INSTITUT NATIONAL DES LANGUES ET CIVILISATIONS ORIENTALES

DEPARTEMENT TEXTES, INFORMATIQUE, MULTILINGUISME

CREATION SEMI-AUTOMATIQUE D'UN THESAURUS DU DOMAINE BANCAIRE ET APPLICATION A LA FOUILLE D'OPINION

MASTER TRAITEMENT AUTOMATIQUE DES LANGUES

PARCOURS INGENIERIE MULTILINGUE

PAR

VIRGINIE POADEY

DIRECTEUR DE MEMOIRE : CYRIL GROUIN

ENCADRANT EXPERT SYSTEM : SONIA COLLADA

18 NOVEMBRE 2016

REMERCIEMENTS

Je tiens à remercier en premier lieu Cyril Grouin, pour tous ses précieux conseils, sa disponibilité et surtout son soutien jusqu'à la fin de la rédaction de ce mémoire.

Je remercie également toute l'équipe d'Expert System et plus particulièrement ma tutrice de stage Sonia Collada pour m'avoir suivi pendant ces 6 mois et avoir partagé ses connaissances avec moi.

Merci à tous les professeurs et camarades du Master TAL de l'INaLCO pour ces deux années incroyables où j'ai énormément appris.

Et enfin, merci à ma famille d'avoir toujours cru en moi en toutes circonstances.

RESUME

La fouille d'opinion devient une approche de plus en plus intéressante pour les entreprises qui souhaitent évaluer la qualité de leurs produits auprès de leurs clients. Notre travail consiste à évaluer des relations, c'est-à-dire des opinions émises précisément sur des objets appelés cibles. Nous créons un thésaurus sur le domaine bancaire afin d'avoir des cibles prédéfinies. Nous nous basons sur un corpus composé d'avis clients récupérés sur internet. Nous abordons deux approches, une à base d'un module d'extraction déjà existant et une autre à base de règles que nous allons développer nous-mêmes. Nous ne pouvons présenter de résultats pour l'approche à base de règles pour cause d'incompatibilité des sorties des différents outils. Nous obtenons un rappel plus faible par rapport à l'état de l'art sur une problématique similaire avec l'approche utilisant le thésaurus. Mais nos résultats globaux sont supérieurs, avec une F-mesure de 70,1%.

Mots-clés : opinion mining, fouille d'opinion, thésaurus, banque, polarité, relations.

TABLES DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

Ce mémoire a été rédigé dans le cadre du stage de fin d'études effectué à Expert System France[1] situé à Paris. Expert System est une entreprise éditeur de logiciels, fondée en 1989 et basée à Modène en Italie. Elle développe et commercialise des logiciels d'analyse sémantique pour la gestion de contenus structurés et non structurés. En 2015, l'entreprise fusionne avec l'entreprise française Temis, éditeur de solutions en fouille de textes présent également en Allemagne, au Royaume-Uni, aux Etats-Unis et au Canada. Ainsi, cette fusion permet à Expert System de s'étendre sur le marché des technologies sémantiques au niveau mondial.

Nous sommes dans une ère où échanger des informations via internet est incontournable. C'est une mine d'or chargée de savoir que les entreprises veulent de plus en plus exploiter. Elles sont soucieuses de ce qui s'y trouvent car elles leur permettent un accès à une grande quantité de données rapidement. Malheureusement, ces données ne sont pas structurées et la plupart du temps, difficilement exploitables telles quelles. La fouille de texte et la fouille d'opinion sont des moyens qu'Expert System appliquent pour rendre ces données utilisables.

Il existe plusieurs manières de hiérarchiser des données, comme à travers un thésaurus, une ontologie ou encore une taxonomie. Ce sont trois concepts distincts, même s'ils sont proches et souvent confondus. Chacun recouvre un domaine de la connaissance, quel qu'il soit, mais une méthode est choisie plutôt qu'une autre en fonction des objectifs à atteindre.

Le dictionnaire Le Trésor de la Langue Française donne la définition suivante d'un thésaurus : « *Langage documentaire fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels* ».

[Dégez & Ménillet, 2001] considèrent qu'un thésaurus est une « *Liste organisée de termes contrôlées et normalisés (descripteurs et non-descripteurs) servant à l'indexation des documents et des questions dans un système documentaire* ».

Cela permet donc de structurer un domaine à travers des descripteurs (ou concepts) qui le définissent. Nous pouvons par exemple avoir un thésaurus sur le domaine médical, littéraire ou bancaire.

Contrairement au dictionnaire avec lequel il est souvent mis en relation, un thésaurus ne contient pas nécessairement les définitions des concepts qui le constituent. La richesse d'un thésaurus vient des relations sémantiques existantes entre ces concepts [Moureau, 1973 ; Roberfroid et Dubois, 2012]. [Moureau, 1973] ajoute par ailleurs une dimension multilingue.

Les relations hiérarchiques entre les concepts se font de par la structure arborescente du thésaurus. Les descripteurs (e.g., « invertébré ») contiennent eux-mêmes des descripteurs spécifiques (ou concepts fils, e.g., « mollusque ») qui vont permettre d'englober cette partie du domaine. Les relations d'équivalence vont avoir un rapport avec la synonymie entre les concepts (e.g. figure 1). Et enfin, nous avons les relations d'association de concepts qui

[1] http://www.expertsystem.com/fr/

permettent d'avoir deux termes ensemble même si a priori, les champs sémantiques diffèrent. Il est aussi possible d'utiliser les relations d'association pour exclure les cas que nous ne voulons pas extraire.

Pour expliciter ces propos, dans un thésaurus (simplifié) sur les animaux, nous aurions :

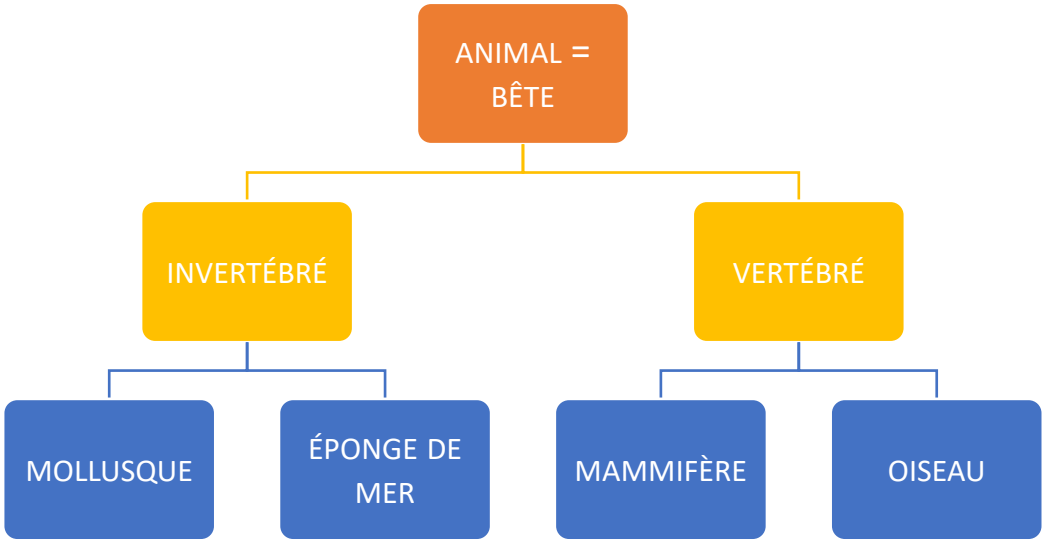

Figure 1 : Schématisation simplifiée d'un thésaurus[2] sur le domaine animalier

Une ontologie est organisée sous forme de classes. Ces classes communiquent entre elles via des relations définies de manière à ce qu'une machine puisse interpréter le lien entre deux instances d'une classe. Friend of a friend[3] (FOAF) est une ontologie qui permet de créer des relations entre les personnes. FOAF est un vocabulaire utilisant la syntaxe RDF/XML du Web sémantique et dans sa version actuelle, il contient 13 classes et 62 propriétés.

[2]Site Motbis : http://www.cndp.fr/thesaurus-motbis/site/mtesp.php? mt=1510&t=frCNDP00880&tlp=2751

[3] http://www.foaf-project.org/

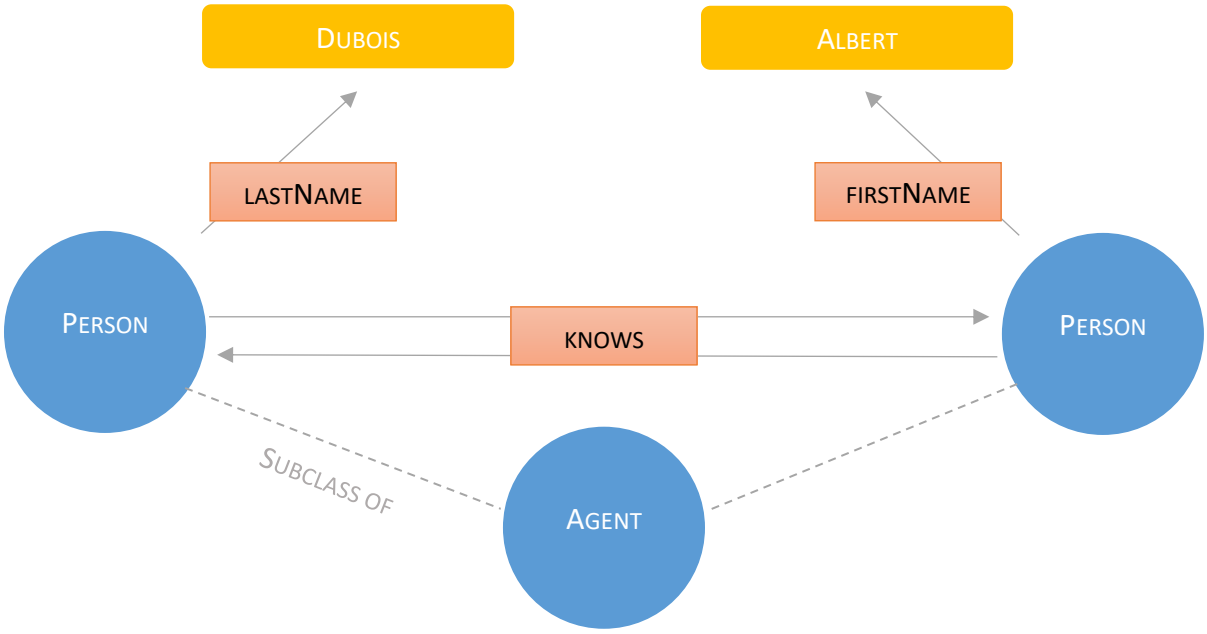

Figure 2 : Représentation schématique du fonctionnement de FOAF

Une ontologie se compose de classes (e.g. « Agent », « Person ») qui contiennent des sous-classes (e.g. « Person »). Les relations entre ces classes sont appelées propriétés (e.g. « knows », « lastName », « firstName »). Ce système permet de s'éloigner du langage naturel en essayant de créer un langage compréhensible pour la machine (pour faire les relations entre les éléments). Toutefois, des noms de classes et de sous-classes explicites sont pratiques pour un (utilisateur) humain.

Le terme « taxonomie » provient du domaine des sciences, qui classifie les organismes vivants en partant du plus générique (qui regroupe tous les éléments qui ont des caractères communs) pour aller au plus spécifique. En linguistique, une taxonomie va plutôt permettre de classifier des contenus via une hiérarchie de catégories appelés groupes. Par exemple, la grande encyclopédie en ligne Wikipédia utilise un système de taxonomie pour regrouper les différentes catégories de ses articles. Chaque article appartient à une catégorie, même ceux qui n'en ont pas puisqu'ils vont dans « non-catégorisé » en attendant d'être affectés à une catégorie.

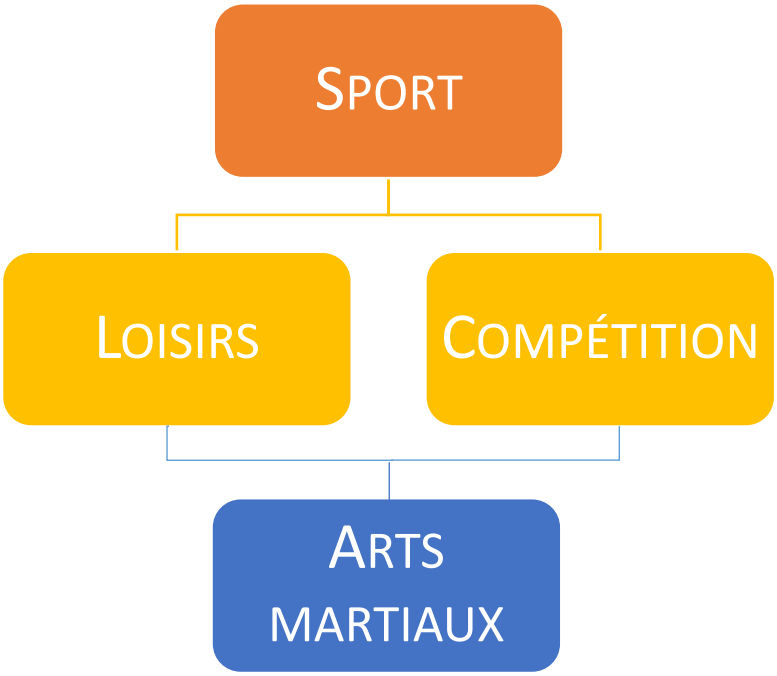

Figure 3 : Représentation schématique d'une taxonomie

Dans une taxonomie, il est possible d'avoir un groupe n'appartenant pas seulement à un groupe mais à plusieurs (e.g. « Arts martiaux ») contrairement au thésaurus qui va se présenter sous forme d'arbre. Un terme peut être présent dans plusieurs branches de cet arbre mais le but étant l'extraction dans un corpus, avoir deux fois la même entrée revient à avoir deux fois la même extraction mais dans deux concepts différents.

Ainsi, une taxonomie va permettre une hiérarchisation des informations sans pour autant utiliser une manière formelle de les décrire (bien que cela soit possible). Un thésaurus va plutôt utiliser les concepts en terme de mots-clefs afin d'effectuer une extraction dans des documents. Une ontologie permet de classer des éléments et établir des relations sémantiques entre ces éléments.

La méthode que nous utiliserons est donc un thésaurus qui, grâce aux concepts, permet d'indexer les documents d'un corpus ou de faire des recherches grâce aux mots clefs contenus dans celui-ci.

Nous avons souhaité travailler sur deux aspects : la mise en place d'un thésaurus du domaine bancaire et la fouille d'opinion pour des entités identifiées avec ce thésaurus.

Tout au long de ce mémoire, nous chercherons à savoir si, en nous focalisant sur les relations qu'impliquent une opinion à sa cible et à travers deux approches différentes, nous obtenons des résultats satisfaisants pour avoir une portée applicative. La première approche utilise un module de fouille d'opinions préexistant. La seconde est une méthode à base de règles que nous développons nous-mêmes. Nous étudierons le domaine de la banque-assurance en exploitant un corpus de verbatim [4] récoltés sur internet. Ces verbatim

[4] Verbatim = Verbum + atim : « mot à mot » ou « mot pour mot ». Transcription fidèle d'une déclaration orale.

contiennent des avis d'internautes sur différentes compagnies. Elles nous permettront dans un premier temps de créer un thésaurus du domaine puis ensuite d'effectuer de la fouille d'opinions. Le thésaurus développé regroupera les concepts existants dans la banque-assurance. La fouille d'opinions permettra, d'un point de vue applicatif, de rassembler des informations précieuses pour ces banques qui souhaitent connaître, entre autres, leur qualité de service auprès de leurs clients.

