Aix-Marseille université

École Doctorale en Mathématiques et Informatique de Marseille

UFR Science

Laboratoire d'Informatique Fondamentale de Marseille (LIF)
Traitement Automatique du Langage Écrit et Parlé (TALEP)

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Informatique

Jérémy TRIONE

Méthodes par abstraction et extraction pour le résumé de conversations orales issues de centres d'appels téléphoniques.

Sophie ROSSET	LIMSI	Rapporteur
Emmanuel MORIN	LS2N	Rapporteur
Géraldine DAMNATI	Orange Labs	Examinateur
Georges LINARES	LIA	Examinateur
Frédéric BECHET	LIF	Directeur de thèse
Benoit FAVRE	LIF	Co-Directeur de thèse

Résumé

Le résumé automatique de document repose généralement sur des méthodes par extraction qui sélectionnent dans le texte des passages pertinents et les juxtaposent pour former un résumé. Ces méthodes sont peu adaptées à la problématique du résumé de conversations orales de part la nature spontanée de celles-ci et l'importance de l'interaction entre les locuteurs. En ne sélectionnant que certains passages, les résumés par extraction ne contiennent qu'un verbatim de ce qui a été dit, et non pas une description synthétique de ce qui s'est passé lors de la conversation. Par exemple, dans le domaine des centres d'appel, il serait souhaitable que les résumés générés renseignent à propos du problème de l'appelant et de comment ce problème a été pris en charge par l'agent ayant traité l'appel. Il n'est pas rare que l'appelant décrive son problème sur plusieurs tours de parole ponctués par des demandes de confirmation ou de reformulation de la part de l'agent, ce qui est difficile à caractériser à l'aide de méthodes extractives lorsque la taille des résumés est fortement contrainte.

Dans un premier temps nous testons et analysons des méthodes de résumés par extraction appliquées à des données orales. Puis nous étudions l'intérêt de l'utilisation de modèles sémantiques dans la tâche de résumé automatique. Enfin nous proposons une méthode de résumé à base de patrons.

Les méthodes de résumé par remplissage de patrons ont montré leur intérêt dans des domaines spécifiques pour le résumé automatique de texte. Dans notre cas, elles permettent de traiter du problème de différence de genre entre les données source (transcriptions de conversations) et la forme des résumés à générer (narration synthétique). Toutefois, elles nécessitent l'écriture manuelle de patrons de résumés et l'annotation manuelle de quantités de données source en concepts à détecter pour remplir ces patrons.

Nos contributions sont les suivantes :

— étude du comportement des méthodes de résumés extractifs sur des données orales ;
— annotation sémantique rapide d'un corpus de conversation orale ;
— l'extraction directe de concepts pertinents à partir des transcriptions pour remplir les patrons, par opposition à une analyse sémantique complète ;
— le transfert des annotations de ces concepts depuis des résumés manuels aux conversations par alignement sémantique, minimisant ainsi le coût d'annotation ;
— la génération dynamique de patrons à partir d'exemples de résumés de référence et des informations détectés dans une conversation ;

— un ensemble d'expériences validant l'approche sur la tâche de génération de synopsis du corpus DECODA, dans le cadre du projet européen SENSEI ("Making Sense of Human - Human Conversation").

Remerciements

Il me sera très difficile de remercier tout le monde car c'est grâce à l'aide de nombreuses personnes que j'ai pu mener cette thèse à son terme.

Je voudrais tout d'abord remercier grandement mes directeurs de thèse, Frédéric Béchet et Benoit Favre, pour leur confiance, leur encadrement, leurs nombreux conseils, et surtout leur soutien dans les moments difficiles. Je remercie également Alexis Nasr d'avoir été là dans les moments de doutes.
Je remercie Sophie Rosset et Emanuel Morin qui ont accepté d'être mes relecteurs.
Je tiens a remercier également Georges Linares pour avoir présidé le jury.
J'exprime évidemment ma gratitude à Géraldine Damnati pour avoir accepté de faire partie de mon jury, et je n'oublie pas non plus Delphine Charlet car sans elles je n'aurais sans doute pas commencé cette thèse.

Je dédie cette thèse à mes parents et ma sœur.

Je tiens aussi à remercier ceux qui ont été là pendant ces 4 ans, qui m'ont apporté leur soutien, leur aide, leur joie, leur gentillesse, leur amitié :

Adrien (alias Coincoin), Bala, Carlos, Elisabeth, Gauthier (alias Chewbi), Hichem, Iskander (alias Baki le BG), Jérémie (alias Taff), Jeremy, José Luis (alias Pépito), Laetitia (alias Bichette), Makki, Manon, Marc (alias Park), Marie, Marie-Hélène, Matthieu, Mickaël (alias Mika), Mokhtar, Olivier, Sébastien D, Sébastien R, Thibault, et tous ceux que j'oublie mais qui sont dans quand même dans mon cœur

Table des matières

Liste des figures

Liste des tableaux

Introduction

L'utilisation de la parole au travers des conversations constitue le moyen de communication le plus naturel et le plus courant, à tel point qu'il tient une place prépondérante dans le paradigme des relations personnelles et professionnelles. L'avènement des nouvelles technologies et le développement des moyens de stockage de l'information permet de conserver une grande partie des flux audios en tout genre, qu'ils proviennent d'émission de radio, d'enregistrement de réunions ou encore de conversations téléphoniques issues de centres d'appels. Ces échanges audios sont très prisés dans le monde de l'entreprise étant donné qu'ils facilitent les relations avec les clients. La circulation des informations entre le client et l'entreprise est ainsi quasi instantanée et les problèmes peuvent être gérés plus rapidement. De façon à répondre à la demande, certaines entreprises se sont spécialisées dans ces échanges, proposant des plateformes de centres d'appels adaptées à chaque domaine.

Sur ces plateformes les appels entrants et sortants peuvent être surveillés en temps réel ou enregistrés pour une révision ultérieure. Le suivi est effectué par des évaluateurs humains sur des échantillons d'appels aléatoires. Leur travail consiste à suivre les indicateurs de qualité des appels et d'efficacité des agents. Ainsi le client de l'entreprise de centre d'appels peut demander un rapport qui peut prendre différentes formes selon la nature de la demande, par exemple, le sujet des appels, les questions que posent les clients, ou encore la capacité de l'agent à répondre. Les technologies d'analyse de langue actuelles n'offrent qu'un support limité. Les analystes de données confrontés à un tel déluge de données doivent être en mesure d'extraire, mais aussi résumer les informations pertinentes à partir de ces données. Par exemple, dans un centre d'appel, des millions de conversations parlées sont traitées quotidiennement pour fournir un soutien aux clients. Cependant, un analyste de centre d'appel cherchant à optimiser certains aspects de l'entreprise ne pourra étudier qu'une fraction de ces données en raison des limites technologiques directement liées au domaine des conversations. En raison de la quantité de données en constante augmentation, seuls de petits corpus sont évalués, ne représentant qu'une infime partie (moins de 1%) des conversations traitées par le centre d'appel.

Les services fournis par les analystes et les évaluateurs humains sont très coûteux, voir même irréalisables en raison de la taille des données ou de la complexité de la tâche. Il est donc indispensable d'offrir des outils de visualisations et de traitements rapides des conversations. Entre autres le résumé de conversation constitue une avancée pour l'utilisateur afin qu'il puisse évaluer rapidement la pertinence d'un document vis-à-vis de l'information recherchée. Les personnes

ciblées par ces outils sont les évaluateurs travaillant dans les centres d'appels, afin de simplifier leur approche au vu de la masse de données disponibles.

Il existe différents types de résumés de conversations, selon leurs cibles et leurs buts. Certains résumés vont s'intéresser à la forme de la conversation, en vérifiant que l'appel s'est bien déroulé et que l'agent à bien suivit les consignes de résolutions de l'appel, ce qui correspond à une forme d'évaluation des agents. Un autre type de résumé va mettre l'accent sur le fond de la conversation, c'est-à-dire sur les informations qui ont pu être échangées durant l'appel. Il s'agit d'un court texte relatant le plus souvent la demande de l'usager d'une part et les démarches de la résolution mises en œuvre d'autre part. Ces résumés sont sujets à de fortes contraintes de taille (entre 5% et 10% selon la taille de la conversation à résumer). Ils doivent être capables de ne retranscrire que les informations importantes de la conversation en une ou deux phrases. Les résumés produits ne doivent pas pour autant rentrer dans la catégorie des titres étant donné qu'ils ne sont pas censés attirer le lecteur, mais plutôt lui donner toutes les informations clés de la conversation. Ils seront appelés synopsis [a] pour la suite notre étude. En effet si on reprend la définition du mot synopsis : "*Bref exposé écrit d'un sujet de film, constituant l'ébauche d'un scénario*" on retrouve les éléments clés qui nous concernent, à savoir *un bref exposé écrit [...] constituant l'ébauche d'un scénario*. Ces synopsis n'ont pas pour but de retranscrire l'intégralité des informations d'une conversation, mais uniquement les éléments importants de celle-ci, ce qui correspondrait au scénario principal d'un film sans prendre en compte les histoires annexes directement liées à ce dernier, mais non indispensables au bon déroulement ou à la compréhension de la trame principale.

Cette thèse a été réalisée au sein du projet SENSEI [b] qui consiste à mettre en place une technologie d'analyse de la conversation capable de :

— Analyser des conversations, à la fois sur leur contenu sémantique, mais aussi sur les dimensions dialogiques et comportementales des participants ;

— Développer des méthodes permettant d'adapter les modèles d'analyse rapidement à la diversité des contenus et des médias véhiculant de nouveaux types de conversations ;

— Générer des rapports de résumés permettant de présenter à un utilisateur, sous une forme synthétique, une collection de conversations entre deux ou plusieurs participants ;

— Évaluer de façon "*écologique*" les technologies développées en concertation

a. Ils ont été créés pour reproduire une tâche naturellement présente dans un milieu professionnel, pour décrire le contenu de conversations et pouvoir retrouver une conversation rapidement. Idéalement ils sont un titre, mais la notion de titre est lourdement chargée et ne correspond exactement aux textes produits.

b. Making Sense of Human - Human Conversation

avec les utilisateurs finaux dans les différents cadres d'études.

Nous nous intéressons à la tâche de résumé automatique appliquée à des conversations parlées provenant de centres d'appels. Le but est de produire un synopsis de conversation lisible qui contient l'ensemble des informations nécessaires à la compréhension de l'appel afin de faciliter la navigation dans un ensemble de conversations. Pour y parvenir, nous sommes confrontés à la nature des données à traiter. Les transcriptions de conversations parlées sont des textes non canoniques dont la syntaxe est rarement proche de celle du français écrit. Le langage parlé introduit aussi des phénomènes qui lui sont propres comme les disfluences, les répétitions ou encore les coupures, qui ne doivent pas apparaître dans le synopsis final. Les principales problématiques auxquelles nous tentons de répondre sont les suivantes :

— Comment résumer une conversation à moindre coût ;
— Comment passer d'un texte non canonique reflétant la langue parlée à un discours rapporté dans le synopsis ;
— Comment identifier les concepts / informations essentielles dans la conversation ;
— Comment rassembler tous ces concepts dans un texte avec de fortes contraintes de taille résumant les échanges de la conversation ;
— Comment évaluer les synopsis ;
— Proposer une approche robuste à l'utilisation de transcriptions automatiques.

Nous nous donnerons comme contrainte d'être le plus indépendants possible du domaine applicatif. Contrairement aux approches utilisant des ontologies spécifiques à un domaine pour pouvoir détecter des concepts dans des textes, nous nous efforcerons d'utiliser des ressources linguistiques génériques telles que FrameNet.

Nous proposons une solution utilisant des traits syntaxiques, dialogiques et sémantiques pour d'identifier les concepts importants d'un document dans le but de remplir des patrons de synopsis générés dynamiquement. La méthode présentée dans cette thèse se déroule en 3 étapes :

1. L'alignement de concepts pertinents entre les synopsis et les transcriptions ;
2. La détection de ces concepts dans les nouvelles conversations ;
3. La réalisation et le remplissage des patrons de synopsis.

L'originalité de la méthode demeure dans le fait d'utiliser un corpus de synopsis pour détecter les concepts importants dans les transcriptions des conversations en identifiant les concepts dans les synopsis directement puis en les reportant dans les conversations associées, constituant ainsi un corpus d'apprentissage pour la phase de détection. Cette approche a pour résultat de diminuer le

risque de mauvaises détections des concepts. La détection des concepts dans une nouvelle conversation se fait alors en entraînant un système à apprendre à reconnaître ces concepts à partir d'un ensemble de traits syntaxiques, dialogiques et sémantiques. Enfin, les patrons qui seront remplis avec les concepts identifiés dans les conversations sont construits à partir de synopsis manuels.

Dans le chapitre 1 notre travail se concentrera sur l'étude des méthodes de résumés déjà existantes, puis dans le chapitre 2 nous testerons et analyserons ces méthodes sur un jeu de données réelles provenant d'un centre d'appel de la RATP afin d'observer l'évolution du comportement des méthodes sur des données orales. Ensuite dans le chapitre 3 nous nous attarderons sur l'intérêt et l'utilisation de la sémantique au service du résumé et plus particulièrement à la compréhension des conversations. Enfin, le chapitre 4 présentera notre approche de résumé automatique par extraction de concepts et recombinaison de patrons.

1. Résumé Automatique de Conversation : état de l'art et problématique de l'étude

Ce chapitre a pour objet de présenter la problématique du résumé de conversations qui est au centre de cette thèse. Nous commencerons par définir les deux concepts clés de *conversation* et de *résumé* avant de donner un état de l'art des principales méthodes ayant été proposées dans la littérature pour répondre à cette problématique. Enfin nous discuterons la notion cruciale d'*évaluation* avant de positionner nos travaux par rapport à l'existant en précisant quelles étaient les questions scientifiques préalables à la conduite de cette thèse.

1.1. Introduction au résumé automatique de documents

Le but d'un système de résumé automatique est de produire un nouveau document condensé à partir d'un ou plusieurs documents sources. Ce document condensé est un résumé informatif contenant les informations dites *importantes* du contenu du ou des documents initiaux. Le résumé ainsi produit peut prendre différentes formes selon la fonction qu'il est censé porter, cela peut être un résumé d'article journalistique, une critique ou encore une bande-annonce de film. La nature des documents utilisés peut être variée qu'ils soient vidéos, audios ou textuels.

Historiquement, le résumé automatique a d'abord été appliqué au texte. La principale approche consistait à extraire des portions de texte, principalement des phrases. En effet, l'approche du résumé par extraction provient d'observations comme celles de LIN 2003 qui montre qu'environ 70% du contenu des résumés textuels produits à la main est extrait depuis des textes existants.

La construction d'un résumé automatique commence par la définition des unités de découpage du document. Il peut s'agir des paragraphes, des phrases, des mots, ou encore des tours de parole dans le cas de document audio. Ensuite la construction du résumé se fait généralement en suivant un schéma composé de 3 grandes étapes (voir la figure 1.1) :

— la représentation des données ;

— la comparaison et la sélection des unités logiques ;

— la génération du résumé.

Figure 1.1. – Étapes de la génération de résumé.

La représentation des données correspond à tous les traits disponibles sur le document source capable de caractériser les phrases. Il peut s'agir de traits lexicaux, syntaxiques, sémantiques, discursifs, tels que les mots, les parties de discours, les dépendances et les rôles syntaxique set sémantiques, les concepts, les entités nommées, les actes de dialogue, les relations rhétoriques. Dans le cadre de résumés audio ou vidéo, d'autres traits liés à la parole (prosodie), à l'accompagnement sonore (jingle) ou encore à l'image peuvent être utilisés.

Une fois les unités logiques représentées à l'aide d'un ensemble de traits spécifiques, celles-ci peuvent être comparées entre elles afin de les classer par ordre d'importance pour le résumé. Cette évaluation de la valeur de l'information peut se faire de différentes façons, de manière purement fréquentielle, par exemple avec les modèles vectoriels avec l'utilisation d'une mesure de similarité cosine, en utilisant des modélisations linguistiques explicites aux niveaux syntaxiques, sémantiques et dialogiques, par des méthodes d'optimisation globale à base de graphe ou de programmation en nombres entiers, ou encore grâce à des méthodes d'apprentissage automatique.

La génération du résumé final se fait alors à partir des phrases sélectionnées dans la phase précédente, soit directement par *sélection* dans l'ensemble des phrases candidates, soit par *compression* ou *abstraction* en modifiant les phrases sélectionnées par rapport à l'objectif visé en terme de résumé. La taille d'un résumé est un facteur important et contraignant. Les résultats obtenus peuvent varier si un résumé peut compter 100, 300 mots ou même plus.

Le résumé par *sélection* ou *extraction* est resté la méthode la plus répandue de par son efficacité et sa simplicité. La sélection parmi les phrases précédemment pondérées peut se faire de façon gloutonne, ou de façon plus précise comme en cherchant une solution optimale étant donné un ensemble de contraintes grâce à l'optimisation globale linéaire en nombre entier et ce jusqu'à atteindre la taille limite imposée du résumé.

En complément de ces méthodes par sélection, d'autres méthodes basées sur de la *compression* ou de la *génération de textes* sont apparues avec LIN et HOVY 2003 ; LE NGUYEN, SHIMAZU, HORIGUCHI et al. 2004 ; KNIGHT et MARCU 2000. Ces approches permettent de se passer de la forme du document initial en n'identifiant que les informations importantes et en les recombinant dans un résumé

final à partir de patron ou de système de génération de texte.

Nous verrons dans ce chapitre certains avantages et inconvénients de ces différentes méthodes dans le cadre du résumé de conversations, en justifiant le choix des méthodes par génération de texte que nous avons fait dans cette étude.

1.2. Conversation et Résumé de Conversation

Le terme "*conversation*" peut être défini de manière très générale comme un échange d'informations entre au moins deux individus portant sur un ou plusieurs sujets précis. Dans cette thèse, nous nous intéresserons aux conversations finalisées, et plus particulièrement à la partie qui concerne l'échange d'informations au sein de la conversation. En effet dans le cadre de la rédaction d'un résumé, la détection des informations importantes et intéressantes est primordiale.

1.2.1. Conversations finalisées

Dans cette étude nous nous intéressons au cas particulier des conversations finalisées (ou "avec but"), en particulier celles provenant des centres d'appels. Ces conversations font intervenir 2 participants ou plus qui ne sont pas au même niveau : l'un est le "demandant", celui qui a besoin d'informations, et l'autre est l'"expert", celui qui est censé renseigner le demandant sur les informations demandées.

Deux types de médias peuvent être utilisés pour de telles conversations :

— le texte, via les conversations se déroulant sur ordinateur (ou smartphone) grâce à des interfaces de discussion en ligne de type "*chat*", type d'interaction étant souvent référencé sous le titre de *Computer-mediated communication (CMC)*.

— la parole, dans le cadre de conversations téléphoniques entre un agent et un utilisateur, au sein de centres d'appels.

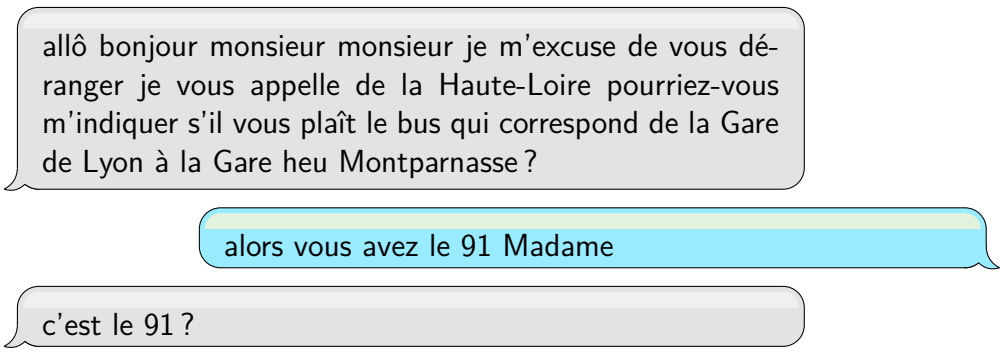
Nous nous intéresserons dans cette étude aux interactions parlées issues de conversations téléphoniques. Même si des similitudes existent entre interactions "chat" et conversations orales, les problèmes posés par chaque média diffèrent, comme étudié dans DAMNATI, GUERRAZ et CHARLET 2016. Dans les deux cas, la langue traitée s'éloigne du langage *canonique* que l'on peut trouver dans des textes journalistiques ou administratifs, cependant les sources de difficultés ne sont pas les mêmes : problèmes d'orthographe, de fautes de frappe, d'agrammaticalité pour les "chats" ; parole spontanée et disfluences, prosodie, qualité de voix et erreurs dues à la transcription automatique pour la parole.

Nous traiterons dans cette thèse des conversations orales, très spontanées et téléphoniques (donc bruités). Les systèmes de transcriptions automatiques ne sont pas très performants dans ce type de conditions. Cependant pour le reste de notre étude des transcriptions manuelles ainsi que des transcriptions automatiques seront utilisées. Les transcriptions automatiques ont été obtenues par le système décrit dans la section 2.3.

D'un point de vue structurel, le découpage d'une conversation ne peut pas se faire grâce à des phrases, puisque la notion même de phrase est difficilement définissable au sein d'une conversation orale, ceci étant dû d'une part à l'absence de ponctuation et d'autre part aux énoncés tronqués, inévitables dans la parole spontanée. Afin de se donner une unité de découpage, le tour de parole sera utilisé pour la suite des travaux. Un tour de parole correspond au laps de temps pendant lequel un interlocuteur s'exprime. Chaque tour de parole sera alors susceptible de contenir une certaine quantité d'information relative à la conversation.

À la différence d'un texte écrit (c'est-à-dire rédigé et réfléchi), une conversation est un échange spontané entre individus, de ce fait les informations au sein de celle-ci peuvent être altérées par de nombreux phénomènes directement liés à la spontanéité. Ce caractère spontané de la conversation introduit du bruit dans les données, il s'agit par exemple de nombreuses répétitions, des changements brusques d'idées, des erreurs de langue et autres (voir la figure 1.2). À cela s'ajoute le fait que ces conversations sont téléphoniques, de ce fait la compréhension devient encore plus bruitée par des problèmes liés à la qualité sonore de la communication. Les conversations étudiées sont tout de même des dialogues finalisés, donc dans un domaine restreint et manipulant des concepts concrets. Un autre aspect facilitant le travail est l'absence de relation sociale entre les appelants, il n'y a pas par exemple de référence à des évènements hors conversations dus à l'historique de leur relation.

Exemple de conversation :



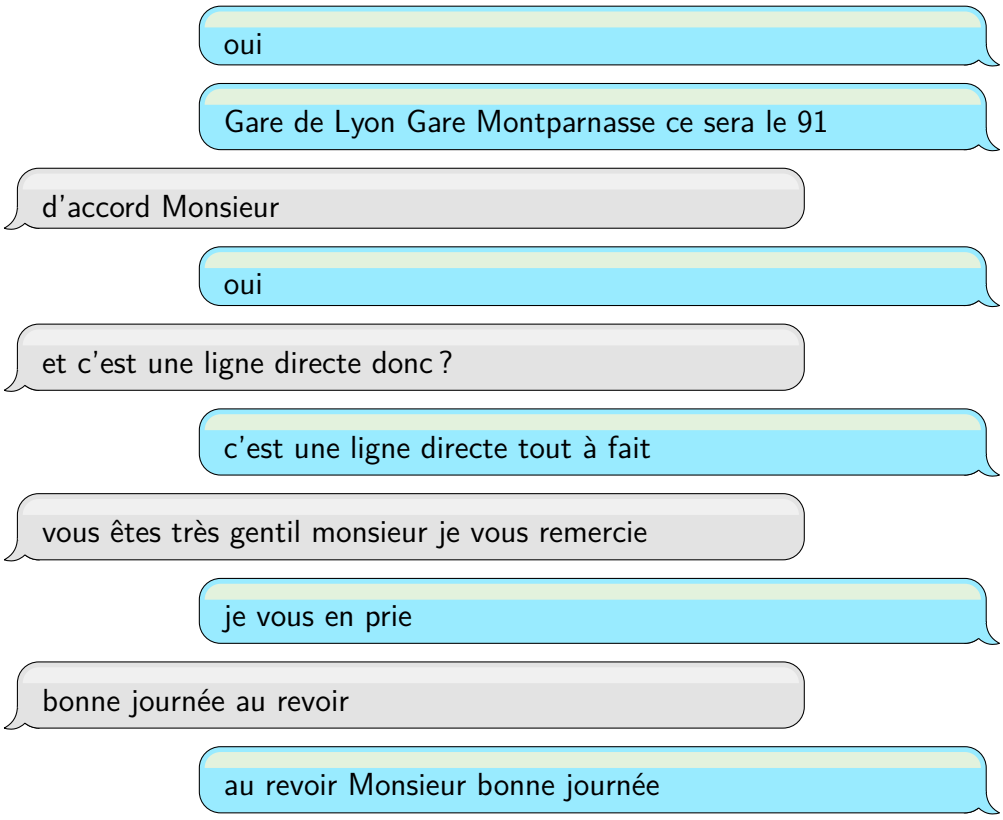


Figure 1.2. – Exemple de conversation.

Dans cette conversation on voit très clairement l'intention de l'appelant dès le premier tour de parole où il expose son problème, à savoir connaître le bus qui se rend d'un endroit précis à un autre. Le problème est rapidement résolu, et mène à un échange de confirmation de l'information "*c'est le 91 - oui [...] ce sera le 91*". L'appel pourrait très bien se terminer ici, mais une nouvelle demande est formulée par l'appelant, puis la conversation se termine par un échange de politesse.

1.2.2. Résumés de conversation

En ce qui concerne la notion de résumé, une définition simple serait : *Forme abrégée du contenu d'un texte, d'un document, d'un film*. Dans le cadre de notre étude, adossée au projet SENSEI décrit en section 2.2, pour une conversation, cette forme abrégée doit contenir l'ensemble des informations clés qui ont été abordées au cours de celle-ci dans une forme extrêmement réduite. Nous appellerons cette forme "*synopsis*" pour la suite de l'étude.

Chaque synopsis doit être capable de retranscrire les informations véhiculées

dans la conversation en un nombre de phrases (ou mots) réduit. Il est important de préciser que la forme des synopsis est textuelle et abstractive, c'est-à-dire que la forme résumée des conversations ne sera pas une nouvelle conversation plus courte ou une sélection des tours de parole les plus pertinents, mais un court texte rappelant les idées abordées à l'instar des synopsis de films. De par ce choix, il est possible que de nouvelles entités apparaissent dans le synopsis par souci de généralisation de l'annotateur (i.e. un itinéraire entre deux rues de Paris, peut être résumé par "itinéraire dans Paris" alors que l'entité "Paris" n'apparaît jamais dans la conversation).

Le principal problème de ces synopsis est directement lié à leur nature. En effet un synopsis est le résultat produit par une personne qui souhaite résumer une conversation, mais cette même conversation pourrait très bien être résumée d'une façon totalement différente par un second individu. Ce phénomène est général à la problématique du résumé, mais il est encore plus marqué dans le cas des résumés de conversations de par leur nature abstractive.

À ce caractère de subjectivité lié à l'individu, on peut ajouter des variantes dans l'orientation d'un résumé. Dans le cadre des centres d'appels on peut identifier deux catégories de synopsis : des synopsis basés sur le contenu sémantique, c'est-à-dire sur le sujet réel de la conversation, et des synopsis basés sur les interactions entre l'usager et le conseiller, cela correspondrait à par exemple privilégier la gestion du temps et l'efficacité du conseiller par rapport au problème même de l'usager. Si ces deux aspects étaient présents dans le projet SENSEI, dans cette étude seuls les synopsis basés sur le contenu sémantique de la conversation seront traités. Ces deux catégories de résumé correspondent à deux cadres applicatifs différents des centres d'appels : déterminer les besoins et les problèmes des usagers dans le premier cas ; améliorer la qualité de service et la formation des opérateurs dans le deuxième cas.

1.3. Représentation des données pour le résumé automatique de conversations

Nous avons vu dans le paragraphe 1.1 que les méthodes généralement utilisées pour le résumé automatique de document pouvaient être présentées selon trois tâches principales : la représentation des données, la comparaison et la sélection d'unités ; la génération du résumé. En commençant par la représentation des données, nous allons présenter dans ce paragraphe quelques méthodes caractéristiques de l'état de l'art du domaine ainsi que leur application au cas particulier du résumé de conversations orales.

Deux grands types de représentation peuvent être utilisés : les représentations sous forme de traits liés à des observations directement extraites du contenu

(audio et textuel) des conversations ; les représentations abstraites provenant d'analyse basée sur des modèles linguistiques, principalement sémantiques et discursifs, des conversations.

1.3.1. Représentations basées sur des observations ou traits

Trois grands types de traits peuvent être utilisés : les traits lexicaux basés sur le contenu textuel des conversations, par exemple les critères fréquentiels tels que TF-IDF ; les traits prosodiques basés sur l'analyse du signal de parole et enfin les traits dialogiques prenant en compte la structure interactive d'une conversation.

1.3.1.1. Traits lexicaux : tf idf

Les travaux sur le résumé automatique ont fait leur apparition dans les années 1958 avec Luhn LUHN 1958 qui base ses expériences sur la fréquence des termes présents dans le texte source afin d'évaluer la pertinence d'une phrase. Cette approche repose sur l'idée qu'une personne aura tendance à répéter certains mots quand elle parle d'un même sujet. La pertinence d'un terme est alors estimée comme étant proportionnelle à sa fréquence dans le document source. Afin d'optimiser la méthode, l'auteur propose de normaliser les termes proches d'un point de vue de l'orthographe afin de s'affranchir de la variabilité des mots, mais aussi l'omission de certains mots (bruits) à l'aide de listes. Les bases du résumé automatique sont alors posées, et cette façon de procéder reste une référence dans le milieu puisqu'une grande partie des systèmes présents est basée sur ce même principe que l'on appelle communément le résumé par extraction.

La fréquence d'un terme n'est cependant pas forcément directement liée à sa pertinence. En effet il existe des cas où certains domaines peuvent partager des termes communs, mais n'apporter que très peu d'informations il s'agit notamment du cas des *mots outils* (i.e *et, ou, il, elle, ...*) qui sont peu discriminants. SPARCK JONES 1972 propose alors de réduire la pertinence de ces termes, et montre que celle-ci est inversement proportionnelle au nombre de documents dans le corpus contenant ce terme. Ainsi le poids d'un terme est calculé comme suit :

$$w_{i,j} = tf_{i,j} \times idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log_2(\frac{N}{n_i}) \tag{1.1}$$

Où $w_{i,j}$ est le poids du terme i dans le document j. $tf_{i,j}$ est la fréquence du terme i dans le document j. $idf_{i,j}$ est la fréquence inverse dans le document, où N, est le nombre de documents dans le corpus et n_i est le nombre de documents dans lequel apparaît le terme i. Le score de la phrase peut alors être calculé à partir de cette mesure, par exemple en sommant les scores de chacun des termes de la phrase.

