RECITAL 2011, Montpellier, 27 juin – 1er juillet 2011

Analyse de l'ambiguïté des requêtes utilisateurs par catégorisation thématique

Fanny Lalleman[1, 2]
(1) CLLE & CNRS, 5, allées Antonio Machado 31058 Toulouse Cedex 9
(2) Orange Labs, 2, Avenue Pierre Marzin 22307 Lannion Cedex
fanny.lalleman@univ-tlse2.fr

Résumé. Dans cet article, nous cherchons à identifier la nature de l'ambiguïté des requêtes utilisateurs issues d'un moteur de recherche dédié à l'actualité, 2424actu.fr, en utilisant une tâche de catégorisation. Dans un premier temps, nous verrons les différentes formes de l'ambiguïté des requêtes déjà décrites dans les travaux de TAL. Nous confrontons la vision lexicographique de l'ambiguïté à celle décrite par les techniques de classification appliquées à la recherche d'information. Dans un deuxième temps, nous appliquons une méthode de catégorisation thématique afin d'explorer l'ambiguïté des requêtes, celle-ci nous permet de conduire une analyse sémantique de ces requêtes, en intégrant la dimension temporelle propre au contexte des news. Nous proposons une typologie des phénomènes d'ambiguïté basée sur notre analyse sémantique. Enfin, nous comparons l'exploration par catégorisation à une ressource comme Wikipédia, montrant concrètement les divergences des deux approches.

Abstract. In this paper, we try to identify the nature of ambiguity of user queries from a search engine dedicated to news, 2424actu.fr, using a categorization task. At first, we see different forms of ambiguity queries already described in the works of NLP. We confront lexicographical vision of the ambiguity to that described by classification techniques applied to information retrieval. In a second step, we apply a method of categorizing themes to explore the ambiguity of queries, it allow us to conduct a semantic analysis of these applications by integrating temporal context-specific news. We propose a typology of phenomena of ambiguity based on our semantic analysis. Finally, we compare the exploration by categorization with a resource as Wikipedia, showing concretely the differences between these two approaches.

Mots-clés : recherche d'information, ambiguïté, classification de requêtes.

Keywords: Information retrieval, ambiguity, classification queries.

1 Introduction

La désambiguïsation lexicale a été appliquée à la recherche d'information avec un succès variable. Le précurseur sur cette thématique (Weiss, 1973) a été suivi par un certain nombre de travaux (Krovetz & Croft, 1992; Stokoe *et al.*, 2003; Sanderson, 2000), où la désambiguïsation était focalisée sur les mots dont les sens étaient répertoriés dans des dictionnaires (Voorhees, 1993). L'enjeu était de réperer de la polysémie, c'est-à-dire différents sens pour une même forme. Les performances des systèmes n'ont pas été à la hauteur des espérances aboutissant à un questionnement sur l'adéquation des dictionnaires pour une telle tâche, et la question de la nature de l'ambiguïté. Les dictionnaires sont peu adaptés au traitement des requêtes, contenant peu de noms propres, ou de termes complexes et la nature de l'ambiguïté des requêtes se limite aux relations sémantiques contenues dans les dictionnaires (homonymie et polysémie). Ces difficultés ont rendu le traitement de l'ambiguïté secondaire en recherche d'information. Mais le développement du web et des moteurs de recherche grand public ont remis en avant la question de l'ambiguïté des requêtes, en particulier le CLIR (Cross-Language Information Retrieval) (Darwish & Oard, 2003). En effet, la traduction est très dépendante de la désambiguïsation, d'autant plus quand le contexte est absent. La taille des requêtes entre également en jeu dans les performances des systèmes de désambiguïsation. Ils se trouvent être relativement performants sur les requêtes dites « longues » provenant des campagnes d'évaluation comme TREC ou CLEF, mais en grande difficulté face des requêtes « courtes » provenant de moteurs de recherche. Actuellement, l'accent est mis sur les données réelles et l'intéractivité avec l'utilisateur. L'enjeu n'est plus de construire un système de désambiguïsation, mais de savoir identifier une requête ambigüe soit pour pouvoir proposer par exemple une présentation adaptée à l'utilisateur, comme une présentation hiérarchique (Hearst, 2009), soit pour améliorer les performances du système en lui-même. Les solutions actuelles préfèrent en présence de données réelles, se servir des logs de requêtes contenant les pages réponses choisies par les utilisateurs pour identifier les requêtes ambigües (Clough *et al.*, 2009). Nous nous situons dans un contexte applicatif spécifique, un moteur de recherche dédié à l'actualité, où l'ambiguïté des requêtes doit être abordée différemment. La base documentaire évolue avec l'actualité et ne conserve pas les documents anciens, il faut donc envisager une méthode différente qui ne se base pas sur les choix des utilisateurs pour essayer d'identifier l'ambiguïté.

Dans ce contexte, nous cherchons à identifier la nature de l'ambiguïté présente dans nos requêtes en les examinant grâce à une tâche de catégorisation. Cette approche doit nous permettre d'étudier la forme que peut prendre l'ambiguïté des requêtes produites dans un contexte applicatif. Dans un premier temps, nous allons passer en revue les différentes formes de l'ambiguïté déjà décrites dans les travaux de TAL, en confrontant la vision « traditionnelle » de l'ambiguïté des requêtes, lexicographique, à une vision produite par des techniques de classification et de clustering appliquées à la recherche d'information. Dans un deuxième temps, nous appliquerons ces techniques de classification pour explorer l'ambiguïté des requêtes dans notre contexte applicatif 2424actu.fr afin d'effectuer une analyse sémantique des requêtes. Cette analyse débouchera sur une typologie de l'ambiguïté dans ce cadre applicatif. Enfin, nous confronterons l'exploration par catégorisation à une ressource comme Wikipédia, en montrant les divergences de ces deux approches et l'importance d'une démarche d'analyse reflétant les particularités du contexte et de la base de texte à traiter.