Nous commencerons d'abord par établir l'état de l'art, puis nous expliquerons comment nous avons constitué le corpus de verbatim et le thésaurus. Nous verrons les différentes méthodes appliquées pour parvenir à extraire les données souhaitées. Nous analyserons ensuite les résultats obtenus et discuterons des améliorations qu'il sera possible d'apporter à ce projet.

1. ETAT DE L'ART

L'expansion rapide d'internet et des nouvelles technologies qui lui sont associées ont permis un essor certain de nombreuses entreprises. Les réseaux sociaux et autres sites internet sont un moyen efficace de partager son avis sur différents types de produits. Les entreprises qui commercialisent ces produits ont alors un accès direct sur les pensées des consommateurs. Plus les clients rapporteront le même genre de critiques (positives ou négatives), plus l'entreprise saura quelles sont les forces et faiblesses de ses produits. Il existe désormais des sites qui regroupent ce type d'informations. La fouille d'opinions ou analyse de sentiments est ce qui permet d'extraire ces avis et lui attribuer une polarité (qui définit si ce qui est extrait est positif, négatif, neutre voire ambigu).

1.1. OPINION – DEFINITION

[Liu, 2010] différencie une opinion (subjective), d'un fait (objectif). Une opinion se présente sous la forme d'un quintuplet, c'est-à-dire qu'il regroupe 5 éléments dépendants les uns des autres : une ***personne*** qui émet une ***opinion*** sur les ***caractéristiques*** d'un ***objet*** à un ***instant*** donné.

L'émetteur s'exprime sur un objet, appelé cible, et cela crée cet ensemble qu'est l'opinion. Une opinion ne peut pas être réduite à un juste un mot. Il y a toujours un contexte donné.

Entre opinions génériques ou spécifiques, la différence se situe sur l'ensemble d'un objet lui-même (générique) ou d'une de ses caractéristiques (spécifique) [Liu, 2010].

« J'aime cet appareil photo » ou « J'aime l'écran tactile de cet appareil photo ».

Plus précisément, les caractéristiques d'un objet peuvent être référées implicitement ou explicitement [Liu & al., 2005].

Par exemple, les critiques d'un appareil photo donneraient « La résolution est parfaite. » ou bien juste « Trop cher. »

Dans le premier cas, « parfaite » est en relation avec « résolution » (caractéristique explicite). En revanche « cher » n'a pas de cible désignée dans le texte mais le prix est indiqué implicitement (caractéristique implicite).

Les caractéristiques d'un objet sont spécifiques à celui-ci. Mais si celui-ci est considéré plus largement, il fait partie d'un domaine dans lequel tous les autres objets qui lui ressemblent peuvent se rassembler (par exemple, l'informatique, la téléphonie, le cinéma, etc).

Les opinions sont également une affaire de contexte [Marchand, 2013]. En effet, certains mots ou certaines expressions ne vont pas garder le même sens selon le sujet abordé. En reprenant la phrase exemple de [Pang & Lee, 2008], « Go read the book » ne va pas s'interpréter de la même manière si le sujet est la littérature ou le cinéma. D'un côté, il y a l'aspect positif du livre où celui-ci était captivant et les autres sont invités à le lire également. De l'autre se trouve une adaptation cinématographique a priori ratée et qui n'est pas à la

hauteur du l'œuvre originale. C'est exactement la même phrase des deux côtés mais le domaine d'utilisation de celle-ci fait que l'opinion exprimée va être différente.

1.2. PRECEDENTS TRAVAUX

De précédentes recherches prennent en compte le contexte en l'associant directement à la fouille d'opinion. [Ruiz-Martínez & al., 2012] ont utilisé la fouille d'opinions dans le contexte journalistique de la finance. Leur module de fouille d'opinions (développé avec le logiciel GATE) a été combiné à une ontologie préexistante. Leur corpus de travail a été créé avec un fil RSS de news du monde de la finance. Des prétraitements sont effectués en amont avec un module d'annotation sémantique qui repère ensuite les entités se trouvant dans l'ontologie. Puis le module de fouille d'opinions analyse le contenu annoté et attribue une polarité positive ou négative au document. Cela crée deux lots de documents, un regroupant les documents classés positifs et un regroupant les négatifs. Les documents de chaque lot sont classés selon le degré de positivité ou négativité qui leur est attribué.

[Chaturvedi & Chopra, 2014] ont également choisi une ontologie sur les services bancaires dans le cadre de leurs recherches. Leur mode d'extraction est similaire à la précédente méthode, les termes bancaires sont extraits d'abord puis vont chercher les opinions exprimées sur ces termes. Les polarités sont extraites en utilisant SentiWordNet3[5]. Les résultats obtenus sont 58% de précision, 78% de rappel et une F-mesure de 67%. Nous prévoyons de nous inspirer de l'ensemble de ces travaux pour notre propre recherche et utiliserons ces derniers résultats comme base de référence pour notre travail.

Dans un premier temps, nous utiliserons un module d'extraction de fouille d'opinion préexistant que nous associerons à un thésaurus créé de manière semi-automatique. Ce module utilise la théorie Appraisal, comme celle proposée par [Gardin, 2009], mais lui ne se focalisait que sur des groupes adjectivaux. Dans un second temps, nous tenterons de recréer un module d'extraction mais en utilisant cette fois-ci un système à base de règles.

La première approche permet d'évaluer la couverture du module existant sur un nouveau corpus. La seconde permet de voir s'il est possible de s'approcher des résultats du module existant dans un temps imparti réduit. Ainsi, cette approche serait substituable au module.

Nous construirons un corpus composé de verbatim d'internautes afin de répondre aux spécificités du domaine bancaire. Nous expérimenterons sur des opinions directes et non des comparaisons qui sont également deux types d'opinion distingués par [Liu, 2012]. Nous ne nous attarderons pas sur la prise en compte des sarcasmes ou de l'ironie, qui relève encore d'un niveau de spécificité plus élevé.

[5] http://sentiwordnet.isti.cnr.it/

2. THESAURUS

Cette première partie consiste à expliquer le processus de création du thésaurus banque-assurance. Il nous amène à décrire d'abord la manière dont nous avons constitué le corpus. Il nous servira également pour la fouille d'opinions.

2.1. CORPUS

Dans un premier temps, nous avons constitué un corpus qui permettait d'obtenir suffisamment d'informations sur le domaine de la banque-assurance. Nous avons donc récupéré des avis d'internautes sur un site web[6] via un script Perl. Nous avons choisi ce site de par l'hétérogénéité des établissements disponibles. Grâce à cela, il était plus probable de trouver des termes génériques au domaine bancaire et non pas à une enseigne en particulier.

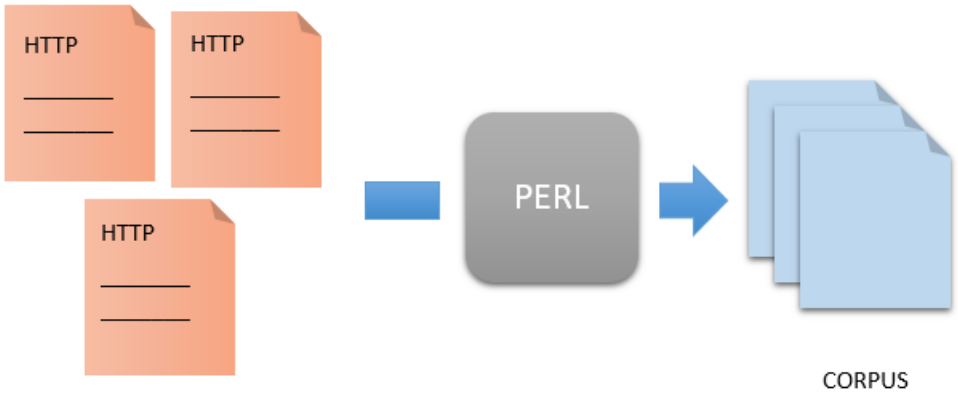

Figure 4 : Processus de création du corpus

2.2. PRETRAITEMENTS

Nous avons d'abord créé une liste des URL menant aux pages des différentes banques sur lesquelles portaient les commentaires. Nous obtenons ainsi 21 liens à parcourir.

[6] http://www.linternaute.com/argent/comparatif/banque/

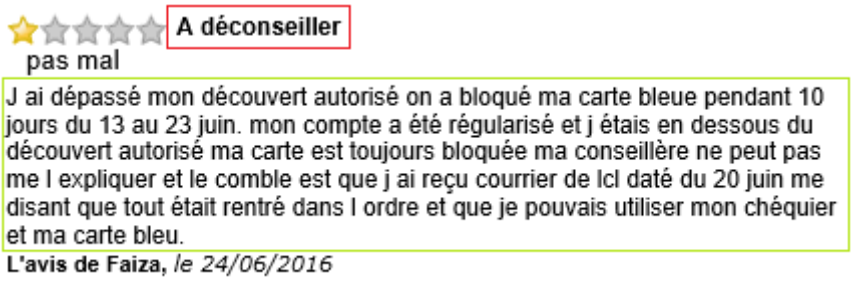

Figure 5 : Exemple de verbatim

Les données que nous voulons récolter ici sont le titre, le commentaire et la banque (e.g. figure 2) dont l'internaute parle.

```
<h1 class="titre_fiche">LCL</h1>
```

Figure 6 : Balise unique dans la page où est indiquée la banque

Etant regroupées dans une balise , nous avons récupéré l'intégralité de cette balise grâce à l'identifiant unique que contient chaque commentaire.

```
<div class="contenu_commentaire" id="commentaire_1729382256910505167">

        <p class="titre"><span class="notation3"><input type="radio"
id="f_note_1729382256910505167" name="f_note_1729382256910505167" score='pas mal' value='1'
READONLY descriptionDefaut = '' classOn='' classOff='' classHover='' classDiv =
'scoreEtoileDivListe' classDescription = '' nullPermis = 'false' checked ><input type="radio"
id="f_note_1729382256910505167" name="f_note_1729382256910505167" score='bien' value='2' ><input
type="radio" id="f_note_1729382256910505167" name="f_note_1729382256910505167" score='très bien'
value='3' ><input type="radio" id="f_note_1729382256910505167" name="f_note_1729382256910505167"
score='excellent' value='4' ><input type="radio" id="f_note_1729382256910505167"
name="f_note_1729382256910505167" score='inoubliable' value='5' ></span>
 <h3>A déconseiller</h3></p>
        <p class="texte_commentaire">
                    <span id="bloc_commentaire_1729382256910505167">J ai dépassé mon découvert
autorisé on a bloqué ma carte bleue pendant 10 jours du 13 au 23 juin. mon compte a été
régularisé et j étais en dessous du découvert autorisé ma carte est toujours bloquée ma
conseillère ne peut pas me l expliquer et le comble est que j ai reçu courrier de lcl daté du 20
juin me disant que tout était rentré dans l ordre et que je pouvais utiliser mon chéquier et ma
carte bleu.</span>

        <br/>
            <span class="auteur">L'avis de Faiza,</span>
            <span class="date">le 24/06/2016</span>

        </p>

    </div>
```

Figure 7 : Structure HTML d'un verbatim

Une fois cet identifiant récupéré grâce à une expression régulière[7], il était possible de parcourir le lien en récoltant les données voulues. Nous avons d'abord procédé à un léger

[7] Expression avec une syntaxe particulière qui permet de reconnaître une chaîne de caractère précise

nettoyage afin de supprimer les retours à la ligne (\n) et remplacer le symbole & par l'entité HTML correspondante *&amp ;*.

Pour les titres et les identifiants, nous avons utilisé deux listes distinctes. La raison pour laquelle nous n'avons pas utilisé de table de hachage est que les données n'y sont pas ordonnées. Or, puisque nous récupérons les titres et les identifiants via deux listes différentes, il était nécessaire que les données soient dans un ordre précis afin de pouvoir les faire correspondre lors de l'enregistrement dans les fichiers de sortie. Pour ce faire, nous utilisons une variable qui va nous servir de compteur et qui va s'incrémenter à chaque nouvel identifiant pour alimenter la liste. Nous enregistrons donc les titres dans une seconde liste mais dont les emplacements correspondent à l'identifiant auquel ils réfèrent. Cela signifie que pour l'identifiant à l'emplacement numéro 10 dans la liste des identifiants, nous aurons le titre correspondant qui sera à l'emplacement numéro 10 dans la liste des titres. Après observation des données, la présence d'un titre n'étant pas requise pour poster un commentaire sur le site, nous constatons que quelques commentaires se trouvent sans titre. Nous avons donc décidé de créer une étape intermédiaire que nous expliquons juste après. Cela permet d'éviter un décalage entre nos deux listes.

Cette étape consiste à créer un patron de remplacement des balises par la balise <sep>. Ainsi, il est plus facile de remplacer ces balises que les originales car ce sont toutes les mêmes. A nouveau, nous récupérons le titre grâce à une expression régulière. Après observation du contenu de la liste de titres, il s'est avéré que quand le titre n'était pas présent, il y avait une double tabulation comme élément de liste (au lieu de mots ou phrases pour les autres).

Une fois cette étape franchie, nous pouvons enfin commencer à enregistrer nos données dans un fichier de sortie.

Le format dans lequel nous enregistrons nos données est le format TMX. C'est un format XML spécifique à l'entreprise.

```
<?xml version="1.0" encoding="UTF-8"?>
<tm xmlns:dc="http://purl.org/dc/elements/1.1/">
  <doc>
    <text>
      <data>The text of the document</data>
    </text>
    <features zone="Authors">
      <ft>/Metadata/Author/John</ft>
      <ft>/Metadata/Author/Jane</ft>
    </features>
  </doc>
</tm>
```

Figure 8 : Exemple de fichier TMX

C'est au moment de l'enregistrement dans le fichier de sortie que la correspondance entre les identifiants et les titres nous importent. Nous utilisons à nouveau le compteur que nous avons remis à 0 pour parcourir les deux listes en même temps. Si le titre correspond à une double tabulation dans la liste, nous affichons [Pas de titre] afin que le contenu soit plus

explicite pour l'humain. Puis nous récupérons le nom de la banque, le commentaire et autour de tous ces éléments, nous ajoutons les balises correspondant au format TMX.

Ce format servira dans l'analyse des résultats (Application p.43). Mais il nous sert également de pivot pour ensuite utiliser un script de conversion[8] qui permet de transformer des fichiers TMX en fichiers LUX. Le format LUX est également un format XML de l'entreprise, davantage optimisé pour le partage des données entre les différents outils.

```
<?xml version="1.0" encoding="UTF-8"?>
<lux>
  <doc>
    <content>
      <text>The text of the document</text>
    </content>
    <metadata>
      <a name="Author" value="John"/>
      <a name="Author" value="Jane"/>
    </metadata>
  </doc>
</lux>
```

Figure 9 : Exemple de fichier LUX

Nous obtenons un corpus de 2471 documents, qui ont chacun une longueur moyenne de 115 mots.

2.3. THESAURUS

2.3.1. METHODES

La création d'un lexique peut se faire manuellement ou automatiquement. Il est également possible de le faire de manière dite semi-automatique.

2.3.1.1. METHODE MANUELLE

Créer un lexique manuellement revient à insérer les entrées soi-même, sans aide informatique pour compléter. Le lexique est créé en utilisant sa propre connaissance du domaine. Ce type de méthode implique d'y consacrer un certain temps car il faut tenter de faire le lexique le plus complet possible. Même en étant spécialiste du domaine, faire ce travail manuellement est fastidieux et long. En ne l'étant pas, il n'est donc pas conseillé d'utiliser ce type de méthode pour créer son propre lexique.