Il existe des indicateurs supplémentaires pour définir la pertinence d'une phrase dans un document, par exemple sa position vis-à-vis de l'ensemble du document (BAXENDALE 1958), ou encore l'identification de mots-clés (EDMUNDSON 1969).

HOVY et LIN 1998 propose un niveau d'abstraction additionnel en s'aidant des concepts évoqués dans les phrases pour mesurer leur similarité. Par exemple pour le mot "voiture", l'ensemble de ses dérivés de même sens comme "automobile", "volant", "quatre roues" vont avoir la même valeur. Ces liens entre les différents mots peuvent être déterminés en utilisant des bases de données de liens sémantiques comme *WordNet* (MILLER 1995).

1.3.1.2. Traits prosodiques

Un résumé de texte par extraction sélectionne les phrases les plus représentatives pour former un résumé. Un résumé abstractif va identifier des concepts et essayer de reconstruire un résumé à partir de ces informations. Dans les deux cas, les résumés produits s'appuient essentiellement sur des données comme le lexique, la syntaxe, ou encore la position et la structure de l'information, mais le résumé de parole peut aussi tirer parti des sources d'informations supplémentaires contenues dans l'audio. Ces informations peuvent se traduire par l'identification du locuteur, mais aussi par l'analyse de l'information acoustique ou prosodique. La prosodie joue un rôle important dans la communication entre deux humains, elle permet de faire passer des informations non linguistiques comme pouvoir mettre l'accent sur un mot ou une partie de phrase importante, ou encore exprimer une intention ou une émotion particulière. Ce nouveau type d'information permet d'obtenir des informations difficilement identifiables à l'échelle des mots.

L'intégration de ces paramètres prosodiques a été principalement menée sur des meetings ou des situations où le style de parole des locuteurs varie. En effet KAZEMIAN, RUDZICZ, PENN et al. 2008 montrent que dans des domaines comme les émissions journalistiques où le style de parole reste constant, l'utilisation de paramètres prosodiques n'apporte pas de meilleurs résultats et peut même avoir tendance à les dégrader.

1.3.1.3. Traits dialogiques

En complément des informations lexicales et prosodiques, d'autres traits liés à la nature interactive d'une conversation ont montré leur intérêt. Ces traits vont décrire le déroulement de la conversation en caractérisant chaque tour de parole par rapport au locuteur, à la position dans la conversation, à la durée, etc.

MASKEY et HIRSCHBERG 2005 proposent une étude empirique sur l'utilité de l'utilisation de différents types *features*, lexical, structurel, prosodique et dialo-

gique, dans la réalisation de résumés extractifs de journaux télévisés. Leur approche est basée sur un système à base d'apprentissage capable de classifier les segments de phrases en fonction d'un ensemble de *feature* présenté dans le tableau 1.1 ci-après :

Type de *feature*	*feature*
Lexical	- noms propres dans une phrase - nombre de noms propres dans la phrase - nombre de mots dans la phrase courante, précédente et suivante
Prosodique / acoustique	- taux de parole - F0 min, max et mean - F0 range - slope - minDB, maxDB, meanDB - slopeDB - durée de la phrase
Structure	- position de la phrase dans le tour de parole - locuteur courant, précédent et suivant - changement de locuteur - position de la phrase dans l'émission - position du tour de parole dans l'émission
discours	- information présente/nouvelle

Table 1.1. – Liste des *features* utilisées par Maskey et Hirschberg 2005.

Les *features* du discours utilisés (i.e information présente/nouvelle) correspondent à une pondération de l'information dans une phrase donnée. Le score de chaque phrase est calculé selon la formule suivante :

$$[htb]S(i) = \frac{n_i}{d} - \frac{s_i}{t-d} \tag{1.2}$$

Ici, n est le nombre de nouveaux noms dans la phrase i, d est le nombre total de noms uniques dans l'émission ; s_i est le nombre de noms de la phrase i ayant été déjà rencontré dans l'émission ; et t est le nombre total de noms dans l'émission.

L'approche est évaluée par rapport à une *baseline* correspondant aux phrases du début de l'émission jusqu'à atteindre une longueur moyenne de 23% de la longueur totale du document source. MASKEY et HIRSCHBERG 2005 montrent ainsi qu'une combinaison de *features* lexicales, prosodiques, structurelles, dialogiques permet d'obtenir une meilleure classification des phrases susceptibles d'être sé-

lectionnées dans un résumé.

De la même façon MURRAY, RENALS, CARLETTA et al. 2006 montre que l'utilisation de *features* liés à l'activité du locuteur, et au discours peut dépasser les approches purement lexicales utilisées en résumé de texte dans la tache de résumé de conversation sous la forme de *meetings* (corpus ICSI meetings).

On notera que les gains gagnés par l'ajout de ces *features* ne sont pas significativement diminués par un fort taux d'erreur mots initiés par l'utilisation de la reconnaissance automatique de la parole dans les transcriptions automatiques.

1.3.2. Représentations basées sur des analyses linguistiques

Les approches basées sur des analyses linguistiques ont pour but de fournir une représentation abstraite des conversations qui permettent de généraliser plus facilement les traitements ultérieurs de sélection et de génération de résumés en effectuant une certaine abstraction par rapport aux traits liés aux observations présentées dans le paragraphe précédent.
Deux grands types de modèles linguistiques ont été utilisés pour le résumé automatique : les modèles sémantiques et les modèles discursifs.

1.3.2.1. Modèles sémantiques pour le résumé automatique

L'analyse sémantique produit une interprétation en sens à partir des mots et de toutes les informations disponibles dans le document source. Cette représentation permet ainsi de comprendre de façon plus générale ce qui est écrit ou dit selon que l'on travaille sur du texte écrit ou des données orales. La sémantique de par sa nature peut devenir un outil indispensable pour identifier les informations dans un document et donc servir directement pour la production de résumé. Généralement l'utilisation de la sémantique intervient lors de la phase de représentation et sert pour la sélection de phrases.

il existe plusieurs modèles sémantiques par exemple AMR (Abstract Meaning Representation), WordNet ou encore FrameNet (ces modèles seront présentés plus en détail dans le chapitre 3).

BELLARE, Anish Das SARMA, Atish Das SARMA et al. 2004 présentent une approche basée sur l'utilisation d'un modèle sémantique WordNet pour le résumé de texte. Leur approche est basée sur l'utilisation de WordNet dans la production de liens entre différentes parties du document. Ainsi les parties les plus pertinentes et présentes peuvent être extraites pour générer un résumé.

BHARTIYA et SINGH 2014 utilisent également *WordNet* pour étiqueter leur texte en rôles sémantiques afin de donner une représentation en sens du document source. Puis utilisent une méthode de compression de phrases pour générer leurs résumés.

LIU, FLANIGAN, THOMSON et al. 2015 proposent une méthode de résumé de texte basée sur l'utilisation de graphes AMR pour représenter l'information au sein d'un document. Cette approche se décompose en 3 parties :

— Créer des graphes AMR pour chaque phrase ;

— combiner et transformer ces graphes en graphes AMR résumés ;

— générer le résumé à partir de ces derniers.

Ce type de représentation est précise et demande généralement beaucoup d'annotations ou de données qui ne sont pas toujours disponibles notamment pour le français.

HAN, LV, HU et al. 2016 utilisent le modèle sémantique *FrameNet* pour identifier les portions de texte importantes, récupérant ainsi le contenu informatif d'un document. Les informations sémantiques sont exploitées dans un graphe où les phrases sont les sommets et les relations sémantiques correspondent aux arêtes. *FrameNet* est utilisé pour calculer la similarité entre phrases lors de la sélection de ces dernières dans la génération de résumé.

En suivant la présentation de la génération de résumé présentée en introduction, GENEST et LAPALME 2013 donnent un schéma d'utilisation de la sémantique dans les résumés de la façon suivante :

— représentation du document via une analyse sémantique ;

— sélection de contenu par identification des éléments les plus pertinents de l'analyse précédente ;

— génération de texte pour créer une séquence de phrase en linéarisant les éléments sémantiques sélectionnés.

1.3.2.2. Modèles discursifs pour le résumé automatique

L'analyse du discours correspond à l'étude des aspects du langage qui dépassent le niveau des phrases isolées. Lorsqu'il est appliqué aux dialogues, il comprend les phénomènes qui expliquent les interactions linguistiques entre les orateurs, leur compréhension mutuelle et leur partage d'informations par exemple. L'un des principaux objectifs dans ce domaine est l'étude des facteurs qui rendent un discours cohérent. La cohérence peut être expliquée en posant des relations entre des clauses, des phrases ou des actes de dialogue, qui organisent les intentions d'un individu (avec des explications, des élaborations, des contrastes par

exemple) ou expliquent les tours de paroles entre les locuteurs (i.e Réponse à une question, reconnaissance d'une proposition ou une affirmation, correction d'une assertion). Un texte cohérent est un texte dans lequel chaque constituant est relié par une relation de discours. Un certain nombre de théories ont été proposées, pour le texte écrit et le dialogue, qui font différentes hypothèses sur les types de relations (produisant ainsi différentes taxonomies des relations discursives) ou de structures résultantes (une chaîne, un arbre, des graphes plus ou moins contraints qui influencent le processus d'interprétation).

Si l'analyse de la structure du discours pour les textes écrits est maintenant bien établie, il y a beaucoup moins de travaux sur l'application de ces théories aux conversations. La structure du discours dans les dialogues diffère sur plusieurs points de celle du texte ; Elle présente différentes relations entre des types d'actes de parole plus variés (i.e questions et directives, qui sont rares en monologue).

Modélisation RST Parmi les modèles d'annotation discursive qui ont été utilisés pour le résumé automatique, on peut citer la théorie de la Structure Rhétorique (RST) formulée par MANN et THOMPSON 1987. Elle permet de décrire l'organisation d'un texte en établissant des relations entre différentes parties du texte source. À l'origine cette idée a été développée dans le cadre de la génération de texte, mais plus tard MARCU 1997 montre que ces idées peuvent être utilisées dans l'analyse du discours ou encore dans le cadre du résumé automatique.

La structure rhétorique est fondée sur deux types d'unité : les noyaux et les satellites. Les noyaux sont considérés comme les parties de texte les plus importantes tandis que les satellites gravitent autour des noyaux pour les compléter, mais ne peuvent exister sans eux. Ainsi les noyaux contiennent les informations essentielles et les satellites les informations additionnelles et complémentaires de celles du noyau. D'une certaine façon, un texte sans satellite resterait cohérent et lisible, alors qu'un texte essentiellement constitué de satellites serait incompréhensible.

En RST la plus petite unité de texte est l'unité élémentaire de discours (EDU), ainsi les EDU adjacentes sont combinées et liées par des relations rhétoriques pour former une unité plus grande. Chacune de ces unités de plus grande taille participe de manière récursive dans les relations de façon à obtenir une structure arborescente hiérarchique couvrant l'ensemble du texte.

MARCU 1999 valide son approche à base d'arbre de discours dans la génération de résumé au moyen d'une étude psycholinguistique en montrant qu'il existe une forte corrélation entre les noyaux et la structure discursive d'un texte et ce que

les lecteurs perçoivent comme les unités les plus importantes dans le texte.

Modélisation par actes de dialogue Dans le cas des conversations, la capacité à modéliser et détecter automatiquement la structure du discours est un point important dans la compréhension du dialogue. Bien qu'il n'y ait peu de consensus sur la façon exacte de décrire la structure du discours, il existe un certain accord sur un premier niveau d'analyse qui consiste à identifier les actes de dialogue (DA). Un acte de dialogue représente le sens d'un énoncé au niveau de la force illocutoire (SEARLE 1968).

Le tableau 1.2 montre un échantillon du genre de structure du discours. Chaque énonciation reçoit une étiquette DA unique (montré dans la colonne 2), tirée d'un ensemble bien défini. Ainsi, les DA peuvent être considérés comme un jeu d'étiquettes qui classe les énoncés selon une combinaison de critères pragmatiques, sémantiques et syntaxiques. Généralement ces catégories de DA sont définies de manière à être pertinentes pour une application particulière. Il existe cependant des systèmes d'étiquetage en DA indépendants du domaine, tels qui l'architecture DAMSL (CORE et ALLEN 1997).

Locuteur	Acte de dialogue	Énoncé
A	yes-no-question	So do you go to college right now ?
A	Abandonned	Are yo-,
B	Yes-answer	Yeah,
B	Statement	it's my last year.
A	Declarative-question	You're a, so you're a senior now,
B	yes-answer	Yeah,
B	Statement	I'm working on my projects trying to graduate
A	Appreciation	Oh, good for you.
B	Backchannel	Yeah.
A	Appreciation	Taht's great.

Table 1.2. – Exemple d'annotation en acte de dialogue (Stolcke, Ries, Coccaro et al. 2000).

Tout en ne constituant pas une compréhension du dialogue dans un sens profond, l'annotation en DA semble être clairement utile à diverses applications. Par exemple, un résumeur de réunion doit garder une trace de ce qui a été dit et à qui, ou encore un agent de centre d'appel doit savoir si on lui a posé une question ou demandé de faire quelque chose.

Un des avantages de l'annotation en actes de dialogue est d'être relativement indépendante d'un contexte applicatif donné, voire même d'une langue donnée.

Ainsi ROSSET, TRIBOUT et LAMEL 2008 proposent une méthode d'annotation en actes de dialogue de conversations parlées en français et en anglais. Pour valider la généralité de la méthode, le modèle est appris sur un corpus français puis testé sur un corpus anglais d'une part, puis testé sur un second corpus français provenant d'un domaine différent. ROSSET, TRIBOUT et LAMEL 2008 obtiennent un taux de détection en actes de dialogue de 86% sur un corpus du même domaine et 77% lorsque la langue ou le domaine change.

De nombreuses méthodes de classification automatique peuvent être employées pour l'annotation automatique en actes de dialogue. Par exemple SALIM, HERNANDEZ et MORIN 2016 comparent des approches de classification automatique des actes de dialogue pour des corpus de conversation écrites en ligne (i.e *chats, forum, courriels*) en se basant sur des traits classiquement utilisés pour les textes écrits (n-grammes, racines, lemmes, étiquettes morpho-syntaxiques informations contextuelles).

Concernant le résumé automatique, ZECHNER 2002 utilise une annotation en DA dans leur système de résumé de conversation couplé à un ensemble d'annotations syntaxiques permettant de classer et sélectionner les phrases les plus pertinentes afin de produire un résumé de dialogue. On notera que les résumés produits ici sont de la même forme que les dialogues sources, il s'agit d'un dialogue raccourci ne contenant que les tours de paroles essentiels à la compréhension globale de la conversation.

Il était prévu de disposer d'annotation en actes de dialogue dans les corpus utilisés pour cette thèse, malheureusement ces annotations n'étaient pas été disponibles lors du développement des systèmes de résumé présentés dans ce document. C'est pour cette raison que les seules informations liées à la structure des dialogues qui sont utilisées sont les observations dialogiques présentées au paragraphe précédent.

1.4. Comparaison et sélection d'unités pour le résumé automatique de conversations

Nous nous intéresserons dans ce paragraphe à plusieurs méthodes ayant été proposées pour la comparaison et la sélection d'unités pour le résumé automatique de document. Même si ces méthodes ont été utilisées principalement pour des systèmes de résumé automatique par extraction, elles peuvent être vues plus généralement comme des méthodes de sélection d'unités pertinentes pouvant servir de base à la production d'un résumé par abstraction ou compression.

1.4.1. Maximal Marginal Relevance (MMR)

Lorsque l'on parle de résumé par extraction, cela signifie de sélectionner une phrase dans un document source en fonction de sa signification. Les phrases d'un document peuvent alors être redondantes entre elles ou complémentaires. CARBONELL et GOLDSTEIN 1998 proposent en 1998 de construire un résumé en prenant en compte la pertinence des phrases tout en limitant leur redondance dans le résumé final.

L'algorithme de Maximal Marginal Relevance (MMR) est un algorithme glouton qui vise à réordonner les phrases d'un document en fonction de leur pertinence vis-à-vis du document dans son ensemble et de leur redondance en fonction des phrases déjà sélectionnées dans le résumé. À chaque itération, l'algorithme détermine la phrase S_i la plus proche du document D, tout en étant la plus éloignée des phrases S_j déjà présentes dans le résumé. La phrase S_i ainsi sélectionnée est ajoutée aux phrases S_j du résumé. L'algorithme s'arrête lorsqu'une condition est remplie, cela peut être un nombre de phrases, un nombre de mots ou un taux de compression du document source.

Le score attribué à une phrase S_i en fonction de son importance et de sa redondance par rapport aux phrases déjà sélectionnées est calculé comme suit :

$$MMR(S_i) = \lambda \times Sim_1(Si, D) - (1 - \lambda) \times Sim_2(S_i, S_j) \tag{1.3}$$

Où $Sim_1()$ et $Sim_2()$ sont des fonctions de similarité cosine, qui ont fait leurs preuves dans le domaine de la recherche dans les documents. Cependant toutes autres mesures de similarité peuvent être utilisées pour ce problème. λ est un hyper paramètre permettant de jouer sur l'influence de la redondance entre le document et le résumé, généralement il est défini empiriquement.

1.4.2. Graphes

MIHALCEA et TARAU 2004 proposent une méthode d'extraction de phrases basée sur des graphes, qui consiste à identifier des segments de texte les plus populaires dans un graphe. L'idée d'utiliser des graphes peut se justifier par la nature des documents traités. En effet les conversations parlées possèdent un certain degré de structure pouvant permettre de générer des graphes et ainsi identifier les tours de paroles les plus pertinents.

Cette approche permet de définir l'importance d'un sommet du graphe par rapport à l'ensemble du graphe en parcourant ce dernier de façon récursive. De ce fait, le document est représenté par un graphe de phrases liées entre elles par des calculs de similarité. Les phrases sont ensuite sélectionnées dans le graphe en fonction des poids des sommets afin d'être extraites pour le résumé. Ainsi

cela permet de simuler une marche aléatoire sur les sommets en utilisant les similarités entre phrases normalisées comme probabilité de passer d'un sommet à un autre.

Un arc entre deux nœuds symbolise la proximité lexicale entre deux phrases. Si une phrase S_i est une succession de mots $w_1^i, w_2^i, ..., w_n^i$, la similarité entre deux phrases est donnée comme suit :

$$Sim(S_i, S_j) = \frac{|S_i \cap S_j|}{\log |S_i| + \log |S_j|} \tag{1.4}$$

où S_i et S_j sont les ensembles de mots des phrases i et j

D'autres mesures de similarité peuvent aussi être utilisées comme la similarité cosine, la plus longue sous séquence commune, etc.

Le résultat est un graphe fortement connecté où chaque arc est pondéré, indiquant ainsi la force de connexion entre les différentes paires de phrases du document. À partir de la pondération des arcs, une pondération des sommets (i.e des phrases) peut être déduite à l'aide de la fonction de pondération suivante :

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \tag{1.5}$$

Où V_i et V_j sont deux sommets du graphe et w_{ij} correspond au poids entre ces deux sommets. d est un hyper paramètre pouvant prendre une valeur comprise entre 0 et 1 (généralement fixé à 0.85 comme présenté dans PAGE, BRIN, MOTWANI et al. 1999). Pour un sommet V_i donné, $In(V_i)$ est l'ensemble des sommets pointant vers ses prédécesseurs et $Out(V_i)$ est l'ensemble des sommets pointant vers ses successeurs.

1.4.3. Optimisation globale linéaire en nombres entiers (ILP)

La sélection des phrases candidates vue jusqu'à présent se fait de façon gloutonne en fonction du score qui lui a été attribué. Ces méthodes consistent à ajouter des phrases issues du texte source sans jamais remettre en question les phrases préalablement sélectionnées. En effet une fois qu'une phrase est sélectionnée pour faire partie du résumé, elle est exclue du processus de recherche pour les suivantes. La sélection de phrases de plus hauts scores n'est jamais remise en cause une fois qu'elle est sélectionnée, et ce même si de nouvelles phrases permettent de mieux couvrir des informations déjà présentes dans les phrases sélectionnées comme le montre la figure 1.3 où la phrase couvre l'ensemble des informations alors que la phrase 1 aura été sélectionnée par le système.

Dans GILLICK et FAVRE 2009, les auteurs proposent de palier à ce problème en

(1)	The device should be white
(2)	The device should be round
(3)	The device should be round and white

Figure 1.3. – Exemple de problème lié à la sélection gloutonne de phrases (Riedhammer, Favre et Hakkani-Tür 2010)

estimant de façon globale la pertinence et la redondance du document source en choisissant d'utiliser la programmation linéaire en nombre entier. La programmation linéaire en nombre entier cherche à maximiser le résultat de la fonction objective en fonction de certaines contraintes de longueur, ce qui permet de déterminer dans l'espace des solutions le résumé optimal. Ils considèrent que chaque phrase est constituée de concepts. Les phrases sont ainsi construites pour que la pertinence de celles-ci soit directement liée aux concepts qui la composent. Ces concepts correspondent à des informations porteuses de sens (par exemple la raison d'un appel, l'avis d'un participant, etc.). Les concepts peuvent être exprimés dans plusieurs phrases, mais leur valeur ne doit pas être comptée en double dans le résumé, cela permet alors de réduire la redondance.GILLICK et FAVRE 2009 se restreignent à des séquences de mots plus courtes par exemple de simples mots, des entités nommées ou des sous-arbres syntaxiques.

Cependant, il est nécessaire de nettoyer le document source des éléments bruités comme les termes fortement redondant et peu porteur de sens (exemple : "en fait"), ou les marqueurs de discours dans le cadre de conversations (ex : "hein ?", "vous savez ?", "oui oui oui", ...).

Cette méthode permet de maximiser la fonction objective en fonction d'une contrainte de longueur donnée. Les contraintes de cohérence liées à la sélection des concepts sont de telles sortes que lorsqu'une phrase est sélectionnée, tous ses concepts associés le sont aussi.

Ainsi si on considère c_i la présence d'un concept i dans le résumé et s_j la présence de la phrase j dans le résumé. Chaque concept unique peut être présent dans une ou plusieurs phrases, et de la même façon chaque phrase peut contenir un ou plusieurs concepts. L'occurrence du concept i dans la phrase j est notée par o_{ij}. Le score du résumé est alors exprimé comme la somme des poids de w_i du concept i dans le résumé. La taille du résumé est limitée par la variable L sur la somme des longueurs l_j des phrases qui le composent. La recherche d'un résumé peut alors être exprimée comme un problème d'ILP défini ci-après :

$$\text{Maximiser : } \sum_i w_i c_i$$

$$\text{Sous la contrainte de : } \sum_j l_j s_j \leq L$$

$$s_j o_{ij} \leq c_i \qquad \forall i, j$$

$$\sum_j s_j Occ_{ij} \geq c_i \qquad \forall i$$

$$c_i \in \{0, 1\} \qquad \forall i$$

$$s_j \in \{0, 1\} \qquad \forall j$$

Avec i le nombre de concepts du document, et j le nombre de phrases du document.

1.4.4. Méthode par apprentissage

KUPIEC, PEDERSEN et CHEN 1995 tout d'abord puis d'autres comme CHUANG et YANG 2000 proposent une approche par apprentissage prédisant la présence d'une phrase ou non dans le résumé final en fonction d'un ensemble de traits.

Tout d'abord les phrases du document initial sont découpées en segments de façon à être capables d'identifier deux informations différentes dans la même phrase. Par exemple la phrase "I love playing tennis because it is exciting" peut être segmentée en "I love playing tennis" et "because it is exciting". La segmentation ainsi produite regroupe les unités basiques utilisées pour la génération du résumé.

Afin de pouvoir apprendre à identifier un segment important, ces derniers doivent être représentés par un ensemble de traits. Les traits utilisés sont divisés en deux catégories, structurelles et non structurelles. Les premiers sont directement liés à la structure du texte (i.e les relations rhétoriques) et les secondes ne le sont pas (i.e les mots du titre). Au final les segments sont représentés par un ensemble de 23 traits, parmi ceux-ci se retrouvent la position du segment dans le texte, la fréquence moyenne de chacun des termes utilisés dans le segment, ou encore le nombre de mots du titre apparaissant dans le segment.

Le but est de sélectionner un nombre limité de segments qui vont apparaître dans le résumé et qui sont capables de représenter le texte original. À l'aide des traits précédemment générés, un algorithme d'apprentissage peut être appliqué pour entraîner le système à résumer un texte. Il s'agit là d'un apprentissage supervisé où le système apprend à reconnaître les traits sur un corpus d'appren-

tissage, et permet ainsi de prédire l'importance d'un segment.

1.5. Génération automatique de résumé de conversations

Une fois les unités (phrases ou tour de parole) sélectionnées et évaluées, la dernière phase dans le processus de résumé automatique consiste à générer le résumé final en intégrant les contraintes applicatives (type et taille des résumés). Si pour les systèmes par extraction cette phase est bien souvent réduite à la simple sélection des n unités les plus pertinentes selon les métriques définies dans la phase précédente, ce n'est pas le cas pour les résumés par abstraction qui nécessite une réelle phase de *génération de texte* pour produire le résumé. C'est ce type de méthode qui permet de répondre aux critères des résumés de conversation de type *synopsis* étudié dans cette thèse.

1.5.1. Résumé par sélection et fusion de phrase

Les méthodes extractives sélectionnent des phrases directement issues du document source afin de créer un court résumé. Cependant cette approche peut mener à de la redondance entre les informations dans les phrases sélectionnées. C'est pourquoi la méthode compressive a été introduite, elle consiste à supprimer une partie des phrases sélectionnées pour limiter ce phénomène. CHEUNG et PENN 2013 ; JING et K. R. MCKEOWN 2000 ont montré que les résumés écrits manuellement sont plus abstractifs, ce qui peut se traduire par des fusions de phrases et/ou de concepts identifiés dans le document source.

BING, LI, LIAO et al. 2015 proposent une méthode de résumé abstractive basée sur la construction de nouvelles phrases à partir d'un ensemble d'unités syntaxiques plus fine appelée phrases nominales / verbales. Leur approche s'appuie sur la construction d'un groupe de concepts ou de faits représentés par des groupes de phrases issus du document source. Les nouvelles phrases produisant le résumé sont ensuite générées en sélectionnant et fusionnant les phrases porteuses d'information tout en satisfaisant les contraintes de constructions des phrases du résumé final.

Les phrases du document source sont décomposées en un ensemble de phrases nominales (NPs) et de phrases verbales (VPs) provenant de verbes objets et symbolisant respectivement de potentiels concepts et faits clés. Ces ensembles de phrases constituent alors les éléments de base à la génération de nouvelles phrases. Un score d'importance est ensuite calculé pour chaque phrase en fonction de la position de l'information. La construction de nouvelles phrases pour le résumé est formulée comme un problème d'optimisation capable de générer un

ensemble de phrases simultanément. Chaque nouvelle phrase est alors composée d'un NP et au moins un VP, sous la contrainte que NP et VP proviennent d'une phrase source différente. Dans le processus de génération de nouvelles phrases, des relations de compatibilités entre NP et VP ainsi que des contraintes sur les résumés sont appliquées conjointement.

1.5.2. Résumé à base de patrons

SAGGION et LAPALME 2000 proposent une méthode de résumé de texte produisant des *résumés* informatifs pour des papiers techniques. Leur approche est basée sur une identification de l'information ainsi que du sujet du document initial, puis un résumé est généré par la suite utilisant un ensemble de patrons prédéfinis en fonction des éléments précédemment identifiés.

Plus tard, OYA, MEHDAD et R. NG 2014 présentent une méthode de résumé à base de patrons. L'objectif est de produire des résumés informatifs, c'est-à-dire à la fois lisible (i.e grammaticalement corrects) et contenant les informations importantes du document à résumer. Le corpus utilisé pour valider leur approche est un corpus de conversations de réunions. L'approche se divise en deux parties, une première générant des patrons à partir de résumés écrits manuellement, et une seconde partie permettant d'extraire les phrases importantes avec lesquelles les patrons seront remplis.

La phase de génération de patron se fait à partir des résumés humains. Ceux-ci sont sélectionnés en fonction de leur nature (ici les phrases contenant un verbe d'action et dont le sujet de ce verbe est un participant de la réunion), puis une annotation en hyperonyme est menée afin de généraliser la phrase extraite (cf figure 1.4).

a	The project manager goes through the minutes of the last meeting.
b	the Marketing Expert discussed his marketing strategy for the project
a	[speaker] goes through [evidence.n.02] of the last meeting.
b	[speaker] discussed [content.n.05] for [act.n.02]

Figure 1.4. – Exemples d'annotation en hyperonyme pour la génération de patrons

Ces patrons sont ensuite regroupés en fonction de la similarité (en terme de sens) des verbes qui les composent, des classes de *verbes racines* sont ainsi créées. Finalement les patrons sont convertis en un graphe connecté où chaque nœud représente un verbe racine, et chaque arête représente le score montrant la similarité entre deux mots en terme de sens.

Participants dominants
Project Manager (speaker)
Industrial Designer (speaker)
Concepts de hauts scores (hyperonymes)
the whole look (appearance.n.01)
the company logo (symbol.n.01)
the product (artifact.n.01)
the outside (region.n.01)
electronics (content.n.05)
the fashion (manner.n.01)

Table 1.3. – Exemples de participants dominants et concepts de hauts scores extraits après la segmentation en thème.

La seconde phase commence par une segmentation en thème au sein des réunions. Pour chaque thème les phrases les plus importantes sont extraites de la même façon que pour la génération de patron, avec une annotation en hyperonyme et un système de score de similarité. La redondance étant gérée en supprimant toutes les phrases étant des sous-ensembles d'autres. Il en résulte un ensemble de participants dominants et de concepts ayant obtenu de hauts scores comme le présente le tableau 1.3.

GENEST et LAPALME 2013 proposent avec leurs schéma d'abstraction un modèle de génération de résumé qui peut inclure des variables à partir des éléments pertinents du document source. Le modèle génère alors une sortie textuelle lorsque le schéma d'abstraction est identifié. Par exemple, un modèle de génération simple lié au modèle suivant : [marcheur(X) marche lieu(Y)], qui pourrait être instancié par : John marche dans le parc. On peut remarquer que cette représentation ressemble fortement à celle que propose *FrameNet*.