2 L'ambiguïté dans les requêtes utilisateurs

2.1 L'ambiguïté des requêtes vue à travers les dictionnaires

L'ambiguïté des requêtes est souvent analysée à la lumière des dictionnaires et autres ressources (WordNet, Wikipédia) (Sanderson, 2000, 2008; Santamaría *et al.*, 2010). Le principe est de rechercher les termes des requêtes dans divers dictionnaires afin de répérer les mots ayant plusieurs sens. Comme en désambiguisation classique, WordNet est la principale source de comparaison, mais il lui est reproché d'être peu représentatif de la diversité présente dans les requêtes des utilisateurs de moteur de recherche. Il est donc utilisé en complément d'une autre ressource, en l'occurrence Wikipédia (Santamaría *et al.*, 2010; Sanderson, 2008; Clough *et al.*, 2009). Par exemple, (Sanderson, 2008) propose d'examiner l'ambiguïté en s'appuyant conjointement sur WordNet et sur les pages de Wikipédia, il collecte dans cette ressource les pages identifiant une forme ambigüe et les différents sujets qui portent le nom de cette forme. Grâce à ces deux types de ressources, Sanderson évalue l'ambiguïté présente dans deux corpus de requêtes provenant de deux moteurs de recherche (1 million de requêtes Live Search et 500 000 requêtes de UK's Press Association). 16% des requêtes Live Search sont estimées ambigües à la lu-

mière des ressources, et 23,6% pour les requêtes du moteur UK's Press Association. Il confirme l'importance du phénomène et suggère que l'ambiguïté varie selon les requêtes étudiées. Par ailleurs, on peut se demander si l'ambiguïté présente en recherche d'information est de la polysémie décrite dans les dictionnaires (*hôte* peut désigner la personne qui reçoit ou celle qui est reçue) ou bien de l'homonymie comme Wikipédia le propose (*éruption* comme nom ou *Eruption* groupe de musique).

D'autre part, il y a deux problèmes liés à cette approche de l'ambiguïté des requêtes. Le premier problème est lié aux ressources. Wikipédia ne couvre que 60% des sens présents dans un moteur de recherche (en anglais) et WordNet seulement 32 %(Santamaría *et al.*, 2010). Le deuxième problème est l'absence de prise en compte de l'environnement spécifique à la tâche de recherche d'information. En effet, l'ambiguïté d'une requête ne peut être envisagée en dehors de la base textuelle interrogée, l'utilisateur cherchant une information présente dans la base qu'il consulte et qu'il anticipe. L'ambiguïté typée dans des ressources comme des dictionnaires et des thésaurus généralistes n'est donc pas forcément adaptée pour caractériser l'ambiguïté effective des requêtes utilisateurs pour une application donnée.

2.2 L'ambiguïté « non classique »

Rompant avec la vision lexicographique de l'ambiguïté des requêtes, certains travaux proposent d'analyser d'autres formes d'ambiguïté. Ils avancent que des requêtes peuvent être ambigues si elles renvoient à plusieurs « sous-domaines » (Zhai *et al.*, 2003) ou à plusieurs « facettes »(Hearst, 2006). Ce type de distinction est à mettre en rapport avec l'utilisation de techniques de classification et de clustering en recherche d'information. Ces techniques permettent de faire apparaître des rapprochements pertinents sans poser d'a priori sur ce qui est recherché. On retrouve ce type de distinction vis à vis de l'ambiguïté des requêtes dans (Song *et al.*, 2009). Ils utilisent une typologie à trois éléments qu'ils ont définie dans le but d'annoter manuellement des requêtes :

1. Type A (requête ambigüe) : requête qui a plus d'un seul sens : « giant » > plusieurs référents Giant Company, Giant (film), San Francisco Giant (équipe de basket ball)
2. Type B (requête « large ») : requête qui couvre plusieurs sujets ou thématiques : « songs » > « song lyrics », « love songs », « download songs »
3. Type C (requête non ambigüe) : requête qui a un sens spécifique et un référent facilement identifiable : « Billie Holiday » (chanteuse jazz)

Les critères ont utilisé pour discriminer les requêtes de type A et C ne sont pas clairement établis. En effet, potentiellement l'exemple donné comme requête non ambigüe, « Billie Holiday » peut être discuté puisqu'il s'avère qu'il existe un album eponyme de Billie Holiday. La distinction entre « ambigüe » et « non ambigu » est donc difficile à établir, laissant apparaître des requêtes « entre les deux ». (Song *et al.*, 2009) les distinguent en créant une catégorie de requêtes contenant un terme générique (requête B). Le phénomène de « généralité » existe déjà dans les conceptions de l'ambiguïté sémantique (Aarts & MacMahon, 2006). Il décrit un phénomène d'ambiguïté différent de la polysémie et de l'homonymie dont le nom varie selon les auteurs : sous-spécification, sens vague, indétermination ou généralité. Cette forme d'ambiguïté décrit un sens général ou inclusif qui peut avoir différentes significations selon le contexte, par exemple « vache » va pourvoir prendre un sens général « animal ruminant » ou un sens spécifique « femelle du taureau ». Il se pose donc la question de la présence de ce type d'ambiguïté « non classique » en recherche d'information.

