pour valider notre méthodologie. Nous étudierons plus particulièrement le point suivant :

– *Quelles sont les conséquences de l'expansion des questions sur la qualité de la classification ?* En effet, comme nous l'avons vu dans la section précédente, la com-

binaison de différentes descripteurs sans l'utilisation de ressources sémantiques améliore légèrement la performance des classifieurs. Nous souhaitons enrichir les questions par des termes sémantiques liés (e.g. *synonymes* et *hyperonymes*).

Lors des premières expérimentations, la classification est effectuée uniquement avec l'intégration de *Projection & Hyperonymes*, dont l'objectif est d'évaluer l'amélioration apportée par cette expansion. Nous limitons le calcul de fréquences à $tf \cdot idf$ et nous n'utilisons que des bi-grammes. Les résultats sont présentés dans le tableau 4.

Avec la méthode *Projection & Hyperonymes*, nous constatons que le pourcentage des questions bien classées est amélioré avec les trois algorithmes de classification, et significativement avec SVM qui est amélioré de 7,4% (de 73,5% à **80,9**%) (cf. table 4).

Algorithmes	**K-NN**	**SVM**	**NB**
Projection & Hyperonymes	78,6	80,9	80,1
Synonymes	78,1	78,6	77,9
Entités Nommées	76,5	76,6	76,3
Sacseq	84,9	**86,7**	85,2

Tableau 4. *Résultats de classification.*

Dans une deuxième série d'expérimentations, nous avons enrichi les questions avec les *synonymes*. Le tableau 4 décrit les résultats de la classification. La précision est encore améliorée avec les trois algorithmes d'apprentissage, et en particulier avec SVM, qui a été amélioré de 5,1% (de 73,5% à **78,6**%).

Le tableau 4 décrit les résultats de la classification lors de l'expansion des questions par les entités nommées. Encore une fois, le résultat avec tous les algorithmes est amélioré, spécialement avec SVM par **3,1%** (de 73,5% à **76,6**%). Les deux premières méthodes d'expansions décrits ci-dessus performent mieux que *entités nommées*.

Dans la dernière série d'expérimentations, nous avons appliqué toutes les étapes d'expansions sémantiques de notre approche SACSEQ. Le tableau 4 affiche les résultats de la classification. La précision est améliorée pour tous les algorithmes d'apprentissage, spécialement avec SVM de **13,2%** passant de 73,5% à **86,7**%.

7. Discussion

Comme nous avons pu le constater lors de nos expérimentations, le fait d'utiliser des aspect sémantiques et syntaxiques pour enrichir les questions a permis de valoriser le contexte des questions. De même, le fait d'étendre les questions par des informations sémantiques et en combinaison des différents descripteurs a amélioré considérablement la performance de classification. Revenons à présent sur les méthodes utilisées pour la classification, nous avons également pu montrer que notre approche

surpasse les approches de classification traditionnelle. Cette différence montre bien que le fait d'enrichir les questions apporte un réel bénéfice dans le cadre de la classification des questions.

Les limites auxquelles nous avons fait face sont dues au plusieurs facteurs. Les analyseurs (*Treetagger*, *Stanford parser* et *IdentFinder*) peuvent produire des résultats incorrects ce qui influence l'extraction des *dépendances grammaticales* et la sélection selon des catégories grammaticales prédéfinies. En plus, le nombre de pages retourné par Google est approximatif, ce qui influe sur la qualité des mots retenus. Aussi, nous avons trouvé une ambigüité concernant l'association des catégories des questions. Par conséquent nous trouvons plusieurs questions qui peuvent être associées à différentes catégories de questions.

Cependant, notre méthode d'expansion était décisive et ciblée, puisque nous cherchions la corrélation entre les mots initiaux et tous les synonymes qui pourraient être utilisés dans un contexte donné. Le fait d'appliquer un filtre $AcroDef_{IM^3}$ sur les synonymes à retenir a également permis de bien cibler la qualité des synonymes pertinents par rapport au contexte traité. Les résultats démontrent l'utilité de notre méthode d'expansion sémantique et les caractéristiques de combinaison des descripteurs.