1.6. Conclusion

Nous avons vu dans ce chapitre les différentes méthodes qui ont pu être employées pour la tâche de résumé de document, et en particulier de conversations. Au niveau de la représentation des données, en complément aux traits provenant d'observations lexicales et dialogiques, nous étudierons dans cette thèse l'apport de modèles linguistiques génériques, en particulier sémantiques, pour permettre la sélection des informations importantes provenant du document source.

Concernant la comparaison et la sélection d'unités pour le résumé, nous comparerons empiriquement, sur les données du projet SENSEI, différentes approches. Tout d'abord en implémentant plusieurs méthodes standard utilisées habituelle-

ment pour le résumé de documents textuels, puis en utilisant des méthodes à base d'apprentissage automatique grâce aux corpus de données disponibles.

Pour la génération des résumés, nous étudierons particulièrement les méthodes abstractives à base de patrons, tout en les comparant aux méthodes par sélection ou extraction classiques. En effet les limites des méthodes par extraction sont directement liées à la nature des documents sources. Étant donné que dans ce type de méthodes le résumé est composé essentiellement de phrases provenant de ce dernier, il est impossible de changer de style syntaxique, ce qui ne correspond pas aux types de résumés qu'un humain produirait. En effet dans le cadre des conversations, il est souvent préférable d'avoir un texte court au style indirect relatant des évènements qui ont eu lieu pendant la conversation plutôt qu'une version raccourcie de celle-ci. De plus l'extraction d'informations dans les conversations ne permet pas de synthétiser des évènements se déroulant sur plusieurs tours de parole et peut ainsi passer à côté de certains points qui pourraient être importants au sein de la conversation.

L'alternative que nous proposons est d'extraire automatiquement les informations nécessaires à la production d'un résumé sans prendre toute la phrase ou le segment de texte qui lui est associé. Le type de méthode abstractive que nous proposons d'étudier est le résumé par remplissage de patrons. Ces méthodes ne sont pas nouvelles, mais elles sont généralement basées sur un important travail manuel pour déterminer la forme des patrons et les entités à récupérer dans les documents cibles pour les remplir. Les contributions de cette thèse portent sur des méthodes permettant d'automatiser dans une certaine mesure ces tâches en se basant à la fois sur des représentations linguistiques génériques et aussi sur des méthodes d'apprentissage utilisant un corpus de résumé pour constituer les patrons. Notre méthode se veut donc une méthode hybride entre méthodes par extraction et méthode par abstraction : l'extraction de phrases a lieu non pas dans les documents sources, mais dans les corpus de résumés, l'abstraction consiste à généraliser les phrases sélectionnées grâce aux entités trouvées dans les documents sources.

2.Cadre de l'étude et expériences préliminaires

2.1. Introduction

Nous allons étudier dans ce chapitre la problématique du résumé de conversation dans le contexte applicatif de cette thèse, au sein du projet européen SENSEI.

En partant de ce qui existe dans la littérature pour le résumé automatique de textes, et en se basant sur quelques heuristiques du dialogue concernant la position de l'information au sein d'une conversation issue d'un centre d'appel, des méthodes classiques utilisées en résumé automatique peuvent être exploitées sur des conversations téléphoniques. Les conversations utilisées sont issues du corpus du projet DECODA. Ce sont des conversations spontanées, provenant de cas concrets et réels (c'est-à-dire non jouées par des acteurs) entre des utilisateurs de la RATP et les conseillers téléphoniques.

Dans un premier temps nous présenterons les conversations et les résumés que nous considérerons dans cette étude, puis nous donnerons une description plus détaillée du corpus utilisé et enfin nous présenterons des résultats de référence en appliquant des méthodes classiques de résumé automatique. Une description des différentes méthodes d'évaluation des résumés sera également présente pour permettre de comprendre comment interpréter la comparaison de systèmes différents.

2.2. Projet SENSEI

2.2.1. Présentation

L'interaction conversationnelle est le paradigme le plus naturel et persistant pour les relations personnelles, sociales et commerciales. De larges quantités de données de ce type sont disponibles au sein de nombreuses entreprises, mais les technologies d'analyse de langue actuelles n'offrent qu'un support limité. Les analystes de données confrontés à un tel déluge de données doivent être en mesure d'extraire, mais aussi résumer les informations pertinentes à partir de ces données. Par exemple, dans un centre d'appel, des millions de conversations parlées sont traitées quotidiennement pour répondre aux besoins des clients. Cependant, un analyste de centre d'appel cherchant à optimiser certains aspects

de l'entreprise ne pourra étudier qu'une fraction de ces données en raison des limites technologiques directement liées au domaine des conversations.

Des problèmes similaires limitent l'analyse des fils de conversations sur les plateformes de réseaux sociaux, un nouveau type de conversation multipartite dans laquelle des centaines de millions de publications de blogs ou commentaires sont générés (e.g Twitter, site web journalistique). Un journaliste souhaitant prendre part à un fil de conversation sera rapidement submergé par la quantité de données produites.

Ces deux types de conversation ont un impact limité sur les auditeurs cibles du fait de leur volume trop important, de leur vitesse de production et de leur diversité (médias, style, contexte social). Le projet SENSEI consiste alors à transmettre une technologie d'analyse de conversation capable de :

— Analyser des conversations, à la fois sur leur contenu sémantique, mais aussi sur les dimensions dialogiques et comportementales des participants ;

— Développer des méthodes permettant d'adapter les modèles d'analyse rapidement à la diversité des contenus et des médias véhiculant de nouveaux types de conversations ;

— Générer des rapports de résumé permettant de présenter à un utilisateur, sous une forme synthétique, une collection de conversations entre deux ou plusieurs participants ;

— Évaluer de façon "*écologique*" les technologies développées en concertation avec les utilisateurs finaux dans les différents cadres d'études.

2.2.2. Le cas des centres d'appels

L'étude de cas des centres d'appels propose de nombreux défis scientifiques et technologiques en terme de compréhension de la langue dans un contexte concret.

Dans les centres d'appels externalisés, les grandes entreprises *sous-traitent* leurs services de relations clients dans les centres d'appels. Les appels entrants et sortants peuvent être surveillés en temps réel ou enregistrés pour une révision ultérieure. Le suivi est effectué par des évaluateurs humains sur de petits échantillons d'appels aléatoires (moins de 1%). Leur travail consiste à suivre les indicateurs de qualité des appels et d'efficacité des agents. Ainsi le client de l'entreprise de centre d'appels peut demander un rapport qui peut prendre différentes formes selon la nature de la demande, comme par exemple, le sujet des appels, les questions que posent les clients, ou encore la capacité de l'agent à répondre. Les services fournis par les analystes et les évaluateurs humains sont très coûteux, voir même irréalisables en raison de la quantité de données ou

de la complexité de la tâche. C'est pourquoi le projet SENSEI propose un ensemble d'analyses destinées aux professionnels travaillant dans les centres d'appels. Ainsi selon la cible de l'évaluation (par exemple le contrôle qualité des appels, l'identification de sujet, ou encore l'évaluation des besoins de formation des agents), ces derniers pourront profiter des différentes catégories de résumés et de rapports générés par les systèmes du projet SENSEI.

Dans ce contexte de développement technologique, produire un résultat directement exploitable par l'industrie est un défi supplémentaire. Parmi les tâches décrites plus haut, nous nous intéresserons principalement à la génération de résumés. Le défi ici est donc d'être capable de produire des résumés de qualité tant sur le fond que sur la forme, ou autrement dit, autant lisibles qu'informatifs.

Nous allons décrire dans le paragraphe suivant le cadre expérimental que nous avons utilisé dans cette thèse à travers le corpus de conversations RATP-DECODA.

2.3. Corpus RATP DECODA

2.3.1. Description

Dans le cadre de cette étude, des conversations issues du corpus RATP-DECODA (BECHET, MAZA, BIGOUROUX et al. 2012) ont été utilisées. Ce corpus regroupe plus de 1500 conversations téléphoniques issues d'un centre d'appels entièrement anonymisées. Chaque conversation est disponible en version audio et textuelle. Pour la version textuelle sont disponibles les transcriptions manuelles, mais aussi automatiques (leur obtention sera décrite dans cette section).

Ces conversations ont été recueillies dans un centre de la RATP [a] sur une période de 2 jours pour se donner une idée de toutes les requêtes qui peuvent se poser. Étant donné qu'elles ont été enregistrées dans un centre d'appel de transport, elles traitent de tous sujets se rapportant de près ou de loin au transport. Cela peut aller de la demande d'itinéraire, aux oublis d'objets en passant par les plaintes de fonctionnement sur le réseau ou encore des demandes de remboursement.

Le tableau 2.1 ci-dessous regroupe les dix sujets les plus courants :

À noter ici que tout découpage futur du corpus respectera cette répartition dans les sujets abordés.

a. Régie Autonome des Transports Parisiens

Raison de l'appel	%
Info trafic	22.5
Itinéraire	17.2
Objets trouvés/perdus	15.9
Souscriptions forfaits	11.4
Horaires	4.6
Billets	4.5
Appels spécialisés	4.5
Aucun sujet particulier	3.6
Nouvel enregistrement	3.4
Information tarifaire	3.0

Table 2.1. – Top 10 des sujets les plus fréquents sur le corpus DECODA.

Étant donné qu'il s'agit de conversations issues de centres d'appel, la durée de dialogue est généralement courte. En effet en moyenne un appel dure entre 55 secondes pour les plus courtes à 16 minutes pour les plus longs. Le tableau 2.2 montre la répartition des conversations en fonction de la durée de l'appel.

Durée (min)	<=1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	>10
# dialogues	597	367	230	139	68	43	27	13	10	6	14

Table 2.2. – Distribution des conversations de DECODA en fonction de leur durée.

En terme de mots, le corpus complet compte 96 103 tours de parole, ce qui correspond à 483745 mots après découpage. Un fait intéressant est que le mot le plus utilisé dans le corpus est le marqueur de discours "*euh*", ce qui symbolise la forte spontanéité des dialogues traités.

Ce corpus a été utilisé dans plusieurs études telles que le projet ANR DECODA et le projet ANR ORFEO. Dans le cadre d'ORFEO, les transcriptions du corpus ont été annotées selon plusieurs niveaux d'analyse linguistique. Ce même corpus est utilisé dans le projet SENSEI, qui est le cadre de cette thèse, et a été enrichi en annotation sémantique comme cela sera présenté dans le chapitre 3. Nous allons présenter dans le paragraphe suivant les annotations syntaxiques qui étaient présentes dans le corpus avant cette étude.

2.3.2. Annotations syntaxiques

Le corpus DECODA a été annoté syntaxiquement sur 5 niveaux :

— Disfluences ;

— Parties de discours (POS) ;

— Découpage en syntagme ;

— Entités nommées ;

— Dépendances syntaxiques.

Le premier niveau (disfluences) correspond à toutes les répétitions (i.e "*le le le*"), tous les marqueurs de discours (e.i "*euh*", "*bien*"), ainsi que tous les faux départs (i.e "*Bonj-*"). Ces annotations ont été réalisées manuellement. Le tableau 2.3 montre les différentes disfluences identifiées en fonction de leur étiquette, ainsi que les plus fréquentes. Les résultats présentés montrent que les marqueurs de discours sont la forme de disfluence la plus présente, intervenant dans plus de 28% des tours de parole.

Disfluence	# occ.	% tours	Formes les plus fréquentes
Marqueur de discours	39125	28.2%	[euh] [hein] [ah] [ben] [hmm]
Répétitions	9647	8%	[oui oui] [non non] [c'est c'est] [le le] [de de]
Faux départs	1913	1.1%	[s-] [p-] [l-] [m-] [d-]

Table 2.3. – Distribution des disfluences annotées manuellement dans le corpus DECODA.

L'annotation des entités nommées a été réalisée manuellement via une interface web. Dans le corpus la majorité des entités nommées correspondent à des noms de rue, des noms de stations de métro ou de bus.

L'annotation en syntagmes syntaxiques et en partie de discours a été réalisée automatiquement grâce à l'outil Macaon (NASR, BÉCHET, REY et al. 2011 ; décrit par la figure 2.1). En même temps une annotation manuelle a été menée sur un sous corpus. Finalement un processus itératif est appliqué, consistant à corriger manuellement les erreurs produites par le système automatique, puis en réentraînant le modèle linguistique sur ce corpus corrigé. La qualité des annotations ainsi proposées est mesurée sur le corpus manuel de référence. Ce processus itératif est répété jusqu'à atteindre un certain seuil (BAZILLON, DEPLANO, BECHET et al. 2012). Le tableau 2.4 montre le taux d'erreur obtenu sur les parties de discours avant et après le processus d'adaptation sur le corpus de référence.

Taux d'erreur partie de discours	baseline	après adaptation
Corpus de référence	21.0%	8.5%

Table 2.4. – Taux d'erreur sur les parties de discours obtenus avant et après adaptation.

Le corpus est aussi annoté en dépendances syntaxiques (NASR, BECHET, FAVRE et al. 2014).

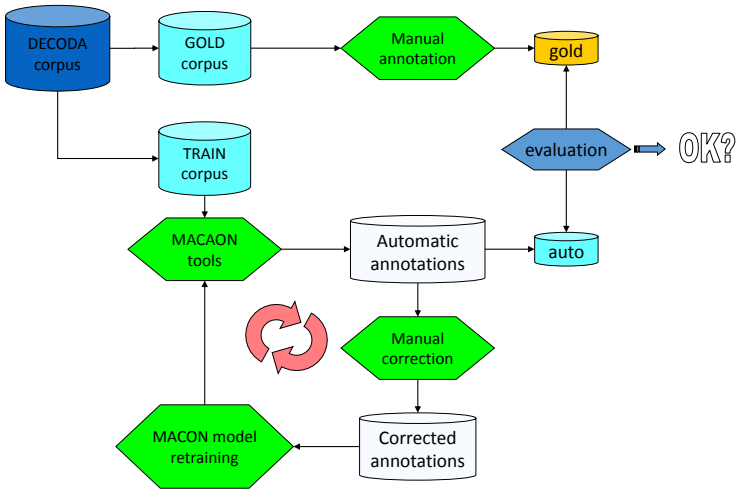


Figure 2.1. – Processus d'annotation de Macaon.

2.4. Un corpus de synopsis de conversation

En complément des diverses annotations disponibles sur les conversations du corpus RATP-DECODA, l'une des principales originalités de ce corpus est de disposer d'un ensemble de résumés, appelés *synopsis*, tels que présentés dans le chapitre précédent. L'essentiel de ces résumés a été produit durant le projet SENSEI.

Afin de limiter ce caractère subjectif dans notre étude, nous nous basons pour notre évaluation sur plusieurs (au moins deux) résumés de référence pour une même conversation. À cela s'ajoute la création d'un guide d'annotation en synopsis (annexe C) pour tout annotateur désirant enrichir le corpus, afin que tous les synopsis produits suivent la même orientation.

Dans cette étude seuls les synopsis basés sur le contenu sémantique de la conversation seront traités.

Pour illustrer ces propos, le tableau 2.5 présente deux exemples concrets de synopsis rédigés par deux individus différents.

Comme on peut facilement le voir, les synopsis de l'annotateur 2 sont syntaxiquement mieux construits que ceux de l'annotateur 1, mais en termes d'informa-

	Annotateur 1	Annotateur 2
Conversation 1	quel bus pour gare de Lyon vers Montparnasse	Demande de renseignement sur une ligne de bus pour aller de de Gare de Lyon à Gare Montparnasse.
Conversation 2	horaires RER E de Meaux à la Gare de l'Est	Demande d'horaires de train de la gare de Maux à la gare de l'Est à une heure donnée

Table 2.5. – Exemples de synopsis produits par deux annotateurs différents.

tions, les deux synopsis sont très similaires.

On notera que l'annotateur 1 est un transcripteur ayant réalisé ces synopsis afin de rechercher rapidement une conversation parmi toutes les conversations traitées. Il s'agit là de la réalisation de résumés qui lui permettait d'avoir les informations essentielles en un coup d'œil, ce qui répond à la définition de synopsis. Cependant ces résumés ne suivent pas forcément tout le temps la même structure dans la démarche de la rédaction, ce qui peut poser problème par la suite à cause d'un manque de régularité.

Afin de constituer un corpus de synopsis suffisant pour pouvoir évaluer les futurs systèmes, un guide d'annotation a été mis en place pour limiter le biais induit par la subjectivité dans la production des résumés de conversations. Un autre moyen de limiter ce facteur de subjectivité et d'obtenir le plus de synopsis possible pour une seule conversation donnée. Ainsi il est plus facile de pénaliser les informations isolées qui auraient attiré l'attention d'un annotateur en particulier.

La création manuelle de ce genre de corpus est très coûteuse à la fois en temps et en argent, c'est pour quoi il est nécessaire d'utiliser des moyens capables d'obtenir une telle ressource pour un coût plus réduit. Le guide d'annotation a pour but d'aiguiller les annotateurs dans leur façon de rédiger les synopsis, et ainsi permet de n'avoir besoin que d'un nombre plus limité d'annotations pour une même conversation. Les indications du guide ne portent que sur la sélection des informations et très peu sur la syntaxe et la rédaction, cela permet de conserver une certaine diversité dans la nature des synopsis tout en conservant l'essentiel des informations voulues.

Ce corpus de synopsis est une opportunité unique pour tester et comparer différentes approches de résumé automatique sur des données de conversation orale. Un certain nombre de méthodes standard sont comparées dans ce chapitre, mais avant de présenter cette étude, il nous faut définir la notion de comparaison entre systèmes de résumés.

2.5. Métrique d'évaluation pour la comparaison de méthodes de résumés automatiques

Évaluer un résumé est une tâche difficile parce qu'il n'existe pas de résumé de référence idéal pour un document donné. Lorsque deux humains rédigent un résumé d'un même document, il est très peu probable que les deux résumés soient identiques. De ce fait pour un même document chacun va décider de mettre en avant les informations qui lui paraissent importantes, nécessaires ou intéressantes à la compréhension du document original.

L'évaluation des résumés générés automatiquement demeure encore aujourd'hui un problème ouvert. Il existe cependant un certain nombre de mesures d'évaluation, que ce soit de façon automatique ou même de façon manuelle en fonction d'un certain nombre de critères (taille, cohérence, lisibilité, contenu).

Dans un premier temps, nous allons voir d'une part des métriques d'évaluation automatiques largement utilisées dans le domaine du résumé, puis nous verrons aussi des formes d'évaluations semi-automatiques et manuelles.

2.5.1. Évaluation automatique

2.5.1.1. Précision, Rappel et F-Mesure

Dans le cas du résumé par extraction le problème consiste à sélectionner un certain nombre de phrases du document source. On peut donc considérer ce problème comme un problème de classification binaire phrase par phrase puisque chaque phrase du texte peut être sélectionnée ou non dans le résumé (acceptation/rejet d'une phrase) et ainsi utiliser des métriques d'évaluation comme la précision, le rappel et la F-mesure. La précision (P) est définie comme le rapport entre le nombre de phrases pertinentes trouvées et le nombre total de phrases sélectionnées dans le résumé de référence. Le rappel (R) est le rapport entre le nombre de phrases pertinentes trouvées et le nombre total de phrases pertinentes dans le résumé de référence. Enfin la F-mesure (F) est la moyenne harmonique entre le rappel et la précision :

$$F = \frac{2 \times P \times R}{P + R} \tag{2.1}$$

L'équation 2.1 permet de pondérer la précision et le rappel de façon égale, mais on peut aussi favoriser l'un ou l'autre de la façon suivante :

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 P + R} \tag{2.2}$$

Où β est un poids permettant de favoriser la précision lorsque $\beta \leq 1$ et de favoriser le rappel lorsque $\beta \geq 1$.

2.5.1.2. ROUGE : Recall-Oriented Undestudy for Gisting Evaluation

La mesure de similarité cosine ne prend pas en compte l'ordre des mots d'un résumé test. En mélangeant l'ordre des mots d'un résumé test, on obtiendra les mêmes résultats qu'avec le résumé test initial, et ce même s'il devient illisible. LIN 2004 propose la méthode Recall-Oriented Undestudy for Gisting Evaluation (ROUGE) qui permet de mesurer la similarité des n-grammes entre un résumé test et un ou plusieurs résumés de référence.

Soit R_{ref} le nombre de résumés de référence. Le score ROUGE est donné par l'équation suivante :

$$ROUGE\_N = \frac{\sum_{s \in R_{ref}} \sum_{ng \in s} C(ng)}{\sum_{s \in R_{ref}} \sum_{ng} N(ng)} \tag{2.3}$$

Où $C(ng)$ est le nombre de mots qui interviennent à la fois dans le résumé test et dans les échantillons de résumés de référence. $N(ng)$ est le nombre de n grammes dans le résumé de référence. Le tableau 2.6 ci-dessous donne un exemple de fonctionnement de la métrique ROUGE.

Phrase	La force est avec toi
ROUGE_1	La, Force, est, avec, toi
ROUGE_2	La-force, force-est, est-avec, avec-toi
ROUGE_SU2	ROUGE_1, ROUGE_2, La-est, la-avec, force-avec, force-toi, /est-toi
ROUGE_SU4	ROUGE_SU2, La-toi

Table 2.6. – Différents découpages selon les variantes de ROUGE

LIN 2004 a montré que ROUGE_1 et ROUGE_2 constituaient de bons indicateurs du jugement humain dans le cas de résumé de conversations.

2.5.1.3. ROUGE_WE : Word Embeddings

ROUGE est une mesure d'évaluation très largement utilisée en résumé automatique, et même s'il a été montré qu'il était relativement proche d'un jugement humain, cette métrique reste limitée par des problèmes liés à la syntaxe du fait de son fonctionnement par comparaison de n-grammes, et est biaisée dans le cas de similarité lexicale (synonymie). J.-P. NG et ABRECHT 2015 proposent une

méthode à base de word embeddings pour franchir cette limite.

Au lieu d'utiliser les mots du résumé, ils utilisent les words embeddings de ceux-ci pour calculer la similarité sémantique. Un word embedding est une fonction W où $W : w \rightarrow \mathbb{R}^n$ et w est un mot ou une séquence de mots. Une fonction de similarité peut alors être définie comme suivant :

$$f_{WE}(w_1, w_2) \begin{cases} 0 & \text{si } v_1\text{ou } v_2 \text{ dans OOV} \\ v_1 \cdot v_2 & \text{sinon.} \end{cases} \tag{2.4}$$

Où w_1 et w_2 sont des mots à comparer, et $v_x = W(w_x)$. Ici OOV correspond à une situation où le mot w n'a pas de word embeddings (il s'agit donc d'un mot inconnu pour le système).

2.5.1.4. METEOR : Metric for Evaluation of Translation with Explicit ORdering

BANERJEE et LAVIE 2005 proposent une métrique appelée METEOR qui est plus souvent utilisée dans le cadre de la traduction automatique, mais qui peut être appliquée au résumé automatique de par sa corrélation avec le jugement humain (LAVIE et AGARWAL 2007). Cette métrique est basée sur l'utilisation de la moyenne harmonique entre la précision et le rappel des uni grammes d'une phrase avec une importance plus grande du rappel sur la précision. Cette métrique apporte aussi certains traits comme la reconnaissance de synonymie ou de racinisation en plus de la reconnaissance de mots plus classiques.

L'algorithme crée un alignement entre deux phrases données (la phrase de référence et la phrase à évaluer). L'alignement se fait au niveau du mot. Un mot ne peut être aligné qu'une seule fois au maximum. L'ordre des mots lors de l'alignement entre les deux phrases n'influe que sur la sélection d'un certain alignement dans le cas ou plusieurs alignements différents sont possibles, auquel cas l'alignement avec le moins de décalage entre les mots est sélectionné.

La précision et le rappel au niveau des uni grammes sont donnés comme suit :

$$Précision = \frac{m}{w_t} \tag{2.5}$$

$$Rappel = \frac{m}{w_r} \tag{2.6}$$

Où m est le nombre d'uni grammes en commun entre la phrase candidate et la référence, w_t le nombre d'uni grammes total de la phrase candidate, et w_r le nombre d'uni grammes dans la référence. La précision et le rappel sont alors combinés pour donner la moyenne harmonique suivante :

$$H = \frac{10PR}{R + 9PR} \tag{2.7}$$

Afin de prendre en compte des segments de mots plus longs que de simples uni grammes, les alignements de n grammes sont utilisés pour calculer un score de pénalité. Ainsi au plus il y a d'alignements non contigus au plus la pénalité est grande. Le score de pénalité est donné ci-dessous :

$$Pénalité = 0.5 \times (\frac{N_m}{N_{ua}})^3 \tag{2.8}$$

Avec N_m le nombre de mots et N_{ua} le nombre d'uni grammes alignés

Le score final de la métrique METEOR est donné par la formule suivante :

$$Score = H \times (1 - Pénalité) \tag{2.9}$$

2.5.2. Évaluation manuelle

2.5.2.1. Pyramid

Une même idée peut être exprimée de façon différente (par exemple : Arthur convoque ses frères d'armes, le roi demande à ses chevaliers de venir), ce qui devient rapidement un problème lors de l'évaluation avec des méthodes automatiques. NENKOVA, R. PASSONNEAU et K. MCKEOWN 2007 proposent une alternative avec une méthode semi-automatique : Pyramid.

Cette méthode consiste à identifier des unités sémantiques (Summarization Content Unit SCU) à partir de l'analyse de plusieurs résumés de références (e.g. écrit par des humains). Les SCU décrivant les mêmes objets ou actions sont regroupées et pondérées en fonction du nombre de leurs occurrences dans les résumés de références. Une pyramide est alors construite en fonction des scores de pondération précédents. Au sommet de la pyramide se trouve les SCUs de plus grands poids, c'est-à-dire ceux qui apparaissent le plus souvent dans les résumés de référence, et en bas de la pyramide, par extension, se trouvent les SCUs de poids les plus faibles, ou autrement dit les SCUs qui ne sont que rarement répétées dans les différents résumés humains. Le score attribué à un résumé dépend directement du nombre d'unités sémantiques qui le composent, considérées comme importantes par les annotateurs (c'est-à-dire en haut de la pyramide). L'identification et l'alignement des SCUs entre les résumés produits par le système à évaluer et les SCUs de référence se font manuellement, ce qui rend le processus long et coûteux.

Ci-dessous un exemple concret d'évaluation avec la méthode pyramid (traduit de R. J. PASSONNEAU, K. MCKEOWN, SIGELMAN et al. 2006) :

A1. L'affaire sur l'espionnage industriel concernant GM et VW a commencé avec l'embauche de Jose Ignacio Lopez, un employé de GM par VW en tant que directeur de production.

B3. Cependant, il a quitté GM pour VW dans des conditions décrites par les juges allemands comme "potentiellement la plus grosse affaire d'espionnage industriel.".

C6. Il a quitté GM pour VW *en mars 1993*.

E1. *Le 16 mars 1993*, [...], Ignacio Lopez De Arriortua, a quitté son travail de responsable des achats chez General Motor's Opel, en Allemagne, pour devenir responsable des achats et directeur de production chez Volkswagen.

F3. *En mars 1993*, Lopez et sept autres directeurs de GM sont passés chez Volkswagen du jour au lendemain.

Ici les SCU ont été identifiées dans les résumés, et sont rappelées ci-dessous avec leurs valeurs et leur poids associé.

SCU1 (w=5) : Lopez a quitté GM pour VW
A1. l'embauche de Jose Ignacio Lopez ... par VW
B3. Il a quitté GM pour VW
C6. Il a quitté GM pour VW
E1. Ignacio Lopez De Arriortua, a quitté sont travail ... chez General Motor's Opel ... pour devenir ... chez Volkswagen
F3. Lopez ... GM ... sont passés chez Volkswagen

SCU2 (w=3) : Lopez change d'employeur en mars 1993
C6. En mars 1993
E1. le 16 mars 1993
F3. En mars 1993

Cette méthode a été développée pour évaluer des résumés abstractifs, mais elle demande toutefois l'intervention de l'homme de façon intensive afin de créer une ressource de référence, et ce même si des outils ont été développés pour faciliter cette tâche. De plus, cette façon d'évaluer les résumés est d'autant plus pertinente avec un grand nombre de résumés de références. Lorsqu'il n'y a qu'un seul résumé de référence, l'importance des SCUs devient uniforme.

2.5.2.2. Évaluation subjective

Dans le cadre de l'évaluation de résumé où il existe un grand nombre de façon différente de dire la même chose, il est souvent difficile de trouver une métrique adaptée. De ce fait proposer une solution manuelle permet de quantifier un jugement humain à l'échelle du résumé. L'avantage majeur de l'évaluation manuelle dans notre cadre, c'est qu'elle reflète l'intérêt applicatif pour un évaluateur.

L'évaluation peut toucher différents aspects du résumé, selon que l'on souhaite mettre l'accent sur la lisibilité ou le sens retranscrit.

Ce type d'évaluation est utilisé dans des campagnes d'évaluation. Dans le cadre de la conférence TAC[b] (DANG et OWCZARZAK 2008), l'évaluation manuelle mesure la qualité de la réponse qu'apporte le résumé d'une part et la lisibilité de ce dernier d'autre part. Chacun de ces deux critères est évalué sur une échelle allant de 1 à 5 définie comme suit :

1. Très pauvre
2. Pauvre
3. Acceptable
4. Bon
5. très bon

Ainsi chaque résumé évalué se voit attribuer un score pour chacun des critères cités précédemment correspondant à la moyenne des scores donnés par l'ensemble des évaluateurs.

Le fait que cette évaluation soit menée par des annotateurs humains rend le processus long et très coûteux surtout s'il est appliqué sur de grands corpus. Il faut aussi prendre en compte le fait que deux évaluateurs différents ne vont pas forcément avoir la même vision et donc la même évaluation du corpus. De la même façon, les questions posées peuvent parfois être ambiguës et l'annotateur peut ne pas comprendre ce qui lui est demandé. Afin de minimiser ce biais humain, il est souvent préférable de définir au préalable un ensemble de règles permettant de guider les évaluateurs dans la même direction et obtenir une évaluation plus homogène. Augmenter le nombre d'annotateurs permet aussi de limiter le biais lié à l'accord inter annotateurs.

b. Text Analysis Conference

2.6. Expériences préliminaires de résumé automatique de conversation

Afin de nous positionner vis-à-vis de l'état de l'art en résumé automatique, nous allons présenter et évaluer un certain nombre de systèmes décrits dans la section 1. En plus de ces différentes méthodes de résumé, nous proposerons aussi des approches heuristiques basées directement sur la nature de notre corpus. Nous donnerons également les éventuelles modifications apportées pour adapter les systèmes à nos données ainsi que les différents paramètres choisis.