[8] Script propriétaire : luxconversion.bat

2.3.1.2. METHODE AUTOMATIQUE

La méthode automatique inclut une aide informatique basée sur des algorithmes statistiques. Ces algorithmes vont calculer les termes ou expressions les plus pertinents pour enrichir automatiquement le lexique à créer. Cette technique fait gagner énormément de temps comparée la méthode manuelle mais implique de faire confiance à l'aide informatique apportée. Il faut également avoir un corpus sur lequel s'appuyer pour pouvoir utiliser cette méthode, c'est-à-dire un corpus sur le domaine. Avec un corpus existant, cela représente un gain de temps. En effet, la création d'un corpus implique de trouver la quantité de données nécessaires et de les formaliser pour pouvoir être exploitées.

2.3.1.3. METHODE SEMI-AUTOMATIQUE

Le thésaurus que nous créons est fait de manière semi-automatique. Cela signifie que nous utilisons une combinaison des deux méthodes expliquées ci-dessus.

Ici, nous appliquons un algorithme statistique sur le corpus pour relever les termes pertinents qui pourraient être ajoutés. Puis nous ajoutons manuellement les termes au thésaurus. Le rôle joué par l'humain ici n'est pas négligeable. Il permet d'avoir une vue globale sur ce que propose les résultats statistiques et de choisir ce qui est pertinent d'ajouter au thésaurus. Cette méthode s'avère être un bon compromis pour gagner du temps tout en ayant une certaine qualité du contenu.

2.3.2. INTEGRATION A L'OUTIL WEBSTUDIO

Luxid Webstudio est une application web à l'aide de laquelle nous avons constitué notre thésaurus. Cela permet de créer aisément, entre autres, un thésaurus ou bien même d'en compléter un existant en le chargeant tout simplement dans un projet. C'est dans cet outil que nous allons télécharger nos fichiers LUX. Pour cela, nous utilisons un corpus d'apprentissage composé de 2251 verbatim. Grâce à ces documents, nous pourrons avoir des suggestions de termes ou de combinaisons de termes à intégrer dans notre thésaurus (e.g. figure 10).

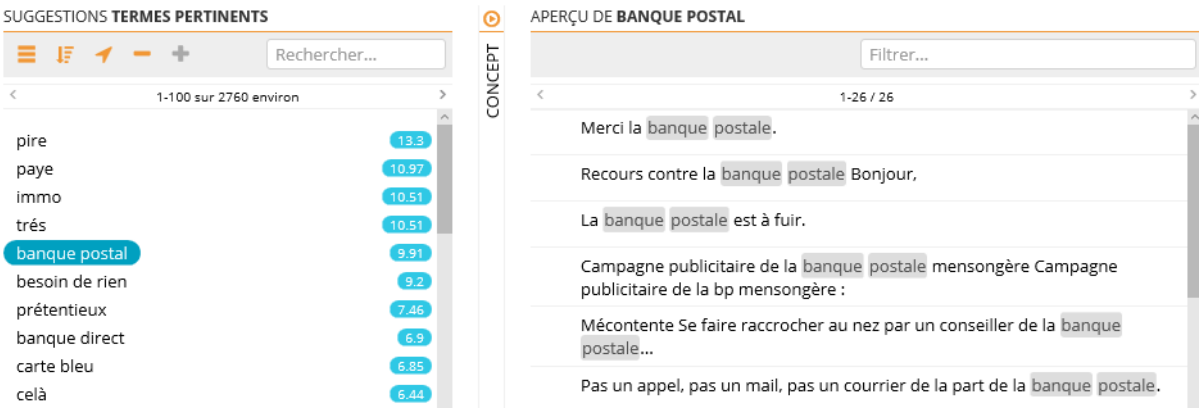

Figure 10 : Suggestions du Webstudio

Sur la figure 10, nous voyons à gauche les suggestions de termes calculées et à droite, la manière dont ils apparaissent dans les documents (surlignés en gris). L'aperçu nous permet de savoir si ajouter ces termes dans notre thésaurus est réellement pertinent ou non. Ces suggestions s'avèrent très pratiques pour avoir une base de travail. En effet, la création du thésaurus se fait d'A à Z sans expertise du domaine.

2.3.3. LES CARTOUCHES DE CONNAISSANCES

Pour obtenir les suggestions de termes dans le Webstudio, une cartouche de connaissance va être appliquée. Une cartouche de connaissance est un module qui permet d'extraire des informations de toutes sortes (termes, catégories, entités nommées, etc.). La cartouche Relevant Term Finder Skill Cartridge® (RTF) est celle utilisée pour obtenir les suggestions. Elle utilise la méthode statistique du TF-IDF (de l'anglais Term Frequency – Inverse Document Frequency).

$$tf \cdot idf_{t,d,D} = tf_{t,d} * idf_{t,D}$$

Figure 11 : Formule du TF-IDF

Cela consiste à prendre le nombre d'occurrences d'un terme *t* dans un document *d* (TF) et de donner un poids à ce terme sous forme de score (plus il est élevé, plus le terme a son importance). Un poids plus important est donnée à un terme qui apparaît beaucoup de fois mais dans un nombre restreint de documents dans le corpus *D*[9] (IDF). Celui-ci est considéré comme plus discriminant et donc plus révélateur. A l'inverse, un terme apparaissant peu de fois dans un seul document ou apparaissant souvent dans plusieurs documents aura moins de poids et aura donc un score moins élevé.

[9] Le corpus utilisé pour IDF provient de Wikipédia.

2.3.4. ENRICHISSEMENT DU THESAURUS

Les concepts peuvent être créés en intégrant les suggestions dans le thésaurus. L'avantage de cette méthode est d'avoir un large éventail de choix de concepts à disposition. Mais la visualisation de chacun des concepts reste assez coûteuse en temps sans connaissance du domaine. Il est également possible d'ajouter les concepts sans passer par la liste de suggestions.

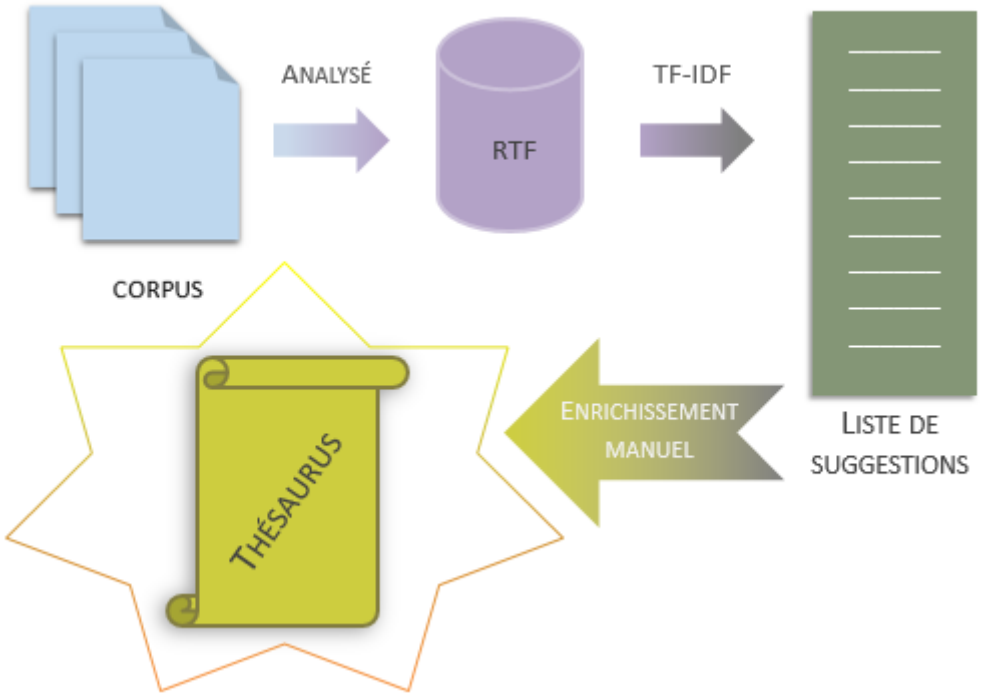

Figure 12 : Etapes de l'enrichissement du thésaurus

Ce schéma reprend tout ce qui a été mis en œuvre pour créer le thésaurus. Toute la partie avant l'enrichissement du thésaurus est automatique. La validation de concepts est la seule étape manuelle de la chaîne. En faisant cela, nous alimentons une cartouche appelée Smart Taxonomy Facilitator Skill Cartridge® (STF). Elle permet d'extraire les termes d'un thésaurus qui lui sont associés. De plus, la STF contient un paramètre de correspondance des termes nommé « fuzzy matching ». Grâce à cela, il est possible d'extraire les termes même s'ils contiennent des fautes d'orthographe ou des permutations de lettres ou de mots (dans le cas de concepts à termes multiples).

Après avoir effectué tout le travail d'enrichissement, nous obtenons 398 concepts rangés dans 9 grandes catégories (e.g. figure 13).

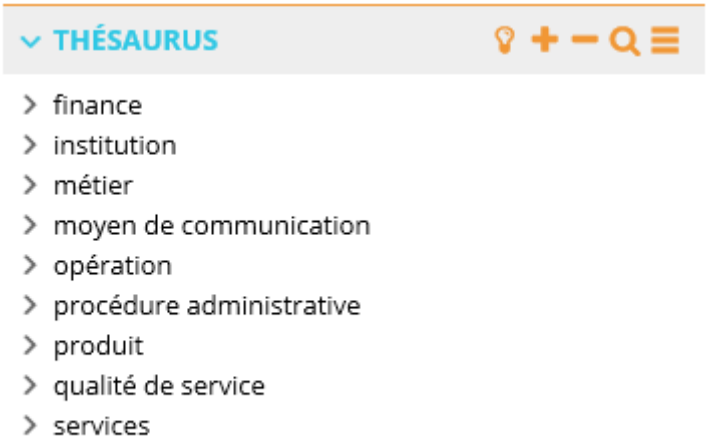

Figure 13 : Vue globale du thésaurus créé

Une fois le thésaurus créé, il nous faut en évaluer la qualité.

2.4. RESULTATS

2.4.1. ANNOTATION WORKBENCH

Annotation Workbench ou AWB est l'outil avec lequel nous analysons la qualité du thésaurus. Nous utilisons un corpus de test composé des 220 documents restants de notre corpus.

Pour évaluer la qualité du thésaurus, il faut vérifier que ce qui est extrait est bon. Dans AWB, il est possible d'assigner plusieurs types de validation (e.g. figure 14).

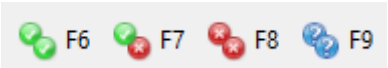

Figure 14 : Possibilités de validation dans AWB

4 types sont différenciés et chacun joue un rôle spécifique. Nous les appellerons par les raccourcis clavier associés pour faciliter la lecture. De gauche à droite, nous avons donc : F6, F7, F8 et F9. Il y a deux choses à prendre en considération quand nous validons dans AWB, le « type » et la « valeur ».

F6 (vrai positif) correspond à un type et une valeur bien extraits (e.g. figure 15). Le nom « conseillers » a bien été extrait dans un bon contexte.

Type	Value	Left context	Normalized value	Original value	Right context
	OK	Incompétence de la bpDes	conseiller	conseillers	qui changent souvent, sans ...

Figure 15 : Extraction correcte validée F6

A l'inverse, F8 (faux positif) n'a ni le type, ni la valeur de corrects. La figure 16 montre que le verbe « conseiller » a été extrait. Or, nous voulons le nom.

Type	Value	Left context	Normalized value	Original value	Right context
	KO	... Pour votre tranquillité je vous	conseiller	conseille	de bien réfléchir avant de vous ...

Figure 16 : Extraction incorrecte validée F8

En revanche, F7 désigne un type correct mais ne va pas avoir une bonne valeur extraite. Comme nous le remarquons sur la figure 17, ce qui a été extrait est bien en rapport avec un conseiller mais l'extraction n'est pas tout à fait ce que nous aurions souhaité.

Type	Value	Left context	Normalized value	Original value	Right context
	KO	... ? Ai demandé à être reçu par un	conseiller personnel	conseiller compétent mais personne	n ? A pu être trouvé pour liquider ...

Figure 17 : Extraction partiellement correcte validée F7

F9 est la valeur par défaut. Cela signifie que l'extraction n'a pas encore été validée par l'annotateur et est en attente de validation. Si l'annotateur a lui-même annoté F9, c'est qu'il ne sait pas quelle validation lui attribuer.

Type	Value	Type	Value
	UNDEFINED		UNKNOWN

Figure 18 : Annotation F9 : à gauche, par défaut (UNDEFINED) et à droite, par l'annotateur (UNKNOWN)

2.4.2. RESULTATS

Nos tableaux de résultats présentent la précision, le rappel et la F-mesure. La précision permet de savoir combien de bonnes extractions il y a eu parmi les extractions faites. Le rappel lui, permet de savoir combien de bonnes extractions il y a eu parmi toutes les extractions qu'il aurait dû y avoir dans le corpus. La F-mesure est une moyenne pondérée de ces deux valeurs.

$$F\text{-}mesure = 2 * \frac{précision * rappel}{précision + rappel}$$

Figure 19 : Calcul de la F-mesure

1625 extractions ont été effectuées par l'outil et 209 manuellement, 1350 ont donc été évaluées. Le tableau 1 présente les résultats stricts de l'évaluation du thésaurus.

Strict		
Précision	**Rappel**	**F-Mesure**
95,6%	88,3%	91,8%

Tableau 1 : Résultats stricts de l'analyse de qualité du thésaurus

Les résultats stricts ne comptent pas les extractions validées F7 comme étant correctes. Mais c'est le cas pour les résultats tolérants (e.g. tableau 2).

Tolérant		
Précision	**Rappel**	**F-Mesure**
97,3%	88,3%	92,7%

Tableau 2 : Résultats tolérants de l'analyse de qualité du thésaurus

2.5. DISCUSSION

Les résultats stricts sont déjà très bons. Plus de 9 extractions sur 10 sont correctes. Mais les résultats tolérants montrent que nous pourrions avoir de meilleurs résultats si nous apportions des améliorations au module d'extraction. Les extractions validées comme partiellement correctes ne sont pas extraites entièrement dans le texte ou à l'inverse sont un peu trop larges.

Left context	Normalized value	Original value	Right context
... de tenue de compte et frais de	débit	débit	. Le conseiller me dit : "suite à un ...
... compte il y a 3mois que des problème	frais de rejet	rejet chéques et frais	rejet prélévement prés de 900e de ...
... cet etablissement est un gouffre	frais de prélèvement	prélèvement abusifs frais	abusif manipulation mystère ! Un ...
... ne la recommande à personne. Des	frais de prélèvement	frais de rejet de prélèvement	alors que le taux de découvert ...
..., donc tous les mois 2*20 ¤ de	frais de découvert	frais de rejet pour un découvert	de... 18¤62 + lettre de relance + ...

Figure 20 : Extractions de concepts partiellement correctes

Les mots surlignés en jaunes sur le figure 20 sont ceux qui auraient dû être extraits. Ce sont des concepts à termes multiples, ce qui rend l'extraction plus difficile à effectuer. La correction de ce type d'erreurs permettrait de se rapprocher des résultats tolérants.

Left context	Normalized value	Original value	Right context
... plus haut et saisir des organismes	téléphone	tels	que l'afnacab, la fnacab, ufc que ...
... des lettres (ça vous savez à peu	prêt	prêt	le faire) et laissez la banque et ...
... que je ne suis visiblement pas	prêt	prête	de recevoir... Pour moi cela ...
... email... Amazing isn't it ?Je ne	conseiller personnel	conseillerais a personne	de mettre tous ses fonds dans cette ...
... bougé et tout ca sans me prévenir (	transfert	transfert	de domicile). De plus, je me suis ...