2.3 La question des entités nommées « ambigües »

L'utilisation de Wikipédia comme ressource pour capter l'ambiguïté des requêtes ou bien désambiguïser (Rahurkar *et al.*, 2008) est principalement due à la possibilité d'exploiter les entités nommées (ENs) qu'elle contient. Or, l'étude des requêtes utilisateurs de AllTheWeb et Altavista avaient mis en évidence que 11-17% des requêtes étaient composées d'un nom propre désignant une personne (Spink *et al.*, 2004). Les entités nommées désignant des lieux sont aussi très présentes. Gan et al. (Gan *et al.*, 2008) montrent dans une étude des logs de requêtes de AOL (2006) qu'il y a environ 38% des requêtes qui contiennent des termes de type « géographique » comme « New York » ou « Kentucky Fried Chicken ». Pour des moteurs de même type comme AOL et Altavista (moteurs généralistes), on arriverait environ à plus de 50% d'entités nommées. L'utilisation d'entitées nommées en nombre par les utilisateurs de moteurs de recherche renforce les phénomènes d'ambiguïté des requêtes, à tel point que

cela a justifié une tâche dédiée à la désambiguïsation des noms de personnes dans un contexte de recherche d'information nommée WePS (Artiles *et al.*, 2007, 2010). En effet, il est courant qu'un même nom propre désigne un grand nombre de personnes. (Artiles *et al.*, 2007) donnaient l'exemple du bureau de recensement américain qui rapporte 90 000 noms portés par 100 millions de personnes. L'homonymie n'est pas la seule cause d'ambiguïté des ENs. Elles forment un ensemble hétérogène composé d'une collection d'expressions linguistiques diverses, réunies sur la base de caractéristiques référentielles communes (Erhmann, 2008), c'est-à-dire que sont rassemblées sous la même étiquette différentes expressions comme les noms de lieux, de personnes ou des dénominations d'organisation, qui par nature, ne manifestent pas les mêmes propriétés. Par conséquent, l'ambiguïté apparaît sous différentes formes : métonymie (« la France a gagné 3-0 »), facettes référentielles (Carla Bruni, chanteuse, ancienne mannequin, épouse du président de la République) ou comme on l'a vu homonymie. Face à une telle variabilité, l'utilisation de ressources pour traiter l'ambiguïté des Entités Nommées a des limites. C'est pourquoi les techniques de clustering se sont développées pour tenter de désambiguïser les ENs (Santamaría *et al.*, 2010; Bernardini *et al.*, 2009).

2.4 La question de l'évaluation de l'ambiguïté

Actuellement, la question de la forme que prend l'ambiguïté des requêtes utilisateurs reste ouverte. En effet, considérer uniquement l'ambiguïté à travers des ressources semble non satisfaisant (Hearst, 2006, 2009). L'utilisation de méthodes inductives reste dans les travaux récents un complément pour augmenter la couverture de Wikipédia. En effet, la question de l'évaluation de l'ambiguïté des requêtes reste entière. D'une part, l'accès aux réponses choisies par l'utilisateur à une requête donnée est difficile, ce sont des informations périssables et détenues par les moteurs de recherche. D'autre part, il n'y a que très peu de tests d'évaluation et il ne sont pas conçus pour des requêtes « réelles » (TREC 7). Cependant, le repérage et l'évaluation de l'ambiguïté ne peuvent être effectués sans savoir au préalable quelle forme prend le phénomène que l'on cherche à contrôler. Il est donc nécessaire de mieux comprendre comment se manifeste l'ambiguïté, en tentant de voir si les ambiguïtés dites non classiques sont bien présentes. Cette tâche implique de prendre en compte un certain nombre de paramètres variables et subjectifs comme les connaissances de l'utilisateur, son intention lorsqu'il effectue une requête. Il faut également considérer le cadre dans lequel il effectue cette recherche, le type de moteur et la base documentaire associée, un moteur généraliste ne semble pas générer le même type d'ambiguïté qu'un moteur dit vertical, dédié à la recherche d'information sur un domaine particulier comme la vidéo ou les publications scientifiques (Sanderson, 2008).

3 Faire émerger l'ambiguïté des requêtes automatiquement à partir de corpus

Le cadre applicatif de ce travail est un plateforme d'actualités vidéos développée par Orange Labs, 2424 actu.fr [1]. Ce site permet de consulter l'actualité française en temps réel, et propose différents types d'accès à l'information : par clustering, par thématiques et par barre de recherche traditionnelle. Le moteur de recherche à la disposition des utilisateurs ne propose pas de présentation de résultats par clusters ou par thématiques, mais par format des sources (vidéo, texte, audio).

La question de l'ambiguïté des requêtes dans un cadre applicatif est une question importante pour plusieurs secteurs de la recherche d'information comme la complexité des requêtes et la prédiction de la difficulté d'une requête (Clough *et al.*, 2009). L'ambiguïté est également un aspect important pour la personnalisation des moteurs de recherche vis à vis des utilisateurs. Notre contexte renforce cet intérêt, l'actualité est un usage quotidien pour beaucoup d'utilisateurs, mais elle a la particularité d'être évolutive. Nous connaissons également des contraintes propres au contexte applicatif [2] qui demandent d'aborder l'ambiguïté des requêtes sous un angle différent. Dans ce but, nous cherchons à identifier l'ambiguïté à partir d'une tâche de catégorisation.

1. http ://www.2424actu.fr

2. Le site est une plateforme qui reccueille temporairement les productions des partenaires, elle ne possède donc pas les contenus et ne peut les conserver pour un usage commercial.