Étant donné que la taille d'un résumé est variable selon les besoins de l'utilisateur, nous considérerons que nos résumés devront faire au maximum 7% de la taille du document original en terme de mot. Ce chiffre correspond à la taille moyenne des résumés qui étaient disponibles sur une partie du corpus RATP-DECODA avant notre étude.

2.6.1. Systèmes état de l'art

2.6.1.1. MMR

L'algorithme de Maximal Marginal Relevance (MMR) est un algorithme glouton qui vise à réordonner les phrases d'un document en fonction de leur pertinence vis-à-vis du document dans son ensemble et de leur redondance en fonction des phrases déjà sélectionnées dans le résumé.

MMR est prévu à la base pour sélectionner des phrases, mais cette notion de phrase n'existe pas lorsque l'on manipule des données issues de la parole. Pour reproduire ce comportement, nous avons choisi de découper les conversations en tours de parole. Chaque tour de parole pourra être alors sélectionné ou non pour constituer le résumé final.

À chaque itération de l'algorithme, la prochaine phrase est sélectionnée en fonction de son importance et de sa redondance par rapport aux phrases déjà sélectionnées. Son score à l'itération t est calculé comme suit :

$$MMR_t(ST_i) = \lambda \times Sim_1(ST_i, C) - (1 - \lambda) \times Sim_2(ST_i, ST_j) \tag{2.10}$$

Où ST_i et ST_j sont respectivement les tours de parole les plus proches de la conversation C et les tours de parole déjà sélectionnés dans le résumé. $Sim_1()$ et $Sim_2()$ sont des fonctions de similarité cosine, et la valeur de λ, qui permet de jouer sur l'influence de la redondance entre ST_i et le résumé, a été fixé empiriquement à 0.7.

L'algorithme sélectionne des tours de parole jusqu'à ce que le nombre de mots dans le résumé dépasse les 7% fixé précédemment.

2.6.2. Heuristiques du dialogue

Pour compléter la baseline, des heuristiques basées sur les conversations ont été ajoutées afin de compléter l'étude et utiliser la structure du dialogue. Les conversations étant des appels provenant de centres d'appels, il existe une certaine structure provenant notamment des instructions données aux opérateurs pour mener l'appel.

Le premier constat est que l'usager appelle le centre pour obtenir des informations sur un sujet bien précis. Il est donc normal de penser que l'information essentielle de l'appel se trouve en début de conversation, dans les tout premiers tours de parole, juste après les échanges de politesse conventionnels ("*bonjour*"). D'autre part, l'opérateur se doit d'écouter l'appelant pour savoir quelle est sa requête. À partir de ces deux constats, il est possible d'établir une première heuristique : le résumé (qui sera noté **Longest@25** par la suite) sera constitué de l'unique tour de parole dont la taille en mot est maximale parmi tous les tours de parole du premier quart de la conversation (en terme de tours de parole).

Dans la même optique que précédemment, nous choisirons comme résumé (noté **Longest**) l'unique tour de parole dont la taille est maximale sur l'ensemble de la conversation. Celui-ci peut symboliser soit une explication détaillée de la requête par l'usager, soit une explication de la réponse à cette requête par l'opérateur. Dans les deux cas, il est aussi possible de révéler une prise d'informations importantes dans la conversation.

Afin d'observer si la fin de la conversation contient ou non des informations utiles, un dernier résumé (noté **Longest>75**) correspondra à l'unique tour de parole dont la taille est maximale parmi les tours de parole contenus dans le dernier quart de la conversation. La fin d'une conversation pourrait très bien représenter un rappel global de ce qu'il s'est dit tout au long de l'appel, mais pourrait aussi contenir la solution apportée ou non par l'opérateur à l'usager.

2.7. Résultats

L'évaluation se fait sur le corpus de test composé de 43 conversations conservant les mêmes proportions en terme de découpage en thème.

Les systèmes proposés sont évalués à la fois sur les transcriptions manuelles et automatiques pour mesurer aussi l'impact de la qualité de la transcription dans

le procédé du résumé de conversations.

LIN 2004 a montré que la métrique ROUGE-2 constituait un bon indicateur du jugement humain dans le cadre du résumé de conversations, c'est pourquoi il s'agit de la métrique choisie ici pour évaluer les résumés entre eux. Le tableau 2.7 regroupe les résultats de cette évaluation. En plus des résumés produits précédemment et pour se donner une idée de leur efficacité, les résultats sont comparés à la référence des résumés produits pas les humains pour l'occasion.

Système	**Transcription**	**ROUGE-2**
Human	-	**0.11848**
MMR	manuelle	0.03145
Longest	manuelle	0.02688
Longest@25	manuelle	**0.04046**
Longest>75	manuelle	0.01102
MMR	ASR	**0.02093**
Longest	ASR	0.01734
Longest@25	ASR	0.01734
Longest>75	ASR	0.01475

Table 2.7. – Résultats des systèmes de référence évalués avec la métrique ROUGE-2.

Le tableau 2.8 montre clairement les limites des méthodes à base d'extraction de tours de parole. En effet le meilleur des résultats reste très loin des humains.

En ce qui concerne l'influence de la transcription, les méthodes à base d'heuristiques sont les plus touchées notamment le résumé qui correspond au tour de parole le plus long du début de la conversation (LB) qui a le meilleur résultat avec les transcriptions manuelles, mais qui est tout juste meilleur que le pire système avec des transcriptions automatiques. Cette différence de résultats s'explique par le nombre d'erreurs ajouté dans les conversations par la transcription automatique.

Il est difficile ici d'obtenir des résultats proches des humains étant donné que les styles d'écriture sont opposés. Le tableau 2.8 montre des exemples de sorties des systèmes évalués ci-dessus.

Comme déjà dit plus haut, les exemples présentés ici montrent clairement la différence de styles d'écriture entre les références rédigées par des humains dans le but de résumer une conversation, et les systèmes évalués qui correspondent à des tours de parole extraits des conversations sans modification.

Système	Transcription	Résumé
Référence 1	-	Demande de renseignements sur le fonctionnement de la ligne 486 suite à un mouvement de grève. Communication du numéro de téléphone de la ligne concernée.
Référence 2	-	infos sur grève ligne bus 486, ligne privée, communication du numéro à joindre.
MMR	manuelle	Donc euh 0825003836. Oui euh bonjour madame voilà je voulais avoir un renseignement concernant les bus euh euh de la ligne 486 euh il semblerait qu'il y ait un mouvement de grève euh ce matin
LA	manuelle	Oui euh bonjour madame voilà je voulais avoir un renseignement concernant les bus euh euh de la ligne 486 euh il semblerait qu'il y ait un mouvement de grève euh ce matin
LB	manuelle	et donc vous vous pensez qu'on peut obtenir des renseignements concernant euh
MMR	ASR	lorsque 0825 est 00 8 36 est bien ça
LA	ASR	oui et rejoint la très bien un un renseignement euh cautionna les 10 et maintenant eu un ligne 4 10 84 15 n'est ça veut dire qu'il y a un mouvement de grève
LB	ASR	oui et rejoit la très bien un un renseignement euh cautionna les 10 et maintenant eu un ligne 4 10 84 15 n'est ça veut dire qu'il y a un mouvement de grève
LE	ASR	il y a encore un peu donc vous passe quand même on a punira des renseignements hein quand même

Table 2.8. – Exemple de résumés produits des systèmes baselines pour une conversation donnée.

À remarquer aussi le fait que souvent MMR ou l'heuristique du plus long tour de parole en début de conversation retournent des résultats similaires.

La différence de résultat pour le MMR entre les transcriptions manuelles et automatiques est principalement due à l'algorithme glouton de sélection de phrases, qui s'arrête s'il dépasse la taille maximale (fixée) du résumé final.

2.8. Conclusion

On a vu dans ce chapitre que la tâche de résumé est une tâche qui peut avoir des applications réelles en dehors du laboratoire (comme le montre le projet SENSEI). Le corpus ainsi utilisé est directement tiré de faits existants, ici des conversations issues d'un centre d'appel de la RATP. Ce corpus a été annoté pour pouvoir tester des méthodes classiques de résumé basé sur l'extraction de tours de parole (étant donné que le découpage en phrase est un problème difficile dans les conversations orales).

Après avoir testé et évalué des méthodes extractives, même si les tours de parole extraits ont du sens, il paraît évident que ces dernières sont peu adaptées à la tâche de résumé de conversation. En effet les informations divisées sur plusieurs tours de parole sont difficilement prises en compte ; le style entre les transcriptions et les synopsis est complètement différent, la conversation un langage parlé non canonique alors que le synopsis correspond à du discours rapporté ; le rôle des locuteurs n'est pas pris en compte et donc n'apporte aucune information supplémentaire. Il est donc nécessaire de se passer du tour de parole et de chercher des solutions un niveau d'abstraction plus élevé s'appuyant sur la sémantique des conversations.

3.La sémantique dans les résumés

L'analyse sémantique produit une interprétation en sens à partir des mots et de toutes les informations syntaxiques disponibles dans un document. De nombreux modèles ont été proposés allant des modèles formels codant une structure sémantique profonde à peu profonde en passant par des modèles envisageant seulement le sujet principal d'un document à partir de ses concepts principaux et entités nommées présentes dans ce dernier.

L'intérêt de produire cette analyse est de pouvoir être capable d'identifier les éléments les plus pertinents d'une conversation afin de s'appuyer sur ces derniers pour générer un résumé (comme le présentent GENEST et LAPALME 2013).

Ce chapitre expose dans une première partie les modèles sémantiques utilisés pour réaliser une annotation sémantique, puis dans une seconde partie présente les données existantes dans le domaine des conversations, une troisième partie décrit l'approche choisie ainsi que les ajouts qui ont été apportés pour correspondre au mieux au corpus, enfin dans une dernière partie nous donnons, commentons et analysons les résultats obtenus vis-à-vis de la tâche de résumé.

3.1. Modèle sémantique

3.1.1. Introduction

Les résumés à base d'extraction de phrases (ou tours de parole) ne donnent pas de résultats satisfaisants dans le sens où ils ne correspondent pas à la définition d'un synopsis tel que défini préalablement. En effet les natures des documents initiaux et finaux sont différentes. Les documents initiaux correspondent à des conversations audios dont la syntaxe et l'écriture ne sont pas destinées à être utilisées pour la lecture, alors que les documents finaux sont des résumés dont le but est d'être lus pour donner un aperçu global de ce qui a été dit durant la conversation. Si leurs natures sont très différentes, les informations que contiennent ces documents sont cependant proches. Les méthodes par extraction ont montré qu'elles étaient capables d'identifier des tours de parole contenant des informations nécessaires à la compréhension de la conversation. Ces informations sont présentes, mais souvent noyées dans la syntaxe du langage parlé. L'idée serait alors d'être capable de comprendre la conversation pour n'en garder que les informations nécessaires à sa compréhension. Pour cela il faut se passer de la forme du langage parlé et exploiter un niveau d'abstraction supplémentaire que constitue la sémantique.

3.1.2. Modèles sémantiques

Cette section présente différents modèles sémantiques, leurs qualités et leurs défauts au vu des données utilisées et leur facilité d'utilisation afin d'annoter un corpus de dialogue en maximisant l'efficacité d'annotation et en minimisant le coût à la fois en temps et en ressources.

Les modèles décrits sont les suivants :

— WordNet

— PropBank

— AMR

— FrameNet

3.1.2.1. WordNet

MILLER 1995 propose une base de données qui lie les différents noms, verbes, adjectifs et adverbes anglais à un ensemble de synonymes qui sont liés sémantiquement et qui permettent de donner une définition d'un mot.

Les phrases porteuses de sens sont composées de mots significatifs, ainsi n'importe quel système voulant prendre en compte le langage naturel comme un humain le ferait, se doit de prendre en compte les informations sur les mots et leur sens. Ces informations sont généralement contenues dans des dictionnaires, mais les entrées dans les dictionnaires évoluent pour convenir aux humains et non aux machines. WordNet propose une solution plus adaptée à l'utilisation de la machine.

Le langage dans WordNet est défini comme un ensemble de mots W eux-mêmes définis par le couple (f, s) où f est la forme lexicale du mot, et s correspond à son sens issu d'un ensemble de sens (plusieurs sens peuvent être attribués à la même forme f). Ce couple est appelé *synset*. Ainsi pour chaque mot une liste de synsets est disponible correspondant à toutes les définitions répertoriées par WordNet. Les synsets permettent aussi de représenter des concepts plus abstraits, de plus hauts niveaux que le simple sens des mots. Ces concepts sont classés sous forme de catégories permettant de classifier les différents éléments d'un thème donné. Ce système de catégorisation représente les relations sémantiques, le tableau 3.1 montre un exemple pour le mot *car* avec ses différents sens en fonction des différents niveaux d'abstraction.

Le dernier niveau (*entity, something*) représente le niveau de concept le plus abstrait et pourrait facilement être le superconcept d'une multitude d'autres concepts plus spécialisés.

niveau d'abstraction	Concept associé
0 (synset)	, car, auto, automobile, machine, motorcar
1	vehicle
2	conveyance, transport
3	intrumentality, intrumentation
4	artifact, artefact
5	object, physical object
6	entity, something

Table 3.1. – Exemple d'entrée dans WordNet pour le mot *car*

La nature de ce dictionnaire permet de regrouper de manière cohérente toutes les composantes d'un univers telles que les mots, les sens ou les concepts. Le lexique de Wordnet est divisé en quatre grandes catégories : les noms, les verbes, les adjectifs et les adverbes.

WordNet (MILLER 1995) contient plus de 118 000 mots et plus de 90 000 sens, ou plus de 166 000 pairs (f, s). Approximativement 17% des mots de WordNet sont polysémiques et 40% ont un synonyme ou plus.

Cette ressource a été développée en anglais, mais des versions de WordNet ont été exportées et développées dans d'autres langues comme le français par exemple (SAGOT et FIŠER 2008).

Une des limitations de WordNet touche la catégorie des verbes, où le système de hiérarchie est beaucoup moins riche avec moins de niveaux d'imbrication des sens. On passe ainsi rapidement d'un concept spécialisé à un concept très général. De plus il s'agit d'un dictionnaire, et ne propose pas comment faire d'analyse de phrases. Enfin la désambigüisation peut s'avérer difficile.

3.1.2.2. PropBank

PropBank (KINGSBURY et PALMER 2002) est une ressource comprenant un million de mots provenant de la partie du *Wall Street Journal* issue du Penn TreeBank II (MARCUS, KIM, MARCINKIEWICZ et al. 1994) avec une structure prédicat-argument pour les verbes, prenant compte des étiquettes de rôle sémantique pour chaque argument du verbe. Afin de rester générique, et pour augmenter la vitesse d'annotation, les étiquettes de rôle ont été définies au niveau des lexèmes. Bien que les mêmes identifiants de rôles aient été utilisés pour tous les verbes (Arg0, Arg1, ..., Arg5), ceux-ci sont destinés à avoir une signification spécifique du verbe.
Ainsi, l'utilisation d'un argument donné doit être cohérente entre les différentes utilisations du verbe en question. Par exemple, l'Arg1 (souligné) dans "John

broke the window", est le même *window* qui est annoté comme Arg1 dans "The window broke", et ce même si *window* est sujet dans une phrase et objet dans l'autre.
Il n'y a aucune garantie qu'un argument sera utilisé de façon uniforme pour tous les verbes.
PropBank organise les mots en lexèmes en utilisant un schéma de désambigüisation gros-grain : deux sens sont considérés différents seulement si leurs étiquettes d'argument sont différentes. Dans PropBank chaque sens est regroupé dans un *frameset*. Les informations de chaque *frame*, y compris les définitions des arguments (Agr0, ... Arg5) spécifiques au verbe, sont définies dans des *frames files* distribués avec le corpus.

Le modèle prédicat argument de PropBank diffère de l'analyse syntaxique en dépendance dans le sens où contrairement à la syntaxe qui cherche à créer des analyses arborescentes, la sémantique de surface considère les verbes de manière indépendante et il en résulte qu'un constituant peut avoir plusieurs pères dans l'analyse (à l'opposé de la syntaxe où le père est unique), la structure créée est donc un DAG. Par exemple, dans la phrase suivante (voir figure 3.1), PropBank utilise l'expression "his dog" comme argument de deux prédicats, "scouted" et "chasing" ; Cette description ne serait pas tolérée dans les modèles traditionnels d'analyse en dépendance, puisque chaque expression ne peut être dépendante que d'une seule autre.

a. His dog **scouted** ahead, chasing its own shadow
b. His dog scouted ahead, **chasing** its own shadow

Figure 3.1. – Exemple de phrase gérée par le modèle PropBank

Un problème de ce modèle touche l'identité de l'étiquette *ARGx* qui est définie de façon globale entre des phénomènes différents. Ce défaut est corrigé par le modèle FrameNet par exemple.

Même si PropBankest est un modèle intéressant qui a l'avantage de disposer d'une grande quantité de données annotées (corpus ontonote (WEISCHEDEL, HOVY, MARCUS et al. 2011), il reste très limité pour le français du fait du manque de données annotées.

3.1.2.3. AMR

AMR est une structure complexe qui n'est plus ancrée dans les mots et qui hérite de l'étiquetage proposé par PropBank.

L'analyse sémantique correspond au problème d'identification et d'association entre une chaine de caractère et sa signification en langage naturel. Abstract Meaning Representation (BANARESCU, BONIAL, CAI et al. 2012 ; DORR, HABASH et TRAUM 1998) est un formalisme sémantique dans lequel le sens d'une phrase est représenté dans un arbre acyclique et orienté. Les nœuds représentent les concepts, et les arrêtes représentent les relations entre ces concepts (la figure 3.2 donne un exemple d'arbre AMR). Le formalisme utilisé est basé sur la logique propositionnelle ainsi que sur les représentations d'évènements néo-Davidsoniennes (DORR, HABASH et TRAUM 1998 ; TERENCE 1990). Bien qu'AMR ne codifie pas les quantifieurs, le temps ou les modalités, l'ensemble des phénomènes sémantiques couvert a été sélectionné avec des applications de langage naturel (en particulier la traduction automatique).

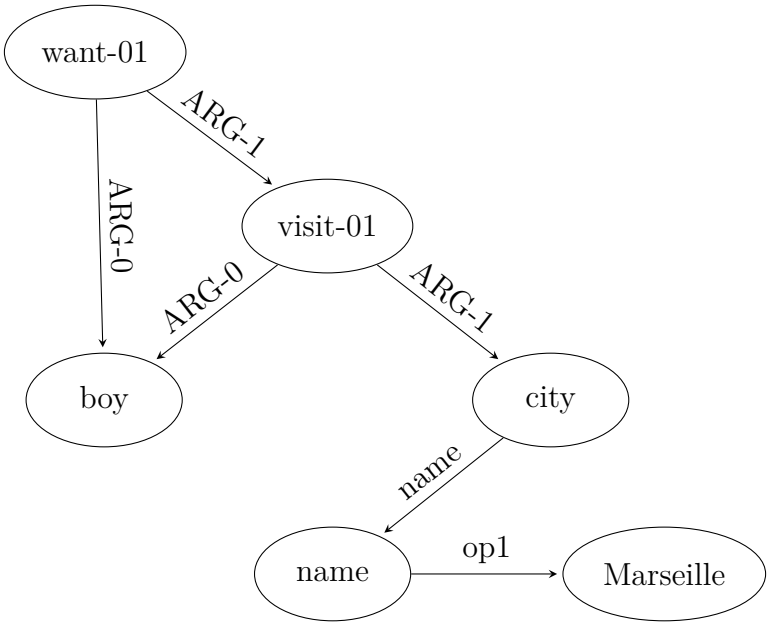
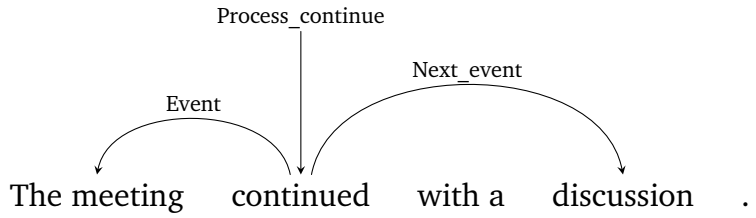


Figure 3.2. – Exemple de représentation en arbre avec le formalisme AMR pour la phrase : "the boy wants to visit Marseille"

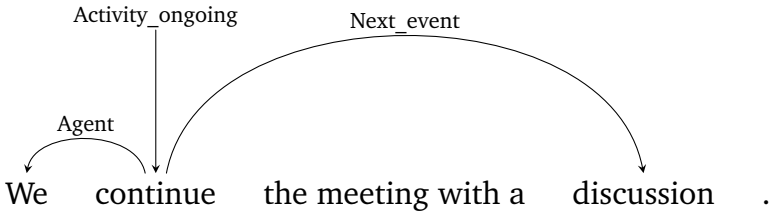
3.1.2.4. FrameNet

Le projet FrameNet regroupe à la fois une théorie de modélisation de la sémantique des phrases et un ensemble de données annotées manuellement. L'unité structurelle utilisée pour représenter la structure sémantique d'une phrase est appelée Semantic Frame (cadre sémantique en français), et elle est fondée sur les travaux de Charles J. Fillmore (FILLMORE 1976). Par exemple, le verbe *continuer* peut évoquer deux situations différentes : une activité qui continue (modélisée par le cadre Process_continue) ou la poursuite d'une activité (Activity_ongoing).

Ces situations, bien que proches, ne représentent pas la même chose : la première concerne une tâche qui se poursuit, alors que la seconde concerne des participants qui continuent une activité. Le cadre sémantique de Process_continue met en jeu essentiellement un évènement (Event), mais peut aussi contenir des circonstances (Circumstances), une manière de continuer (Manner), ou encore l'évènement suivant (Next_subevent), comme dans l'exemple suivant :



Le cadre sémantique de Activity_ongoing nécessite absolument de définir l'activité (Activity) et un agent (Agent) effectuant cette activité. Bien entendu ce cadre sémantique peut aussi contenir des indicateurs de circonstances (Circumstances), d'une manière de continuer (Manner), ou encore de l'évènement suivant (Next_subevent), comme dans l'exemple suivant :



Les deux situations évoquées précédemment (à savoir Process_continue et Activity_ongoing) sont ce que nous appellerons des cadres (frames en anglais) et les intervenants (Event, Agent,...) sont appelés rôles (frame elements). Les mots qui pourront évoquer ces cadres (comme *poursuivre*) sont appelés déclencheurs (lexical units) du cadre. Les rôles sont spécifiques à un cadre. Le mot ou l'expression qui joue un rôle dans un cadre est appelé acteur. Le nombre de rôles peut varier selon les cadres. Certains rôles peuvent ne pas être réalisés pour une occurrence de cadre donnée, le jeu de rôles instanciés peut donc varier d'une occurrence de cadre à une autre. Lorsqu'un déclencheur est associé à un cadre, on dit que le déclencheur déclenche le cadre, qui est alors instancié.

Notons également qu'un mot ne peut déclencher qu'un seul cadre alors qu'un mot peut occuper un rôle dans plusieurs cadres. De plus, un mot peut être à la fois déclencheur et acteur.

3.1.2.5. Conclusion

Parmi les modèles sémantiques décrits, WordNet ne convient pas à notre tâche étant donné qu'il ne propose pas d'analyse de texte, et la désambigüisation peut être problématique. PropBank permet de faire de l'analyse grâce aux corpus annotés, et dispose d'une grande quantité de données, mais très peu en français, de plus il ne traite pas les noms, et ne nomme pas les arguments de manière globale. AMR est une structure complexe héritant directement de l'étiquetage de PropBank, conservant ses défauts, et n'est pas un modèle très répandu. Enfin FrameNet de par sa généricité, sa facilité à s'adapter à un domaine bien défini, et sa capacité à produire des analyses sémantiques de textes, semble être le modèle le plus prometteur pour notre tâche.

Pour la suite de notre étude nous avons donc choisi d'utiliser le modèle FrameNet pour les raisons évoquées précédemment d'une part et d'autre part car ce modèle fait partie intégrante du projet SENSEI. La suite de cette section montre les efforts fournis pour permettre l'annotation sémantique de notre corpus, ainsi que l'utilité de cette dernière vis-à-vis des synopsis existant.

3.2. De l'annotation syntaxique à l'annotation sémantique

3.2.1. Introduction

L'annotation sémantique est une tâche subjective à cause de sa complexité et qui peut rapidement se révéler assez subjective. Deux humains peuvent comprendre un même message différemment et donc proposer une annotation différente. De plus le processus d'annotation manuelle en cadre et rôle sur un corpus comme RATP-DECODA coûte très cher. Pour pallier à ces problèmes, l'annotation syntaxique peut être utilisée ainsi que des ressources sémantiques externes afin de produire une annotation sémantique rapide et à faible coût tout en restant significative.

Dans cette section nous allons présenter une méthode d'annotation rapide semi-automatique. Une alternative consistant à adapter un système générique déjà existant n'est pas raisonnable, car il n'en n'existe pas (au moment des expériences) pour le français. Un transfert sur des textes traduits pourrait être envisagé, mais la différence de style entre les conversations orales et les données sur lesquelles sont entrainés ces systèmes donne de mauvais résultats comme cela a été étudié dans le projet SENSEI.

3.2.2. SEMAFOR

Comme prémisse à la présentation de notre approche, nous proposons de regarder quelques exemples caractéristiques de l'approche à base de transfert, notamment avec le système SEMAFOR qui est une référence sur l'anglais. Les expériences sont menées sur des traductions manuelles d'un petit ensemble de DECODA.

L'annotation sémantique d'un corpus en français est souvent un problème du fait du manque de ressource dans ce domaine, notamment en ce qui concerne les données issues du langage parlé. En revanche les modèles sémantiques développés pour l'anglais sont plus courants. Une solution serait d'adapter ces modèles afin de les rendre utilisables sur des données en français.

L'idée est de tirer parti d'un système en anglais sur des données du corpus à traiter dans cette langue. Les données annotées sur la version traduite sont ensuite transférées vers la version originale du corpus (i.e en français).

Pour tester cette approche, une partie du corpus DECODA a été traduite en anglais. Les méthodes de traduction automatique sont basées sur des données structurées et syntaxiquement correctes comme des textes de journaux ou des livres. Lorsqu'il s'agit de données non canoniques comme le discours parlé, les erreurs de traduction peuvent atteindre les 40% en terme de taux d'erreur mots, d'autant plus lorsqu'il s'agit d'expressions idiomatiques spécifiques de la langue utilisée. Afin de limiter au maximum l'impact des erreurs de traduction, une traduction manuelle est proposée sur un sous corpus pour tester cette approche.

Les conversations préalablement traduites ont été annotées avec SEMAFOR. La figure 3.3 présente des exemples de tours de parole après traduction.

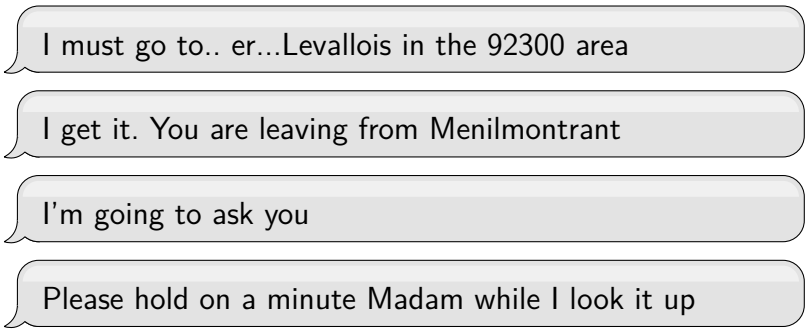


Figure 3.3. – Exemples de tours de parole traduits en anglais

Les figures 3.4 à 3.7 présentent des résultats de l'annotation sémantique d'un tour de parole avec SEMAFOR, ainsi que l'annotation attendue.

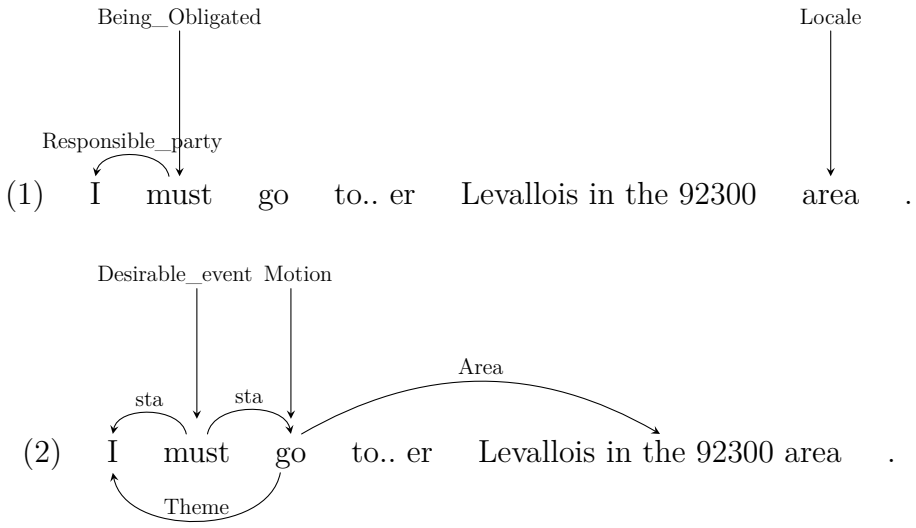


Figure 3.4. – Comparaison d'annotation entre SEMAFOR (1) et l'oracle (2) (*sta : State_of_affairs*)

Dans cet exemple le tour de parole contient du bruit lié à la nature spontanée de la conversation. L'analyse sémantique ne peut plus s'appuyer sur la syntaxe. Le cadre *Being_obligated* est juste, mais n'apporte que peu d'information supplémentaire dans la tâche d'identification d'éléments nécessaires à la compréhension de la conversation. En revanche *Locale* donne l'information de l'existence d'un lieu.

