TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Apprentissage Automatique de Paraphrases pour l'Amélioration d'un Système de Questions-Réponses

Florence Duclaye (1 et 2), Olivier Collin (1), François Yvon (2)
(1) France Télécom R&D
2 avenue Pierre Marzin
22307 Lannion Cedex
{florence.duclaye,olivier.collin}@rd.francetelecom.fr
(2) GET/ENST et LTCI, CNRS URA 820
46 rue Barrault
75624 Paris Cedex 13
{fduclaye,yvon}@enst.fr

Mots-clefs – Keywords

Questions-Réponses, Apprentissage Automatique, Acquisition de Paraphrase
Question Answering, Machine Learning, Paraphrase extraction

Résumé - Abstract

Dans cet article, nous présentons une méthodologie d'apprentissage faiblement supervisé pour l'extraction automatique de paraphrases à partir du Web. À partir d'un seule exemple de paire (prédicat, arguments), un corpus est progressivement accumulé par sondage du Web. Les phases de sondage alternent avec des phases de filtrage, durant lesquelles les paraphrases les moins plausibles sont éliminées à l'aide d'une procédure de clustering non supervisée. Ce mécanisme d'apprentissage s'appuie sur un système de Questions-Réponses existant et les paraphrases apprises seront utilisées pour en améliorer le rappel. Nous nous concentrons ici sur le mécanisme d'apprentissage de ce système et en présentons les premiers résultats.

In this paper, we present a nearly unsupervised learning methodology for automatically extracting paraphrases from the Web. Starting with one single instance of a pair (predicate,arguments), a corpus is incrementally built by sampling the Web. Sampling stages alternate with filteringstages, during which implausible paraphrases are filtered out using an EM-based unsupervised clustering procedure. This learning machinery is built on top of an existing question-answering system and the learnt paraphrases will eventually be used to improve its recall. We focus here on the learning aspect of this system and report preliminary results.

1 Introduction

Les systèmes de Questions-Réponses (Voorhees, 1999) nécessitent des outils efficaces et sophistiqués de traitement automatique des langues, capables de traiter la variabilité linguistique des questions et des réponses: une même signification peut, en effet, être véhiculée par de multiples structures lexico-syntaxiques. Cette variabilité est une source de difficultés dans la plupart des applications de traitement automatique des langues.

Cette variabilité se manifeste au niveau syntaxique, où elle prend, par exemple, la forme de transformations régulières de la voix actice à la voix passive. Un traitement plus systématique de ce phénomène nécessite néanmoins des connaissances sémantiques, comme celles disponibles dans des réseaux sémantiques (Miller *et al.*, 1990). Le bénéfice de telles ressources est toutefois limité car (i) les relations de synonymie qu'elles contiennent sont définies hors de tout contexte d'usage; (ii) la synonymie implique une notion de la paraphrase qui est beaucoup trop restreinte pour l'application visée. La réponse à une question est en effet souvent exprimée à l'aide de termes qui ne sont que faiblement (par ex. métaphoriquement) liés à ceux de la question. Ainsi l'expression "X a causé Y" peut être considérée comme sémantiquement similaire à "la responsabilité de Y est attribuée à X" dans le contexte des Questions-Réponses (Lin & Pantel, 2001). Au lieu d'essayer de compléter manuellement ces ressources statiques, nous avons choisi d'exploiter les avantages d'une approche fondée sur des corpus et d'apprendre de telles équivalences de manière automatique. Nous utilisons le terme de paraphrase pour faire référence à ces relations, bien que la définition du terme adoptée ici soit surtout focalisée sur deux types de phénomènes linguistiques : les paraphrases linguistiques et les dérivations sémantiques. (Fuchs, 1982) décrit les paraphrases comme des phrases dont le sens linguistique dénotatif est équivalent. Les dérivations sémantiques sont des phrases dont le sens est préservé mais dont la structure lexico-syntaxique est différente (ex : *AOL a acheté Netscape* / *l'acquisition de Netscape par AOL*). Le corpus utilisé pour acquérir les paraphrases est le Web. Cette utilisation du Web comme corpus offre plusieurs avantages (Grefenstette, 1994). (i) Les informations qu'il contient sont d'une grande variété et d'une grande redondance, une même information pouvant apparaître sous de multiples formes. Notre algorithme d'apprentissage repose fortement sur cette propriété. (ii) Le Web contient des informations contextuelles pouvant contraindre la portée de la relation de paraphrase. En outre, comme notre système de Questions-Réponses utilise le Web comme unique source d'information, il est important d'extraire les formulations d'un concept qui sont les plus fréquemment utilisées sur le Web. Cette stratégie n'est pas sans difficultés: la réduction du niveau de bruit dans les données extraites est en particulier un problème important. Le mécanisme d'apprentissage que nous proposons est capable d'acquérir automatiquement de multiples formulations d'une relation sémantique donnée à partir d'*un seul exemple*. Cette donnée de départ consiste en un exemple de la relation sémantique visée, où l'expression linguistique (formulation) de la relation et le couple d'arguments ont tous deux été identifiés. Ce type de données est directement fourni par notre système de Questions-Réponses, mais il est également largement disponible dans les dictionnaires. Étant donné cet exemple positif, notre mécanisme d'apprentissage envoie de manière répétitive des requêtes sur le Web et utilise alternativement les formulations connues pour acquérir des nouveaux couples d'arguments, et les couples d'arguments connus pour trouver de nouvelles formulations. Ce mécanisme se décompose en deux étapes: d'une part la recherche de paraphrases potentielles de la relation sémantique et d'autre part la validation de ces paraphrases en se basant sur des comptages de fréquences et sur l'algorithme d'Estimation-Maximisation (EM).