3.1 L'ambiguïté examinée par une tâche de catégorisation

Le but de cette expérimentation n'est pas de reproduire un processus de recherche d'information, mais d'étudier la forme que peut prendre l'ambiguïté des requêtes produites dans un contexte applicatif. L'utilisation d'une ressource nous paraît peu appropriée pour cette tâche. En effet, comme nous l'avons vu dans (2.2), il existe des ambiguïtés que l'on peut qualifier de non-classiques et qui ne sont pas recensées dans une ressource existante. L'enjeu est donc de faire émerger à la fois l'ambiguïté « classique » telle que la polysémie ou l'homonymie, et les ambiguïtés « non classiques » et de pouvoir observer leurs manifestations.

Cette étude préliminaire utilise les moyens disponibles dans le cadre de cette application, une double catégorisation des documents « cibles » des requêtes. Notre but est de parvenir à transposer une catégorisation externe à une catégorisation interne. Pour cela, nous nous servons de documents qui ont la particularité d'être classifiés à deux niveaux : chaque document (ou news) appartient à un cluster et à une catégorie thématique. Les catégories thématiques sont héritées de l'AFP et elles sont au nombre de six : *économie* (questions économiques), *international* (actualités hors de France), *société*, *politique*, *cultures* (musique, sciences, art, people) et *sport*. Elles forment un étiquetage que nous allons utilisés. L'hypothèse est que les catégories, en nombre réduit et correspondant à un classement adapté pour les news, vont donner une approximation des domaines présents dans l'actualité et un premier point d'entrée sur le comportement des requêtes. Nous nous appuyons donc sur le corpus que nous avons constitué au fil des jours à partir des actualités quotidiennes du site (tableau 1). Le corpus de documents est la collection de documents sous forme textuelle disponible pour les utilisateurs du site d'actualité pour une période donnée en français. Les sources sont hétérogènes : audio retranscrit, dépêches AFP, articles de journaux, retranscription de journaux télévisés. Les documents proviennent des différents partenaires du site (AFP, Le Monde, Le Point, L'Express, France Télévision, Paris Match, etc.). La notion temporelle structure le corpus, il est partitionné en huit sous-corpus.

L'expérimentation consiste à catégoriser les requêtes pour pouvoir observer leur distribution sur plusieurs domaines. Les requêtes d'une période temporelle sont projetées sur la base textuelle correspondant à la même période temporelle, et si elles apparaissent dans un document, elles héritent de la catégorie thématique du document. Les catégories sont pondérées selon la fréquence d'apparition. Nous effectuons également un filtrage des catégories attribuées : une catégorie est considérée seulement si elle représente plus de 10% des textes où la requête apparaît. Le filtre permet de limiter l'apparition de catégories résiduelles, il a été choisi de façon arbitraire. Le corpus de requêtes utilisé contient les requêtes utilisateurs du site 2424actu des huit derniers mois (de Mai 2010 à Décembre 2010). Il totalise 487 231 requêtes non dédoublonnées (pour environ 30 700 requêtes uniques). Ce corpus de requêtes est découpé en partition temporelle, chaque partition est organisée en fonction de la fréquence des requêtes [3]. Nous utiliserons dans cette expérimentation les 49 requêtes les plus fréquentes de chaque sous-partition du corpus c'est-à-dire ayant une fréquence supérieure à 100 pour la période considérée, ce qui fait au total 391 requêtes. Préalablement, les requêtes contenant plusieurs mots et ne formant pas des termes sont volontairement exclues de la catégorisation comme par exemple « boue hongrie », alors qu'une requête comme « festival de cannes » va être catégorisée. Le parti pris est que ces requêtes contenant plusieurs termes sont moins affectées par l'ambiguïté (Sanderson, 2008).

Mai2010	Juin2010	Juillet2010	Août2010	Sept2010	Octobre2010	Nov2010	Déc2010
23 521	26 782	15 773	19 543	17 634	22 822	16 015	11 096

TABLE 1 – Corpus de documents (news 2424actu)

3.2 Analyse de l'ambiguïté des requêtes

Nous cherchons à savoir si la catégorisation thématique est un bon procédé pour explorer la manifestation de l'ambiguïté et donc si une pluri-catégorisation est synonyme d'ambiguïté. Pour étudier cette question, nous allons procéder tout d'abord à une analyse des requêtes qui n'ont été rattachées qu'à une seule catégorie ce qui va permettre de regarder si l'ambiguïté est présente malgré un rattachement unique. Puis dans un deuxième temps, l'analyse se focalisera sur les requêtes rattachées plusieurs catégories thématiques.

3. A noter, que les sous-partitions pour les mois de septembre et novembre ne sont pas complètes.

Nous avons effectué une évaluation sur le corpus de requêtes les plus fréquentes (mai à décembre), soit 391 requêtes. Ce corpus contient 70% de requêtes contenant une EN ou étant une EN. Parmi ces 391 requêtes, 35% ne sont pas catégorisées, leur recherche ne donnant pas de résultats dans la base textuelle. La répartition entre les requêtes mono-catégorielles et pluri-catégorielles est la suivante : 54% sont mono-catégorielles et 46% pluri-catégorielles (au moins deux catégories).

3.2.1 Les requêtes rattachées à une seule catégorie

La répartition des requêtes qui donnent lieu à un classement thématique unique varie selon les corpus-tests (environ 67% pour le sous-corpus de décembre contre 27% dans le sous-corpus de mai). Ces requêtes sont à 80% des entitées nommées, ce qui ne représente pas l'ensemble des requêtes contenant des ENs (40% environ sont rattachées à plusieurs catégories). Par exemple, « miss france » va être catégorisée exclusivement en *cultures* tout comme les requêtes « prince william » ou « audrey pulvar », « nicolas dupont-aignan » sera catégorisé en *économie*. Ces requêtes contiennent des noms propres complets ce qui aide à l'identification, mais il y a également des requêtes mono-mots qui n'ont qu'une seule catégorie comme « vogica » en *international*. Les requêtes mono-catégorielles ne contenant pas d'entitée nommée comme « neiges » en *société* ou « agriculture » en *économie* sont moins nombreuses (environ 20%). Nous observons également qu'une grande partie de ces classements uniques s'opère dans les catégories *international* (environ 40% des requêtes mono-catégorisées) et *cultures* (environ 20%). Ces catégories sont vastes et englobent énormément de sujets.