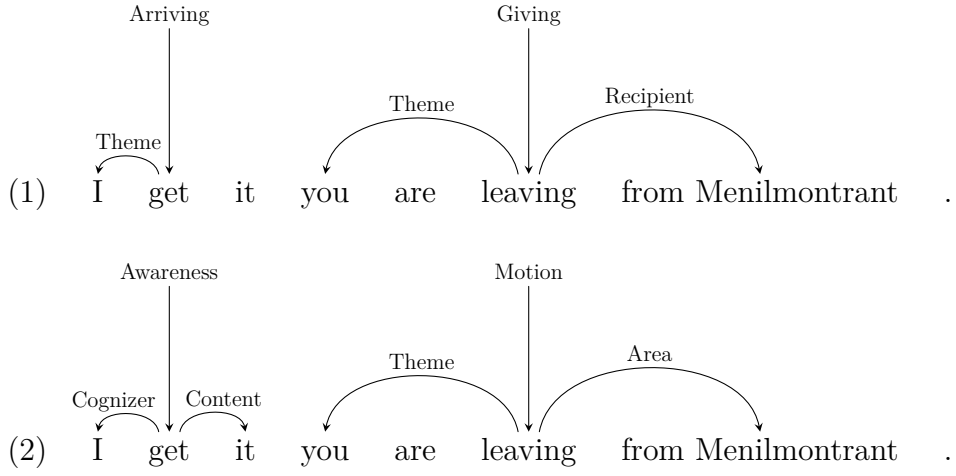


Figure 3.5. – Comparaison d'annotation entre SEMAFOR (1) et l'oracle (2)

Ici *get* déclenche le cadre *Arriving* alors qu'il devrait être associé au cadre *Awareness*, cette erreur est due à la nature spontanée de la conversation. Il s'agit

d'expression qui sont très largement utilisées à l'oral, mais que très peu à l'écrit d'où le fait que l'analyseur se trompe très facilement. Cette erreur se retrouve avec le second cadre, où ces tournures de phrases ne sont pas courantes dans les textes écrits.

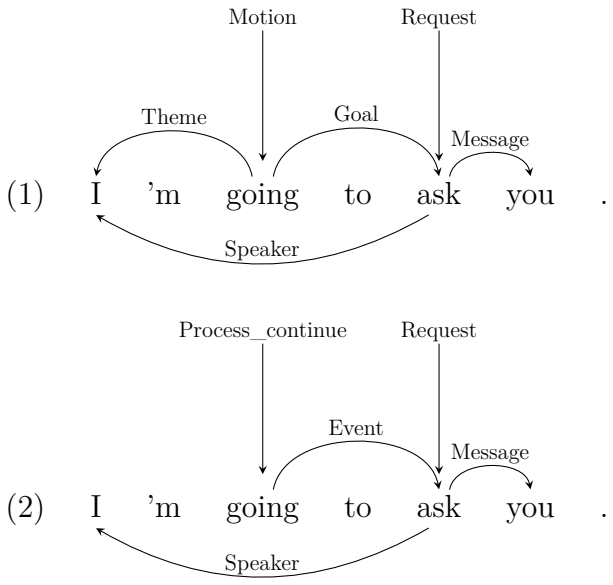


Figure 3.6. – Comparaison d'annotation entre SEMAFOR (1) et l'oracle (2)

Cet exemple est un cas typique d'ambigüité où le verbe *go* peut déclencher plusieurs cadres différents et le système ne parvient pas à prendre la bonne décision. Au lieu de *Motion* ici le cadre correct serait *Process_continue*. En revanche *Request* a bien été choisi pour représenter la requête de l'interlocuteur.

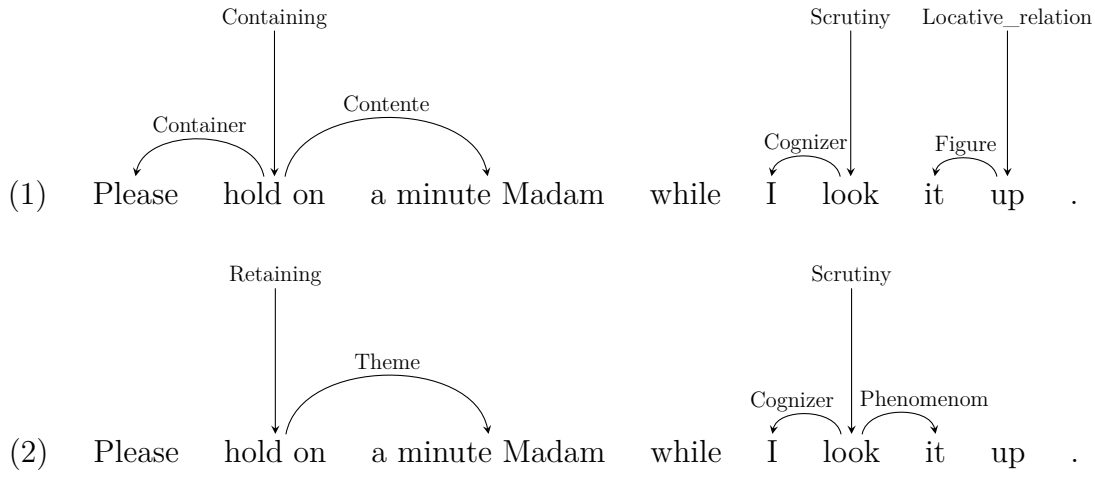


Figure 3.7. – Comparaison d'annotation entre SEMAFOR (1) et l'oracle (2)

Dans ce cas de figure on retrouve le problème lié aux différentes tournures de phrases utilisées dans les conversations où *hold on* est associé à un contenant au lieu de représenter l'action de garder la ligne téléphonique active, ce qui donne lieu à une annotation surprenante où *Please* devient le conteneur et *a minute* le contenu. En ce qui concerne le cas de *look up*, FrameNet ne possède pas de sens adapté, le sens le plus proche est celui de *Scrutiny* qui correspond à l'action de rechercher un objet perdu.

L'annotation produite par SEMAFOR correspond à une annotation de texte écrit. Tous les phénomènes syntaxico-sémantiques liés au langage spontané ne sont pas couverts, et de ce fait génèrent des frames qui ne sont pas adaptés. L'ajout d'erreurs de traduction supplémentaires réduit aussi la qualité des annotations. Il est nécessaire de produire un système adapté à ce type de donnée et capable de fournir une annotation rapide, peu coûteuse, et suffisante tout en restant significativement acceptable pour répondre aux problématiques d'extraction d'informations vu précédemment.

Ces résultats préliminaires sur un petit sous-ensemble du corpus tendent à confirmer que cette approche n'est pas facilement adaptable à notre tâche. Pour refuser complètement cette méthode, il faudrait procéder à une évaluation complète, mais cela nécessite d'annoter les données en frames, autant produire cette annotation sur la langue source.

3.2.3. Contributions

Cette section décrit une proposition de méthode d'annotation sémantique basée sur le modèle FrameNet. Cette méthode se veut rapide et peu coûteuse tout

en restant la plus précise possible pour la tâche de résumé.

Cette annotation s'inspire du projet ASFALDA (CANDITO, AMSILI, BARQUE et al. 2014) qui propose une adaptation française de FrameNet. En développement au moment de rédiger ce document, le modèle compte 106 cadres différents répartis en 9 domaines. Chaque cadre est associé à un groupe d'unités lexicales (LU) qui sont susceptibles de déclencher une occurrence d'un cadre dans le texte source. Le projet ASFALDA traite de documents journalistiques, les annotations et analyseurs créés dans le cadre de ce projet ne pourront donc pas être réutilisés directement.

3.2.3.1. Définitions des cadres

La première étape pour annoter un corpus avec FrameNet consiste à détecter les potentiels déclencheurs et générer des hypothèses de cadres pour chaque détection (voir 3.1.2.4). Sur le corpus RATP-DECODA 188 231 déclencheurs possibles ont été identifiés parmi 94 définitions de cadre tirées du projet ASFALDA.

L'annotation proposée correspond à une approche semi-supervisée basée sur les unités lexicales du texte. Pour chaque unité lexicale, on cherche parmi les sorties de l'analyseur syntaxique les dépendances possibles (par exemple *sujet / objet*) pour chaque déclencheur. Cette recherche est ensuite affinée à l'aide de contraintes sémantiques basées sur les unités lexicales données, et en considérant le domaine restreint du corpus.

Avant d'annoter à proprement dit le corpus, il faut déterminer un ensemble de déclencheurs et de cadre en accord avec le domaine. La nature spontanée du corpus et les thèmes précis qui y sont abordés conduisent à adopter comme déclencheurs les 200 occurrences de verbes les plus fréquentes. Ce choix se justifie du fait que les conversations traitées sont pour la plupart courtes et fortement orientées dans un but précis (par exemple une demande d'itinéraire), ce qui favorise l'utilisation de verbes d'action. À noter que les entités nommées sont aussi annotées dans le corpus.
Après avoir analysé les déclencheurs les plus fréquents pour chaque thème de conversation, les cadres ont pu être classés en 7 domaines différents :

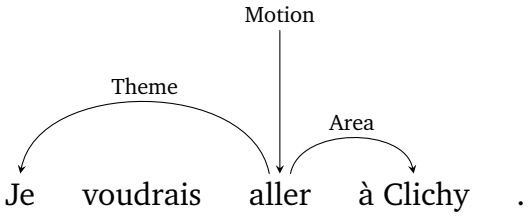
— **Déplacement**.

  Les cadres du domaine *Déplacement* mettent en avant tous les déplacements d'objets ou d'individus vers une cible. Le cadre peut être complété par des informations comme le véhicule utilisé, le lieu de départ, le chemin emprunté, ou encore l'heure à laquelle le déplacement a eu lieu, ce qui nous donne des informations capitales dans le cas de demande d'itinéraire par

exemple.

Cadres les plus utilisés dans ce domaine : Motion, Path_shape, Ride_vehicle, Arriving.

Exemple :

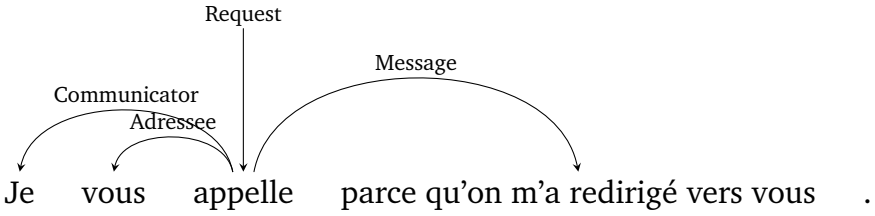


— **Communication**.

Les cadres du domaine *Communication* impliquent un intervenant transmettant un message à un récepteur. Étant donné que notre corpus est basé sur des appels téléphoniques, ces cadres sont particulièrement importants pour décrire la structure de la conversation.

Cadres les plus utilisés dans ce domaine : Communication, Request, Communication_response.

Exemple :

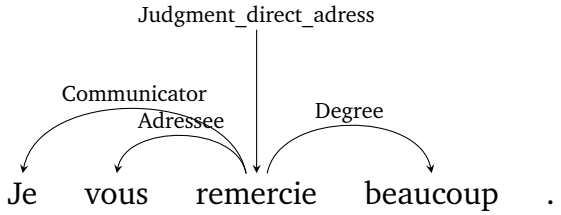


— **Sentiments**.

Les cadres de ce domaine mettent en relation un intervenant et un récepteur. Dans notre cas ces détections de Sentiments peuvent évaluer le comportement des interlocuteurs dans la conversation, ou apprécier dans une certaine mesure le degré de politesse de l'échange.

Cadres les plus utilisés dans ce domaine : Judgment_direct_address, Desiring.

Exemple :

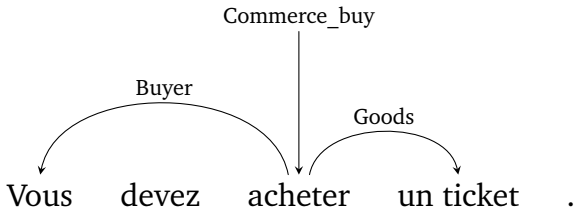


— **Commerce**.

Les cadres de ce domaine impliquent un acheteur, des biens et parfois un vendeur. Ces cadres sont très présents dans les conversations traitant de la tarification des transports, des remboursements, ou des frais en général.

Cadres les plus utilisés dans ce domaine : Commerce_buy, Commerce_pay.

Exemple :

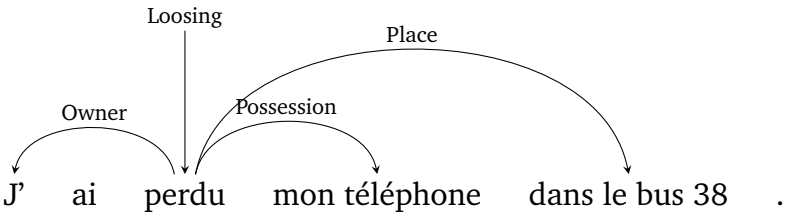


— **Action**.

On appelle ici les cadres Action tous les cadres qui impliquent une action liée à une personne. Ces cadres sont fréquents dans les conversations traitant des malheurs de l'appelant.

Cadres les plus utilisés dans ce domaine : Loosing, Giving, Intentionally_affect.

Exemple :



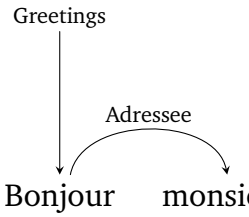
Comme dit précédemment tous ces cadres sont déclenchés par les 200 verbes les plus fréquents dans le corpus. Cependant, FrameNet n'est pas limité aux

verbes. L'annotation peut être étendue avec des déclencheurs non verbaux, ainsi deux types de cadres spécifiques ont été ajoutés et spécialement créés pour les besoins de l'annotation :

— **Salutations**.

Ce cadre est utilisé pour détecter l'ouverture et la fermeture de l'appel. Étant donné une conversation du corpus il ne peut y avoir qu'une seule ouverture et une seule fermeture, de ce fait le même cadre est utilisé dans les deux cas ("Bonjour", "au revoir"). Ces cadres sont utiles dans la structuration de la conversation.

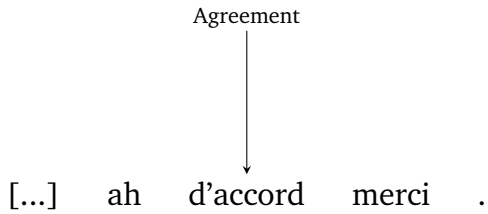
Exemple :



— **Accords**.
Les cadres liés à l'entente entre les interlocuteurs sont cruciaux dans le contexte des conversations. En effet être capable de détecter une réponse positive ou négative après une question booléenne donne un niveau de compréhension élevé du dialogue.

Les cadres de ce domaine se réfèrent à toutes les marques d'entente : *oui, bien sûr, non*.

Exemple :



Pour l'annotation chaque déclencheur peut faire intervenir un des 181 cadres disponibles (listés en annexe .10). Ces cadres sont une combinaison des cadres issus du projet ASFALDA, mais aussi de FrameNet adapté au français et enfin des cadres ont été créés pour répondre au domaine du corpus (par exemple le cadre *Hello*).

L'annotation de ces cadres n'est pas toujours évidente et des cas ambigus se présentent souvent même pour des phrases dites simples. L'exemple présenté sur la figure 3.8 illustre un cas d'ambigüité.

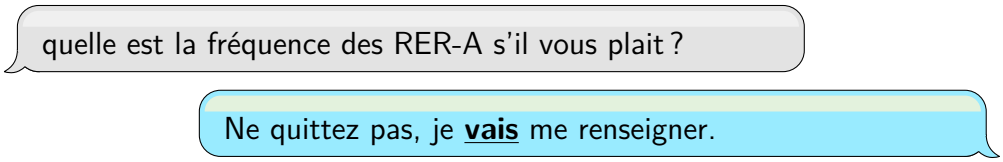


Figure 3.8. – Cas d'ambigüité dans l'annotation en cadre

Ici le verbe *aller* peut déclencher plusieurs cadres différents comme *motion*, *Attempt*, ou encore *Process_continue*. Dans ce cas précis le cadre déclenché par le verbe *aller* ne définit pas un déplacement de personne, mais une action (ici l'action de chercher), c'est aussi un marqueur de futur proche. Cette ambigüité peut être limitée en définissant un ensemble de règles permettant de couvrir et désambigüiser ces situations.

3.2.3.2. Règles de désambigüisation

Une fois le processus de sélection terminé, des hypothèses de cadres peuvent être déterminées à partir de la liste des déclencheurs disponible. Après avoir sélectionné le cadre, les rôles sémantiques sont déduits des dépendances syntaxiques. Un déclencheur ne peut générer plusieurs cadres différents, mais ne peut être associé qu'à un seul dans l'annotation. Si plusieurs cadres sont possibles pour un même déclencheur, c'est qu'il s'agit d'un cas ambigu, un système à base de règles est alors appliqué afin de ne conserver que le cadre le plus approprié au contexte. Une partie des ambigüités est éliminée par le domaine restreint du corpus (par exemple, dans le cas de de *perte* il sera toujours question d'une *perte d'objet* et jamais la *perte d'un proche*), les ambigüités restantes sont traitées par des règles utilisant la syntaxe. L'analyse des ambigüités a montré qu'elles touchaient principalement les 5 verbes les plus fréquents (comme le verbe *aller*).

L'écriture de ces règles de désambigüisation est faite de façon empirique. Pour chaque verbe ambigu, des exemples illustrant les différentes situations ont été sélectionnés afin d'identifier des moyens de séparer les cas pour associer le bon sens au verbe dans son contexte.

6 règles ont été nécessaires pour désambigüiser les cas rencontrés. Parmi ces règles 4 concernaient le cadre *motion*, qui correspond au cadre le plus ambigu. Le tableau 3.2 montre les déclencheurs concernés et les règles associées.

Déclencheur	Règle
Aller	Déclencheur + non verbe = cadre *Motion*
Aller	Déclencheur + verbe = cadre *Action*
Passer	Déclencheur + groupe propositionnel = cadre *Motion*
Changer	Déclencheur + groupe propositionnel = cadre *Motion*

Table 3.2. – Liste des règles de désambigüisation pour le cadre *Motion*

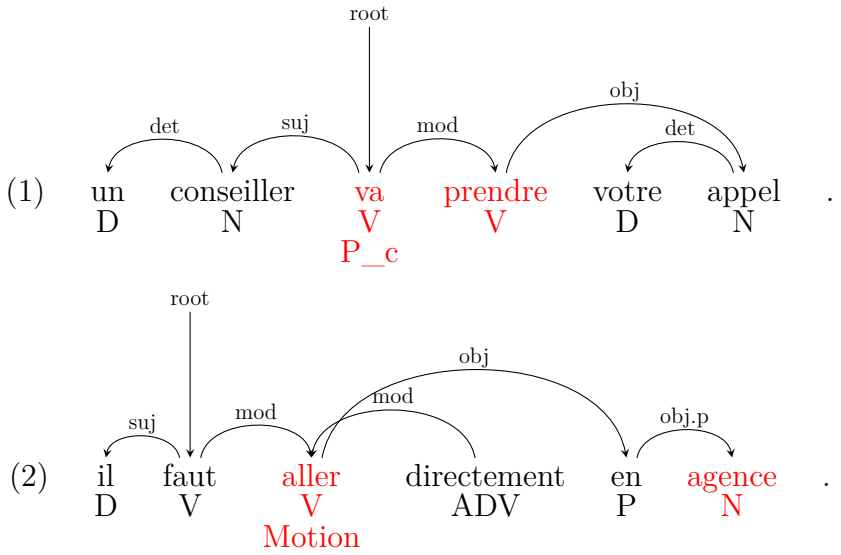


Figure 3.9. – Exemple d'application de la règle donnée dans le tableau 3.2 (*P_c : Process_continue)*

L'exemple 3.9 illustre le caractère ambigu du verbe *aller*. Ce verbe est très fréquent en français, en particulier dans les conversations spontanées. Il s'agit d'un verbe polysémique qui peut avoir comme signification le déplacement d'objet ou de personne, mais aussi la description d'une action en cours (*il est en train de faire quelque chose.*). Une façon de différencier ces deux sens est de se référer à la syntaxe. Si le verbe *aller* est suivi d'un autre verbe il s'agit d'une action en cours, alors que s'il est suivi d'un objet il symbolisera un déplacement (3.9).

Pour chaque règle proposée, une vérification sur un corpus de référence est menée afin de confirmer les ambigüités correctement résolues, et ainsi ne conserver que la règle la plus efficace. Ce procédé à l'avantage d'être rapide, il a été appliqué sur les hypothèses de cadres précédemment produites, après quelques itérations, les 6 règles produites ont permis de retirer la plupart des ambigüités sur les verbes les plus fréquents.

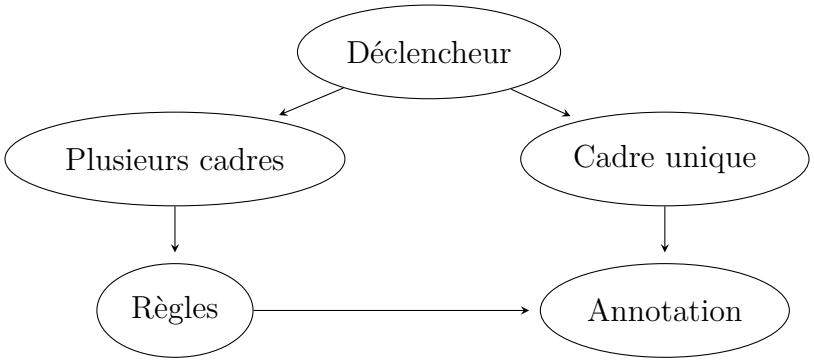


Figure 3.10. – Schéma du processus de désambigüisation

La figure 3.10 décrit le processus de désambigüisation. Le processus de sélection de cadres consiste maintenant à vérifier pour chaque déclencheur s'il est porteur d'ambigüité ou non. Si c'est le cas, la ou les règles associées lui sont appliquées. Sinon le cadre est annoté avec la valeur correspondante dans le dictionnaire.

3.2.4. Évaluation

Un sous corpus de référence a été défini et annoté manuellement en cadres sémantiques pour évaluer le système précédemment décrit (voir la figure 3.10). Le corpus de référence est une sélection de 21 conversations sélectionnées de façon à couvrir et respecter la distribution en thème de l'ensemble du corpus pour garder un maximum de cadres différents. Ces conversations ont été annotées par un annotateur unique. Les tableaux 3.3 et 3.4 donnent une représentation de la distribution des cadres sur ce sous-corpus, tout en comparant l'annotation manuelle et automatique.

Le tableau 3.3 montre qu'en moyenne un tour de parole contient au moins un déclencheur. Ces résultats mettent en évidence que l'annotation automatique à tendance à prédire plus de déclencheurs que l'humain, et trouve plus de variabilité dans les cadres choisis. Le tableau 3.4 permet de retrouver les thèmes principaux du corpus au travers des cadres prédits, et conforte les décisions prises dans le choix du découpage du sous-corpus d'évaluation. Les cadres les plus présents dans l'annotation manuelle et automatique sont les cadres *Agreement* et *Hello* ce qui représente la structure de l'appel (qui est donc présent dans l'ensemble des appels) alors que les cadres comme *Request, motion* ou *Commerce_buy* donnent des indications quant au thème et aux raisons des appels.

La qualité de l'annotation automatique sur le sous-corpus de référence est présentée dans le tableau 3.5. Il y a différentes façons d'évaluer ces résultats, de

	Annotation manuelle	Annotation automatique
Cadres par conversation	23.67	31.33
Cadres par tour de parole	0.97	1.24
Cadre différents	26	37

Table 3.3. – Distribution des cadres dans le sous-corpus de référence

Annotation manuelle		Annotation automatique	
Noms du cadre	Occ	Noms du cadre	Occ
Agreement	161	Agreement	216
Hello	95	Hello	95
Judgment_direct_address	59	Motion	45
Motion	33	Communication	34
Request	21	Judgment_direct_address	27
Waiting	20	Desiring	20
Awareness	18	Awareness	19
Communication	15	Intentionally_affect	16
Losing	14	Possibly	12
Commerce_buy	9	Waiting	11

Table 3.4. – 10 cadres les plus utilisés dans le sous-corpus de référence.

façon très précise en vérifiant les déclencheurs, les cadres choisis, les rôles sémantiques et les remplisseurs de rôles, ou de façon moins stricte en ne regardant que les cadres. L'évaluation menée ici est opérée sur le choix des cadres, ou en d'autres termes, si un déclencheur a bien été trouvé et s'il est associé au bon cadre sémantique ou non. Les résultats de cette évaluation sont donnés dans le tableau 3.5.

	Rappel	Précision	F-mesure
Annotation automatique	83.33	94.54	88.58

Table 3.5. – Évaluation de l'annotation automatique sur le sous-corpus de référence.

Ces résultats sont satisfaisants au niveau de la précision étant donné que 94.5% des cadres prédits sont corrects. En ce qui concerne le rappel, le score obtenu s'explique par le fait qu'un sous-ensemble de déclencheurs a été utilisé pour l'annotation (200 verbes les plus fréquents), mais aussi par le fait que l'annotation manuelle a été orientée, dans le sens ou une même ancre dans le même contexte ne va pas forcément toujours déclencher de frame, cela va dépendre de l'utilité de ce frame pour la conversation estimée par l'annotateur. Une solution

pour améliorer ces résultats serait d'une part d'augmenter le nombre de déclencheurs possibles, sans forcément se limiter aux verbes, et d'autre part d'affiner et agrandir l'annotation manuelle.

Néanmoins comme le montre le tableau 3.6, sur les cadres les plus fréquents, et/ou les plus porteurs de sens dans nos conversations, les résultats sont en général supérieurs à la moyenne globale.

Cadre	Rappel	Précision	F-mesure
Losing	100	100	100
Hello	98.81	96.43	97.60
Waiting	92.86	96.43	94.61
Motion	94.71	89.07	91.80
Request	90	87.5	88.73
Locating	83.33	93.75	88.23

Table 3.6. – Évaluation de l'annotation automatique sur le sous-corpus de référence pour des cadres sémantiques spécifiques.

Le cadre *Losing* apparaît dans 5 conversations de test et est le cadre le mieux annoté étant donné qu'il est systématiquement trouvé et bien annoté par rapport au corpus de test. Ces résultats ne sont pas inattendus vu que le cadre "Losing" est très peu ambigu dans le contexte des transports, et les ancres qui lui sont associées sont souvent réutilisées (*prendre, oublier, égarer* ...). Le cadre spécialement créé pour ce corpus *Hello* est aussi très bien représenté. Il s'agit là aussi d'un résultat attendu et qui signifie aussi que plus de 97% des conversations commencent par un "Bonjour" ce qui traduit une certaine forme de structure, mais aussi de politesse. Les résultats les plus intéressants sont les annotations du cadre *Motion*. En effet avec presque 92% de f-mesure cela signifie que les règles de désambigüisations utilisées sont utiles et marchent relativement bien. Enfin en bas de tableau la cadre *Locating* qui tout de même atteint 88% de f-mesure, n'est pas toujours bien identifié. Ce phénomène, retrouvé dans les résultats globaux, est principalement dû au fait que les ancres qui déclenchent ces cadres sont variées et la sélection de déclencheurs choisie pour la tâche ne couvre pas la totalité des cas rencontrés.

Une fois ces annotations produites, il est maintenant nécessaire de déterminer leur utilité et leur lien vis-à-vis des synopsis disponibles et plus généralement par rapport à la tâche de résumé.

3.2.5. FrameNet et synopsis

Cette partie décrit la place qu'occupent les cadres sémantiques dans les résumés. Dans quelle mesure ces cadres se retrouvent dans les résumés produits par

les humains, et ces informations sont-elles redondantes entre les résumés, les conversations et les annotations produites.

3.2.5.1. Cadres sémantiques et résumés

Les cadres sémantiques se retrouvent fréquemment dans les conversations dans les sens où ils ont été choisis à partir des conversations elles-mêmes afin d'en extraire le sens et de les comprendre. Le but final est d'utiliser ces cadres afin de récupérer les informations qu'ils portent et de les transposer sous forme de synopsis. Afin de conforter cette approche, il est nécessaire de regarder leur présence dans les synopsis.

Là figure 3.11 montre comment les situations de liens entre un tour de parole de la conversation et un synopsis correspondant peuvent apparaître.

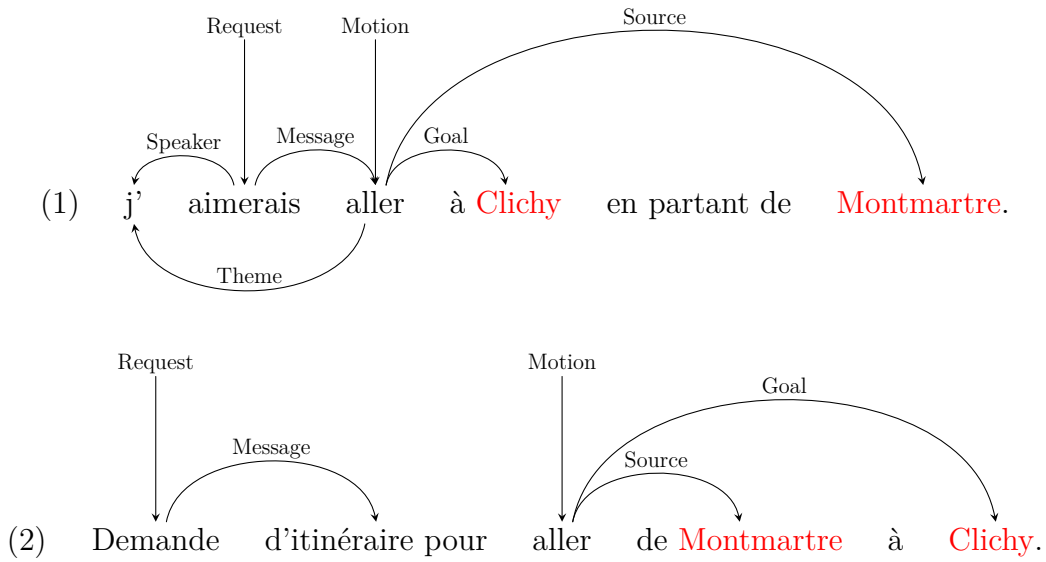


Figure 3.11. – Exemple d'utilisation des cadres sémantiques ((1) : tour de parole, (2) : synopsis, en rouge : entités nommées annotées)

L'exemple 3.11 montre bien l'utilité que peuvent avoir les cadres sémantiques dans la génération de synopsis. Les cadres présents dans la conversation induisent des informations qui seront présentent dans le résumé. Le cadre *Request* sous-entend que la conversation correspond à une demande particulière, ici une demande pour se déplacer, cette idée se retrouve directement dans le résumé sous la forme d'une *demande d'itinéraire*. Le déplacement est ensuite spécifié par le cadre *Motion* qui donne notamment les informations de point de départ et d'arrivée de la demande, ce qui se traduit dans le résumé par *l'itinéraire* entre

les deux lieux identifiés par des entités nommées. Cet exemple met en avant une force de l'utilisation de cadres sémantiques, il s'agit du fait que les entités nommées sont spécifiques et précises, l'une est le lieu de départ et l'autre la destination, cette attribution de rôle n'aurait pas été possible avec de l'extraction de texte.