Cet article présente à la Section 2 les travaux techniques de l'état de l'art ayant influencé notre

approche, ainsi que les travaux de recherche liés à l'apprentissage automatique de paraphrases. La Section 3 décrit ensuite les détails du fonctionnement de notre système. Avant de conclure, la Section 4 présente quelques résultats expérimentaux obtenus qui permettent de mettre en évidence l'intérêt de notre approche.

2 État de l'art

2.1 Apprentissage automatique de paraphrases

Comme les paraphrases peuvent être utilisées dans de nombreux contextes et applications, leur apprentissage peut être réalisé à l'aide de diverses méthodologies. (Barzilay & McKeown, 2001) distinguent trois méthodes différentes pour la collecte des paraphrases. La première est leur collecte manuelle, la seconde est l'utilisation de ressources linguistiques existantes et la troisième est l'extraction de mots ou d'expressions similaires en se basant sur un corpus. De ces trois méthodes, la première est sans doute la plus facile à implémenter, mais probablement la plus fastidieuse et la plus longue.

Les ressources linguistiques tels que les dictionnaires peuvent s'avérer utiles pour la collecte ou la génération de paraphrases. Par exemple, (Kurohashi & Sakai, 1999) utilise un dictionnaire construit manuellement pour reformuler des groupes nominaux ambigus en groupes verbaux. Ces ressources peuvent être utiles pour des besoins de désambiguïsation, mais en l'absence d'information contextuelle supplémentaire, les relations de synonymie qu'elles contiennent doivent être utilisées avec précaution. De plus, elles sont souvent considérées comme peu adaptées aux traitements automatiques (Habert *et al.*, 1997). (Torisawa, 2001) propose une méthode basée sur l'algorithme d'Estimation-Maximisation pour sélectionner les constructions verbales servant à paraphraser certaines expressions.

Enfin, certains travaux menés dans le domaine de l'extraction basée sur un corpus de mots ou d'expressions similaires s'appuient sur l'hypothèse distributionnelle de Harris, selon laquelle les mots apparaissant dans le même contexte tendent à avoir des sens similaires. Partant de ce postulat, (Barzilay & McKeown, 2001) et (Akira & Takenobu, 2002) travaillent sur un ensemble de corpus alignés et utilisent des informations contextuelles basées sur des similarités lexicales pour extraire des paraphrases. De manière similaire, (Lin & Pantel, 2001) utilise un algorithme non supervisé pour la découverte de règles d'inférence à partir de textes. Au lieu d'appliquer l'hypothèse harrissienne aux mots, les auteurs l'appliquent à des chemins dans les arbres de dépendance d'un corpus analysé.

2.2 Extraction d'informations par bootstrapping

Des travaux récents menés en extraction d'informations nous fournissent des approches intéressantes, pouvant être adaptées au problème de l'apprentissage automatique de paraphrases. Ainsi, (Riloff & Jones, 1999) décrit un système d'extraction d'informations reposant sur un mécanisme de bootstrapping à deux niveaux. Le niveau de "bootstrapping mutuel" construit alternativement un lexique et des patrons d'extraction contextuels. Le niveau de "meta-bootstrapping" ne conserve que les cinq meilleurs nouveaux termes extraits durant une itération d'apprentissage, avant de poursuivre avec le boostrapping mutuel. L'auteur parvient ainsi à réduire la quantité

de termes invalides trouvés en appliquant les patrons d'extraction.