8. Conclusion

Une nouvelle méthode pour enrichir automatiquement les questions par des synonymes, et des hyperonymes tout en conservant le contexte sémantique a été présenté. Différents caractéristiques de représentation des questions sont examinées. Leur influence sur la performance des classifieurs a été déterminée. Les expériences ont été menés sur des données réelles. Cela a démontré l'utilité de notre méthode pour améliorer l'efficacité de la classification.

Les perspectives à ce travail sont nombreuses. Tout d'abord, notre méthode dépend de la qualité et du nombre de questions dans le corpus d'apprentissage. Nous tenons à étudier la relation entre la performance de la classification et la taille du corpus d'apprentissage. Deuxièmement, dans le présent document, nous nous sommes concentrés sur l'utilisation des machines d'apprentissage pour la détection de la catégorie des questions. Nous prévoyons d'étendre la première étape de la classification par l'application d'un ensemble de règles construites manuellement qui couvrent les 6 grandes catégories de questions. Troisièmement, nous allons étendre ce travail pour intégrer un système de question réponse interactif. Enfin, nous proposons de compléter ce travail exploratoire dans un système de question réponse dans le contexte de la recherche d'informations dans un corpus de documents structurés.

9. Bibliographie

Bikel D. M., Schwartz R., Weischedel R. M., « An Algorithm that Learns What's in a Name », *Machine Learning*, vol. 34, n° 1-3, p. 211-231, 1999.

Downey D., Broadhead M., Etzioni O., « Locating Complex Named Entities in Web Text », *in* M. M. Veloso (ed.), *IJCAI*, p. 2733-2739, 2007.

Hacioglu K., Ward W., « Question classification with support vector machines and error correcting codes », *NAACL '03 : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Association for Computational Linguistics, Morristown, NJ, USA, p. 28-30, 2003.

Kangavari M., Ghandchi S., Golpour M., « A New Model for Question Answering Systems », *World Academy of Science, Engineering and Technology*, vol. 32, p. 536-543, 2008.

Khardon R., Roth D., Valiant L. G., « Relational Learning for NLP using Linear Threshold Elements », *IJCAI '99 : Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 911-919, 1999.

Kosseim L., Yousefi J., « Improving the performance of question answering with semantically equivalent answer patterns », *Data Knowledge Engineering*, vol. 66, n° 1, p. 53-67, 2008.

Krishnan V., Das S., Chakrabarti S., « Enhanced answer type inference from questions using sequential models », *HLT '05 : Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Morristown, NJ, USA, p. 315-322, 2005.

Marneffe M.-C., Manning C. D., « The Stanford typed dependencies representation », *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, 2008.

Plamondon L., Lapalme G., Kosseim L., « The QUANTUM Question Answering System at TREC-11 », *Proceedings of The Tenth Text Retrieval Conference*, p. 157-165, 2003.

Quinlan J. R., « Induction of Decision Trees », *Machine Learning*, vol. 1, n° 1, p. 81-106, 1986.

Roche M., Prince V., « *croDef* : A Quality Measure for Discriminating Expansions of Ambiguous Acronyms », *in* B. N. Kokinov, D. C. Richardson, T. Roth-Berghofer, L. Vieu (eds), *CONTEXT*, vol. 4635 of *Lecture Notes in Computer Science*, Springer, p. 411-424, 2007.

Salton G., *The SMART Retrieval System—Experiments in Automatic Document Processing*, Prentice-Hall, Inc, Upper Saddle River, NJ, USA, 1971.

Salton G., Buckley C., « Term-Weighting Approaches in Automatic Text Retrieval », *Inf. Process. Manage.*, vol. 24, n° 5, p. 513-523, 1988.

Saxena A., Sambhu G., Subramaniam L., Kaushik S., « IITD-IBMIRL System for Question Answering Using Pattern Matching, Semantic Type and Semantic Category Recognition », *The Fourteenth Text REtrieval Conference (TREC 2007)*, 2007.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Conference on New Methods in Language Processing*, 1994.

Xin L., Dan R., « Learning question classifiers », *Proceedings of the 19th international conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 556-562, 2002.

Xin L., Dan R., « Learning question classifiers : the role of semantic information », *Nat. Lang. Eng.*, vol. 12, n° 3, p. 229-249, 2006.

Zhang D., Lee W. S., « Question classification using support vector machines », *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 26-32, 2003.