3.2.2 Les requêtes rattachées à plusieurs catégories

Nous avons effectué une analyse basée sur les 115 requêtes pluri-catégorisées, nous cherchons à faire émerger des phénomènes identifiables afin de construire une typologie des requêtes manifestant de l'ambiguïté. Les requêtes renvoyant à plusieurs catégories thématiques sont majoritairement des entitées nommées comme pour les requêtes mono-catégorielles, mais on trouve également 40% de requêtes ne contenant pas d'EN.

Parmi les requêtes pluri-catégorielles, il existe des requêtes porteuses d'ambiguïté de type classique (polysémie et homonymie). Elles renvoient à plusieurs catégories qui présentent des pondérations similaires comme par exemple « royal » (tableau 2). C'est une requête ambigüe et la catégorisation fait ressortir deux aspects intéressants, d'une part une catégorisation en *politique* et d'autre part une catégorisation en *sport*. La catégorisation en *politique* de « royal » renvoie à « Ségolène Royal », contrairement à la catégorisation en *sport* qui renvoie l'adjectif « royal » présent dans un certain nombre de stades sud-africains utilisés lors de la Coupe du monde de football, par exemple « stade Royal Bafokeng de Rustenburg ». L'ambiguïté lexicale dans nos requêtes est essentiellement de l'homonymie. En effet, le seul exemple de polysémie trouvé dans notre corpus est la requête « tabac » (tableau 3). Elle peut désigner plusieurs choses (le lieu où l'on vend du tabac, le tabac comme plante, ou comme cigarettes) et c'est aussi un nom que l'on retrouve fréquemment dans l'expression « passer à tabac ». Le contexte d'actualité contient plusieurs sens de « tabac » : lieu de vente, tabac comme cause de maladie, cigarettes, tabac comme plante cultivée ou encore passage à tabac. La polysémie est extrêment rare dans notre corpus.

Thématiques	int	pol	spr
royal	27	**63**	**40**

TABLE 2 – La requête « royal » catégorisée en juin

Thématiques	int	eco	soc
tabac	11	**27**	**25**

TABLE 3 – La requête « tabac » au mois de novembre

Nous observons un autre type de requêtes pluri-catégorielles avec des fréquences d'apparition supérieures aux requêtes précédentes (requêtes homonymiques), comme par exemple « société » (tableau 4), « obama » ou « sarkozy ». Ces requêtes sont difficiles à interpréter même à l'aide du contexte des documents cibles parce qu'elles balaient un champ très large. Ainsi on voit par exemple qu'une requête comme « sarkozy » (tableau 5), va ouvrir vers un

grand nombre de catégories thématiques. Cette requête est ambigüe car elle a besoin d'être spécifiée, d'être complétée. L'observation ne permet pas d'identifier un lien interprétable en termes d'ambiguïtés entre la requête et une catégorie. Nous les rapprochons du type de requêtes dites « larges » (type B) décrites par (Song *et al.*, 2009), ce sont des requêtes génériques.

Thématiques	eco	int	pol	soc
société	**114**	**64**	**45**	**84**

TABLE 4 – La requête "société" (mai)

Thématiques	int	pol	soc
sarkozy	196	**544**	182

TABLE 5 – La requête « sarkozy » (juin)

Enfin, nous distinguons un troisième type de requêtes pluri-catégorielles, relativement proches des requêtes « génériques » du point de vue des catégorisations. Elles sont cependant différentes car elles sont porteuses d'ambiguïté référentielle. Par ambiguïté référentielle, nous désignons une requête qui n'a pas de référent fixe et qui potentiellement peut désigner deux ou plusieurs référents. Par exemple, « intempéries » paraît peu ambigu dans un contexte d'actualités, pourtant la catégorisation fait ressortir deux catégories (*international* et *société*). De nombreuses requêtes comme « otages » ou « ministre » (tableau 6) sont ambigües d'un point de vue référentiel, « ministre » désigne une fonction mais aussi un grand nombre de personnes occupant cette fonction. Ces requêtes sont fortement dépendantes du contexte général et par conséquent sujettes à des variations de sens et de référents, dans des périodes temporelles plus ou moins longues.

Thématiques	eco	int	pol	soc
ministre	**311**	**474**	**571**	**415**

TABLE 6 – La requête « ministre » en novembre

3.2.3 Analyse de la dimension diachronique de l'ambiguïté des requêtes

L'analyse de requêtes a fait apparaître un facteur de variation : la dimension diachronique. En effet, certaines requêtes changent de catégorisation selon les périodes temporelles comme « éruption » ou « cannabis ». Le cas de la requête « éruption » est intéressant car la requête désigne un même phénomène, mais à chaque fois un volcan différent, ce qui va donner la catégorie *société* au mois de décembre pour l'éruption du Piton de la Fournaise, l'éruption du mois de novembre du volcan Merapi en Indonésie sera classée comme *international,* alors que l'éruption islandaise en mai est catégorisée en *économie*. La mono-catégorisation variable s'applique également à certaines entités nommées comme la requête « delarue » consécutivement catégorisée en *cultures* puis en *société,* ce qui permet de réperer un changement, et potentiellement une ambiguïté si deux catégories co-occurrent dans la même période temporelle. On observe que sur les 50 requêtes les plus fréquentes chaque mois, seules quelques unes persistent au cours des huit mois : « afghanistan », « haïti », « israël » ou encore « pakistan ». Elles ont pour caractéristique principale d'être pour la plupart des entitées nommées (environ 70% des requêtes « durables », ce qui est comparable à la proportion dans l'ensemble du corpus). Ces requêtes désignent toutes des pays et s'avèrent être pluri-catégorisées. Nous allons analyser plus en détail deux requêtes à l'actualité très riche « pakistan » et « haïti », pour illustrer la manifestation de l'ambiguïté dans le temps.