La technique DIPRE (Dual Iterative Pattern Relation Extraction), présentée dans (Brin, 1998) est aussi une méthode de bootstrapping, utilisée pour l'acquisition de paires (auteur, titre) à partir d'un corpus de documents du Web. À partir d'une collection d'exemples de tels exemples, l'auteur construit des patrons d'extraction utilisés pour collecter de nouvelles paires (auteur, titre). À leur tour, ces paires sont recherchées dans le corpus et sont utilisées pour construire de nouveaux patrons d'extraction, et ainsi de suite.

Enfin, (Collins & Singer, 1999) décrit une méthode de reconnaissance d'entités nommées capable d'apprendre à partir d'un faible nombre de données de supervision, en construisant en parallèle deux classifieurs utilisant des ensembles disjoints d'attributs.

3 Description du système d'apprentissage de paraphrases

3.1 Fonctionnement global du système

Notre algorithme d'inférence de paraphrases commence son apprentissage à partir d'un unique exemple positif et utilise un mécanisme de bootstrapping à deux niveaux. Cet exemple de départ est, par exemple, une réponse automatiquement calculée par notre système de Questions-Réponses. Dans notre modèle, un exemple est représenté comme l'association d'une formulation linguistique f d'un prédicat avec son couple d'arguments a. Par exemple, la relation d'auteur pourrait être représentée ainsi : f="être l'auteur de", a=("Melville", "Moby Dick"). L'identification des paraphrases repose sur un modèle de décision probabiliste dont les paramètres sont estimés de manière presque non supervisée. L'estimation repose sur un algorithme de clustering fondé sur l'algorithme EM (voir la Section 3.2): il prend en entrée une matrice contenant les fréquences de cooccurrence d'un ensemble de formulations F et des couples d'arguments correspondants A, mesurées dans le corpus C.

Notre corpus initial C_i contient un unique exemple de départ exprimant la relation sémantique visée, représenté comme la cooccurrence d'une formulation f_i et d'un couple d'arguments a_i. Avec ces données de départ, nous souhaitons construire un nouveau corpus C contenant potentiellement beaucoup plus d'exemples de la relation sémantique visée. Cette tâche est réalisée en utilisant indépendemment f_i et a_i pour formuler des requêtes sur le Web. Les documents trouvés sont parcourus pour y trouver de nouvelles formulations et paires d'arguments intéressantes, qui sont utilisées successivement pour produire de nouvelles requêtes, ces dernières étant à leur tour utilisées pour extraire plus d'arguments et de formulations... Durant cette étape, nous avons donc besoin de (i) générer des requêtes et traiter les documents trouvés afin de (ii) extraire de nouvelles formulations et de nouveaux couples d'arguments. De plus amples détails concernant ces procédures sont donnés dans la Section 3.3.

La qualité des paraphrases extraites dépend beaucoup de notre capacité à maintenir le corpus suffisamment *focalisé sur la relation sémantique visée*: pour ce faire, les phases d'acquisition sont entrecoupées d'étapes de filtrage reposant sur notre clustering à base d'EM. Le filtrage est en effet une étape cruciale pour assurer la convergence de cette procédure. L'architecture globale de notre système est représentée sur la figure 1.