Figure 21 : Exemple de faux positifs

Les faux positifs extraits (e.g. figure 21) sont dus à une catégorie grammaticale qui n'est pas celle recherchée. La mise en place d'un système permettant de choisir la catégorie souhaitée permettrait d'éviter ces extractions et donc améliorer les résultats.

2.6. CONCLUSION

Les résultats de qualité du thésaurus sont très satisfaisants. En effet, nous avons de bonnes extractions dans plus de 9 cas sur 10. Tout le processus pour créer le corpus de verbatim est une étape essentielle dans la création du thésaurus. Sans celui-ci, il aurait été plus difficile de compléter son contenu. La méthode des suggestions apportée par le TF-IDF est une bonne approche pour enrichir le thésaurus. Cela étant dit, le thésaurus peut être complété. Nous ne sommes pas experts du domaine et il est possible que quelques concepts aient pu nous échapper dans l'élaboration de celui-ci. Ajouter un moyen d'extraire un concept en fonction de sa catégorie grammaticale diminuerait l'extraction de faux positifs.

Toutes ces modifications pourraient améliorer l'ensemble des résultats pour se rapprocher des résultats tolérants voire même les dépasser.

3. FOUILLE D'OPINION

La seconde partie de notre travail porte sur la fouille d'opinions dans le domaine bancaire. Nous avons tenté deux approches, une approche avec un module d'extraction existant et une approche à base de règles développées pas nous-mêmes. Aborder ces deux approches différentes nous permettra d'en faire une comparaison et voir quelles améliorations il est possible d'apporter à chacune d'elles.

3.1. APPROCHES

Il existe deux types d'approche principales dans la détection d'opinion : les approches symboliques et les approches par apprentissage. La combinaison de ces deux types forme les approches hybrides.

3.1.1. APPROCHE SYMBOLIQUE

Les approches symboliques utilisent un lexique ou dictionnaire contenant des termes qui ont déjà des polarités (positive, négative, neutre ou ambigüe) attribuées. A ces termes peuvent être définis différents attributs comme par exemple indiquer l'intensité de la polarité émise par le terme.

Ces dictionnaires sont parfois préexistants et donc réutilisables. Mais il est tout à fait possible de créer soi-même son propre lexique via un corpus ou même manuellement. Nous utilisons un lexique déjà composé de milliers de termes ou expressions d'évaluation, de sentiments, d'opinions à quoi sont associés différents attributs (catégorie grammaticale, polarité, modification de polarité, type de jugement, type d'opinion). Les récents travaux cités précédemment nous indiquent que les cibles sont extraites avant l'opinion. [Liu & al., 2005] utilisent une méthode de détection de mots avec leur polarité et recherchent ensuite les cibles associées. Notre module fonctionne de la même manière. En extrayant les opinions avec leur(s) cible(s), nous obtenons des relations (entre cibles et opinions). C'est sur cet aspect que nous travaillerons car nous ne souhaitons pas classifier des documents dans leur ensemble. Comme [Chaturvedi & Chopra, 2014], nous prendrons comme unité minimale la phrase pour dégager ces relations.

3.1.2. APPROCHE PAR APPRENTISSAGE

Les approches par apprentissage reposent des observations statistiques en corpus de caractéristiques. Plusieurs algorithmes permettent un apprentissage statistique tels que Naive Bayes, SVM (Support Vector Machine) ou encore KNN (K Nearest Neighbours).

3.2. MODULE D'EXTRACTION DE FOUILLE D'OPINION PREEXISTANT

3.2.1. FONCTIONNEMENT

Le module s'appelle Opinion Mining Skill Cartridge® que nous simplifierons par OMSC. Il utilise un lexique contenant plus de 7000 termes et expressions d'opinions (français et anglais) sur lesquels sont attribuées des polarités. Il peut fonctionner seul ou bien en combinaison avec un thésaurus. Dans notre travail, le thésaurus du domaine bancaire va permettre d'avoir des cibles prédéfinies pour le module. Puis des règles d'extraction repèrent les relations d'opinion basées sur des patrons morphosyntaxiques. La théorie de l'Appraisal est adaptée pour chercher ce qui est déclencheur de l'opinion [Bloom & al, 2007] [Martin & White, 2005].

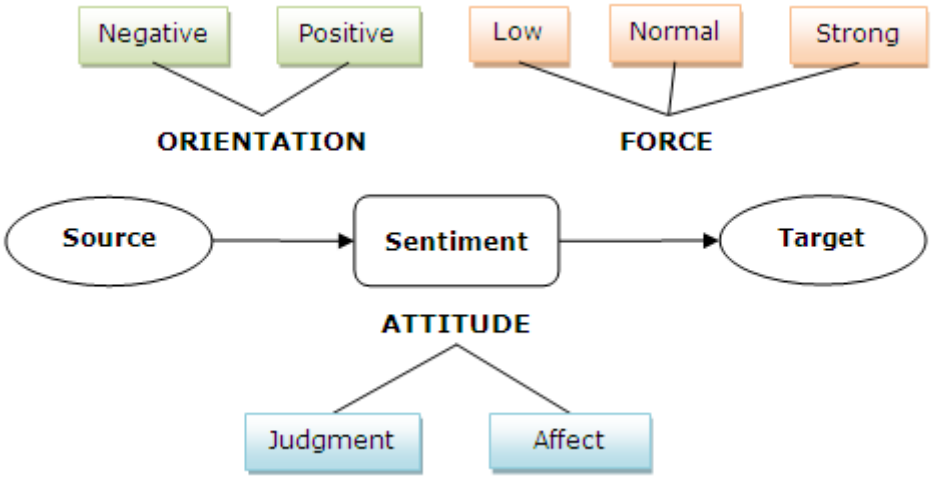

Figure 22 : Adaptation de la théorie Appraisal pour le module de fouille d'opinion

Il faut rappeler que l'OMSC fonctionne de la manière dont [Liu & al., 2005] utilisent leur système de fouille d'opinion. Il détecte une opinion pour laquelle il attribue une polarité et recherche ensuite la cible associée. La combinaison de l'OMSC avec le thésaurus que nous avons créé permet d'avoir un lot de cibles prédéfinies. Celles-ci seront extraites même si elles ne sont pas en relation avec une opinion. Toutes les autres cibles trouvées ne faisant pas parties du thésaurus seront considérées comme cibles dites potentielles. Un code couleur permet de se repérer dans la visualisation de l'extraction :

- Vert : positif
- Rouge : négatif
- Bleu : neutre
- Rose : ambigu
- Orange : cible extraite sans opinion

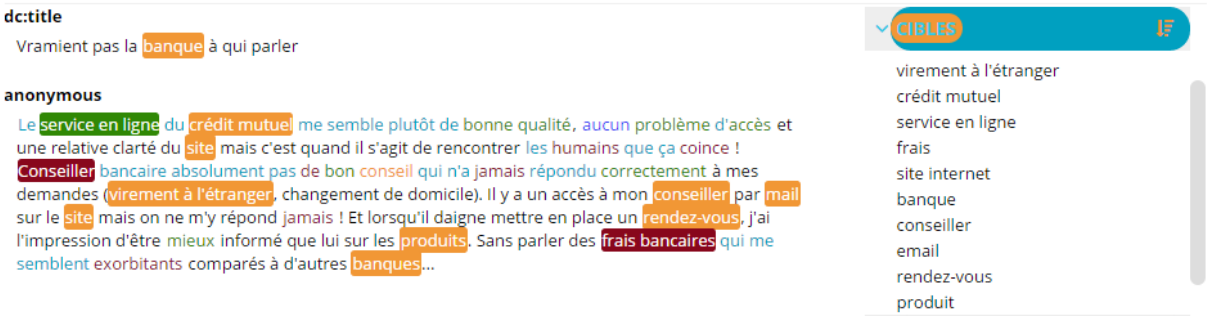

Figure 23 : Code couleur de l'OMSC

La figure 23 illustre le code couleur de l'OMSC visible à travers les cibles sélectionnées. Les cibles colorées en rouge et en vert sont celles avec des polarités qui leur ont été attribuées. Uniquement les termes du thésaurus repérés dans le texte sans relation avec une opinion seront surlignés en orange. Les autres couleurs peuvent être aussi bien pour une cible que pour une opinion.

3.2.2. COUPLAGE DU THESAURUS ET DU MODULE DE FOUILLE D'OPINION

Le couplage avec le thésaurus du domaine bancaire nous donne la possibilité d'extraire des relations entre une opinion et sa cible (e.g. figure 24).

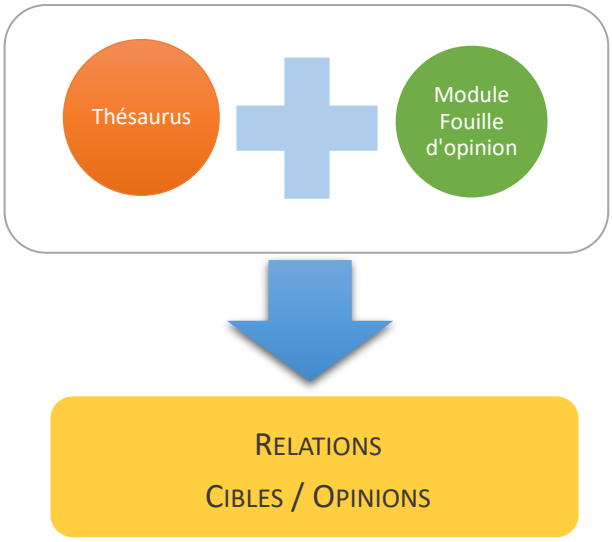

Figure 24 : Couplage du thésaurus et du module de fouille d'opinions

L'évaluation des relations permet d'avoir plus de détails sur ce que les clients aiment ou n'aiment pas dans leur banque. Prendre le contenu du document dans son ensemble et dire si c'est positif ou négatif n'apporte pas d'informations sur ce qui donne ce résultat. Aller voir un petit peu plus en profondeur, c'est pouvoir détecter quels sont les points forts et les points faibles de ces banques à travers des thématiques comme les services, les tarifs, etc.

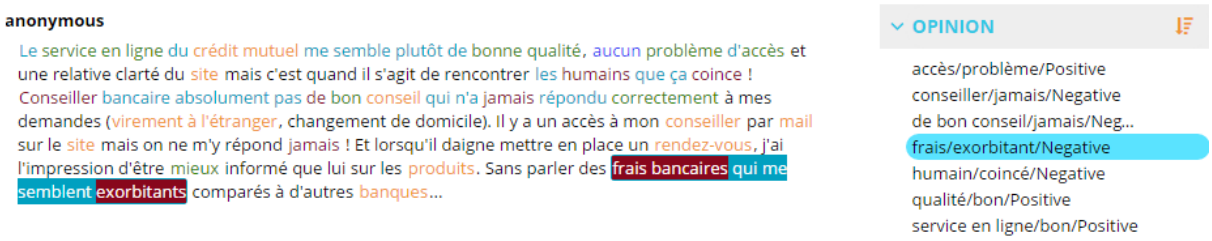

Figure 25 : Extraction d'une relation avec l'OMSC

La figure 25 montre plusieurs relations extraites par le module dont une surlignée. Le contenu dans le texte est visible sur la partie gauche et les éléments de la relation rassemblés sur la partie droite. Ces derniers sont l'objet de notre évaluation.

3.2.3. Mode de validation

Le mode de validation sert à expliciter de quelle manière nous évaluons les extractions effectuées par le module. Nous avons pris en compte trois critères (opinion, polarité, cible) pour évaluer les extractions réalisées par le module. Nous avons ainsi défini 8 cas en fonction des combinaisons de ces trois critères.

Cas	Opinion	Polarité	Cible	Validation
1	✓	✓	✓	Correct
2	✓	x	✓	Partiellement correct
3	x	✓	✓	Incorrect
4	x	x	✓	Incorrect
5	✓	✓	x	Partiellement correct ou Incorrect
6	✓	x	x	Partiellement correct ou Incorrect
7	x	✓	x	Incorrect
8	x	x	x	Incorrect

Tableau 3 : Mode de validation adopté

L'opinion étant extraite en premier, nous nous sommes basés sur cela pour valider ou non une extraction. C'est pour cette raison que nous attribuons partiellement correct sur une opinion qui n'a pas trouvé sa cible. A l'inverse, si une opinion est mal extraite, nous considérons que c'est incorrect. Etant donné que le premier élément n'est pas correct, le reste ne le sera pas non plus (e.g. tableau 3 cas 3, 4, 7 et 8). Dans les cas 5 et 6, deux situations peuvent se présenter :

- soit la cible est mal identifiée (partiellement correct)
- soit il n'y a pas de cible à identifier et une relation qui n'a pas lieu d'être est extraite (incorrect)

3.2.4. ANALYSE DE QUALITE ET RESULTATS

Nous avons effectué l'analyse de qualité de la cartouche OMSC sur 500 documents[10] de notre corpus choisis au hasard. Nous n'avions pas besoin de corpus de test spécifique puisque la cartouche existait déjà. Cela signifie qu'aucun de nos documents n'a été utilisé pour mettre en place le module. Seules les extractions des relations sont concernées par les résultats rapportés. Voici ceux obtenus une fois l'analyse de qualité terminée :

Strict		
Précision	**Rappel**	**F-Mesure**
70,3%	69,8%	70,1%

Tableau 4 : Résultats stricts de l'analyse de qualité du module de fouille d'opinions

Tolérant		
Précision	**Rappel**	**F-Mesure**
89,7%	69,8%	78,5%

Tableau 5 : Résultats tolérants de l'analyse de qualité du module de fouille d'opinions

Il faut bien prendre en compte le fait que ces résultats ne peuvent être dissociés du mode de validation suivi. En suivant un autre mode de validation, les résultats obtenus seraient certainement différents comme par exemple si nous nous étions basés sur les cibles et non pas sur les opinions.

Sur les résultats stricts (e.g. tableau 4), nous observons une qualité globale de 70,1% avec un rappel et une précision équilibrés. Les résultats tolérants (e.g. tableau 5) apportent des

[10] Le temps de stage imparti ne permettait pas d'en faire davantage

informations supplémentaires sur la précision. Pour rappel, les résultats tolérants prennent en compte les extractions validées comme partiellement correctes. Ce qui signifie que nous pourrions avoir presque 9 extractions sur 10 effectuées par l'outil qui pourraient être correctes si nous améliorions le module.