La requête « pakistan » (tableau 7) a une catégorisation très variable. La catégorie *international* est majoritaire sur l'ensemble du corpus. Mais nous observons plusieurs variations au niveau des catégories majoritaires selon le sous-corpus considéré. La première variation en juin (catégorisation en *politique*), correspond dans notre corpus à l'apparition d'une affaire impliquant le Pakistan en tant qu'Etat : la vente de sous-marins au Pakistan par la France aurait servi de moyens de financements à un ancien premier ministre. La catégorie *international*, correspond à un emploi du Pakistan comme lieu, servant à localiser principalement des attentats et des opérations de la CIA contre les talibans. La catégorisation ne permet malheureusement pas d'identifier strictement les deux emplois de ce mot. La deuxième variation en novembre fait ressortir une catégorisation en *société*, celle-ci renvoyant à un

emploi de Pakistan comme lieu, en l'occurrence lieu qui produit des filières de combattants pour Al-Quaïda et des « potentiels auteurs d'attentats ». On identifie donc plusieurs significations pour la requête « Pakistan », à la fois lieu géographique, état-nation et lieu de formation de combattants. On observe que la requête « haïti » (tableau 8) a le même type de comportement. La pluri-catégorisation montre deux emplois possibles du mot « haïti » : comme référence au tremblement de terre (catégorisée en *cultures* et *international*) ou comme lieu. A noter cependant, la requête au mois d'octobre qui est mono-catégorielle (*international*). Il s'avère que cela est du à une focalisation sur un évènement unique en effet pendant cette période, Haïti a été touché par une épidémie de choléra. La requête a pris une signification différente. Ces deux exemples manifestent deux types particuliers d'ambiguïté, décrit par (Lecolle, 2007) sous le terme de « polysignifiance ». La polysignifiance désigne le fait qu'un nom de lieu habité présente des valeurs sémantico-référentielles différentes, renvoyant à la fois au lieu, mais aussi aux habitants et à l'institution qui le gouverne. Cette propriété des noms de lieux leur permet de pouvoir porter différentes significations, comme par exemple « Outreau » qui a pris la valeur d'erreur judiciaire en supplément de sa valeur locative, ou « Tchernobyl » qui désigne désormais une catastrophe nucléaire. Cette maléabilité du nom de lieu décrite par (Lecolle, 2007) ouvre une gamme large de possibilités, mais aussi de problèmes évidents si les différentes valeurs ne peuvent être discriminées et qu'elles apparaissent dans des contextes identiques. La polysignifiance semble difficile à appréhender par le biais de ressources lexicographiques ou de bases de connaissances, les différentes valeurs ne sont pas prises en compte, seule la fonction de localisation est retenue dans le cas des noms de lieu. La catégorisation thématique nous permet d'observer une partie de ce phénomène, en mettant en évidence les différentes significations de requêtes comme « haïti » ou « pakistan ».

Thématiques	clt	eco	int	pol	soc
Mai		9	**22**	14	14
Juin			8	**24**	
Juillet			**67**		
Aout			**474**		
Sept			**82**		
Oct			**91**	10	
Nov			31		**68**
Dec			**47**		

TABLE 7 – Catégorisation de la requête « pakistan »

Thématiques	clt	eco	int	soc	spr
Mai	**1**		**1**	**1**	
Juin		1	**3**	2	
Juillet	**17**				5
Aout			**50**		
Sept			**9**	1	
Oct			**102**		
Nov			**146**		
Dec			**55**	34	

TABLE 8 – Catégorisation de la requête « haïti »

3.3 Vers une typologie de l'ambiguïté des requêtes

L'ensemble de ces descriptions et analyses nous conduisent à proposer une synthèse sous forme de typologie, afin de rassembler les différentes formes que peut prendre l'ambiguïté dans notre contexte applicatif. Nous distinguons au final deux types de requêtes « ambigues » à la suite de notre analyse : les requêtes qui manifestent de l'ambiguïté « classique » et celles qui manifestent de l'ambiguïté « non classique » (tableau 9). Les requêtes du premier type contiennent de l'homonymie. La polysémie est quasi-absente de notre corpus de requêtes. Les requêtes du deuxième type contiennent trois types de manifestations de l'ambiguïté : les requêtes « polysignifiantes », les requêtes « pluri-référentielles » et les requêtes « génériques ». Nous avons en effet considéré les requêtes « polysignifiantes » comme étant des requêtes ambigües car la propriété de polysignifiance vaut potentiellement pour les noms de