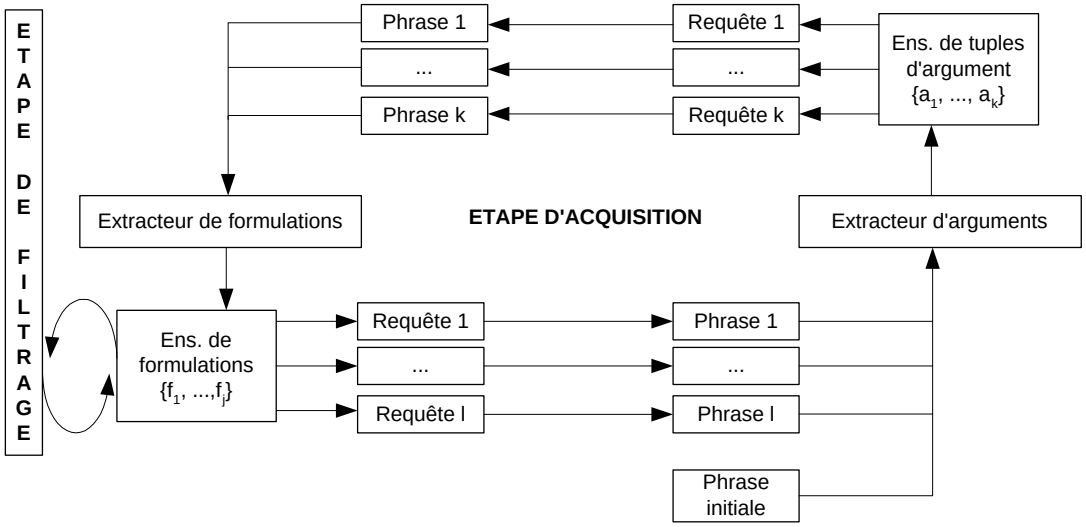

Figure 1: Système d'apprentissage automatique de paraphrases

3.2 Filtrage par l'algorithme d'Estimation-Maximisation

Le filtrage consiste à distinguer les paraphrases valides de la relation sémantique de départ des paraphrases qui sont incorrectes. Ce filtrage intervient à deux endroits de notre mécanisme d'apprentissage: (i) pour identifier les formulations qui vont déclencher une nouvelle série de requêtes et donc une nouvelle itération; (ii) pour sélectionner les paraphrases qui seront finalement conservées (voir la Figure 1). Cette étape revient à classifier chaque formulation du corpus comme 1 (paraphrase valide) ou 0 (paraphrase invalide), en se basant sur des données de cooccurrence entre couples d'arguments et formulations. Ce problème de partitionnement en 2 classes est faiblement supervisé car nous ne disposons initialement que d'un seul exemple étiqueté (positif): la formulation de départ. Il est possible d'utiliser pour ce problème des algorithmes de clustering utilisant des des données de cooccurrence de type EM (Hofmann & Puzicha, 1998). Nous considérons donc que chaque phrase (consistant en une formulation f et ses arguments a) est générée par le modèle stochastique suivant:

$$P(f,a) = \Sigma_{s\in S} P(f,a|s)P(s) \quad (1)$$
$$= \Sigma_{s\in S} P(f|s)P(a|s)P(s) \quad (2)$$

où S est l'ensemble des relations sémantiques exprimées par des phrases de notre corpus. Nous considérons également que notre corpus ne contient que deux relations sémantiques, dont les valeurs sont soit $S = 1$, signifiant qu'une phrase donnée exprime la même relation sémantique que la phrase de départ, soit $S = 0$, signifiant que la phrase exprime une autre relation sémantique (non spécifiée). Étant donné ce modèle, les formules de réestimation se dérivent facilement (Hofmann & Puzicha, 1998). Elles sont présentées dans la Table 1, où $N()$ dénote la fonction de comptage.

Ce modèle nous permet d'incorporer des connaissances durant la phase d'initialisation, où nous utilisons les valeurs suivantes : $P(S = 1|f_i, a_i) = 1$ et $P(S = 1|f_i, a) = 0.6, \forall a \neq a_i$ dans l'équation (3). Toutes les autres valeurs de $P(S|F, A)$ sont égales à 0.5. EM est ensuite

E-Step

$$P(s|f,a) = \frac{P(s)P(f|s)P(a|s)}{\sum_i P(s_i)P(f|s_i)P(a|s_i)} \quad (3)$$

M-Step

$$P(a|s) = \frac{\sum_{f\in F} N(f,a)P(s|f,a)}{\sum_{a\in A}\sum_{f\in F} N(f,a)P(s|f,a)} \quad (4)$$

$$P(f|s) = \frac{\sum_{a\in A} N(f,a)P(s|f,a)}{\sum_{f\in F}\sum_{a\in A} N(f,a)P(s|f,a)} \quad (5)$$

$$P(s) = \frac{\sum_{f\in F}\sum_{a\in A} N(f,a)P(s|f,a)}{\sum_{f\in F}\sum_{a\in A} N(f,a)} \quad (6)$$

Table 1: Formules de réestimation pour EM

lancé jusqu'à convergence des paramètres maximisés. Dans notre cas, cette convergence est généralement atteinte au bout de 10 itérations.