3.2.5. DISCUSSION

Face à ces résultats, nous voulions savoir quels types d'erreurs revenaient le plus souvent. L'ensemble des 406 erreurs relevées dans les 500 documents sont réparties dans le graphique suivant :

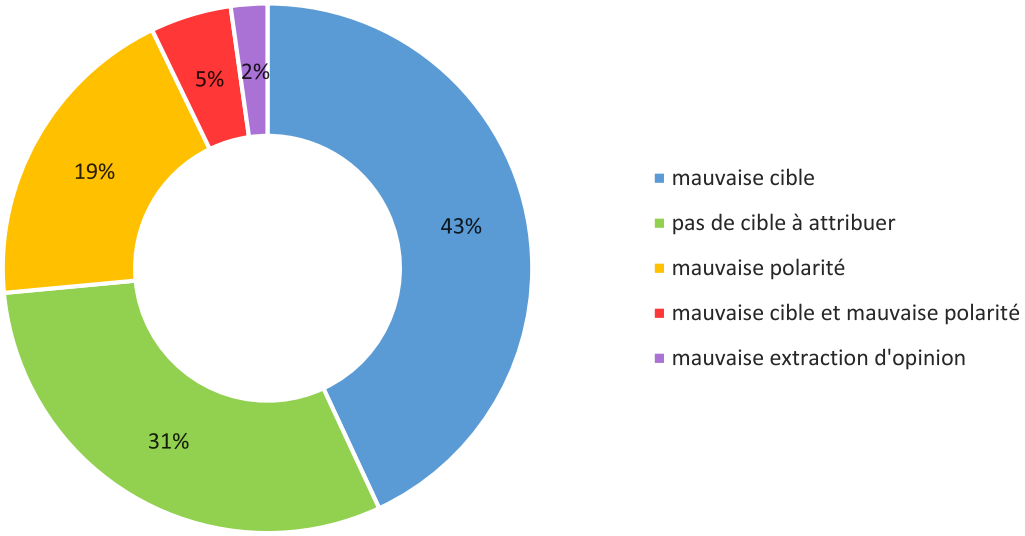

Figure 26 : Taux des erreurs des extractions avec OMSC

La majorité des mauvaises extractions sont dues à des mauvaises attributions de cibles. Il faudrait revoir les patrons syntaxiques sur lesquels est basée la recherche de cibles pour tenter d'en améliorer l'extraction.

Un tiers des erreurs est tout de même dû à des relations qui n'avaient pas lieu d'être (parties verte et violette sur le graphique). Les fautes de frappe ou d'orthographe donnent lieu à des extractions erronées (faux positifs), par exemple, "pur" au lieu de "pour" (e.g. figure 27).

Left context	Normalized value	Original value	Right context
... complexités de hsbc. Fr, mais	chiffre/pur/Positive	un seul chiffre pur	vus montrer l'incroyable ...

Figure 27 : Extraction due à une faute de frappe

Parfois, nous avons des relations extraites alors qu'elles n'avaient pas lieu d'être comme sur la figure 28.

Left context	Normalized value	Original value	Right context
Fuyez !Commençons par le seul	intérêt/vite/Positive	intérêt de cette banque, ça ira plus vite	: les taux d'intérêts de leurs ...

Figure 28 : Relation extraite

Mais si nous prenons dans l'ensemble ce que nous avons validé comme partiellement correct (c'est-à-dire les parties bleue, jaune et rouge sur le graphique), nous obtenons les deux tiers des erreurs. Ces deux tiers représentent ce que nous obtenons comme différence entre les résultats stricts et tolérants.

Les mauvaises polarités attribuées semblent être ce qui serait le plus simple à améliorer. En effet, la polarité est un attribut de l'opinion qu'il suffit de modifier. Comme nous l'avons vu dans l'état de l'art, en fonction du domaine, la polarité peut changer [Marchand, 2013].

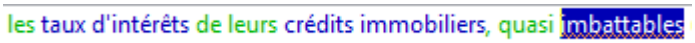

Figure 29 : Exemple d'un terme pouvant changer de polarité dans le domaine bancaire

La figure 29 présente « imbattables » de manière positive. Mais dans le domaine sportif, ce même terme aurait plutôt une polarité négative. Il faudrait donc prendre en compte cet aspect pour modifier les polarités de certaines opinions.

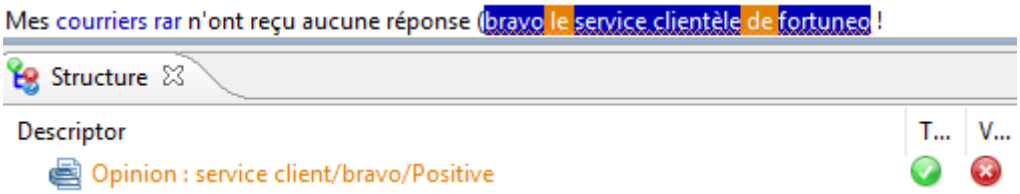

Figure 30 : Ironie validée comme partiellement correcte

Le module ne prenant pas en compte l'ironie, celle-ci a été validée comme étant une mauvaise polarité (cas n°5 du mode de validation). Cela ne signifierait pas pour autant que nous pourrions gagner la totalité des 19% d'erreurs de cette catégorie.

3.2.6. CONCLUSION

Cette première approche utilise un module de fouille d'opinion couplé à un thésaurus du domaine bancaire. L'évaluation de la qualité du module portait sur les relations entre une opinion et sa cible. Nous obtenons une F-mesure de 70,1% avec un rappel et une précision équilibrés. Par rapport aux résultats de référence, la précision ainsi que la F-mesure sont plus élevées avec notre méthode.

3.3. DEVELOPPEMENT DES REGLES D'EXTRACTIONS

Cogito Studio est un logiciel créé par Expert System qui est un environnement de développement de catégorisation et d'extraction linguistiques. L'objectif est d'essayer d'approcher les résultats du module existant en évaluant des relations recréées à travers des règles d'extraction. Dans le temps imparti, l'hypothèse est qu'il sera probablement très difficile d'atteindre le niveau de performance de l'Opinion Mining Skill Cartridge (OMSC) car avec Cogito Studio, nous partons de zéro.

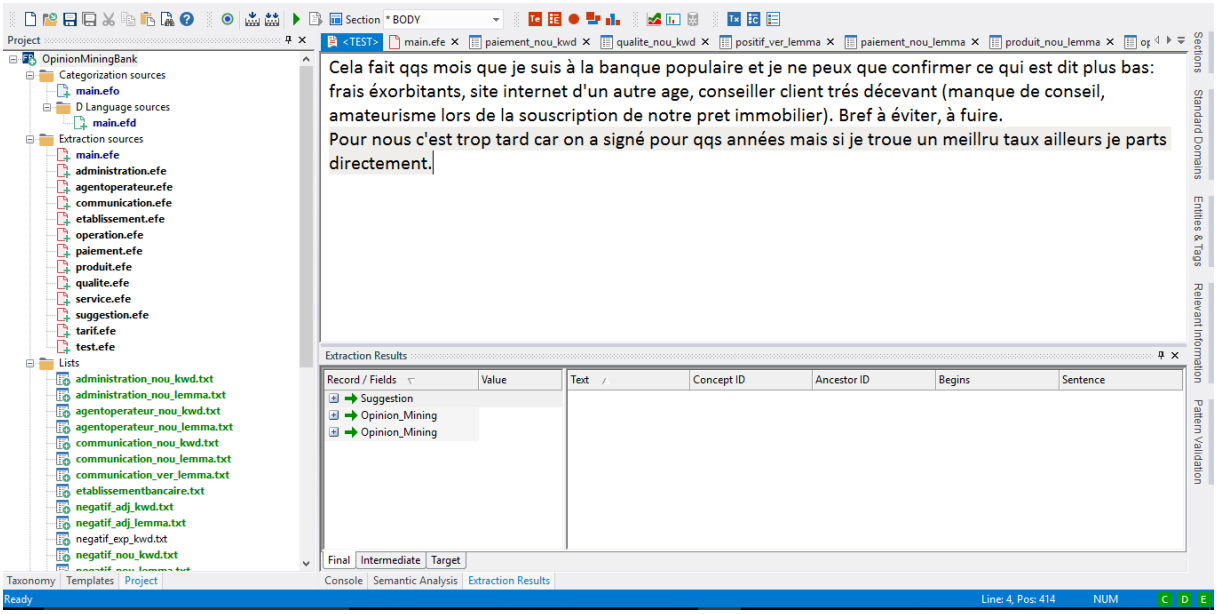

Figure 31 : Interface de Cogito Studio

3.3.1. RESEAU SEMANTIQUE

Cogito Studio utilise un réseau sémantique multilingue propriétaire Expert System. Ce réseau multilingue se sous-divise en réseaux monolingues. Un réseau monolingue se distingue par la langue dont celui-ci fait l'objet. Par exemple, réseau allemand représente la langue allemande. Les réseaux italien et anglais sont à ce jour les seules langues achevées. Les réseaux français, allemand, espagnol, chinois, coréen, russe, portugais et japonais sont en cours de développement.

3.3.2. ANALYSE SEMANTIQUE

L'analyse sémantique d'une phrase se fait sur 5 niveaux et commence par la catégorisation grammaticale des plus petites unités (e.g. « Atom » figure 32). Un réseau monolingue s'appuie sur la catégorie grammaticale des termes pour leur attribuer un sens.

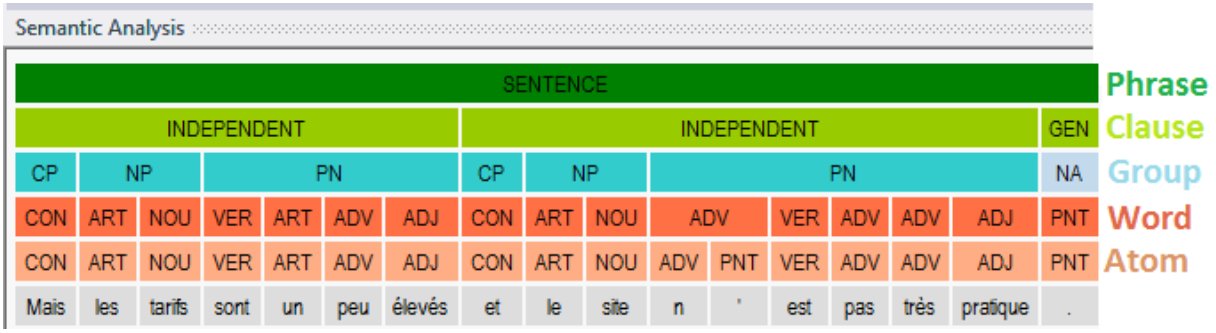

Figure 32 : Analyseur sémantique de Cogito Studio

Mais beaucoup d'erreurs ont été observées sur le sens donné aux termes d'une phrase. Ces erreurs sont dues à une mauvaise analyse syntaxique du français [11]. L'étiquetage morphosyntaxique fait partie d'un programme propriétaire qu'il n'est pas possible de modifier. Il a donc fallu s'adapter en connaissant les lacunes du système.

3.3.3. LES LISTES DE TERMES

Pour le développement des règles d'extraction, nous avons utilisé des listes pour les termes bancaires et d'opinions. Pour créer ces listes, nous avons repris des termes présents dans le thésaurus et le module de fouille d'opinion existant. Ces listes sont construites manuellement, en insérant les termes les uns après les autres. Pour qu'elles soient plus faciles à créer, l'ensemble des termes du thésaurus ou de fouille d'opinion n'ont pas été reportés. Par exemple, au lieu de reprendre l'ensemble des termes faisant référence à des frais bancaires, seul « frais » a été gardé. Si un terme est présent dans un réseau monolingue, il est possible de n'inscrire que le lemme dans la liste. Si cela n'est pas le cas, il faut créer une liste de mots-clefs. Cela signifie qu'elle contient le terme avec toutes ses variantes possibles, que ce soit un nom, un adjectif ou un verbe. Pour éviter toute confusion, nous normalisons les extractions en catégorisant les listes. Par exemple, si « frais de découvert » ou « frais de

[11] Le département Recherche et Développement d'Expert System France (chargé de l'enrichissement du réseau français) a alors recensé les types d'erreurs effectués par le réseau français. Sur plus de 18000 mots dans un corpus de 200 documents, le taux d'erreurs de 46% a été relevé sur les mots pleins (exclusion des mots outils type pronom, articles, etc). Cela signifie que presqu'un mot sur deux n'est pas bien analysé. 61% des erreurs sont des mauvaises désambigüisations bien que les termes soient présents dans le réseau français.

gestion » sont trouvés dans un verbatim, ils seront normalisés sous « tarification ». Nous avons créé 10 catégories de ce type :

- Administration
- Agent/opérateur
- Communication
- Etablissement
- Opération
- Paiement
- Produit
- Qualité de service
- Service
- Tarification

Ces catégories permettent d'avoir le minimum de listes avec des mots-clefs pour gagner du temps sur le développement des règles. Les listes de termes ont été classées en fonction de leur catégorie grammaticale pour faciliter le développement des règles, que ce soit pour les termes bancaires ou pour les termes d'opinion. Les opinions n'ont que les polarités positives et négatives. Les expressions du français comme « être à la hauteur » ou « faire un carton » ne sont pas présentes dans le réseau français et demande d'avoir des règles un peu plus élaborées pour être extraites. Seuls les noms, les adjectifs et les verbes font partie des listes d'opinion. Il semblait nécessaire d'observer les résultats sur des extractions impliquant deux mots entre eux avant de vouloir extraire des expressions.

Ainsi, 23 listes ont été créées dont 15 pour les termes bancaires (229 termes au total) et 8 pour les termes d'opinion (1031 au total).

3.3.4. DEVELOPPEMENT DES REGLES D'EXTRACTION

Le réseau sémantique servira à extraire les termes à travers la catégorie grammaticale souhaitée. Par exemple, « conseiller » peut être soit un nom, soit un verbe.

En indiquant la catégorie grammaticale à extraire, les mauvaises extractions sont a priori exclues. Les mauvaises analyses de l'outil restent néanmoins possibles mais cela n'a aucun lien avec les règles.

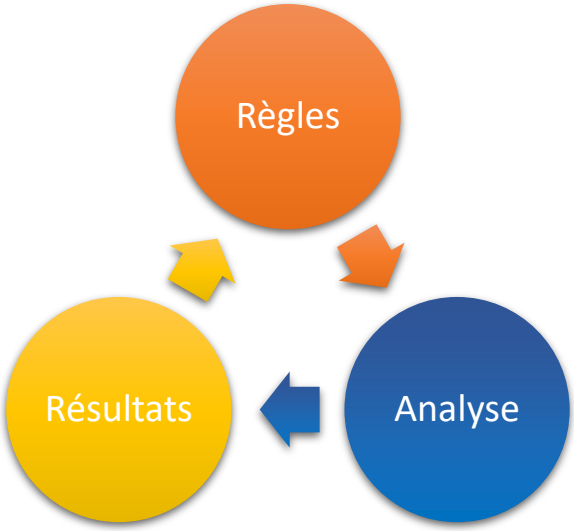

Figure 33 : Processus de développement des règles d'extraction

Nous avons suivi ce processus tout au long du développement des règles d'extraction (e.g. figure 33). Un panneau de test permet de vérifier la portée des règles. Il est possible d'écrire une ou plusieurs phrases voire insérer un verbatim de notre corpus pour tester si la configuration choisie extrait correctement ou non. En fonction des résultats, les règles sont modifiées ou maintenues.

```
SCOPE SENTENCE
{
/*tarif - neg + neg adj
exemple : tarifs élevés */
   IDENTIFY(Opinion_Mining)
   {
      @@Tarif[LEMMA(LOAD "tarification_nou_lemma.txt") + TYPE(NOU)]
      <0:3>
      !SYNCON(111602,6926,111640,111099,111365) // # 111602: peu, pas beaucoup
      <0:2>
      !TYPE(PNT)
      <0:5>
      @@Negatif[LEMMA(LOAD "negatif_adj_lemma.txt") + TYPE(ADJ)]
   }
}
```

Figure 34 : Exemple de règle d'extraction

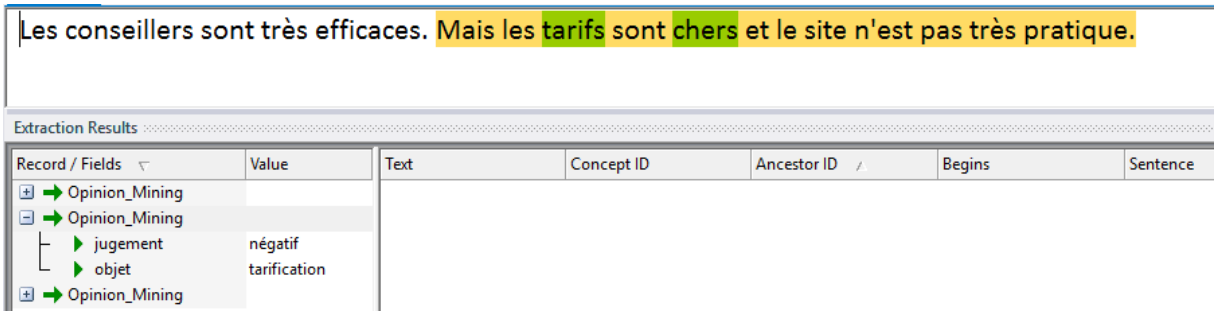

Figure 35 : Extraction dans le panneau de test

La règle de la figure 34 permet d'extraire les informations visibles sur la figure 35. Mais certaines règles peuvent se chevaucher, il est donc nécessaire de faire des exclusions (e.g. figure 34 « !SYNCON ») . C'est notamment valable sur les négations pour éviter les polarités différentes sur la même relation comme sur la figure 36.