Cependant en pratique l'analyse des synopsis montre que la majorité des cadres trouvés dans les conversations ne se retrouve pas dans les synopsis directement. En effet cela s'explique par le style de texte largement différent et le manque de déclencheurs dans les synopsis. Par exemple le cadre *Request* va, selon l'annotateur, se retrouver systématiquement dans le synopsis étant donné que les conversations étudiées correspondent à des appels ayant pour but une demande d'information, mais le déclencheur ne se retrouvera pas forcément dans la conversation où la requête sera définie implicitement par l'appel du client.

Même si les cadres ne se retrouvent pas directement dans les synopsis comme espéré, l'annotation sémantique n'est pas inutile pour autant. Cette annotation reste un outil relativement efficace dans la recherche et l'extraction de concepts pertinents dans la génération de résumés. L'identification des déclencheurs permet de comprendre ce qui à pu se dire durant l'appel. Par exemple une conversation où les cadres *Hello, Motion, Waiting,* et *Commerce_buy* symbolisent une conversation qui traite d'un déplacement de personne (très probablement un itinéraire étant donné le contexte du corpus), accompagné d'une demande de tarif.

3.3. Conclusion

Nous avons présenté dans cette section une méthode d'annotation rapide pour un corpus de conversations parlées issues de centres d'appels. Ce procédé semi-supervisé utilise l'annotation en dépendance syntaxique en conjonction avec un modèle sémantique FrameNet. Un module de décision à base de règle permet de filtrer et supprimer les ambigüités lors de la génération de cadres. Le thème restreint et la nature de notre corpus d'étude nous permettent de générer cette annotation sémantique de façon rapide et peu coûteuse, tout en gardant un indice de confiance convenable pour l'utiliser dans la suite de nos travaux.

Cette couche d'annotation supplémentaire nous permet de nous pencher sur la question de la génération de résumé par abstraction. En effet l'utilisation de cette annotation en sens nous permet de comprendre une conversation donnée et d'en extraire les informations nécessaires à sa compréhension et ainsi de pouvoir générer des synopsis. Pour la génération de ces synopsis nous avons décidé d'explorer une méthode à base de remplissage de patrons comme présentée par

GENEST et LAPALME 2013 mais dont la partie générique (*annotation sémantique*) et la partie dépendante du domaine (patron) sont factorisées.

On se propose alors d'explorer cette voie et de généraliser cette approche en résumé par remplissage de patrons dans le chapitre suivant.

4.Résumé par abstraction

4.1. Introduction

Le résumé automatique de document repose généralement sur des méthodes par extraction qui sélectionnent dans le texte des passages pertinents et les juxtaposent pour former un résumé. Ces méthodes sont peu adaptées à la problématique du résumé de conversations orales de par la nature spontanée de celles-ci et l'importance de l'interaction entre les locuteurs. En ne sélectionnant que certains passages, les résumés par extraction ne contiennent qu'un verbatim de ce qui a été dit, et non pas une description synthétique de ce qui s'est passé lors de la conversation.

Par exemple, dans le domaine des centres d'appels, il serait souhaitable que les résumés générés renseignent du problème de l'appelant et comment ce problème a été pris en charge par l'agent ayant traité l'appel. Il n'est pas rare que l'appelant décrive son problème sur plusieurs tours de parole ponctués par des demandes de confirmation ou de reformulation de la part de l'agent, ce qui est difficile à caractériser à l'aide de méthodes extractives lorsque la taille des résumés est fortement contrainte.

Les méthodes de résumé par remplissage de patron ont montré leur intérêt dans des domaines spécifiques pour le résumé automatique de texte WHITE, KORELSKY, CARDIE et al. 2001. Dans notre cas, elles permettent de traiter du problème de différence de style d'écriture entre les données sources (transcriptions de conversations) et la forme des résumés à générer (narration synthétique). Le résumé par remplissage de patron se fait en deux étapes principales : la détection des informations, et le remplissage de patrons (cf section 1.5.2). Les annotations en concepts, et la production de patrons peuvent être longues et coûteuses si elles sont réalisées manuellement.

Nous proposons d'explorer des méthodes pour le résumé par remplissage de patrons de conversations qui nécessitent moins de supervision de la part d'un expert humain. Nos contributions sont les suivantes :

— l'extraction de concepts à partir de traits issues de l'analyse sémantique et dialogique de surface des conversations, par opposition à une analyse sémantique complète ;

— le transfert des annotations de ces concepts depuis des résumés manuels aux conversations par alignement sémantique, minimisant ainsi le cout d'annotation ;

— la génération dynamique de patrons à partir d'exemples de résumés de référence et des informations détectées dans une conversation ;

— un ensemble d'expériences validant l'approche sur la tâche de génération de synopsis du corpus DECODA.

Le reste de cette section est organisé de la manière suivante : la section 4.2 décrit notre méthode d'extraction d'information et de génération de résumés, la section 4.6 décrit le cadre expérimental et les résultats de notre étude, et enfin la section 4.7 discute ces résultats et dresse une conclusion.

4.2. Méthode générale

Étant donné une conversation il est souvent assez facile pour un humain de trouver les informations importantes et de créer un patron pour rassembler ces informations et ainsi créer un résumé de l'appel. On appelle alors synopsis un tel résumé.

L'approche pour le résumé de conversations proposée dans cette section est basée sur le remplissage de patrons textuels avec des parties variables remplies lors de l'analyse des différents tours de parole des conversations. On appelle patron (ou template) un texte possédant des parties variables. Les textes générés à partir de tels patrons et des informations extraites des conversations sont les synopsis représentant les résumés recherchés dans cette thèse. Les patrons standards sont généralement écrits à la main une fois pour toutes et ne peuvent alors pas être modifiés pour convenir aux spécificités de la conversation traitée. L'originalité de l'approche proposée vient du fait que l'extraction des variables se fait directement sur un corpus de synopsis. Le remplissage des parties variables des patrons en découle.

L'écriture de patron étant un processus long et coûteux pour pouvoir couvrir un nombre de cas suffisants, on peut alors tirer parti du fait que les conversations se ressemblent et essayer de réutiliser leurs synopsis comme patrons pour d'autres conversations.

Cette étape de production de patrons utilise deux types de séquences de mots extraits du corpus de résumé : des *variables de patron* représentant les concepts spécifiques à la conversation traitée, et des séquences de mots génériques qui peuvent s'appliquer à différentes situations. La figure 4.1 donne un exemple de patron pour des conversations issues d'un corpus enregistré dans un centre d'appel.

Exemples de résumés de conversations
- *Écharpe* oubliée dans *bus 140*, mais rappeler à *onze heures* quand le service sera ouvert - La cliente a oublié *son écharpe* dans le *bus 140* à *Colombes le soir précédent*. L'agent répond que le service n'est pas ouvert avant *11h*. - Demande de renseignement sur la perte d'une *écharpe* dans le *bus 140*. Attendre l'ouverture du service et rappeler plus tard.
Patron possible
Le client à oublié $OBJET dans le $TRANSPORT. Objet non retrouvé, rappeler plus tard.

Table 4.1. – Exemple de patron pour le thème objet perdu. Les variables de patron sont $OBJET et $TRANSPORT.

L'approche consiste donc en plusieurs étapes :

— La détection de concept : les transcriptions de conversation et les synopsis associés de l'ensemble d'entrainement sont analysés afin de détecter les variables de patron correspondant aux concepts pertinents pour caractériser les conversations. La liste des concepts utilisés est en lien direct avec le domaine du corpus cible ;

— Génération de phrases de patron : toutes les phrases des synopsis du corpus sont généralisées en remplaçant les valeurs des concepts par leur étiquette générique afin de produire des phrases de patron. Des exemples de ces patrons sont visibles dans le tableau 4.2 pour trois thèmes différents : *Itinéraire*, *Navigo* et *Objet_perdu.* ;

Thème	Patrons
Itinéraire	Demande d'itinéraire (en $TRANSPORT) ? de $FROM à $TO (sans utiliser $NOT_TRANSPORT) ?.
Navigo	Demande de (reçu‖remboursement‖tarifs) pour $CARD_TYPE. Le client doit se rendre en agence à $ADRESS.
Objet_perdu	$ITEM perdu dans $TRANSPORT (à $LOCATION) ? (vers $TIME) ?. (Objet retrouvé, à récupéré à $RETRIEVE_LOCATION ‖ Objet non retrouvé.)

Table 4.2. – Exemple de patron pour les thèmes Itinéraire, Navigo et objet perdu.

— Liaison des concepts : cette tâche consiste à relier les concepts trouvés dans le résumé aux mêmes concepts détectés dans la conversation associée. Un classifieur est entrainé à prédire, pour tous les concepts détectés dans une conversation, s'ils apparaîtront dans le résumé correspondant ou non.

Une fois les patrons et le classifieur de liaison obtenus, le processus de résumé d'une nouvelle conversation se fait comme suit :

— Détection de concepts pertinents : Les concepts sont tout d'abord transférés de l'annotation des synopsis vers les conversations associées, puis le classifieur de liaison apprend à détecter les concepts *pertinents* dans la transcription de la conversation. Un concept est dit *pertinent* s'il est susceptible d'apparaître dans le synopsis de la conversation.

— Sélection de phrases de patron : cette étape consiste à choisir dynamiquement les phrases de patron, parmi toutes les phrases disponibles, en fonction des concepts détectés dans l'étape précédente.

— Génération de synopsis : toutes les phrases de patron sont remplies avec les valeurs des concepts identifiés dans la conversation ; puis ces phrases sont ordonnées pour produire le synopsis final.

Ces étapes sont décrites dans la figure 4.1 et détaillées dans les sections suivantes.

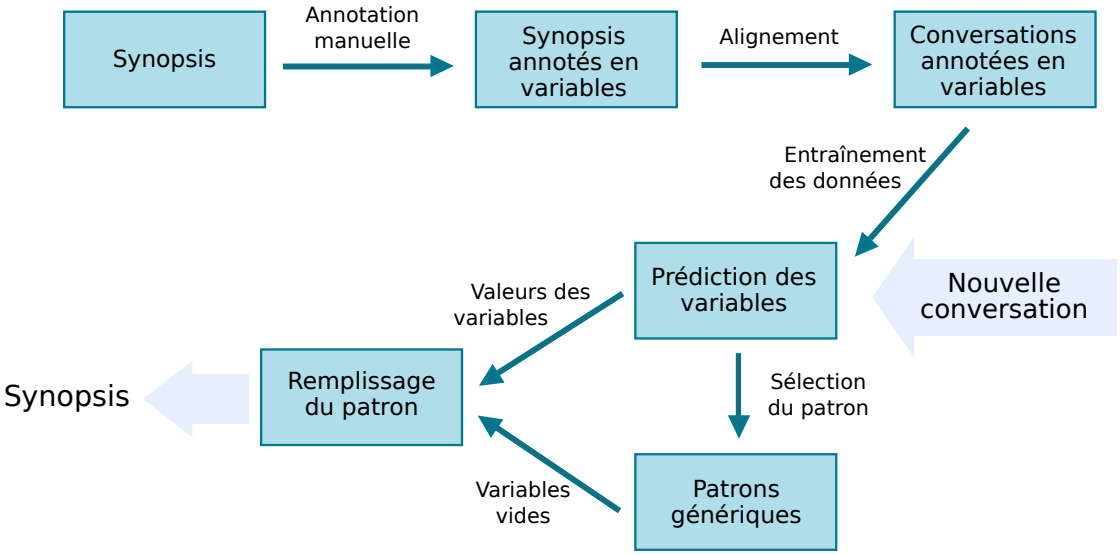


Figure 4.1. – Schéma de l'approche pour le résumé par génération de patrons. Une cohorte de patrons de phrases est générée depuis un corpus d'apprentissage, puis des patrons sont sélectionnés en fonction des variables qui ont été détectées dans le document traité. Le résumé est généré par remplissage de ces patrons et leur juxtaposition.

4.3. Cadre expérimental

Nous utilisons dans cette étude le corpus RATP-DECODA présenté dans la section 2.3.

Comme mentionné dans le chapitre 2 un synopsis correspond à une succession d'évènements entre l'appelant (le client) et l'agent (ou plusieurs agents), il doit contenir une description des besoins ou problèmes du client et comment ces problèmes ont été résolus. Le tableau 4.3 montre quelques exemples de synopsis issus du corpus DECODA BECHET, MAZA, BIGOUROUX et al. 2012.

Conversation	Ann.	Résumé
20091112-0042	1	Est-ce que les bus 172 et 186 circulent ? Non, trafic interrompu à cause de la grève du dépôt de Vitry.
20091112-0042	2	Demande d'information sur la circulation des bus 172 et 186. Grosse perturbation sur ces lignes à cause d'une grève. Plainte de l'appelant.
20091112-0604	1	Itinéraire banlieue, ayant essayé par le site ratp mais pas convaincue du trajet proposé. Récapitulatif du trajet par l'appelant.
20091112-0604	2	Demande d'itinéraire pour aller à la gare de Chilly-Mazarin en partant de la gare de Fontenay-sous-bois. Prendre le RER A en direction de Saint Germain en Laye jusqu'à la gare de Lyon, puis prendre le métro 14 direction Olympiade, descendre à bibliothèque, enfin prendre le RER C jusqu'à Chilly Mazarin. Communication de la durée du voyage et de la fréquence de passage.

Table 4.3. – Exemple de résumés issus du corpus DECODA. Pour chaque conversation, les résumés de deux annotateurs (Ann.) sont donnés.

Ces synopsis montrent aussi une certaine structure dans leur construction, ils présentent le ou les problèmes rencontrés par l'utilisateur, et décrivent les propositions de solutions que l'agent introduit. La construction de ces synopsis par les humains peut se traduire par un remplissage de patron avec les informations importantes qui auront été sélectionnées tout au long de la conversation. Il s'agit là du cœur de l'approche : détecter les informations pertinentes d'une conversation et créer des patrons de synopsis capables d'accueillir ces informations afin de constituer un synopsis final.

Il est évident qu'il est impossible de créer un patron unique pour chaque cas rencontré. En revanche les régularités retrouvées dans les synopsis écrits manuellement permettent de proposer un certain nombre de patrons capables de

généraliser un groupe de cas. L'écriture de ces patrons doit être faite de telle sorte qu'un patron maximise le nombre de synopsis couverts.

Pour rendre les patrons à la fois génériques et modifiables, chaque synopsis a été défini comme une expression régulière. Ces expressions régulières comprennent des regroupements, des quantifieurs (représentés par un point d'interrogation pour zéro ou une fois, et l'étoile de Kleene pour les répétitions), des alternatives, et des variables (de la forme *dollar suivi du nom de la variable en majuscule*). Chaque patron peut alors générer plusieurs synopsis différents en fonction des informations recueillies dans la conversation, ce qui permet d'augmenter le nombre de conversations couvertes par un même patron.

Le tableau 4.4 donne des exemples de patrons sous forme d'expressions régulières.

Thème	Patron
Planification	Demande d'horaire (en $TRANSPORT) ? de $FROM à $TO.
Itinéraire	Demande d'itinéraire (en $TRANSPORT) ? de $FROM à $TO (sans prendre $NOT_TRANSPORT) ?. (Prendre la $LINE (en direction de $TOWARDS) ? de $START_STOP à $END_STOP)*.
Carte Navigo	Demande de (justificatif | remboursement | reçu | d'informations) pour $CARD_TYPE. Le client doit se rendre à $ADRESSE.
Objets perdus	$ITEM perdu dans $TRANSPORT (à $LOCATION) ? (vers $TIME) ?. (Objet retrouvé, et à récupéré à $RETRIEVE_LOCATION | Objet non retrouvé).
Tarif	Demande de tarif pour se rendre de $FROM à $TO. Le cout est de $BUY.
Trafic	Demande d'information sur l'état de $TRANSPORT. (La fréquence de passage est de $FREQUENCY | Trafic interrompu à cause de $ISSUE | Impossible de donner d'information à cause de $ISSUE)

Table 4.4. – Exemples de patrons créés manuellement en utilisant un formalisme de langage régulier.

Au total 175 synopsis ont été annotés en variables de patrons utilisant la méthode décrite au-dessus. Parmi ces synopsis 17 types de variables différents ont pu être créés (cf tableau 4.5) afin de rester génériques tout en gardant une certaine forme de précision dans la construction du synopsis de base.

Les variables de patron jouent un rôle capital dans les patrons, ce sont elles qui portent l'information de la conversation. Le tableau 4.5 montre les variables utilisées et leur couverture sur les synopsis annotés.

Variable	% Synopsis	Variable	% Synopsis
$TRANSPORT	42.29	$TO	27.43
$FROM	25.14	$CARD_TYPE	25.14
$INFO_TARGET	22.29	$ITEM	20.0
$ISSUE	18.29	$LINE	8.0
$LOCATION	5.14	$BUY	4.57
$TOWARDS	2.86	$END_STOP	2.86
$TIME	2.29	$NOT_TRANSPORT	2.29
$START_STOP	0.57	$FREQUENCY	0.57
$RETRIEVE_LOCATION	1.14		

Table 4.5. – Répartition des variables de patrons dans les synopsis.

Comme le domaine restreint du corpus pouvait le laisser penser, certaines variables sont très répandues dans les conversations. La variable $TRANSPORT par exemple apparaît dans presque la moitié des synopsis annotés.

Il est important de noter que la quasi-totalité des variables utilisées ($FROM, $START_STOP, $END_STOP, $TOWARDS, $RETRIEVE_LOCATION, $LOCATION et $TO) sont des variables de type entités nommées, ce qui complique la tâche de différenciation lors de l'annotation par prédiction.

Une fois les patrons définis à partir des synopsis, il est nécessaire de les aligner avec les conversations qu'ils représentent afin de faire la correspondance entre les concepts exprimés dans les transcriptions de conversation et les variables des patrons.

Cette étape est décrite dans la section suivante.

4.4. Détection des variables de patrons

Cette partie consiste à trouver un candidat dans une conversation capable de prendre la place d'une variable de patron dans un synopsis. Une hypothèse serait que les variables de patrons sont des entités nommées. Comme vu précédemment, dans de nombreux domaines, elles sont très largement présentes et jouent un rôle particulier pour lier un résumé à la réalité qu'il décrit. Mais une détection simple d'entités nommées n'est pas suffisante. En effet dans le cas d'une conversation dont le thème est une demande de trajet d'un point A à un point B, il

faut pouvoir être capable de déterminer parmi les deux lieux lequel correspond à l'adresse de départ et lequel est celle d'arrivée. Les variables à détecter doivent être *pertinentes* et *précises* pour un patron donné.

D'autre part, certaines informations nécessaires pour remplir les patrons ne sont pas des entités nommées, mais, par exemple, des objets génériques, des actions ou encore des situations. C'est le cas par exemple du scénario de perte d'objets, dans lequel le type d'objet perdu, et le fait que l'objet ait été retrouvé ne sont pas des entités nommées. Pour couvrir ces types de variables, une solution serait d'avoir recours à d'autres types d'annotations comme l'annotation en cadre sémantique (BAKER, FILLMORE et LOWE 1998), ou encore d'avoir recours à un système d'apprentissage automatique capable de retrouver directement les variables dans la conversation sans passer par une représentation intermédiaire. C'est cette dernière méthode qui sera explorée.

L'approche consiste à aligner les variables de patron des synopsis avec les transcriptions de conversations afin d'entrainer un classifieur permettant de prédire ces variables dans de nouvelles conversations dans le but de produire un résumé, comme décrit dans la figure 4.2.

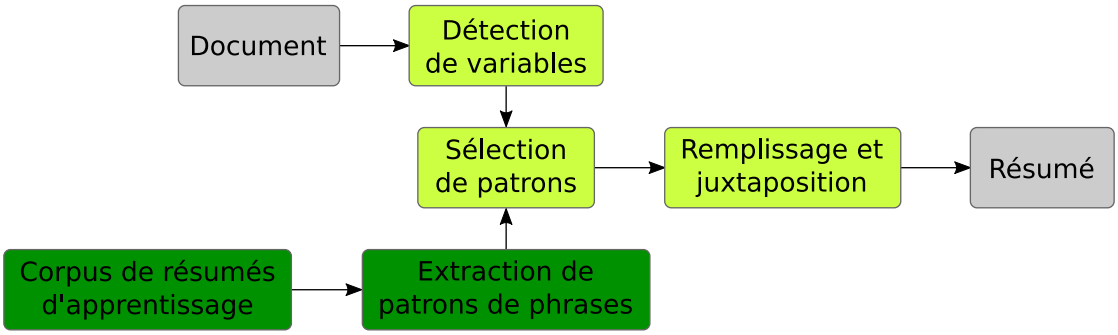


Figure 4.2. – Processus de détection des variables de patrons.

Contrairement à une tâche classique de recherche de concepts dans une conversation, l'objectif ici n'est pas de classifier l'ensemble des variables potentielles de la conversation, mais plutôt d'identifier uniquement les variables (*concepts*) qui seront présentes dans le synopsis final. Il s'agit d'une double classification qui répond à deux problèmes :

— s'agit-il d'une variable de patrons ?

— cette variable sera-t-elle présente dans le synopsis final ?

4.4.1. Annotation par propagation depuis les résumés

Le principe de la phase de propagation est de retrouver dans les transcriptions d'une conversation les séquences de mots qui correspondent à des variables annotées dans les synopsis. Pour effectuer ces alignements, un alignement du contexte de chaque variable est effectué avec les séquences de mots des documents sources, annotées syntaxiquement pour obtenir un contexte de correspondance plus riche. Les annotations syntaxiques sont générées par la chaine de traitement de Macaon adapté au traitement de l'oral (BAZILLON, DEPLANO, BECHET et al. 2012) décrit dans la section 2.3.2.

L'alignement synopsis-conversation suit l'algorithme suivant :

— Les transcriptions de conversations sont tout d'abord analysées syntaxiquement avec Macaon et tous les groupes nominaux sont extraits ;

— Chaque variable d'un résumé annoté est comparée à l'ensemble des groupes nominaux de la transcription de la conversation correspondante grâce à une fonction de coût utilisant une distance de Lenvenstein ;

— Le groupe nominal de plus faible coût est associé à la variable correspondante.

Cette phase d'alignement des concepts entre les synopsis et les conversations ne s'appuie pas sur l'annotation en cadres sémantiques proposée dans le chapitre 3, car comme vu dans la section 3.2.5.1, la majorité des cadres sémantiques des conversations ne se retrouve pas dans les synopsis, du fait du style d'écriture de ces derniers, ne favorisant pas l'apparition de déclencheurs.

L'algorithme 1 présente une version plus détaillée de la fonction de coût utilisée.

Pour ce qui est de l'approximation de la correspondance entre les mots, des choix de normalisation ont été effectués, il s'agit de mettre tous les mots en minuscule et de supprimer les accents. Ces changements permettent d'appliquer une distance de Levenstein basée sur les caractères. Cet alignement entre les variables de synopsis et les conversations a permis d'identifier 316 variables dans les conversations sur un total de 380 annotées dans les synopsis, soit un taux de 83.16% d'alignement sur 141 conversations.

Les variables de patron non alignées sont dans la majorité des cas dus à des erreurs d'annotation (par exemple des erreurs d'orthographe dans les noms de rues), à des variables trop générales qui ne peuvent être détectées au niveau des mots ou encore à une absence de correspondance entre les variables et la transcription (c'est le cas lorsque l'annotateur a trop généralisé les évènements lors de la rédaction du synopsis, par exemple avec l'introduction de nouvelles

Algorithm 1 Alignement synopsis-conversation

```
1: procedure LIER_SYNOPSIS()
2:     for Variable d'un synopsis annoté do
3:         for Phrase de la conversation associée do
4:             Aligner la variable avec la conversation en utilisant une distance de
   Lenvenstein avec une fonction de coût.
5:
6: procedure FONCTION_SUBSTITUTION()          ▷ coût d'une substitution
7:     if correspondance des mots approchée then return 0
8:     if mot ∈ Stopwords then return 0.1
9:     if Autres mots then return ∞
10:
11: procedure FONCTION_INSERTION()            ▷ Insérer un mot du synopsis
12:     if mot ∈ Stopwords then return 1
13:     if Autres mots then return ∞
14:
15: procedure FONCTION_SUPPRESSION()  ▷ Insérer un mot de la conversation
16:     if mot ∈ Stopwords then return 0.1
17:     if Autres mots then return ∞
```

informations comme "*itinéraire dans Paris*" alors que Paris n'est jamais cité dans la conversation).

Une fois les conversations annotées en variables, l'objectif est d'étendre ces annotations sur de nouvelles conversations pour un coût raisonnable.

4.4.2. Prédiction dans de nouvelles conversations

L'étape précédente a permis de créer un corpus associant des variables de patron et des expressions de concepts (sous forme de groupe nominal) dans les transcriptions de conversation. Le corpus peut alors être utilisé pour prédire directement le type de variable de chaque groupe nominal. Une nouvelle fois la structure syntaxique de la conversation donnée par la chaîne de traitement Macaon (BAZILLON, DEPLANO, BECHET et al. 2012) est utilisée afin d'obtenir 10 traits lexicaux et syntaxiques sur lesquels reposent le système d'apprentissage.

Afin de tester l'impact des différentes méthodes d'apprentissage, 3 classifieurs seront testés :

— Icsiboost (FAVRE, HAKKANI-TÜR et CUENDET 2007)

— Liblinear (FAN, CHANG, HSIEH et al. 2008)

— Perceptron multicouche (TRIONE, FAVRE et BÉCHET 2016)

Les données d'entrainement sont produites par l'annotation précédente en variables de patrons dans les conversations. Ainsi pour chaque groupe nominal de la conversation, le classifieur est entrainé à prédire le type de variable parmi les 17 disponibles (cf. tableau 4.5) plus 1 type NULL pour les groupes nominaux qui ne sont pas des variables pertinentes. Les classifieurs reposent donc sur ensemble de traits classés en 3 grandes catégories :

1. Traits syntaxiques :
   1.1. forme fléchie ;
   1.2. partie de discours ;
   1.3. lemme ;
   1.4. partie de discours ;
   1.5. étiquette de dépendance ;
   1.6. sac de n grammes de mots ($n \leq 3$) ;
   1.7. sac de n grammes de parties de discours ($n \leq 3$) ;
   1.8. longueur en mots.
2. Traits discursifs :
   2.1. nombre d'occurrences du type d'entité nommée depuis le début de la conversation ;
   2.2. nombre d'occurrences du lemme nommé depuis le début de la conversation ;
   2.3. thème de la conversation [a] ;
   2.4. position relative du groupe nominal dans la conversation ;
   2.5. rôle du locuteur (appelant ou agent).
3. Traits sémantiques :
   3.1. type d'entité nommée ;
   3.2. nom du cadre sémantique ;
   3.3. lemme du déclencheur ;
   3.4. étiquette de dépendance des éléments de cadre.

Il est bon de noter que les traits liés à la conversation et au discours ne sont pas des traits conventionnels pour l'étiquetage en entités nommées ou en concepts. Ils permettent par exemple de discriminer les lieux de départ et d'arrivée.

Les détails expérimentaux liés à l'entrainement des classifieurs sont donnés dans la section 4.6.

Une fois les nouvelles conversations annotées en variables de patrons, il est possible de construire le synopsis final.

a. parmi les thèmes annotés dans DECODA

4.5. Génération de synopsis

Cette section s'intéresse aux différentes façons d'obtenir un patron dynamique pour générer un synopsis.

L'approche proposée repose sur l'utilisation des synopsis annotés en variables de patrons déjà disponibles. L'idée est que parmi tous les synopsis préalablement annotés il existe une combinaison de synopsis ou de fragments de synopsis (par exemple des phrases issues des synopsis) capable de décrire les informations détectées dans une nouvelle conversation.

Ainsi pour obtenir ce corpus de patron général, chaque synopsis s'est vu modifié de la façon suivante :

— pour chaque synopsis annoté du corpus d'apprentissage, les valeurs des variables ont été remplacées par leur type (exemple : *le bus 144* devient *$TRANSPORT*) ;

— les synopsis ont été découpés en fragments indépendants selon les frontières de phrases (voir exemple de découpage dans le tableau 4.6).

Patron	Patron découpé
Demande de tarif pour se rendre de $DÉPART à $ARRIVÉE. Le cout est de $COUT.	- Demande de tarif pour se rendre de $DÉPART à $ARRIVÉE. - Le cout est de $COUT.

Table 4.6. – Exemple de découpage de patron

Cette approche permet d'avoir facilement un large corpus de fragment de patrons, combinables entre eux pour en créer de nouveaux. Ce choix de fragmenter les patrons se justifie aussi par le fait que dans DECODA il est possible qu'une conversation puisse aborder plusieurs thèmes (par exemple *objet perdu*, puis *itinéraire*). Le remplissage et la sélection des patrons sont alors dictés par les variables détectées dans la conversation et non l'inverse. Ainsi la génération du synopsis se fait en déterminant une combinaison de fragments de patrons maximisant la couverture en variables détectées, sous la contrainte qu'un fragment ne peut être exploité que s'il est saturé, et une variable ne peut être utilisée que par un seul fragment. Le système se laisse tout de même la possibilité d'utiliser des fragments de patrons non saturés dans le cas où aucune saturation n'est possible.

L'implémentation repose sur une sélection gloutonne des fragments de patrons. Le système itère sur la population de fragments satisfaisant les contraintes jusqu'à trouver une couverture complète des variables détectées, ou jusqu'à arri-

ver à l'ensemble vide des candidats acceptables.

L'avantage d'un tel remplissage est que ce sont les variables qui définissent directement le patron, permettant de mieux coller aux diverses situations se déroulant dans les conversations. Mais cela peut aussi être un inconvénient, car il se peut que le thème de la conversation ne soit pas retranscrit dans le patron à cause d'une mauvaise détection de variables importantes.

À titre comparatif, une méthode plus classique de résumé par patrons créés manuellement sera explorée. Cette méthode (que l'on appellera *méthode par patron manuel*) consiste à écrire des patrons statiques, et aller chercher dans la conversation la ou les valeurs décrites dans le patron choisi. Afin de limiter l'impact de n'avoir qu'un seul choix de patron, le même découpage en *phrases / fragments* de patron est proposé (voir l'annexe B).

Cette dernière méthode présente deux inconvénients majeurs, le premier, les patrons sont figés et ne peuvent donc pas forcément s'adapter à la particularité d'une conversation, le second, est lié à la qualité du détecteur de variable qui peut laisser des variables vides dans le patron, ce qui impacte la qualité linguistique du résumé généré.