lieux habités dans un contexte d'actualité. Les noms de lieux sont très présents dans notre corpus de requêtes en particulier parmi les plus fréquentes (18,6% des requêtes de notre corpus test contiennent un nom de lieu). L'ambiguïté référentielle est avant tout observée sur des noms et non pas des entités nommées. Les utilisateurs en formulant ce type de requêtes semblent faire confiance au contexte immédiat (par exemple « inondations »), mais lorsque l'actualité comporte plusieurs évènements auxquels peut référer cette requête, l'ambiguïté naît. Les requêtes que nous avons qualifiées de « générique » réfèrent à un objet, un évènement ou un domaine vaste et riche, contenant différentes « facettes »(Hearst, 2006). Ainsi, la requête « roman polanski » illustre ce cas de requêtes manifestant plusieurs facettes lors de la catégorisation, celle-ci fait apparaître deux catégories *cultures* et *international*, mettant en avant ses différents rôles dans l'actualité. Ces observations questionnent les conceptions qui considèrent le nom propre en tant que « désignateur rigide » (Kripke, 1980), parce qu'il désigne le même objet dans tous les mondes où cet objet est présent, étant alors univoque. Or, les ressources crées pour contenir diverses informations sur les entités nommées sont construites sur ce modèle, les variations contextuelles ne sont pas prises en compte. C'est pourquoi nous allons à présent regarder si ces formes d'ambiguïté sont présentes dans une ressource comme Wikipédia, l'encyclopédie en ligne, utilisée pour construire de nombreuses bases de connaissances comme DBPédia [4].

Requêtes avec ambiguïté « classique »	Requêtes avec ambiguïté « non classique »		
Homonymie	Polysignifiance	Ambiguïté référentielle	Généricité
voile, younes, corée	*haïti, pakistan, irak*	*éruption, otages,*	*sport, sarkozy*
royal, tabac	*afghanistan, xynthia*	*ministres, inondations*	*obama, gouvernement*

TABLE 9 – Typologie de l'ambiguïté des requêtes 2424 actu

3.4 Comparaison avec Wikipédia

Nous avons mis au jour de formes d'ambiguïté qui ne sont pas décrites dans les dictionnaires « traditionnels » comme la polysignifiance. Nous proposons alors de comparer notre classification à une ressource, afin de montrer la différence entre une procédure exploratoire comme la nôtre et une procédure de comparaison, entre une ressource et des requêtes. Nous avons comparé manuellement une partie de notre corpus (98 requêtes) avec l'encyclopédie en ligne Wikipédia. Deux aspects sont examinés :

– est-ce que la requête considérée est présente dans Wikipédia ?
– si la requête est une entrée de page d'encyclopédie, est-ce qu'elle renvoie vers une page d'homonymie ou de désambiguisation ?

Une page d'homonymie ou de désambiguïsation dans Wikipédia est simplement une page qui répertorie les différents sujets et articles partageant un même nom. Par exemple « éruption » renvoie vers une page qui recense un certain nombre d'« éruption » :

– *une **éruption** volcanique en géologie ;*
– *une **éruption** cutanée ou rash en médecine*
– *une **éruption** solaire, un phénomène très énergétique se produisant à la surface du Soleil.*
– *un morceau de guitare électrique du groupe américain Van Halen, **Eruption**.*
– *un groupe disco **Eruption***

Nous avons comparé l'annotation à partir de Wikipédia à la classification thématique, pour mesurer l'éventuel décalage. L'annotation a été effectuée sur les 97 requêtes du corpus test et seulement 61 requêtes sont annotées et comparées. La comparaison porte donc sur un nombre réduit de requêtes (tableau 10). 23 requêtes sont considérées comme non ambigües selon Wikipédia et mono-catégorielles, mais 7 requêtes sont mono-catégorielles et considérées comme ambigües par l'encyclopédie. En effet, Wikipédia recense parfois plus de sens qu'il n'existe dans notre contexte spécialisé, par exemple, Johnny Hallyday peut aussi être un cascadeur selon l'encyclopédie, l'actualité ne connaît que le chanteur. La situation s'inverse lorsqu'on considère l'accord entre la catégorisation et Wikipédia pour la pluri-catégorisation comme le montre le tableau (10). Un certain nombre de requêtes est identifié comme n'étant pas ambigües par Wikipédia, et pluri-catégorisées dans notre contexte (19 requêtes). Ces requêtes n'ont pas de pages de désambiguïsation. Parmi ces requêtes on retrouve « Haïti », « Afghanistan », « facebook », « gouvernement », « Carla Bruni », etc. Le problème étant que Wikipédia ne recense pas des ambiguïtés aussi fines de ce type, il se limite aux ambiguïtés de type homonymie. Des aspects comme la polysignifiance

4. http ://dbpedia.org/About

ou des variations propres au contexte ne figureront pas dans l'encyclopédie aussi riche soit-elle. La question de l'inadéquation d'une ressource pour capter des phénomènes qui créent de l'ambiguïté mais qui ne relèvent pas de la polysémie ou de l'homonymie se pose.

Comparaison	Mono-catégories	Pluri-catégories	**Total**
Un seul sens dans Wikipédia	23	12	**35**
Désambiguïsation dans Wikipédia	7	19	**26**