Une fois les paramètres appris, nous utilisons ce modèle pour décider si une formulation f est une paraphrase valide en nous basant sur le rapport entre $P(S=1|f)$ et $P(S=0|f)$, calculé comme suit: $r = \frac{P(S=1)P(f|S=1)}{P(S=0)P(f|S=0)}$. Étant donné que $P(S=1)$ est fortement sur-estimée dans notre corpus (qui est précisément focalisé autour de tels exemples), la règle de décision utilisée impose que ce rapport soit supérieur à un seuil pré-défini $\theta >> 1$. De manière alternative, nous avons également considéré des scénarios dans lesquels ces probabilités ne servent qu'à ordonner les candidats paraphrases, que ce soit pendant les différentes étapes de filtrage ou même lors de la décision finale. Ceci rendant finalement notre approche moins dépendante de la validité des hypothèses sous-jacentes au modèle probabiliste utilisé, dont certaines sont discutables:en particulier à la condition d'indépendance exprimée en 2, ou encore l'hypothèse que seules deux relations sémantiques sont représentées dans notre corpus.

3.3 Procédure d'acquisition automatique

L'outil utilisé pour la phase d'acquisition est un système de Questions-Réponses, fonctionnant ici comme un outil d'extraction d'informations. Ce système est constitué de deux composants principaux: le premier transforme une question en entrée en une requête sur le Web et déclenche la recherche; le second analyse les pages retournées (plus précisément les extraits de page) et y cherche des réponses potentielles, par appariemment de patrons d'extraction pré-définis. La requête et les patrons d'extraction sont dérivés de la question de départ à l'aide de règles. Les détails concernant ce système de QA et les procédures d'analyse linguistique impliqués à chaque étape du traitement sont donnés dans (Duclaye *et al.*, 2002). En mode "extraction d'information", l'étape de construction de requête est supprimée, la requête étant déduite des couples d'arguments (ou de formulations). La phase d'analyse utilise des patrons d'extraction très généraux, construits à partir des arguments (ou formulations) en cours de traitement. Supposons, par exemple, que nous recherchons des paraphrases, la paire d'arguments courante étant égale à ["Melville", "Moby Dick"]. Ces arguments seront tous deux utilisés comme mots-clés, et deux patrons d'extraction sont utilisés pour faire les extractions dans les documents trouvés : "Melville [verb] Moby Dick" et "Moby Dick [verb] Melville". Dans cet exemple, il est

nécessaire qu'un verbe apparaîsse entre les deux mots-clés pour être extrait. Ce verbe sera considéré comme une paraphrase potentielle de la formulation de départ. Pour chaque requête, seuls les N premiers documents retournés par le moteur de recherche sont pris en compte. Les paires (arguments, formulations) ainsi extraites sont accumulées, itération après itération, pour constituer un corpus sur lequel des statistiques utilisées lors du filtrage sont calculées. Ce processus itératif d'acquisition de formulations et de couples d'arguments, combiné avec celui de validation/filtrage, converge et se termine quand aucune nouvelle formulation n'est trouvée.

4 Résultats expérimentaux

Les expériences décrites dans cette section ont été réalisées sur 18 phrases initiales, représentant 12 relations sémantiques différentes. La Table 2 présente quelques exemples de relations, ainsi que les formulations et couples d'arguments choisis. Pour chacune de ces phrases, la procédure d'apprentissage décrite à la Section 3 a été lancée sur une itération. Les résultats présentés ici ont été obtenus en prenant les N=1000 premiers résumés retournés par le moteur de recherche.

achat de	"acheter"	AOL; Netscape
auteur de	"écrire"	Melville; Moby Dick
inventeur de	"inventer"	Gutenberg; imprimerie
assassinat de	"assassiner"	Oswald; Kennedy

Table 2: Exemples de relations avec leurs formulations et couples d'arguments

Les paraphrases extraites ont été vérifiées manuellement et classées comme valides ou invalides. Dans cette application, le succès peut être mesuré comme la précision moyenne des paraphrases extraites, qui devraient à terme être ajoutées au système de Questions-Réponses. Le rappel, par contre, n'est pas important car nous souhaitons simplement trouver les paraphrases les plus fréquentes. Le taux de sélection représente le pourcentage de formulations classées comme valides par notre système. Rappelons que la décision de classer une formulation comme une paraphrase valide ou invalide est basée sur le rapport entre $log(P(S = 1|f))$ et $log(P(S = 0|f))$, appelé θ. Les taux de sélection et les résultats de précision pour différentes valeurs de θ sont donnés dans la Table 3.