Les synopsis pouvant être générés grâce aux différents processus décrits plus haut, il est maintenant nécessaire de les évaluer entre eux et par rapport à la référence pour se donner une idée de leur pertinence.

4.6. Évaluation

Cette partie compare plusieurs variantes de l'approche proposée dans ce chapitre, ainsi que les baselines décrites dans le chapitre 2. Le corpus utilisé reste le corpus de conversations RATP-DECODA.

4.6.1. Cadre expérimental

Les expériences sont menées sur 141 conversations provenant du corpus RATP-DECODA. Pour rappel, ces conversations ont été manuellement annotées en synopsis. Il en résulte un total de 381 synopsis, eux-mêmes manuellement annotés en variables de patrons typées. Les expériences sont effectuées d'une part sur des transcriptions manuelles avec une annotation linguistique de référence (syntaxique et sémantique) et d'autre part sur des transcriptions générées automatiquement avec le système du LIUM (LAILLER, LANDEAU, BÉCHET et al. 2016 ; avec un taux d'erreur mot de 35%) avec une annotation linguistique produite automatiquement par la chaine de traitement Macaon (BAZILLON, DEPLANO, BECHET

et al. 2012).

Le corpus est découpé en 71 conversations destinées à l'entrainement des modèles, 27 pour le développement et l'optimisation des paramètres d'apprentissage, et 43 conversations pour le test. On notera que ce corpus est de petite taille pour une tâche d'apprentissage.

4.6.1.1. Détails des classifieurs

Dans un premier temps une variable était considérée comme valide si elle recevait un score de probabilité supérieur à un seuil de 0.1 choisi arbitrairement. Dans le cas où, pour la même variable, plusieurs valeurs dépassent ce seuil, seule la valeur avec le plus haut score est conservée.

Dans un second temps ce seuil n'est plus fixé arbitrairement, mais est fixé de façon à maximiser le score ROUGE-2 sur le corpus de développement (avec les transcriptions manuelles).

Icsiboost Le classifieur obtient un rappel de 62% et une précision de 43% soit une f-mesure de 51% sur la tâche de classification en type de variable.

La figure 4.4 montre les variations des scores ROUGE-2 en fonction du seuil de score de probabilité du classifieur. Les courbes présentent les résultats pour les transcriptions manuelles et automatiques. Dans le cas présent il est clair que les transcriptions manuelles obtiennent de meilleurs résultats, c'est pourquoi le seuil choisi correspond au plus haut score pour la version manuelle, soit un seuil de 0.021.

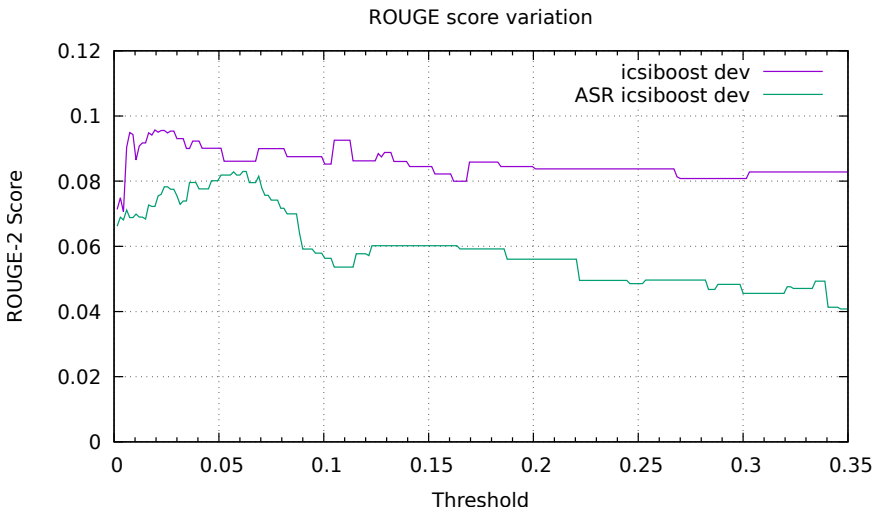


Figure 4.3. – Variation du score ROUGE-2 en fonction du seuil pour Icsiboost.

Liblinear La figure 4.4 montre ici les résultats pour le classifieur Liblinear. Contrairement à Icsiboost, les courbes représentant les résultats pour les transcriptions manuelles et automatiques sont beaucoup plus similaires. De ce fait le choix du seuil de 0.086 correspondant au score maximum pour la transcription manuelle est sensiblement le même que pour la transcription automatique. En d'autres termes, la qualité de la transcription importe peu sur les résultats produits par ce classifieur.

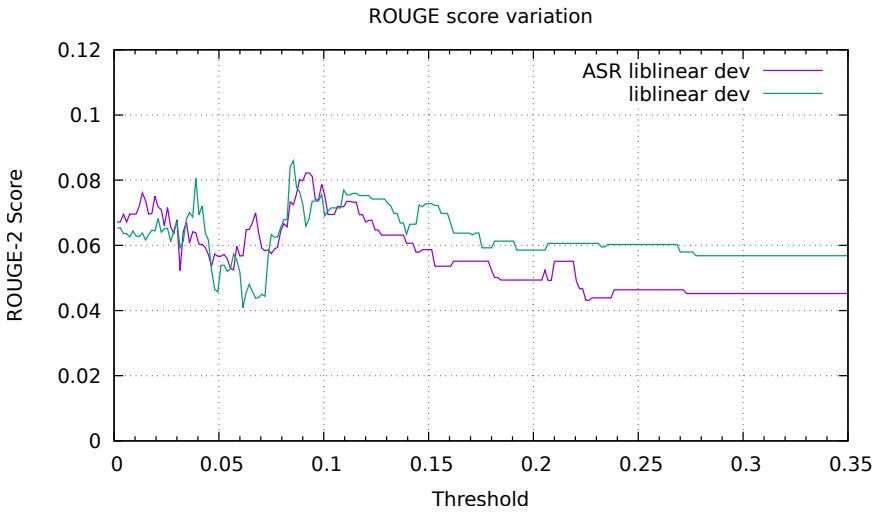


Figure 4.4. – Variation du score ROUGE-2 en fonction du seuil pour Liblinear.

Réseau de neurones Le réseau de neurones a été implémenté avec Chainer [b] selon les paramètres suivants :

— 1 couche cachée ;

— ReLu activations ;

— 4 époques d'entrainement ;

— Aucun dropout.

Ces paramètres ont été déterminés à partir d'un ensemble de configurations différentes visant à maximiser la précision sur le corpus de développement.

La figure 4.5 montre qu'avec un seuil différent pour chaque version des transcriptions (manuelle et automatique) le système serait capable d'obtenir des résultats sensiblement identiques. Cependant la démarche de se référer à la transcription manuelle comme référence force le choix d'un seuil de 0.02 sensiblement moins bon pour les transcriptions automatiques.

b. http ://chainer.org

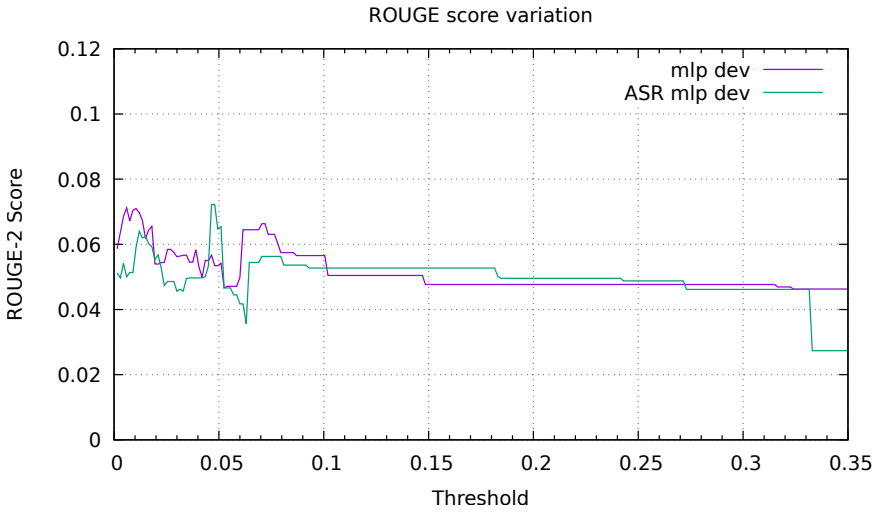


Figure 4.5. – Variation du score ROUGE-2 en fonction du seuil pour mlp.

ROUGE-2 étant la métrique la plus largement utilisée dans le domaine du résumé automatique, elle constituera la métrique utilisée ici pour évaluer les différents systèmes.

Bien que cette métrique ait ses limites, notamment lorsqu'il s'agit d'évaluer des résumés abstractifs, elle permet toutefois d'obtenir des résultats indicatifs de la ressemblance lexicale entre les résumés produits et ceux de référence. Cependant elle ne favorise pas les méthodes par abstraction, car ne traite pas les paraphrases. En revanche il est possible d'exploiter les représentations vectorielles des mots à l'aide de *word embeddings* afin d'essayer de tirer parti de ces cas de synonymie. Les systèmes présentés seront donc évalués avec ROUGE-WE (cf section 2.5.1.3) pour déterminer si cette métrique est plus significative que la métrique de ROUGE sur les mots.

4.6.2. Expériences

Les différents systèmes comparés sont les suivants :

— **Topline** : Patron rédigé manuellement et rempli avec les valeurs de référence, manuellement aussi ;

— **Humain** : La moyenne des performances obtenues en évaluant chacun des synopsis de référence (i.e écrit manuellement) par rapport aux autres références ;

— **Patrons figés** : Patrons rédigés manuellement et remplis par les variables prédites par le système ;

— **Recombiné** : L'approche définie dans ce chapitre, basée sur la recombinaison de synopsis généralisés ;

— **MMR** : maximal marginal relevance ;

— **Longest** : Le plus long tour de parole de la conversation ;

— **Longest@25** : Le plus long tour de parole dans le premier quart de la conversation.

Les résultats sont résumés dans le tableau 4.7. Les systèmes sont définis et entrainés comme décrit plus haut.

Système	Transcription	Variables	ROUGE-2	0.1	seuil
Topline	manuelle	Référence	**0.20491**	-	-
Humain	-	-	0.11848	-	-
Patrons figés	manuelle	Icsiboost		0.06467	0.06590
	manuelle	libLinear		0.03735	0.06835
	manuelle	MLP		0.02041	0.04793
Recombiné	manuelle	Icsiboost		0.05524	**0.08777**
	manuelle	libLinear		0.06382	0.08390
	manuelle	MLP		0.02041	0.04793
MMR	manuelle	-	0.03145	-	-
Longest	manuelle	-	0.02688	-	-
Longest@25	manuelle	-	0.04046	-	-
Patrons figés	ASR	Icsiboost		0.02347	0.05948
	ASR	libLinear		0.02921	0.03883
	ASR	MLP		0.01775	0.02015
Recombiné	ASR	Icsiboost		0.04756	**0.08671**
	ASR	libLinear		0.05445	0.08100
	ASR	MLP		0.02303	0.04033
MMR	ASR	-	0.02093	-	-
Longest	ASR	-	0.01734	-	-
Longest@25	ASR	-	0.01734	-	-

Table 4.7. – Résultats ROUGE-2 obtenus pour chacun des systèmes.

Tout d'abord les résultats détaillés dans le tableau 4.7 montrent clairement que les méthodes *abstractives* produisent de meilleurs résultats que les approches *extractives* ce qui conforte l'approche adoptée et le fait que la forme des synopsis joue un rôle prédominant pour la tâche. Ensuite la méthode décrite plus haut à base de recombinaison de fragments de patrons obtient globalement des résultats significativement meilleurs que la méthode basée sur l'utilisation de patrons figés écrit manuellement. Ces résultats étaient attendus étant donné qu'en

recombinant des morceaux de phrases provenant de différents synopsis, le système peut couvrir un plus large panel de situations que ne pourrait traiter un simple patron par thème. Les résultats semblent aussi dépendre grandement de la qualité des variables prédites par les classifieurs. En effet la topline est toujours loin devant avec les valeurs de références introduites dans des patrons écrits à la main. En revanche le système ayant obtenu le meilleur score se retrouve très proche, en terme de score ROUGE, des résumés humains. Ceci peut s'expliquer par le fait que les humains ont tendance à diverger lorsqu'ils écrivent des synopsis même s'il s'agit d'une même conversation. De plus il faut garder en tête que chaque résumé humain a été évalué sur une référence de moins que les systèmes, ce qui peut altérer les résultats obtenus.

Il semble aussi que le choix du classifieur influe sur la qualité des variables identifiées. Icsiboost et le CRF semblent obtenir des résultats relativement similaires, alors que le réseau de neurones MLP n'est clairement pas aussi bon que les deux premiers cités. Ces mauvaises prédictions sont probablement dues à la petite taille du corpus d'apprentissage. En effet avec si peu de données pour s'entrainer, le réseau de neurones a énormément de difficulté à trouver une généralisation pour prédire convenablement les variables de patrons (à noter aussi que le réseau a été optimisé sur le corpus de développement pour la tâche).

L'ajout d'un seuil de décision a été aussi bénéfique. Il y a clairement des améliorations visibles entre la version avec le seuil fixé arbitrairement à 0.1 et la version avec le seuil optimisé sur le corpus de développement, à la fois sur les transcriptions automatiques et manuelles.

Le choix des *features* est aussi un facteur important. Le tableau 4.8 présente les résultats obtenus en utilisant différentes combinaisons de *fetures* parmi celles décrites dans la section 4.4.2 avec icsiboost.

Transcriptions	*features*	ROUGE-2
Manuelle	Syntaxiques	*0.07025*
	Syntaxiques + dialogiques	0.07820
	Syntaxiques + sémantiques	0.07422
	Syntaxiques + dialogiques + sémantiques	**0.08777**
ASR	Syntaxiques	*0.06555*
	Syntaxiques + dialogiques	0.06136
	Syntaxiques + sémantiques	0.08185
	Syntaxiques + dialogiques + sémantiques	**0.08671**

Table 4.8. – Résultats ROUGE-2 obtenus pour chacun des jeux de *features* sur le système Icsiboost.

On peut observer que l'ajout respectivement de *features* dialogiques et séman-

tiques permet d'améliorer significativement les scores ROUGE-2 par rapport à l'utilisation de représentations syntaxiques uniquement. Ce gain de performance est d'autant plus important lorsque les jeux de *features* sont combinés tous ensemble. L'utilisation de ces représentations génériques traduit une meilleure compréhension de la conversation, ce qui tend à prouver l'utilité des ressources dialogiques et sémantiques dans la génération de résumés.

Une observation intéressante concerne les résultats obtenus sur les transcriptions automatiques avec les annotations linguistiques automatiques car il semble que les erreurs n'aient que peu d'effet sur les résumés produits. Cela provient que les variables de patrons sont généralement répétées plusieurs fois dans la conversation à la fois par le client et l'agent. Ainsi, même si une portion du message n'est pas correctement transcrite, il est probable que l'entité perdue soit présente ailleurs dans une autre portion qui elle sera correctement transcrite.

4.6.3. Analyse

La mesure ROUGE, même si elle a le mérite de pouvoir comparer facilement de nombreux systèmes entre eux, sous réserve de disposer de résumés références, ne permet pas de rendre compte de l'utilité des résumés produits, et atteint vite ses limites tout particulièrement dans le cadre de la production de résumé abstractif. En effet ROUGE ne prend pas en compte certains phénomènes propres au résumé abstractif comme le changement de style d'écriture, ou encore la synonymie, mais suppose que les quelques résumés de références sont représentatifs des termes possibles. C'est pourquoi ROUGE Word Embeddings J.-P. NG et ABRECHT 2015 a été développé. Il utilise le même procédé d'évaluation que le ROUGE classique, mais en prenant en entrée des embeggings de mots et non les mots eux-mêmes. Ainsi il serait possible grâce à cette représentation vectorielle des mots de palier aux problèmes cités au-dessus.

Les représentations vectorielles des mots ont été apprises sur l'ensemble du corpus de DEOCDA avec l'outil Word2Vec [c] proposé par la suite TensorFlow.

Cependant après avoir évalué les différents systèmes avec la métrique basée sur les *words embeddings* ROUGE-WE, il s'est avéré que les deux métriques étaient relativement bien corrélées, comme en témoigne la figure 4.6, ce qui ne donne pas de nouvelle vision sur l'efficacité de résumés produits. Ces résultats décevants de l'utilisation des *words embeddings* viennent peut-être du fait qu'il aurait fallu entrainer ces derniers sur un bien plus gros corpus.

c. https ://www.tensorflow.org/tutorials/word2vec

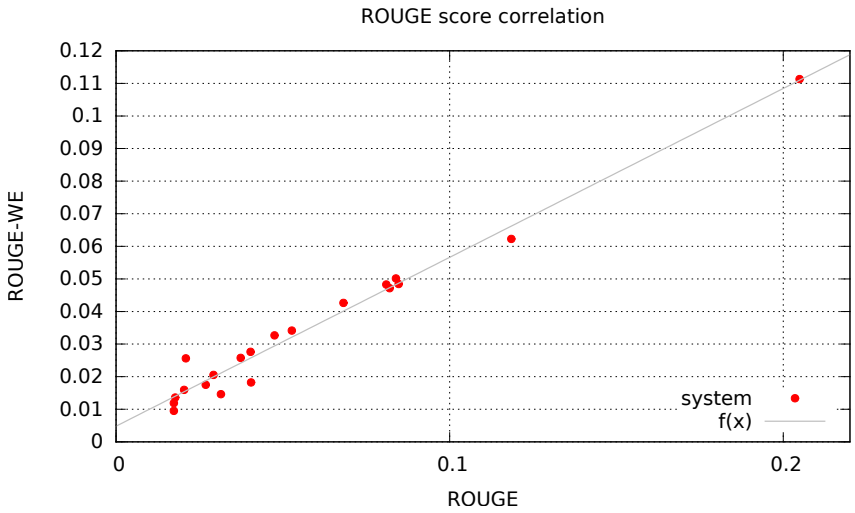


Figure 4.6. – Corrélation des scores ROUGE-2 et ROUGE-2-WE.

C'est pourquoi la métrique ROUGE est limitée, et ne répond pas à des questions comme :

— les résumés produits sont-ils lisibles ?

— sont-ils compréhensibles ?

— sont-ils utiles ?

Pour répondre à ces questions, il convient d'analyser subjectivement les résumés produits afin d'en déterminer les forces et les faiblesses.

A cet effet, dans un premier temps des exemples de synopsis sont donnés, d'abord des synopsis pour lesquels le système à produit un résultat considéré comme acceptable par les juges humains, puis des synopsis problématiques illustrant les limites du système actuel. Puis dans un second temps les résultats d'une évaluation manuelle seront donnés, pour se faire une idée de l'avis des humains sur les synopsis produits par le système.

Ci-dessous sont présentés quelques exemples de synopsis générés par le système. Dans ces exemples les variables de patrons détectées par le système sont écrites en gras :

Synopsis acceptables

- "Demande de renseignements sur l'état du trafic de **le RER** en prévision d'une grève. Demande de renseignements au sujet de **la grève ou pas**."
- "Demande de renseignements sur la perte d'un **téléphone** dans **le autobus**."
- "Demande d'itinéraire pour se rendre à **Dantzig** en partant de **Place Gambetta**. Demande d'informations sur **la grève** sur le **un autobus**."

Mis à part quelques soucis de syntaxe comme *de le* à la place de *du* ou *le un* à la place de *l'* les synopsis présentés ici sont syntaxiquement corrects, facilement lisibles et donnent une bonne approximation de ce qui a pu se dire dans la conversation. Il est important de noter que lorsque les variables de patrons sont bien détectées, le choix des patrons à remplir se fait sans connaître le thème général, or dans les exemples présentés le thème du patron sélectionné correspond aux variables identifiées. Il s'agit là de résultats encourageants, considérant les données initiales et la tâche.

Cependant le système n'étant pas parfait il génère aussi des synopsis aberrants :

Synopsis problématiques

- "Demande de renseignements sur la perte d'un **mardi à mercredi à une heure du matin** avec **bouteille thermos** dans le **la ligne quatre** en **au Montparnasse**. L'objet est retrouvé au **mardi et mercredi**."
- "Demande d'informations sur l'état du trafic du **RER B** en passant par **Bagneux**. Demande d'itinéraire pour aller à la gare de Chilly-Mazarin en partant de la **Châtelet**."
- "Demande de renseignements sur d'éventuels **horaires** sur **les RER B**. Demande du cout de **horaires** et remarque que cela devrait être signalé dans le message d'accueil.
- "Demande de renseignements sur l'état du trafic de **l'autobus** en prévision d'une grève. Objet retrouvé **un téléphone portable** avec **nom**, pas retrouvé."

Parmi les erreurs possibles, il y a les erreurs directement liées à la détection des variables comme dans le premier exemple qui présente un cas concret où les variables identifiées par le système ne sont pas les bonnes. En effet ici *mardi à mercredi à une heure du matin* a été associé à un objet perdu, de même que *mardi et mercredi* a été identifié en tant que lieu, en revanche *la ligne quatre* à bien été trouvée comme transport.

Un autre type de problème a aussi fait son apparition, il s'agit d'erreurs dans le choix du patron, et plus particulièrement dans le patron lui-même. En effet si on regarde le deuxième exemple, la notion de *gare de Chilly-Mazarin* est introduite par le patron et non par les variables, il s'agit là d'erreurs d'annotation lors de la création du corpus de fragments de patrons, on retrouve ce problème dans le troisième synopsis avec *et remarque que cela devrait être signalé dans le message d'accueil* qui ne correspond à rien de la conversation initiale. Ces problèmes liés au corpus de patrons sont souvent dus au fait que les patrons choisis par le système sont trop spécifiques d'une conversation en particulier ou tout simplement

que le patron n'a pas été correctement *nettoyé* des notions spécifiques. Ces erreurs peuvent cependant être corrigées en modifiant directement le corpus de patrons, ou en améliorant la phase de choix des fragments de patrons.

Enfin une autre source d'erreur touche le choix des fragments de patrons. Le dernier exemple montre ce type d'erreur où les variables détectées sont cohérentes entre elles, ici il s'agit d'un téléphone, d'un nom et d'un autobus, mais le choix glouton des fragments de patrons par le système fait que la couverture idéale ne correspond pas au thème et donc à la conversation, ce qui induit un synopsis faux. Un moyen de corriger ces erreurs serait de modifier l'algorithme glouton de sélection des fragments de patrons pour une méthode plus efficace.

L'analyse de ces exemples montre aussi les limites de la mesure utilisée dans le sens où la lisibilité des résumés, entre autres, n'est pas mesurée. C'est pourquoi afin de se donner une idée de l'efficacité des résumés produits dans un contexte réel, une évaluation manuelle est présentée dans la section suivante.

4.6.4. Évaluation subjective

Comme vu précédemment il existe de nombreuses façons différentes de dire la même chose dans le cadre des résumés. Cependant il est facile pour un humain de déterminer si un résumé est bien construit tant dans le fond que dans la forme. C'est pourquoi une évaluation manuelle permet de se donner une idée de la justesse et l'efficacité des résumés produits.

L'avis d'une seule personne pourrait biaiser les résultats de l'évaluation étant donné que les avis peuvent varier entre les individus. Ainsi plus le nombre de personnes impliquées dans l'évaluation est grand, plus l'évaluation elle-même aura du sens. 9 personnes ont participé à l'évaluation des résumés proposés. Ce nombre est estimé satisfaisant pour négliger le biais de l'avis personnel des individus.

L'évaluation se fait via une interface où plusieurs résumés sont donnés à évaluer après écoute de la conversation :

— 2 résumés écrits par des humains ;

— 1 résumé provenant de méthodes baselines extractives (MMR) ;

— 1 résumé à base de patron fixe ;

— 1 résumé à base de patron recombiné.

Afin de ne pas donner de structure vis-à-vis de l'ordre de proposition des résumés, ces derniers sont présentés dans un ordre aléatoire. Cet ordre aléatoire permet de limiter l'impact de la nature du résumé qui pourrait influencer l'avis

de l'évaluateur.

En ce qui concerne l'évaluation à proprement parlé des résumés, 2 aspects sont concernés : la syntaxe et la sémantique au travers de 3 questions :

— Dans quelle mesure le résumé est syntaxiquement bien construit ?

— Dans quelle mesure le résumé relate des évènements de la conversation ?

— Dans quelle mesure le résumé propose une vérité alternative ?

Ces 3 questions permettent de déterminer d'une part si le résumé évalué est syntaxiquement bien construit ou autrement dit tout simplement lisible, d'autre part si le résumé contient les informations suffisantes à la compréhension de ce qui a pu se dire dans la conversation. Enfin la dernière question répond à la question de l'introduction d'une vérité alternative, c'est-à-dire si le résumé propose des informations qui n'ont jamais été évoquées dans la conversation. Cette notion correspond au problème déjà identifié lors de l'analyse des résumés à problèmes des patrons mal annotés.

L'évaluation se fait en terme de score. Chaque évaluateur donne à chaque question pour chaque résumé un score allant de 0 à 10, ou 0 signifiant *le résumé répond négativement à la question posée* et 10 *le résumé répond positivement à la question posée*. Le choix de ne pas donner de réponse binaire se justifie par le fait que la réponse n'est pas forcément évidente, ce qui laisse le choix à l'évaluateur de pondérer sa réponse (il en résulte qu'une réponse proche de 5 signifie que l'évaluateur ne sait pas vraiment répondre à la question).

Les résultats de cette évaluation sont donnés dans le tableau 4.9.

Système	Syntaxe	sémantique	contrevérité
Humain 1	6.91	8.93	0.45
Humain 2	9.57	9.76	0.19
MMR	3.29	4.77	2.13
Patrons fixes	5.17	5.23	7.22
Patrons recombinés	6.77	5.29	7.57

Table 4.9. – Résultats de l'évaluation manuelle des résumés.

Les résultats de syntaxe (i.e lisibilité du synopsis) sur les résumés à base d'extraction étaient attendus dans le sens où il s'agit de français parlé et donc différents de la langue écrite. Cela se traduit par des scores relativement bas en terme de construction du résumé et en récupération d'informations. En revanche le résultat de *2.13/10* obtenu en vérité alternative est surprenant même s'il est bas.

La construction des résumés par extraction est telle qu'aucune information extérieure ne peut être intégrée. Ceci peut être dû à une mauvaise compréhension des évaluateurs sur la question posée.

Un résultat intéressant concerne la différence entre les deux annotateurs humains. Si les deux ont un score similaire en terme d'informations que contiennent les résumés, ils sont largement plus différents au niveau de l'écriture. L'annotateur 1 n'a reçu qu'un score de *6.91/10* là où un score proche de *10/10* était attendu comme pour l'annotateur 2.

Les seuls synopsis qui proposent une vérité alternative significative sont les résumés produits par les systèmes abstractifs. Ceci s'explique facilement du fait de leur génération. Pour rappel les synopsis sont produits à partir de fragment de plusieurs synopsis provenant de l'ensemble des données annotées disponibles, lesquels sont sélectionnés à partir des concepts identifiés dans les conversations. Ainsi plusieurs types d'erreurs peuvent être à l'origine de l'apparition de nouvelles notions. Tout d'abord une mauvaise identification des concepts de la conversation. Par exemple un lieu de départ étant identifié en tant que lieu d'arrivée mènera à la production d'un synopsis bien construit, mais qui proposera une action n'intervenant par dans la conversation initiale. Ces erreurs sont communes aux synopsis produits à la fois avec les patrons fixes et dynamiques. Ensuite une seconde catégorie d'erreur est liée directement aux fragments de patrons et ne touche que les patrons dynamiques. Du fait de leur construction, ces fragments peuvent contenir des informations rémanentes après annotation. L'apparition de ces informations est principalement due à des erreurs d'annotation. Enfin un troisième type d'erreur a été identifié touchant également les deux types de synopsis (avec patrons fixes et dynamiques), il s'agit de la *surgénération* de concepts menant à la production d'un synopsis comportant trop d'informations. Il arrive dans certains cas où le synopsis produit est correct dans une première partie, puis dans une seconde partie diffère de la réalité du fait du trop grand nombre de concepts identifiés. Chaque concept menant à la sélection d'un fragment de patron, plus leur nombre est important plus le risque de produire un synopsis comportant des erreurs est grand.

Les solutions qui pourraient être apportées sont assez évidentes, il faudrait une meilleure détection des concepts dans les conversations, une annotation des synopsis plus précise afin de retirer complètement toutes informations pouvant mener à l'introduction de données dérivant de la conversation initiale.

En terme de sémantique, les synopsis produits à partir de patrons fixes et dynamiques sont très proches, ce qui s'explique par le fait que les informations porteuses de sens dans le résumé sont données par le même système. En revanche au niveau syntaxique l'utilisation de patrons recombinés semble apporter une

amélioration de *1.6* point par rapport aux synopsis générés à partir de patrons fixes. Le résultat obtenu de *6.77/10* est d'autant plus satisfaisant que les synopsis écrits par l'humain 1 obtiennent un score très proche de *6.91/10*. Ces résultats montrent que les synopsis ainsi générés sont proches de ceux d'un humain en terme de lisibilité. Cependant en ce qui concerne le contenu, les synopsis écrits manuellement sont loin devant avec quasiment *4* points d'avance.

4.7. Conclusion

Dans cette section nous avons présenté une méthode de résumé abstractif de conversations par recombinaison de patrons. L'originalité de cette approche réside dans le fait d'utiliser les concepts des synopsis seulement puis de les aligner avec les conversations, limitant ainsi le risque d'erreur de détection. Cette façon de procéder permet aussi de ne se concentrer que sur les concepts estimés utiles par les annotateurs, ce qui renforce le système lors de la détection.

La production de patrons se base aussi sur un ensemble de synopsis annotés, permettant au système de sélectionner une combinaison de fragments de patrons répondant au remplissage imposé par la détection des informations de la conversation.

L'analyse sémantique présentée dans le chapitre 3 a permis de spécifier certains traits lors de la phase de détection des concepts aidant à l'identification de ces derniers, mais n'est pas présente dans la production de patrons. Un approche plus profonde de la production de patrons pourrait tenir compte de cette annotation comme dans l'exemple présenté dans la section 3.2.5.1 (figure 3.11).

Le système présenté obtient un score ROUGE-2 sensiblement proche du score obtenu par les humains individuellement (mais évalué sur un synopsis de moins). Un résultat encourageant concerne l'utilisation de transcriptions automatiques ou manuelles qui ne semble pas influer grandement sur la qualité des synopsis produits.