TABLE 10 – Comparaison de la catégorisation avec Wikipédia

4 Conclusion et Perspectives

L'expérimentation montre que l'ambiguïté peut se manifester à différents niveaux et sous différentes formes. La catégorisation thématique des requêtes a été un premier pas pour observer la diversité de l'ambiguïté et la complexité d'un contexte applicatif comme le nôtre. L'actualité est un domaine riche et continuellement en mouvement, c'est également un enjeu important d'améliorer l'accès au contenu des news. La constitution des corpus a été un challenge, il a fallu réunir à la fois les requêtes utilisateurs et les corpus de news où les utilisateurs avaient effectué leurs recherches. Bien qu'il ne soit pas possible d'assimiler la pluri-catégorisation à de l'ambiguïté de manière équivalente, la catégorisation nous a permis de mettre en évidence différentes manifestations de l'ambiguïté dans nos requêtes, à différents niveaux (durables ou variables). Les requêtes ambigües ne relevant pas de la polysémie ou de l'homonymie s'avèrent très mouvantes thématiquement, faisant apparaître une granularité plus fine, illustrant des facettes ou des évènements. Par ailleurs, la typologie proposée a fait apparaître certaines phénomènes qui ne peuvent être repérés par des ressources classiques. Les dictionnaires ne recensent pas les variations de points de vue et de signification ponctuelles. On voit donc l'intérêt d'utiliser des méthodes de type catégorisation ou clustering dans le repérage de l'ambiguïté. Cela pose la question dans un deuxième temps de l'intérêt de l'utilisateur, que va-t-il gagner à percevoir ces différents facettes ? Du point de vue de la méthode utilisée dans cette expérimentation, il apparaît qu'il serait intéressant de la compléter et de la comparer à d'autres types de méthodes automatiques de classification. La catégorisation par thématique a été utilisée parce que celle-ci offrait une lisibilité des résultats obtenus, c'est un outil grossier mais éclairant. Mais l'ambiguïté référentielle ou l'ambiguïté causée par trop de généralité, gagnerait à être observée à travers une représentation de l'information différente. La double catégorisation de la base textuelle, catégorisation thématique et clustering, nous permet de débuter l'analyse de requêtes classifiées, et de mettre en relation une requête avec un ou plusieurs clusters de documents dégageant ainsi des groupements pertinents.

Dans la perspective d'un repérage de l'ambiguïté des requêtes, nous pensons que la combinaison des moyens de classification avec d'autres indices pourrait être fructueuse, comme par exemple la reformulation de requêtes. Mais c'est une mesure à ne pas dissocier du comportement utilisateur, le cadre applicatif joue sur la reformulation ainsi que le mode de consommation (dans le cas de l'actu un « mur » interactif permet de consommer de l'actualité sans faire de requêtes). Cette expérimentation mérite donc d'être complétée. Elle constitue un préalable à un potentiel repérage de l'ambiguïté et à la mise en place d'aide pour l'utilisateur. Le repérage est un outil pour avancer dans la connaissance de la complexité des requêtes, ou pour assurer une meilleure expansion de requête. Le traitement de l'ambiguïté en elle-même peut passer par la proposition d'aide à l'utilisateur (Hearst, 2009). Notre contexte nous permet de développer une réponse de ce type à la manifestation de l'ambiguïté. La présentation des résultats peut gagner à l'injection de techniques de clustering et de classification, mais elle doit être maîtrisée et faire sens pour l'utilisateur. De ce fait, la mise en place de tests utilisateurs est indispensable.

Références

AARTS B. & MACMAHON A. (2006). *The Handbook of English Linguistics*. Oxford : Blackwell.

ARTILES J., BORTHWICK A., GONZALO J., SEKINE S. & AMIGÓ E. (2010). Weps-3 evaluation campaign : Overview of the web people search clustering and attribute extraction tasks. In *CLEF (Notebook Papers/LABs/Workshops)*.

ARTILES J., GONZALO J. & SEKINE S. (2007). The semeval-2007 weps evaluation : Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (Semeval-2007)*.

BERNARDINI A., CARPINETO C. & D'AMICO M. (2009). Full-subtopic retrieval with keyphrase-based search results clustering. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, p. 206–213, Washington, DC, USA : IEEE Computer Society.

CLOUGH P., SANDERSON M., ABOUAMMOH M., NAVARRO S. & PARAMITA M. L. (2009). Multiple approaches to analysing query diversity. In *SIGIR*, p. 734–735.

DARWISH K. & OARD D. W. (2003). Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, p. 338–344, New York, NY, USA : ACM.

ERHMANN M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation.* PhD thesis, Paris VII.

GAN Q., ATTENBERG J., MARKOWETZ A. & SUEL T. (2008). Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, p. 49–56, New York, NY, USA : ACM.

HEARST M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM*, **49**, 59–61.

HEARST M. A. (2009). *Search User Interfaces*. Cambridge University Press.

KRIPKE S. A. (1980). *Naming and Necessity*. Harvard University Press.

KROVETZ R. & CROFT W. B. (1992). Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, **10**, 115–141.

LECOLLE M. (2007). Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas Outreau. *Corpus*, (6), 101–125.

RAHURKAR M. A., ROTH D. & HUANG T. S. (2008). Which "Apple" are you talking about ? In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, p. 1197–1198, New York, NY, USA : ACM.

SANDERSON M. (2000). Retrieving with good sense. *Information Retrieval*, **2**(1), 45–65.

SANDERSON M. (2008). Ambiguous queries : test collections need more sense. In *SIGIR*, p. 499–506.

SANTAMARÍA C., GONZALO J. & ARTILES J. (2010). Wikipedia as sense inventory to improve diversity in web search results. In *ACL*, p. 1357–1366.

SONG R., LUO Z., NIE J.-Y., YU Y. & HON H.-W. (2009). Identification of ambiguous queries in web search. *Information Processing and Management*, **45**(2), 216–229.

SPINK A., JANSEN B. J. & PEDERSEN J. (2004). Searching for people on web search engine. *Journal of Documentation*, **60**(3), 266–278.

STOKOE C., OAKES M. P. & TAIT J. (2003). Word sense disambiguation in information retrieval revisited. In *SIGIR*, p. 159–166.

VOORHEES E. M. (1993). Using wordnet to disambiguate word senses for text retrieval. In *SIGIR*, p. 171–180.

WEISS S. F. (1973). Learning to disambiguate. *Information Storage and Retrieval*, **9**(1), 33–41.

ZHAI C. X., COHEN W. W. & LAFFERTY J. (2003). Beyond independent relevance : methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, p. 10–17.