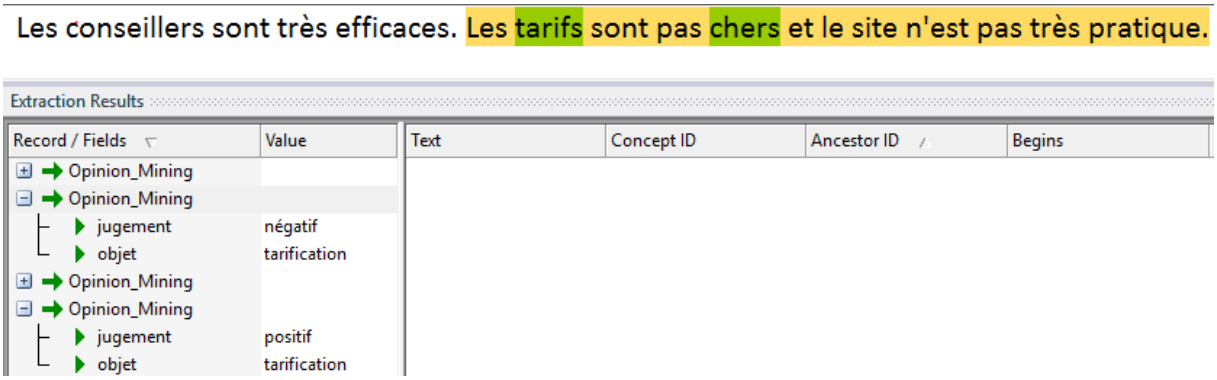

Figure 36 : Extraction de la même relation avec des polarités différentes

Les extractions sont normalisées en sortie. « Tarifs » et « chers » sont respectivement normalisés en « tarification » et « négatif ». De cette manière, la catégorie extraite est tout de suite identifiable avec la polarité attribuée à celle-ci.

Au total, 542 règles ont été développées sur toutes les catégories avec chaque liste de termes d'opinions.

3.3.5. AMELIORATIONS DES EXTRACTIONS AVEC UN SCRIPT DE CORRECTION

Nous avons cherché à augmenter nos chances d'extraction en passant par une correction orthographique des termes bancaires.

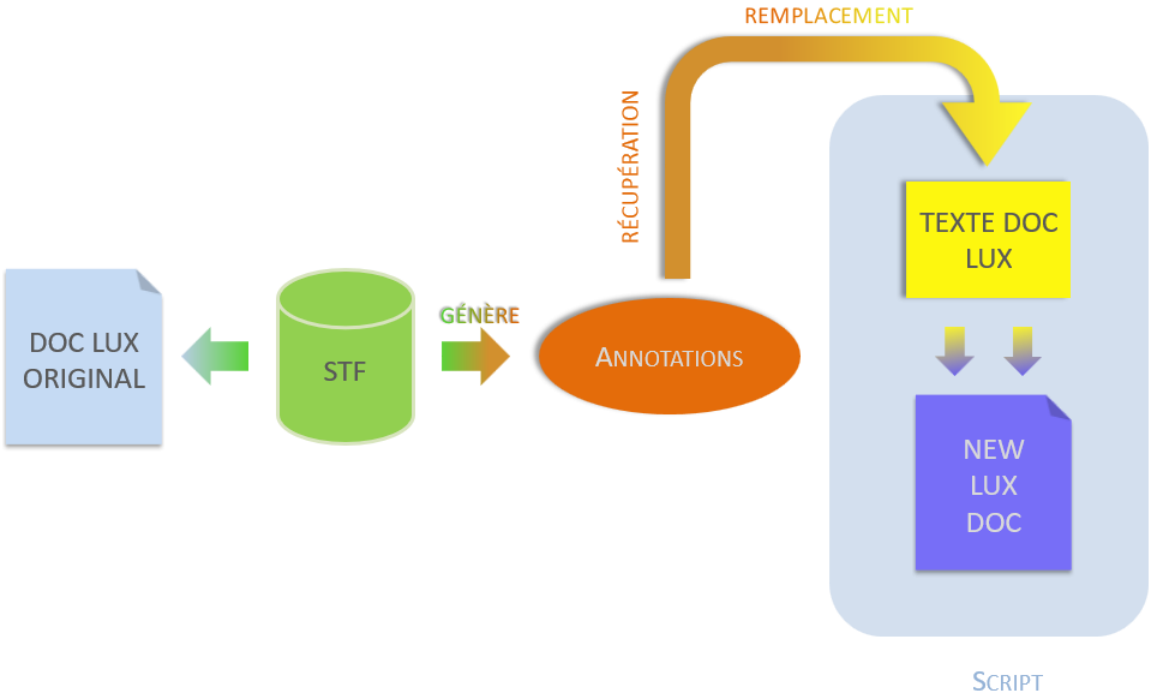

Figure 37 : Processus de correction des annotations

Le thésaurus créé dans la première partie est réutilisé pour faire fonctionner la correction orthographique.

Les annotations générées par la STF du thésaurus vont être récupérées pour remplacer les termes bancaires du texte quand ceux-ci sont reconnus. Les listes étant créées depuis le thésaurus, les mots corrigés font partie du vocabulaire contenu dans ces listes. Les extractions avec les règles développées se feront alors sur le texte corrigé.

3.3.6. CONVERSION DE FICHIERS

Après correction du texte, il a fallu modifier la structure XML du document LUX pour correspondre aux sorties des figures 23 et 25. Cela a été effectué à travers un second script récupérant les données nécessaires pour mener à bien ces modifications.

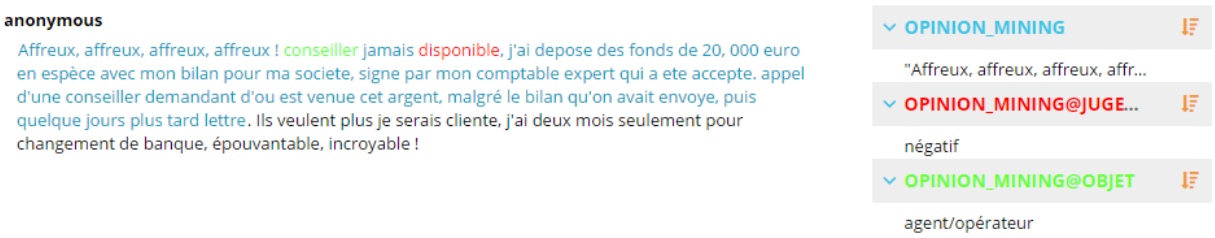

Figure 38 : Sortie des règles d'extraction

3.3.7. ANALYSE DE QUALITE ET RESULTATS

Toutes ces étapes mènent à l'évaluation de la qualité des règles d'extraction. Malheureusement, dû à une incompatibilité des sorties entre les deux suites d'outils et malgré les solutions mises en place, il n'a pas été possible d'effectuer cette analyse de qualité.

Pour rappel, nous souhaitions comparer les deux approches et leurs résultats. Les sorties obtenues avec les règles ne nous permettent pas de comparer les résultats entre eux.

Avec l'OMSC, chaque relation entre une opinion et sa cible faisait l'objet de sa propre extraction. Ici, en essayant de recréer le même système, toutes les extractions sont regroupées si elles appartiennent à la même phrase. Les figures 39 et 40 démontrent que les relations ne sont plus maintenues.

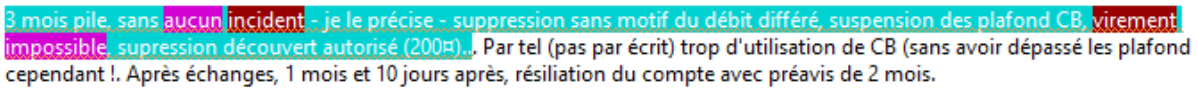

Figure 39 : Extrait d'un verbatim contenant une extraction multiple obtenue avec les règles d'extraction

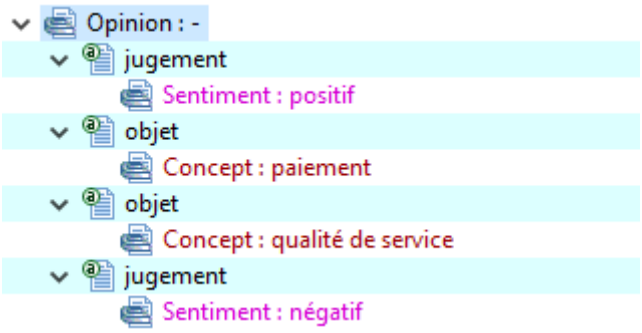

Figure 40 : Relations extraites provenant de la phrase de la figure 39

Sur la figure 40 sous « Opinion », plusieurs éléments sont présents là où il n'y en avait que deux avec l'OMSC. Ces éléments sont toutes les extractions qui ont été effectuées dans la même phrase. La partie « Opinion » ne doit regrouper qu'une opinion et une cible pour permettre une évaluation de la relation. Avec ces multiples éléments, il n'est plus possible de savoir quelle relation est validée sous « Opinion ».

Actuellement, l'analyse du problème est en cours afin de poursuivre le travail et obtenir des résultats.

3.3.8. DISCUSSION

Nos listes étant classées en fonction des catégories grammaticales, l'amélioration de l'analyseur apporterait de meilleurs résultats d'extraction. Le réseau français complété

permettrait d'extraire les termes avec le sens souhaité. Il contiendrait aussi des expressions utilisables pour la fouille d'opinion comme « être à l'écoute » par exemple. Les listes de termes d'opinion ne contiennent aucune expression car composée de plusieurs termes. Nous souhaitions aborder le problème avec des termes simples et en visualiser les résultats avant de prendre en compte des expressions complexes.

3.3.9. CONCLUSION

N'ayant pas de résultats sur le module à base de règles créées, il n'est pas possible d'affirmer que l'utilisation de cette méthode n'est pas optimale pour la fouille d'opinion. Néanmoins, les limites rencontrées lors du développement (absence d'expression d'opinion par exemple) montrent qu'il ne fait pas concurrence au module existant à ce stade.

4. APPLICATION

4.1. LUXID INFORMATION ANALYTICS (LIA)

LIA est une plateforme en ligne permettant de restituer des résultats et d'effectuer des analyses sous divers aspects comme des graphiques ou des tableaux. Cette plateforme est utilisable par un client souhaitant faire des recherches sur ses propres données. Nous avons utilisé LIA pour montrer des applications concrètes sur notre corpus avec le thésaurus et le module de fouille d'opinions existant[12].

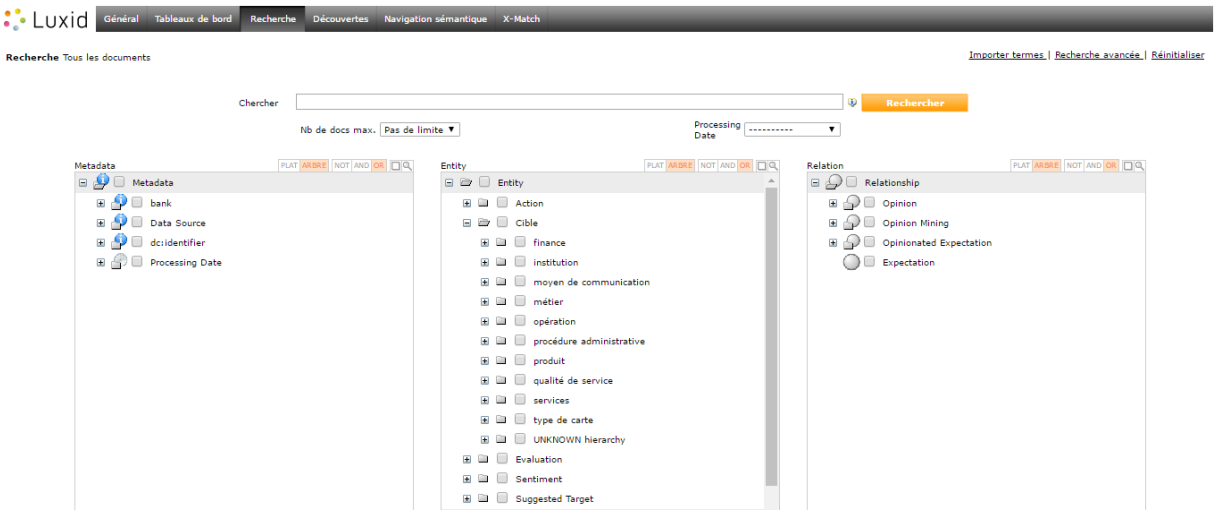

Figure 41 : Luxid Information Analytics

4.2. GRAPHIQUES ET ANALYSES SUR LE THESAURUS

LIA permet d'effectuer des graphiques qui analysent les données et métadonnées du corpus. La figure 42 regroupe l'ensemble des grandes catégories du thésaurus en fonction du nombre d'occurrences apparaissant dans le corpus. Toutes les sous-branches de l'arbre sont prises en compte dans ce regroupement. Les verbatim relatent le plus souvent des institutions (1800 occurrences) suivies de près par les produits (1700 occurrences).

[12] Les règles développées ne sont pas compatibles avec LIA

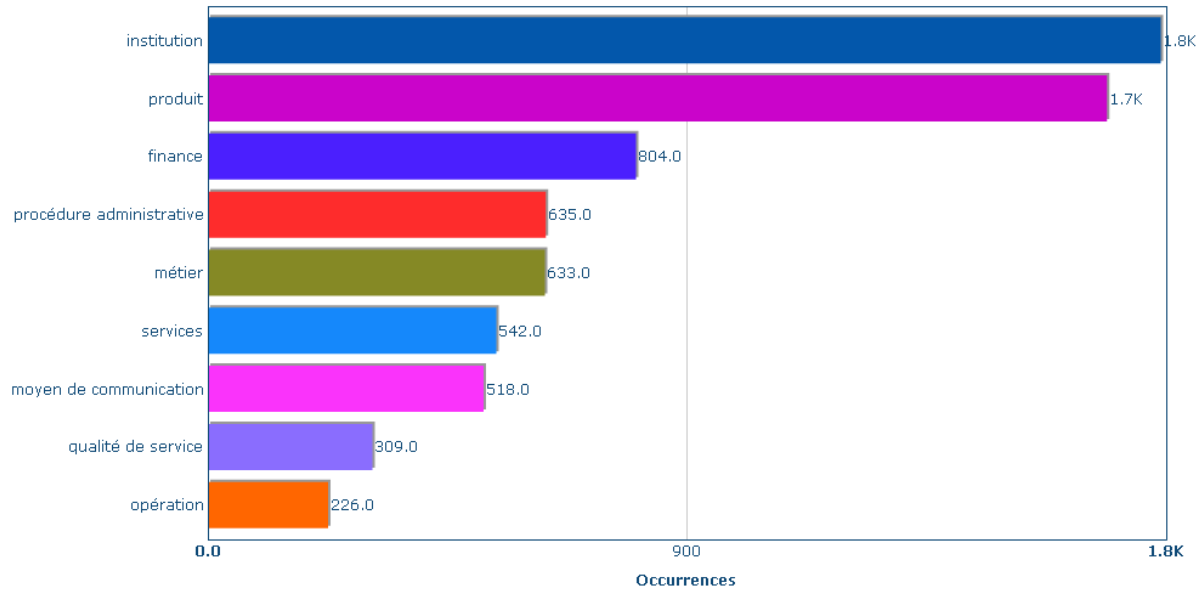

Figure 42 : Nombre d'occurrences pour chaque catégorie de concepts

Les métadonnées du corpus ne concernent que la banque dont fait l'objet un verbatim et un numéro d'identifiant unique de document. La figure 43 présente un graphique se basant sur la métadonnée comprenant les différentes banques. Pour chaque banque, le nombre de documents y faisant référence est indiqué par ordre décroissant.