θ	7	25	48	117	186	232
Taux de sélection	44.0%	29.8%	23.9%	14.2%	10%	9.4%
Précision	42.9%	47.3%	47.3%	54.9%	66.6%	65.4%

Table 3: Résultats expérimentaux

Dans ces expérimentations, la meilleure précision moyenne atteinte est de 66.6%, quand $\theta = 186$. Effectuées sur plusieurs relations sémantiques, ces expérimentations ont montré que le taux de précision peut varier de manière importante d'une relation sémantique à une autre: il peut atteindre 100% pour certaines relations, et descendre jusqu'à 6% pour d'autres. Ces résultats peuvent paraître faibles. Cela est dû en partie à la quantité variable de données extraites du Web pour les relations sémantiques. Le fait d'appliquer le même seuil θ à toutes les relations sémantiques n'est certainement pas la meilleure méthode. De plus, la majorité des formulations

classées à tort comme de bonnes paraphrases sont thématiquement liées à la formulation de départ et ne peuvent donc être considérées comme totalement mauvaises.

Comme indiqué dans la Table 3, l'augmentation des valeurs de θ provoque la diminution du taux de sélection et l'augmentation de la précision. La tendance générale est que plus θ augmente, plus la quantité de formulations classées comme mauvaises paraphrases augmente, de sorte que finalement, seule la formulation de départ est conservée comme valide. Augmenter θ n'est donc pas suffisant pour améliorer la précicion moyenne des paraphrases extraites. Il est nécessaire de trouver un équilibre entre le taux de sélection et la précision des paraphrases extraites.

Une autre stratégie de filtrage consiste à conserver les k meilleures formulations à chaque itération ($k = 5, 10, ...$). La Table 4 porte sur la première itération de la relation d'achat et compare les taux de précision obtenus pour différents seuils de k. La deuxième colonne représente les taux de précision dans l'ensemble (de taille k) des formulations classées comme paraphrases valides. La troisième colonne représente les taux de précision dans l'ensemble (de taille toujours k) des formulations classées comme paraphrases invalides. Il est intéressant de noter que les formulations classées comme paraphrases invalides ont globalement une meilleure précision que celles classées comme valides. Dans de prochaines expérimentations, on pourrait envisager d'utiliser ces formulations classées comme paraphrases invalides comme exemples négatifs d'apprentissage.

	Formu. classées en paraph. valides	Formu. classées en paraph. invalides
Classe entière	39.6%	85.2%
k=5	60%	80%
k=10	80%	80%
k=15	73.3%	73.3%

Table 4: Taux de précision en fonction de k, pour la relation d'achat

Des expériences complémentaires ont été conduites sur plusieurs itérations, en ne conservant que les k=5 meilleures formulations à chaque itération. La Table 5 montre les résultats obtenus pour la relation d'achat, après cinq itérations d'apprentissage. On note que la précision augmente entre la première (60%) et la cinquième (80%) itération. Des expériences similaires sont en cours pour d'autres relations sémantiques.

Itération.	Formulations classées comme paraphrases valides
1	racheter, acquérir, acheter, utiliser, recevoir
2	racheter, acquérir, acheter, reprendre, absorber
3	racheter, acheter, acquérir, qui racheter, devenir
4	racheter, acheter, acquérir, absorber, grouper
5	racheter, acheter, reprendre, devenir, acquérir

Table 5: Résultats sur cinq itérations d'apprentissage pour la relation sémantique d'achat

5 Conclusions et perspectives

Dans cet article, nous avons présenté une méthodologie faiblement supervisée pour l'apprentissage automatique de paraphrases, commençant avec un unique exemple positif d'apprentissage. En

utilisant une stratégie de validation basée sur l'algorithme EM, nous pouvons filtrer les paraphrases potentielles invalides extraites durant les phases d'acquisition. Non seulement ces paraphrases sont utiles pour améliorer les résultats de notre système de Questions-Réponses, mais les couples d'arguments acquis pourraient également être utilisés pour d'autres besoins que l'apprentissage de paraphrases, comme la construction de lexiques sémantiques. Dans cette optique, l'étape de filtrage pourrait aussi bien être appliquée aux couples d'arguments acquis.

Au-delà de résultats expérimentaux prometteurs, obtenus dans un scénario relativement simple, de nombreuses améliorations portant sur les phases d'acquisition et de validation sont actuellement envisagées. Concernant l'étape de filtrage, les développements concernent principalement (i) une variante consistant à conserver des informations sur les valeurs des paramètres du modèle stochastique entre deux étapes successives de filtrage; (ii) l'utilisation de stratégies incrémentales les paraphrases potentielles qui seront utilisées dans de nouvelles requêtes pour augmenter le corpus d'exemples; (iii) l'utilisation d'autres algorithmes de filtrages, exploitant des proximités distributionnelles entre la formulation d'origine et les autres formulations trouvées sur Internet. Le but recherché étant d'obtenir un maximum d'exemples différents, tout en gardant le corpus en expansion suffisamment focalisé sur la relation sémantique en cours d'examen.

Concernant la phase d'acquisition, nous projetons d'apprendre des paraphrases multilingues, ainsi que des structures plus complexes de formulations (comme les nominalisations). Nous projetons également d'utiliser des informations de contexte automatiquement apprises, afin d'améliorer la qualité des requêtes soumises au moteur de recherche: l'idée est d'extraire, dans le voisinage lexical des paraphrases identifiées comme valides, des termes discriminant permettant (i) de raffiner et/ou de varier les requêtes et (ii) de qualifier plus finement le contexte (thématique) dans lequel les relations de paraphrases sont valides. Il apparaît en effet clairement que de nombreuses relations de paraphrases de valent que dans un contexte bien défini, qu'il est essentiel de pouvoir décrire.

Basé sur une stratégie d'apprentissage indépendante de la langue, notre système d'apprentissage de paraphrases sera intégré au système de Questions-Réponses. Notre système fonctionnera alors comme un composant indépendant du module de QA et apprendra des paraphrases à partir des réponses fournies par le système de QA. Son intégration ne nécessite en fait que peu de développements nouveaux, dans la mesure où notre système de Questions-Réponses intègre déjà des règles de paraphrasage entrées manuellement. Il ne s'agit donc que d'automatiser ce processus d'ajout de règles de paraphrasage des questions et des réponses. Ceci nous permettra d'évaluer dans un contexte applicatif notre méthodologie et de mesurer les améliorations apportées par les paraphrases extraites.

Références

AKIRA T. & TAKENOBU T. (2002). Automatic disabbreviation by using context information. In *Proceedings of the NLPRS Workshop on Automatic Paraphrasing : Theories and Applications*.

BARZILAY R. & MCKEOWN K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceeding of the 39th Annual Meeting of the Association for Computational Linguistics*, p. 50–57, Toulouse.

BRIN S. (1998). Extracting patterns and relations from the world wide web. In *Proceedings of WebDB Workshop at EDBT*.

COLLINS M. & SINGER Y. (1999). Unsupervised models for named entity classification. In *Proceedings of the Workshop on Empirical Methods for Natural Language Processing*.

DUCLAYE F., FILOCHE P., SITKO J. & COLLIN O. (2002). A polish question-answering system for business information. In *Proceedings of the Business Information Systems Conference*, Poznan.

FUCHS C. (1982). *La Paraphrase*. Presses Universitaires de France.

GREFENSTETTE G. (1994). *Explorations in Automatic Thesaurus Discovery*. Boston: Kluwer Academic Publishers.

HABERT B., NAZARENKO A. & SALEM A. (1997). *Les linguistiques de corpus*. Armand Colin, Paris.

HOFMANN T. & PUZICHA J. (1998). *Statistical Models for Co-occurrence Data*. Rapport interne AI. 1625, MIT, AI Lab.

KUROHASHI S. & SAKAI Y. (1999). Semantic analysis of japanese noun phrases : a new approach to dictionary-based understanding. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, p. 481–488.

LIN D. & PANTEL P. (2001). Discovery of inference rules for question-answering. In *Natural Language Engineering*, volume 7, p. 343–360.

MILLER G., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. (1990). Introduction to wordnet: An on-line lexical database. In *Journal of Lexicography*, volume 3, p. 234–244.

RILOFF E. & JONES R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence*.

TORISAWA K. (2001). A nearly unsupervised learning method for automatic paraphrasing of japanese noun phrases. In *Proceedings of the NLPRS 2002 workshop on Automatic Paraphrasing : Theories and Applications*, Tokyo.

VOORHEES E. (1999). The TREC-8 question answering track report. In *Proceedings of TREC-8*.