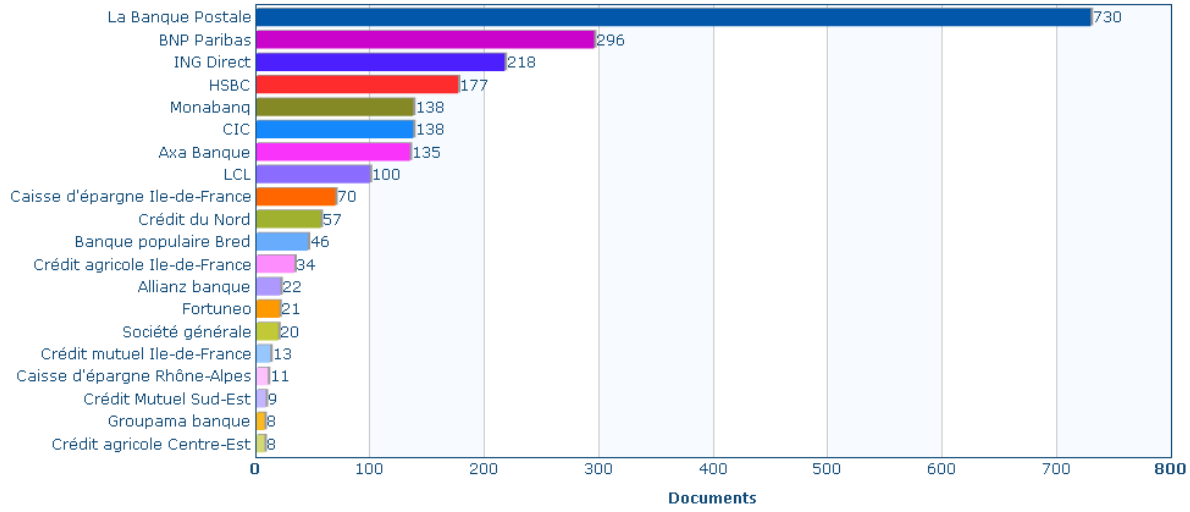

Figure 43 : Nombre de documents en fonction des banques

La Banque Postale est la banque faisant l'objet d'une majorité de verbatim. Les données de ces deux graphiques peuvent être croisées dans un tableau (e.g. tableau 6).

Filtrer / Cible / bank	finance	institution	moyen de communication	métier	opération	procédure administrative	produit	qualité de service	services
Allianz banque	5	13	4	6	2	7	13	4	3
Axa Banque	34	58	25	22	6	30	47	21	25
Banque populaire Bred	16	24	9	13	2	10	14	3	15
BNP Paribas	93	165	49	81	31	67	128	37	62
Caisse d'épargne Ile-de-France	15	27	8	22	2	9	23	13	8
Caisse d'épargne Rhône-Alpes	5	6	2	5	1	3	7	1	5
CIC	40	71	25	51	8	21	46	21	18
Crédit agricole Centre-Est	4	4	1	2		4	5	4	2
Crédit agricole Ile-de-France	8	13	6	6	3	4	12	7	8
Crédit du Nord	18	34	2	21	3	7	22	12	8
Crédit mutuel Ile-de-France	3	7	5	4		4	4	2	2
Crédit Mutuel Sud-Est	4	4	2	2	2	1	2	3	
Fortuneo	6	12	3		2	1	11	2	6
Groupama banque		7	2	4	1	3	3	2	3
HSBC	30	88	43	36	10	37	52	26	55
ING Direct	62	115	49	23	22	51	93	34	47
La Banque Postale	160	356	143	192	97	186	308	151	162
LCL	24	60	18	29	7	17	37	14	25
Monabanq	35	65	17	30	22	27	64	8	22
Société générale	7	13	4	6	3	4	7	1	7

Tableau 6 : Vue croisant les documents entre les catégories du thésaurus et les banques

En croisant ces données, nous remarquons des corrélations avec les résultats des graphiques. La Banque Postale est la banque regroupant le plus de documents dans chaque catégorie. Ce résultat paraît logique puisque 730 documents sont sur cet établissement.

4.3. GRAPHIQUES ET ANALYSES SUR LA FOUILLE D'OPINION

Le but de ce module est d'explorer ce que les clients apprécient ou non dans leur établissement bancaire. La figure 44 donne un aperçu du nombre d'occurrences par polarité, sans se préoccuper des relations.

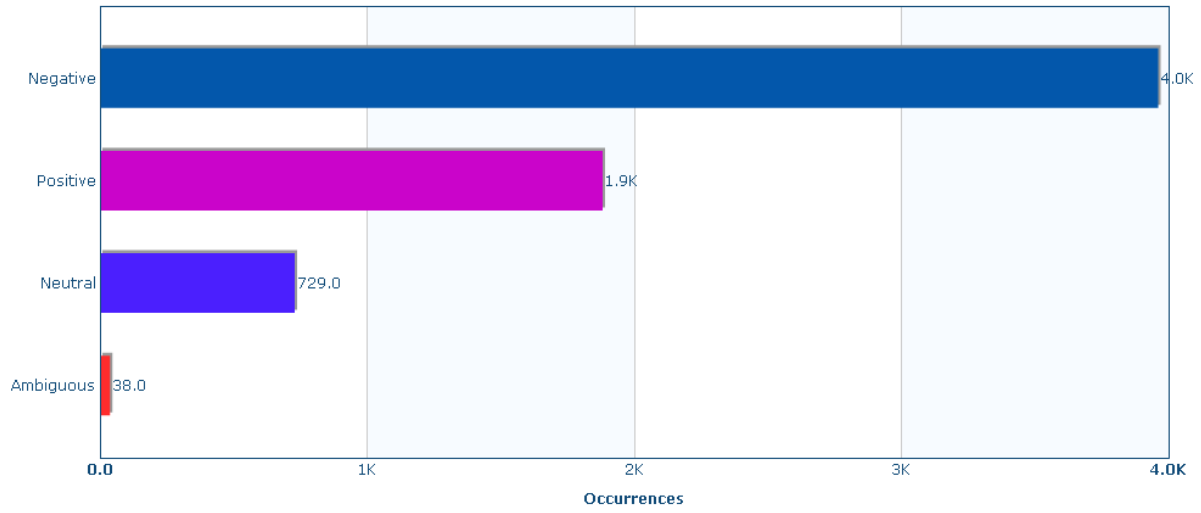

Figure 44 : Nombre d'occurrences par polarité

Globalement, le contenu est négatif avec environ 4000 termes. Mais en comparant avec la figure 45, ces 4000 termes sont rassemblés dans 1200 documents. Cela signifie que la proportion de contenus négatifs sur un document est en moyenne de 3 alors que pour les contenus positifs elle est de 2.

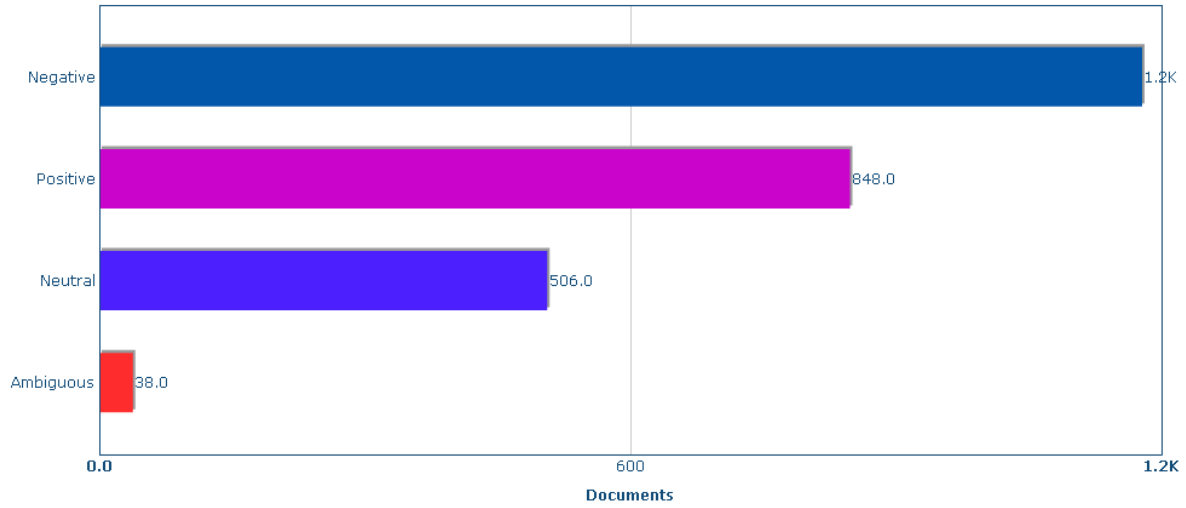

Figure 45 : Nombre de documents par polarité

Le terme « banque » est en général associé à la polarité négative (147 fois dans le corpus). Les relations ne sont pas transposées ici mais le graphique permet d'avoir une vue globale concernant le contenu négatif.

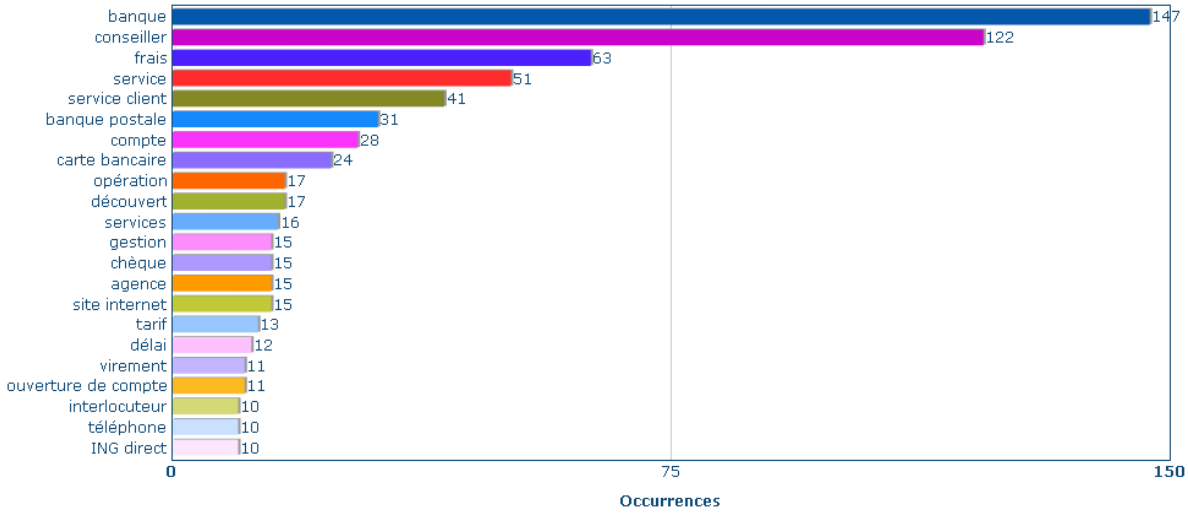

Figure 46 : Nombre d'occurrences des termes à polarité négative

Retrouver la Banque Postale n'est pas une surprise, puisque de nombreux documents sont sur cette banque. Les produits bancaires ne sont pas en seconde position comme sur la figure 42. Les conseillers font également l'objet du mécontentement des clients, toutes banques confondues, ainsi que les frais et les services.

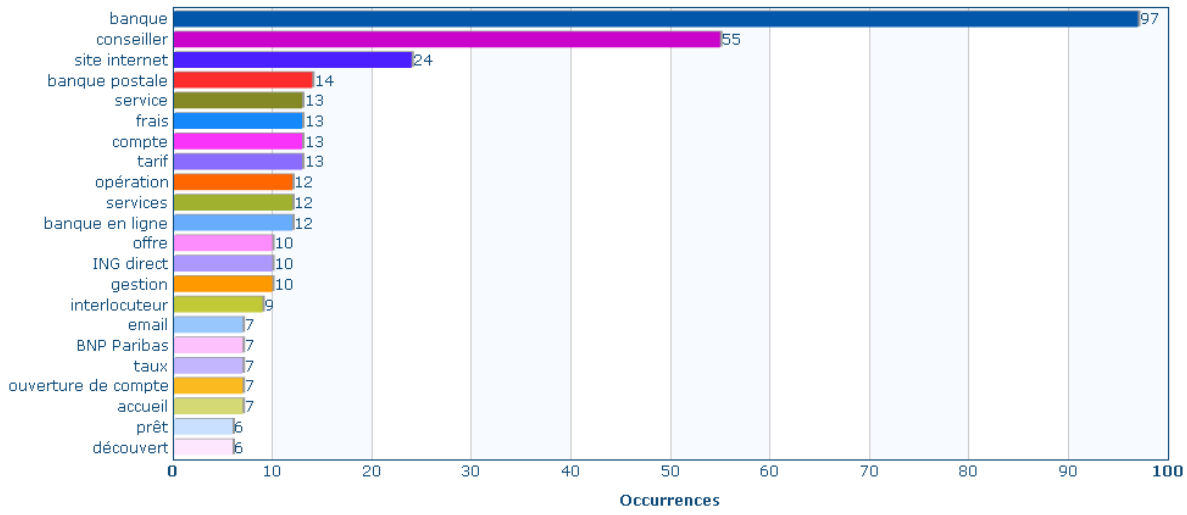

Figure 47 : Nombre d'occurrences des termes à polarité positive

En comparant les figures 46 et 47, « banque » et « conseiller » ont tous deux la même position en terme d'apparition dans le corpus. En observant plus attentivement, beaucoup de concepts se retrouvent dans ces deux graphiques. Cela suppose qu'un client est soit ravi soit mécontent de sa banque, en suivant les mêmes critères.

	Axa Banque	BNP Paribas	CIC	ING Direct	La Banque Postale	LCL	Monabanq	HSBC
Negative	68.2 %	66.5 %	59.0 %	55.2 %	62.0 %	66.7 %	57.1 %	63.4 %
Positive	31.8 %	33.5 %	41.0 %	44.8 %	38.0 %	33.3 %	42.9 %	36.6 %

Tableau 7 : Pourcentages des documents positifs et négatifs pour chaque banque

Le tableau 7 représente une sélection de banques parmi celles qui disposent d'au moins 100 documents afin d'avoir un échantillon avec le plus de données possibles. Pour chaque banque, le tableau affiche le ratio entre documents positifs et négatifs contenus dans le corpus. Aucune banque n'a un pourcentage positif supérieur au négatif. Pour la Banque Postale, les résultats ont l'air de concorder avec les graphiques précédents. Toutefois, cette banque ne cumule pas uniquement des commentaires négatifs, un peu plus d'un tiers des documents contiendrait des opinions positives.

dc:title :
Satisfait

anonymous :
Je suis à la banque postale depuis près de 2 ans et j'en suis très satisfait.

Que ce soit les tarifs des services (comparés aux autres banques) l'écoute, **la confiance ou des taux de prêts**, c'est vraiment bien. Bien sûr il y a bien eu une ou 2 fois un quiproquo avec un(e) téléconseiller mais à bien y regarder parfois on ne se rend pas compte comment on est soi même certains jours(les torts peuvent être partagés) . D'une manière générale, **la bp est aussi très bien** coté téléconseil même si je préfère (quand j'ai le temps) me déplacer pour un entretien avec ma conseillère bp attitrée. **Le site de la banque postale est de mieux en mieux** et opérationnel. **Je regrette absolument pas d'avoir rejoint la banque postale**, une banque de qualité et compétitive.

Figure 48 : Verbatim positif sur la Banque Postale

dc:title :
De + en + nulle

anonymous :
Les services ne sont plus du tout ce qu'ils étaient.

J'ai quitté la caisse d'epargne pour des manquements sévères et ai soldé tous mes comptes pour ouvrir à la banque postale où j'étais relativement satisfaite jusqu'au récent changement de mon conseiller financier qui a été muté ailleurs laissant **la place à une personne pas tout compétente** qui ne renseigne le client que pour lui ponctionner des frais rentables pour la banque mais pas du tout **un placement intéressant pour le client**. **Les avis sont même mensongers du fait** que les placements proposés ne tiennent pas leur promesse de rentabilité. De plus, les frais bancaires sont de + en plus chers, les dates de valeur pour les encaissements de + en + longues. D'autre part, **les investigations dans la vie personnelle des clients sont très désagréables**.

Figure 49 : Verbatim négatif sur la Banque Postale

Les thèmes comme la banque, les conseillers ou les frais ressortent dans les deux documents des figures 48 et 49. Cela présente d'un côté quelqu'un de satisfait et de l'autre, quelqu'un de mécontent. Ces verbatim sont des contenus concrets amenant aux résultats des figures 46 et 47.

Les relations entre une opinion et sa cible sont la base de notre travail. Elles sont également observables à travers un graphique récapitulatif.

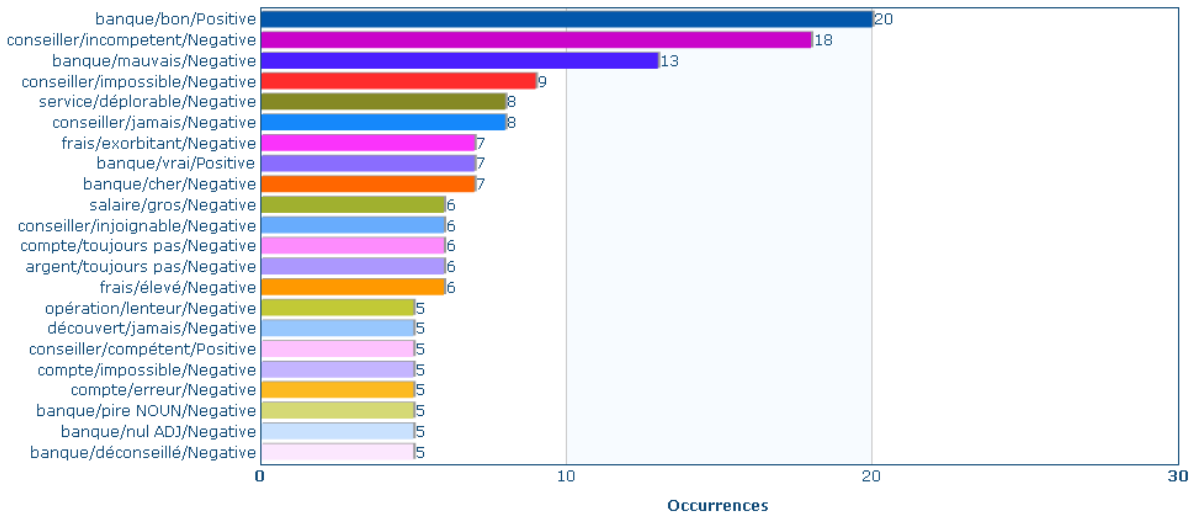

Figure 50 : Relations

La relation majoritaire de la figure 50 est sur le terme « banque », ce qui n'est pas surprenant à la vue des graphiques précédents. Bien que sa polarité soit positive, la plupart des relations sur ce graphique sont négatives, les suivantes étant sur le terme « conseiller » puis encore « banque ». La figure 50 présente les relations dans le détail mais corrèle avec les résultats de la figure 46.

4.4. CONCLUSION

La visualisation des données sous forme graphique permet d'avoir un aperçu global du contenu. Le filtrage est un avantage pour rapidement vérifier le nombre de documents ou d'occurrences, que ce soit sur un terme ou sur toute une catégorie. Croiser plus de deux types de données devient vite incompréhensible car le tableau mélange toutes les informations.

Cette mise en application donne un aperçu rapide et efficace pour la banque qui souhaite obtenir des données sur son établissement, que ce soit pour savoir de quoi parlent ses clients en général ou leur opinion sur un aspect en particulier.

DISCUSSION

La comparaison des résultats non disponibles pour le moment ne signifie pas que c'est le cas pour les deux approches.

D'un point de vue global, la complétude du module existant donne un avantage à cette méthode. Néanmoins, cela ne signifie pas qu'elle est parfaite. Il semble nécessaire de revoir les polarités attribuées au lexique et la recherche des cibles (e.g. figure 26). Il faut surtout se demander s'il est réellement envisageable d'utiliser un lexique générique pour une utilisation spécifique (ici, le domaine bancaire). Les règles d'extraction développées permettent cette focalisation car elles ne sont valables que sur ce domaine.

Avoir une méthode de résolution des anaphores serait un moyen d'extraire davantage de relations. Par exemple, dans l'extrait suivant « Mon conseiller est super. Il est toujours très à l'écoute et facilement joignable ! », l'ensemble porte sur le même conseiller. Mais dans la seconde phrase, « il » n'est pas un terme bancaire. Résoudre cette anaphore en faisant le lien avec « conseiller » apporterait des informations supplémentaires.

Il arrive parfois que les verbatim ne contiennent pas de phrases entières comme « Nulle et chère ». Le domaine permet de savoir de quoi il est question (la banque) alors il pourrait être intéressant d'ajouter une cible implicite pour créer d'autres relations.

Toutes ces propositions n'affirment pas que les résultats seraient meilleurs mais elles pourraient apporter plus d'informations dans l'ensemble.

CONCLUSION

Notre travail constituait à mettre en place un thésaurus du domaine bancaire et effectuer de la fouille d'opinions utilisant ce thésaurus.

La méthode semi-automatique adoptée pour la création du thésaurus nous a permis d'obtenir des résultats de qualité élevés (91,8%).

Nous avons choisi d'étudier la fouille d'opinions à travers deux approches différentes, une avec un module de fouille d'opinions existant et une autre à base de règles que nous avons développées. Notre travail s'est focalisé sur les relations entre cibles et opinions. L'évaluation des relations sur un domaine spécifique entraîne des contraintes comme avoir un lexique le plus complet possible (aussi bien de fouille d'opinions ou bien du domaine). Un lexique seul ne fait pas la qualité d'un module. Les règles d'extractions associées doivent aussi être optimisées en fonction de la langue étudiée. Le module existant utilisé dans notre travail se basent sur des patrons syntaxiques pour chercher la cible mais n'extrait pas la bonne cible dans presque 1 cas sur 2 des erreurs relevées. Malgré ces erreurs, le résultat de qualité global sur le module existant est de 70,1% de F-mesure (avec un rappel et une précision équilibrés) et est meilleur que la base de référence (67%). La comparaison entre les résultats de nos deux approches n'est pas possible car l'approche à base de règles n'a pas permis de fournir des sorties satisfaisantes.

Proposer un module de fouille d'opinion associé à un domaine permet ainsi d'éviter autant que possible les problèmes liés au contexte. L'utilisation d'un module universel impliquerait de prendre en compte tous les domaines pour exclure toute ambiguïté.

Pour améliorer les extractions du module existant, revoir les patrons syntaxiques intégrés afin de réduire le taux de mauvaises cibles trouvées serait une première solution. Reprendre les polarités du lexique d'opinion et même envisager d'introduire des polarités sur les concepts du thésaurus permettrait de diminuer le taux des mauvaises polarités attribuées.

BIBLIOGRAPHIE

[Bloom & al., 2007] Bloom, Kenneth, Navendu Gard, and Shlomo Argamon. "Extracting Appraisal Expressions." pp. 308-315. Rochester, NY: Proceedings of NAACL HLT 2007, 2007.

[Chaturvedi & Chopra, 2014] Chaturvedi, Deepshikha, and Shalu Chopra. "Customers Sentiment on Banks." *International Journal of Computer Applications*, 2014: 6p.

[Dégez & Ménillet, 2002] Dégez, Danièle, and Dominique Ménillet. *Thésauroglossaire des langages documentaires.* Edited by Paris : ADBS éd. Paris: Bulletin des bibliothèques de France (BBF), 2002.

[Francart, 2013] Francart, Thomas. "Ontologie, Thesaurus et Taxonomie sur le web de données." *Sparna.* Décembre 7, 2013. http://blog.sparna.fr/2013/12/07/ontologie-thesaurus-taxonomie-web-de-donnees/ (accessed Juin 2016).

[Gardin, 2009] Gardin, Pierre. "Application de la théorie de l'Appraisal à l'analyse." *MajecSTIC 2009.* Avignon, 2009.

[Liu, 2010] Liu, Bing. "Handbook of Natural Language Processing." Edited by N. Indurkhya and F. J. Damerau, pp. 1-9. Chicago, Illinois: Chapman & Hall/CRC/Taylor & Francis Group, 2010.

[Liu, 2012] Liu, Bing. "Sentiment Analysis and Opinion Mining." pp. 25-28 & pp. 58-62. Morgan & Claypool Publishers, 2012.

[Liu & al., 2005] Liu, Bing, Minqing Hu, and Cheng Junsheng. *Opinion Observer : Analyzing ans Comparing Opinions on the Web.* Document. Illinois, Chicago, 2005.

[Marchand, 2013] Marchand, Morgane. "Fouille d'opinion : ces mots qui changent de polarité selon le domaine.", CEA, LIST & LIMSI-CNRS, Orsay, 2013.

[Martin & White, 2005] Martin, J.R., and P.R.R. White. *The Language of Evaluation: Appraisal in English.* London & New York: Palgrave Macmillan, 2005.

[Moureau, 1973] Moureau, Magdeleine. "Principe et développement d'un thesaurus." pp. 5-29. Bulletin des bibliothèques de France (BBF), 1973.

[Pang & Lee, 2004] Pang, Bo, and Lillian Lee. *A Sentimental Education: Sentiment Analysis Using Subjectivity.* Department of Computer Science, Cornell University, Ithaca, NY: Proceedings of the ACL, 2004.

[Pang & Lee, 2008] Pang, Bo, and Lillian Lee. *Opinion Mining and Sentiment Analysis.* Vol. 2, pp. 38-46. Ithaca, NY: Foundations and Trends® in Information Retrieval, 2008.

[Roberfroid & Dubois, 2012] Roberfroid, Akémi, and Marie-Laurence Dubois. "Un thésaurus : Comment ? Pourquoi ?" *Etopia.* Décembre 2012. www.etopia.be/spip.php?article215

[Ruiz-Martínez & al., 2012] Ruiz-Martínez, Juana María, Rafael Valencia-García, and Francisco García-Sánchez. "Semantic-Based Sentiment analysis in financial news." Facultad de Informática, Universidad de Murcia, Espinardo (Murcia), 2012, 14p.

ANNEXES

Thésaurus du domaine bancaire

- finance
  - billet
  - bourse (1 sous-niveau)
  - capital
  - cotisation
  - dette (1 sous-niveau)
  - devise (1 sous-niveau)
  - fonds
  - frais (1 sous-niveau)
  - hypothèque
  - interdit bancaire
  - montant
  - plafond (1 sous-niveau)
  - solde du compte
  - tarif
  - taux (1 sous-niveau)
- institution
  - agence
  - banque
  - centre financier
  - établissement (2 sous-niveaux)
  - siège
- métier
  - agent

- assureur
- banquier
- chargé de clientèle
- chargé de compte
- chef d'agence
- conseiller (1 sous-niveau)
- courtier (1 sous-niveau)
- directeur (1 sous-niveau)
- gestionnaire de compte
- guichetier
- médiateur
- notaire
- responsable (1 sous-niveau)
- stagiaire
- standardiste

moyen de communication

- courrier (1 sous-niveau)
- email
- fax
- sms
- téléphone (3 sous-niveaux)

opération

- débit (1 sous-niveau)
- dépôt (1 sous-niveau)
- encaissement
- paiement
- prélèvement (1 sous-niveau)

remboursement (1 sous-niveau)

retrait

transfert (1 sous-niveau)

versement

procédure administrative

accord de principe

autorisation de prélèvement

bordereau (1 sous-niveau)

clôture de compte

contrat (1 sous-niveau)

convention de compte

déclaration (1 sous-niveau)

délai (1 sous-niveau)

demande (1 sous-niveau)

dossier (1 sous-niveau)

gestion

iban

justificatif (1 sous-niveau)

numéro de compte

ouverture de compte

préavis

procédure (1 sous-niveau)

rejet (1 sous-niveau)

relance

relevé bancaire

renégociation

rib

- signature (1 sous-niveau)
- type de documents

produit

- action bancaire
- produit assurance (1 sous-niveau)
- produit bancaire (4 sous-niveaux)
- produit financier

qualité de service

- accueil
- geste commercial
- guichet
- interlocuteur
- relation client
- rendez-vous
- service
- suivi client

services

- application bancaire
- borne automatique
- consultation de compte
- déblocage
- découvert (1 sous-niveau)
- distributeur
- gestion en ligne
- intérêt
- procuration
- sécurité (1 sous-niveau)

service (2 sous-niveaux)

site internet

smartphone (1 sous-niveau)

web (1 sous-niveau)

Tableau complet des pourcentages de documents positifs et négatifs pour chaque banque (documents provenant du site de l'internaute)

	Allianz banque	Axa Banque	Banque populaire Bred	BNP Paribas	Caisse d'épargne Ile-de-France	Caisse d'épargne Rhône-Alpes	CIC	Crédit agricole Centre-Est	Crédit agricole Ile-de-France	Crédit du Nord	Crédit mutuel Ile-de-France	Crédit Mutuel Sud-Est	Fortuneo	Groupama banque	HSBC	ING Direct	La Banque Postale	LCL	Monabanq	Société générale
Negative	45.5 %	68.2 %	76.9 %	66.5 %	66.7 %	55.6 %	59.0 %	75.0 %	68.4 %	62.1 %	71.4 %	50.0 %	50.0 %	57.1 %	63.4 %	55.2 %	62.0 %	66.7 %	57.1 %	60.0 %
Positive	54.5 %	31.8 %	23.1 %	33.5 %	33.3 %	44.4 %	41.0 %	25.0 %	31.6 %	37.9 %	28.6 %	50.0 %	50.0 %	42.9 %	36.6 %	44.8 %	38.0 %	33.3 %	42.9 %	40.0 %