La tâche d'évaluation automatique des résumés étant une tâche difficile et subjective, une évaluation manuelle est souvent nécessaire pour apporter un nouveau regard sur les résumés produits. L'évaluation que nous avons menée tend à conforter l'idée que sémantiquement parlant les synopsis produits peuvent être proches de ceux d'un humain, mais en revanche, ils restent généralement moins bien construits, donc moins lisibles, et introduisent des contrevérités beaucoup plus souvent.

Conclusion

Ces dernières années, on a pu constater que les quantités de données disponibles étaient de plus en plus importantes, et que les entreprises de centres d'appels étaient de plus en plus prisées. Dans ce contexte, de nouvelles problématiques sont apparues, notamment sur le traitement des données audio de ces plateformes d'appels. Les évaluateurs humains ne pouvant plus suivre avec le flux constant et massif de conversations enregistrées, et les outils existants n'étant pas adaptés à ce genre de données, il était nécessaire de proposer une nouvelle approche pour permettre de traiter ces données en grand nombre et de façon rapide. Cette gestion de l'information est devenue un enjeu industriel, scientifique et économique. Des entreprises se sont spécialisées dans les échanges entre les clients et des opérateurs spécialisés, sous la forme de centres d'appels. Ces plateformes de communication doivent fournir des évaluations à la demande de l'employeur. Ces évaluations sont manuelles et ne sont menées que sur une infime partie des conversations à leur disposition. Des outils de navigation parmi ces conversations apporteraient un plus sur la qualité et la rapidité des évaluations menées.

La gestion de l'information de masse peut-être gérée par des approches de résumés automatiques. On s'aperçoit que le résumé automatique évolue au fil du temps et de l'évolution du partage de l'information. Si au début des années 50 il était principalement question de texte, aujourd'hui les documents sont aussi vidéos et audios. Les méthodes par extractions ont fait leurs preuves sur des documents textuels, mais montrent rapidement leur limite lorsqu'elles sont appliquées à des conversations du fait de la nature du langage utilisé. En effet il s'agit de langue parlée, qui comporte de nombreuses disfluences, erreurs de langue, répétitions, etc. De plus le style direct utilisé dans les dialogues ne correspond pas au résumé le plus souvent écrit au passif. De simples extractions de tours de parole ne suffisent pas à produire un résumé compréhensible et utilisable. Il est nécessaire de développer des méthodes de résumés abstractifs ne dépendant pas ou peu de la nature de la langue du document source.

Une approche raisonnable pour se passer de la nature de la langue est d'exploiter un niveau abstrait de celle-ci à travers l'utilisation de la sémantique. La sémantique dans les conversations correspond au fait de donner un sens à ce que chaque partie a pu échanger avec l'autre. Même si cette tâche d'annotation sémantique s'appuie principalement sur la syntaxe, on a pu montrer qu'il était possible de produire une annotation de façon rapide sur un corpus de conversations dont le thème était restreint, ce qui aide à la résolution des problèmes de désambigüisation.

Nous proposons dans le chapitre 4 une approche de résumé abstractif par recombinaison de patrons. Cette approche se divise en 3 étapes, l'alignement des concepts des synopsis avec ceux de la conversation ce qui permet de limiter des erreurs d'identifications dans les conversations, la détection de concept au sein d'une conversation après avoir appris un modèle capable de reconnaître ces informations en fonction d'un ensemble de features, et enfin la production et le remplissage des patrons, générés à partir du corpus de synopsis annotés.

L'approche proposée a montré qu'elle était beaucoup plus adaptée à la tâche qu'un modèle extractif en passant d'un score ROUGE-2 de 0.0315 à 0.08390 sur les transcriptions manuelles et de 0.02093 à 0.08471 sur les transcriptions automatiques, se rapprochant du score des humains de 0.11848. L'utilisation de patrons recombinés pour la génération de synopsis par opposition à des patrons manuels a aussi été bénéfique et propose une amélioration significative à 95% autant sur les transcriptions manuelles qu'automatiques. Les résultats obtenus témoignent de la robustesse de la méthode adoptée vis-à-vis de l'utilisation de transcriptions automatiques.

L'évaluation des résumés est une tâche subjective, il n'existe pas de métrique parfaite, et même si une évaluation en score ROUGE-2 donne une bonne estimation des informations recueillies dans le synopsis, elle ne renseigne pas de la qualité globale de ce dernier comme la lisibilité ou l'introduction de fausses vérités. Une évaluation manuelle complémentaire a été menée, et tend à confirmer les hypothèses établies lors de l'analyse d'erreur.

Perspectives

Nous allons explorer les perspectives de ce travail dans le but d'apporter des pistes visant à renforcer l'approche.

À court terme, une des premières remarques que l'on peut faire touche la quantité de données disponibles. Même si le corpus DECODA comporte plus de 1500 conversations toutes transcrites, seules 200 d'entre elles ont été annotées en synopsis par au moins 2 annotateurs différents. Augmenter le nombre de synopsis par conversation et de façon globale sur le corpus pourrait apporter de meilleurs résultats sur plusieurs aspects du travail présenté :

— La détection de concepts dans les transcriptions ;
— La génération de patrons ;
— L'évaluation.

L'augmentation du nombre de synopsis annotés permettrait d'aligner plus de concepts, ce qui augmenterait de cette façon le nombre d'exemples disponibles lors de l'apprentissage pendant la phase de détection des concepts. Il en résulterait un système d'apprentissage plus performant et des identifications de concepts plus régulières et plus fines. Le nombre de synopsis disponible influe aussi sur le choix dans les fragments de patrons à sélectionner, augmentant la couverture des éventuels nouveaux cas se présentant. Enfin l'augmentation du nombre de synopsis de référence, notamment pour une même conversation permettrait de raffiner l'évaluation et de proposer des évaluations de type *Pyramid* nécessitant un grand nombre de résumés de références.

Au niveau de la méthode proposée, le point des patrons reste à améliorer. D'une part la génération de fragments de patrons dépend directement de la qualité des synopsis et de la façon dont les concepts ont été identifiés puis retirés. En effet comme nous avons pu le voir lors de l'évaluation subjective (section 4.6.4), certains fragments de patrons contiennent toujours des informations relatives à la conversation à laquelle ils sont associés. Une vérification plus en profondeur de ces annotations permettrait de limiter le facteur de contrevérité au seins des synopsis produits. D'autre part, la méthode de choix des fragments de patrons définie par les concepts détectés dans les conversations, se fait actuellement de façon gloutonne, et certains fragments de patrons semblent plus prisés que d'autres du fait de leur ordre d'apparition. Une approche plus réfléchie à base de graphes ou de pondération des fragments permettrait de faire des choix plus cohérents.

En ce qui concerne l'évaluation, nous n'avons pas encore assez de recul vis-à-vis de l'évaluation manuelle et des questions qui ont été posées aux évaluateurs, de ce fait il est possible qu'elles n'aient pas été assez claires pour tout le monde, ce qui peut influencer les résultats. Le nombre d'évaluateurs permettrait aussi de limiter ce biais de compréhension.

L'approche que nous avons présentée a été développée et testée pour répondre à une problématique spécifique directement liée aux centres d'appels téléphoniques, mais cette approche pourrait très bien être utilisée pour traiter d'autres domaines comme les médias sociaux, les articles scientifiques ou encore les articles de journaux. Le seul pré requis essentiel est d'avoir un corpus de résumés de référence, à partir de quoi toute la chaîne de production de synopsis se base. Ces domaines restent dans un format textuel, mais il serait intéressant de voir le comportement de la méthode à base de recombinaison de patron sur des formats comme la vidéo en ce basant sur des séquences de type "*zapping*" ou *journaux télévisés* composés de reportages.

Nous avons montré que l'approche marchait pour des résumés à forte contrainte de taille. Pour ce qui est des résumé plus long, il est possible que la méthode ne soit pas adaptée directement du fait de trop grand nombre d'informations importantes à identifier et collecter pour remplir un patron. La phase de sélection de fragments de patrons est la plus problématique pour les résumés longs, étant donné le manque de structure entre les fragments. Une solution envisageable pour parvenir à produire ce genre de résumés serait de définir une certaines structure entre les différents fragments de patrons sous la forme d'un graphe au sein duquel les concepts détectés dans le document initial définiraient un chemin.

Bibliographie

[1] Collin F BAKER, Charles J FILLMORE et John B LOWE. « The berkeley framenet project ». In : *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics. 1998, p. 86–90 (cf. p. 86).

[2] Laura BANARESCU, Claire BONIAL, Shu CAI et al. « Abstract meaning representation (AMR) 1.0 specification ». In : *Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle : ACL*. 2012, p. 1533–1544 (cf. p. 60).

[3] Satanjeev BANERJEE et Alon LAVIE. « METEOR : An automatic metric for MT evaluation with improved correlation with human judgments ». In : *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. T. 29. 2005, p. 65–72 (cf. p. 47).

[4] Phyllis B BAXENDALE. « Machine-made index for technical literature : an experiment ». In : *IBM Journal of Research and Development* 2.4 (1958), p. 354–361 (cf. p. 23).

[5] Thierry BAZILLON, Melanie DEPLANO, Frederic BECHET et al. « Syntactic annotation of spontaneous speech : application to call-center conversation data. » In : *LREC*. 2012, p. 1338–1342 (cf. p. 42, 87, 88, 91).

[6] Frederic BECHET, Benjamin MAZA, Nicolas BIGOUROUX et al. « DECODA : a call-centre human-human spoken conversation corpus. » In : *LREC*. 2012, p. 1343–1347 (cf. p. 40, 83).

[7] Kedar BELLARE, Anish Das SARMA, Atish Das SARMA et al. « Generic Text Summarization Using WordNet. » In : *LREC*. 2004 (cf. p. 25).

[8] Divyanshu BHARTIYA et Ashudeep SINGH. « A Semantic Approach to Summarization ». In : *arXiv preprint arXiv :1406.1203* (2014) (cf. p. 26).

[9] Lidong BING, Piji LI, Yi LIAO et al. « Abstractive multi-document summarization via phrase selection and merging ». In : *arXiv preprint arXiv :1506.01597* (2015) (cf. p. 34).

[10] Marie CANDITO, Pascal AMSILI, Lucie BARQUE et al. « Developing a french framenet : Methodology and first results ». In : *LREC-The 9th edition of the Language Resources and Evaluation Conference*. 2014 (cf. p. 67).

[11] Jaime CARBONELL et Jade GOLDSTEIN. « The use of MMR, diversity-based reranking for reordering documents and producing summaries ». In : *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, p. 335–336 (cf. p. 30).

[12] Jackie Chi Kit CHEUNG et Gerald PENN. « Towards Robust Abstractive Multi-Document Summarization : A Caseframe Analysis of Centrality and Domain. » In : *ACL (1)*. 2013, p. 1233–1242 (cf. p. 34).

[13] Wesley T CHUANG et Jihoon YANG. « Extracting sentence segments for text summarization : a machine learning approach ». In : *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2000, p. 152–159 (cf. p. 33).

[14] Mark G CORE et James ALLEN. « Coding dialogs with the DAMSL annotation scheme ». In : *AAAI fall symposium on communicative action in humans and machines*. T. 56. Boston, MA. 1997 (cf. p. 28).

[15] Géraldine DAMNATI, Aleksandra GUERRAZ et Delphine CHARLET. « Web Chat Conversations from Contact Centers : a Descriptive Study. » In : *LREC*. 2016 (cf. p. 18).

[16] Hoa Trang DANG et Karolina OWCZARZAK. « Overview of the TAC 2008 Update Summarization Task. » In : *TAC*. 2008 (cf. p. 50).

[17] Bonnie DORR, Nizar HABASH et David TRAUM. « A thematic hierarchy for efficient generation from lexical-conceptual structure ». In : *Conference of the Association for Machine Translation in the Americas*. Springer. 1998, p. 333–343 (cf. p. 60).

[18] Harold P EDMUNDSON. « New methods in automatic extracting ». In : *Journal of the ACM (JACM)* 16.2 (1969), p. 264–285 (cf. p. 23).

[19] Rong-En FAN, Kai-Wei CHANG, Cho-Jui HSIEH et al. « LIBLINEAR : A library for large linear classification ». In : *Journal of machine learning research* 9.Aug (2008), p. 1871–1874 (cf. p. 88).

[20] Benoit FAVRE, Dilek HAKKANI-TÜR et Sebastien CUENDET. *Icsiboost*. 2007 (cf. p. 88).

[21] Charles J FILLMORE. « Frame semantics and the nature of language ». In : *Annals of the New York Academy of Sciences* 280.1 (1976), p. 20–32 (cf. p. 60).

[22] Pierre-Etienne GENEST et Guy LAPALME. « Absum : a knowledge-based abstractive summarizer ». In : *Génération de résumés par abstraction* 25 (2013) (cf. p. 26, 36, 56, 78).

[23] Dan GILLICK et Benoit FAVRE. « A scalable global model for summarization ». In : *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*. Association for Computational Linguistics. 2009, p. 10–18 (cf. p. 31, 32).

[24] Xu HAN, Tao LV, Zhirui HU et al. « Text Summarization Using FrameNet-Based Semantic Graph Model ». In : *Scientific Programming* 2016 (2016) (cf. p. 26).

[25] Eduard HOVY et Chin-Yew LIN. « Automated text summarization and the SUMMARIST system ». In : *Proceedings of a workshop on held at Baltimore, Maryland : October 13-15, 1998*. Association for Computational Linguistics. 1998, p. 197–214 (cf. p. 23).

[26] Hongyan JING et Kathleen R MCKEOWN. « Cut and paste based text summarization ». In : *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics. 2000, p. 178–185 (cf. p. 34).

[27] Siavash KAZEMIAN, Frank RUDZICZ, Gerald PENN et al. « A critical assessment of spoken utterance retrieval through approximate lattice representations ». In : *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM. 2008, p. 83–88 (cf. p. 23).

[28] Paul KINGSBURY et Martha PALMER. « From TreeBank to PropBank. » In : *LREC*. Citeseer. 2002, p. 1989–1993 (cf. p. 58).

[29] Kevin KNIGHT et Daniel MARCU. « Statistics-based summarization-step one : Sentence compression ». In : *AAAI/IAAI* 2000 (2000), p. 703–710 (cf. p. 17).

[30] Julian KUPIEC, Jan PEDERSEN et Francine CHEN. « A trainable document summarizer ». In : *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1995, p. 68–73 (cf. p. 33).

[31] Carole LAILLER, Anaïs LANDEAU, Frédéric BÉCHET et al. « Enhancing the RATP-DECODA corpus with linguistic annotations for performing a large range of NLP tasks ». In : *Proceedings of LREC*. 2016 (cf. p. 91).

[32] Alon LAVIE et Abhaya AGARWAL. « METEOR : An automatic metric for MT evaluation with high levels of correlation with human judgments ». In : *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics. 2007, p. 228–231 (cf. p. 47).

[33] Minh LE NGUYEN, Akira SHIMAZU, Susumu HORIGUCHI et al. « Probabilistic sentence reduction using support vector machines ». In : *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics. 2004, p. 743 (cf. p. 17).

[34] Chin-Yew LIN. « Improving summarization performance by sentence compression : a pilot study ». In : *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*. Association for Computational Linguistics. 2003, p. 1–8 (cf. p. 16).

[35] Chin-Yew LIN. « Rouge : A package for automatic evaluation of summaries ». In : *Text summarization branches out : Proceedings of the ACL-04 workshop*. T. 8. Barcelona, Spain. 2004 (cf. p. 46, 53).

[36] Chin-Yew LIN et Eduard HOVY. « The potential and limitations of automatic sentence extraction for summarization ». In : *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*. Association for Computational Linguistics. 2003, p. 73–80 (cf. p. 17).

[37] Fei LIU, Jeffrey FLANIGAN, Sam THOMSON et al. « Toward abstractive summarization using semantic representations ». In : (2015) (cf. p. 26).

[38] Hans Peter LUHN. « The automatic creation of literature abstracts ». In : *IBM Journal of research and development* 2.2 (1958), p. 159–165 (cf. p. 22).

[39] William C MANN et Sandra A THOMPSON. « Rhetorical structure theory : Description and construction of text structures ». In : *Natural language generation*. Springer, 1987, p. 85–95 (cf. p. 27).

[40] Daniel MARCU. « From discourse structures to text summaries ». In : *Proceedings of the ACL*. T. 97. Citeseer. 1997, p. 82–88 (cf. p. 27).

[41] Daniel MARCU. « Discourse trees are good indicators of importance in text ». In : *Advances in automatic text summarization* 293 (1999), p. 123–136 (cf. p. 27).

[42] Mitchell MARCUS, Grace KIM, Mary Ann MARCINKIEWICZ et al. « The Penn Treebank : annotating predicate argument structure ». In : *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics. 1994, p. 114–119 (cf. p. 58).

[43] Sameer MASKEY et Julia HIRSCHBERG. « Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization ». In : *Ninth European Conference on Speech Communication and Technology*. 2005 (cf. p. 23, 24).

[44] Rada MIHALCEA et Paul TARAU. « TextRank : Bringing order into texts ». In : Association for Computational Linguistics. 2004 (cf. p. 30).

[45] George A MILLER. « WordNet : a lexical database for English ». In : *Communications of the ACM* 38.11 (1995), p. 39–41 (cf. p. 23, 57, 58).

[46] Gabriel MURRAY, Steve RENALS, Jean CARLETTA et al. « Incorporating speaker and discourse features into speech summarization ». In : *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics. 2006, p. 367–374 (cf. p. 25).

[47] Alexis NASR, Frederic BECHET, Benoit FAVRE et al. « Automatically enriching spoken corpora with syntactic information for linguistic studies. » In : *LREC*. 2014, p. 854–858 (cf. p. 43).

[48] Alexis NASR, Frédéric BÉCHET, Jean-François REY et al. « Macaon : An nlp tool suite for processing word lattices ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Systems Demonstrations*. Association for Computational Linguistics. 2011, p. 86–91 (cf. p. 42).

[49] Ani NENKOVA, Rebecca PASSONNEAU et Kathleen MCKEOWN. « The pyramid method : Incorporating human content selection variation in summarization evaluation ». In : *ACM Transactions on Speech and Language Processing (TSLP)* 4.2 (2007), p. 4 (cf. p. 48).

[50] Jun-Ping NG et Viktoria ABRECHT. « Better summarization evaluation with word embeddings for rouge ». In : *arXiv preprint arXiv :1508.06034* (2015) (cf. p. 46, 97).

[51] Tatsuro OYA, Yashar MEHDAD et Raymond NG. « A template-based abstractive meeting summarization : Leveraging summary and source text relationships ». In : (2014) (cf. p. 35).

[52] Lawrence PAGE, Sergey BRIN, Rajeev MOTWANI et al. « The PageRank citation ranking : bringing order to the web. » In : (1999) (cf. p. 31).

[53] Rebecca J PASSONNEAU, Kathleen MCKEOWN, Sergey SIGELMAN et al. « Applying the pyramid method in the 2006 Document Understanding Conference ». In : *Proc. DUC'06*. 2006 (cf. p. 49).

[54] Korbinian RIEDHAMMER, Benoit FAVRE et Dilek HAKKANI-TÜR. « Long story short–global unsupervised models for keyphrase based meeting summarization ». In : *Speech Communication* 52.10 (2010), p. 801–815 (cf. p. 32).

[55] Sophie ROSSET, Delphine TRIBOUT et Lori LAMEL. « Multi-level Information and Automatic dialog Act Detection in Human-Human Spoken Dialogs ». In : *Speech Communication* 50.1 (2008) (cf. p. 29).

[56] Horacio SAGGION et Guy LAPALME. « Concept identification and presentation in the context of technical text summarization ». In : *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*. Association for Computational Linguistics. 2000, p. 1–10 (cf. p. 35).

[57] Benoît SAGOT et Darja FIŠER. « Building a free French wordnet from multilingual resources ». In : *OntoLex*. 2008 (cf. p. 58).

[58] Soufian SALIM, Nicolas HERNANDEZ et Emmanuel MORIN. « Comparaison d'approches de classification automatique des actes de dialogue dans un corpus de conversations écrites en ligne sur différentes modalités ». In : *23ème Conférence sur le Traitement Automatique des Langues Naturelles*. 2016 (cf. p. 29).

[59] John R SEARLE. « Austin on locutionary and illocutionary acts ». In : *The philosophical review* (1968), p. 405–424 (cf. p. 28).

[60] Karen SPARCK JONES. « A statistical interpretation of term specificity and its application in retrieval ». In : *Journal of documentation* 28.1 (1972), p. 11–21 (cf. p. 22).

[61] Andreas STOLCKE, Klaus RIES, Noah COCCARO et al. « Dialogue act modeling for automatic tagging and recognition of conversational speech ». In : *Computational linguistics* 26.3 (2000), p. 339–373 (cf. p. 28).

[62] Parsons TERENCE. *Events in the semantics of English : A study in subatomic semantics*. 1990 (cf. p. 60).

[63] Jérémy TRIONE, Benoit FAVRE et Frédéric BÉCHET. « Beyond utterance extraction : summary recombination for speech summarization ». In : *Interspeech 2016* (2016), p. 680–684 (cf. p. 88).

[64] Ralph WEISCHEDEL, Eduard HOVY, Mitchell MARCUS et al. « OntoNotes : A large training corpus for enhanced processing ». In : *Handbook of Natural Language Processing and Machine Translation. Springer* (2011) (cf. p. 59).

[65] Michael WHITE, Tanya KORELSKY, Claire CARDIE et al. « Multidocument summarization via information extraction ». In : *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics. 2001, p. 1–7 (cf. p. 79).

[66] Klaus ZECHNER. « Automatic summarization of open-domain multiparty dialogues in diverse genres ». In : *Computational Linguistics* 28.4 (2002), p. 447–485 (cf. p. 29).

ANNEXES

A. Cadres sémantiques utilisés

Abandonment	Accuracy	Activity_pause
Activity_prepare	Activity_resume	Adjusting
Agreement	Agree_or_refuse_to_act	Amalgamation
Arriving	Assessing	Assistance
Attaching	Attempt	Avoiding
Awareness	Becoming	Becoming_a_member
Becoming_aware	Being_in_effect	Being_in_operation
Borrowing	Breaking_apart	Breaking_off
Breathing	Bringing	Building
Bungling	Canceling	Categorization
Causation	Cause_change	Cause_change_of_strength
Cause_harm	Cause_motion	Cause_to_experience
Cause_to_perceive	Certainty	Change_accessibility
Change_event_time	Change_operational_state	Change_position_on_a_scale
Chatting	Choosing	Closure
Coming_to_be	Commerce_buy	Commerce_collect
Commerce_pay	Commerce_sell	Commitment
Communication	Communication_response	Complaining
Compliance	Conferring_benefit	Contacting
Containing	Contingency	Contrition
Control	Cotheme	Deciding
Defending	Departing	Deserving
Desirable_event	Desiring	Difficulty
Duration_description	Duration_relation	Emitting
Emphasizing	Emptying	Erasing
Estimating	Event	Evidence
Existence	Expend_resource	Expensiveness
Experiencer_focus	Experiencer_obj	Explaining_the_facts
Feeling	Filling	Forgiveness
Forming_relationships	Getting	Give_impression
Giving	Givinig	Grasp
Halt	Having_or_lacking_access	Hello
Hiding_objects	Hiring	Impact
Ingestion	Intentionally_affect	Intentionally_create
Judgment	Judgment_direct_address	Justifying
Labeling	Leadership	Lending
Locale_closure	Locating	Location_in_time
Losing	Making_arrangements	Memory
Motion	Name_conferral	Offering
Operating_a_system	Opinion	Participation

Path_shape	Perception_active	Performers_and_roles
Placing	Possession	Possibility
Posture	Practice	Predicting
Preference	Prevarication	Process_continue
Process_end	Processing_materials	Process_start
Questioning	Receiving	Redirecting
Reliance	Removing	rentraire
Repayment	Replacing	Reporting
Request	Required_event	Reserving
Reshaping	Residence	Resolve_problem
Respond_to_proposal	Ride_vehicle	Run_risk
Scrutiny	Self_motion	Self_otion
Sending	Sign	Similarity
Simultaneity	Spelling_and_pronouncing	Statement
Storing	Studying	Subscribing
Success_or_failure	Sufficiency	Surpassing
Taking_sides	Telling	Text_creation
Theft	Topic	Transfer
Trap	Triggering	Using
Using_resource	Verification	Wagering
Waiting	Warning	Work
Working	Negation	

Table .10. – Cadres disponibles pour l'annotation automatique.

B. Fragments de patrons figés

variable détectée	fragment de patrons
FROM	l'appelant voudrait se rendre de $FROM
TO	jusqu'à $TO
ITEM	le client à perdu $ITEM
PASS_TYPE	$PASS_TYPE
TRANSPORT	dans le $TRANSPORT
LOCATION	à $LOCATION
LINE	Le conseillé lui dit de prendre $LINE
TOWARDS	en direction de $TOWARDS
END_STOP	jusqu'à $END_STOP
RETRIEVE_LOCATION	le conseillé lui dit de se rendre au $RETRIEVE_LOCATION pour récupérer son dû.
NOT_TRANSPORT	sans prendre $NOT_TRANSPORT
START_STOP	de $START_STOP
TIME	à $TIME
FREQUENCY	pour $FREQUENCY
DURATION	pendant $DURATION
ADDRESS	à $ADDRESS
ISSUE	à cause de $ISSUE
BUY	pour $BUY

Table .11. – Patrons écrit manuellement et découpés.

C. Guide d'annotation en synopsis SENSEI

Summary / Synopsis Guide (DECODA corpus inspired version)

Why do we need this document ?

Making a summary is a very subjective task. Two different persons won't systematically pick up the same information, or the same level of detail, and then won't write the same summary.
Our systems needs these human summary to learn how to automatically generate them in a good way, and to be evaluated as well. Here we there are two options :

— The first option is to hire a lot of annotator and let them write their summaries in a free way. And then make a study on all the resource collected to find a good shape for a summary. But this method a very expensive

— The second option is to write a guide to lead the annotators in a certain way of annotation to get some similarity in the summaries generated. In this way we introduce some guideline depending on the corpus to be annotated (previously studied). The annotator are not free anymore, but then only a few can annotate the corpus. Because of that kind of guide, all the summaries generated will come with some similarity and become more usable for our systems

DECODA Corpus inspired version.

First of all, let's define here what's we call a synopsis. A synopsis is a summary of the call taken from a call center. This call mostly involved two persons, the caller (here someone who wants something from the call center) and the adviser (the one who answer in the call center and give some solution to the caller).

After some studies we concluded that this synopsis should be no longer than 7% of the original call (in terms of words). We are aware that the annotator won't spend their time counting the words... So we probably include this length limit in the next version of the interface. Speaking of the interface, right now, it's just a simple and minimalist interface that just provide the conversation (spoken and written) and a box to fill with the synopsis. (Everything should be revamp to a better version).

Go back to the synopsis, this is a pretty subjective things to do, that's why we'll try in this guide to give some line to follow.
First we can distinguish two kind of synopsis :

— The semantic oriented synopsis : that focused more on the content of the conversation than the end of it

— The “structure” oriented synopsis : That focused more on the way the adviser treat the caller and his call

For the rest of this guide, only the semantic oriented synopsis we’ll be considered (the “structure” oriented synopsis could be developed in an other guide).

Syntax, length, way to write.

You have to keep in mind that you are limited in term of length. Moreover we don’t really need to get some really good language, what we need the most is the information !

Here’s some tips :

— Only pick the important information in the call

— The shorter your sentence is the better your synopsis will be

— Do not hesitate to make a “telegraphic style” sentence (e.g. “Route request in Paris center”)

— Try to sum the long exchange with a simple action, or if there isn’t any good information in it don’t even pick it up in the summary

Practical cases.

Nothing’s better than a good example. Here’s three of them translated from french with some comments on how we ended here :

Example 1

- Hello - Hello - Hello - eh it’s Mrs [name removed] eh I was calling you because I lost my scarf, where hmm in the bus hmm 140, when - Yes - I left at Colombes [name] yesterday night between - Yes you need to call later madam, around 11am, it’s too early, it’s not open yet. - At 11 - 11am yes, ok ? - Ok - See you later.

Synopsis 1
Needs information about a scarf lost in the bus 140. Wait the service opening and call later.

Synopsis 2
/ Maybe there is an error made by the annotator here about the bag / Bag lost in the bus 140, but call at 11am when the service will be open.

Comments :
Here's the call is pretty clear, the caller just want to know if there is any news about her loss. Then both annotators ended with the result of the call (e.g. call later when the service is opened). We could argue a bit on it due to the length of the synopsis, but it's a pretty important information and the length of the original conversation is pretty short. . .

Example 2

- Hello
- please
- hello ?
- Yes, hello
- hello
- I call you because I would like to have some information, when I'm at the Javel's station
- yes
- of the RER [the name of the train], is there a bus that could drive me closer to the Vauthier's street at Boulogne [name of the city] without taking the metro
- Vauthier's street at Boulogne-Billancourt
- yes
- just a moment, I'm looking for it
- thanks
- Madam
- yes
- you have the 72, you cross the Mirabeau's bridge to the Mirabeau's stop in way to Saint-Cloud's park.
- yes
- and you stop at the Reine Jean-Jaurès'road's stop
- Wait, at the end of the Mirabeau's bridge I take the 72
- 72
- to Saint-Cloud and you stop at the Reine Jean-Jaurès'road's stop
- Ok
- Ok thank you
- Good bye
- Good bye
- thank you

Synopsis 1
Needs for a connection by bus between the exit of the RER and a street in Boulogne.

Synopsis 2
Needs for a connection by bus between the RER and the Vaulthier's street at Boulogne.

Synopsis 3
Information on a potential bus connecting the Javel's station of the RER and a street in Boulogne.

Comments :

As you can see here, the three annotators picked up the same informations, only the syntax is slightly different. Some annotator are more accurate about the name of the street (i.e Vaulthier's street / street in Boulogne).

Example 3

- Hello - Yes hello madam, eh I'd like to have some information - Yes - I'd like to know if the bus center strike at Vitry [city] will continue tomorrow or if it was just today ? - No it's, it's just today, tomorrow everything is normal - Ok thank you very much - My pleasure - goodbye - Have a good day, goodbye

Synopsis 1
Needs informations about the renewal of the strike. Normal traffic in prevision.

Synopsis 2
Renewal of the bus center strike at Vitry [city], no tomorrow the traffic is normal.

Comments :

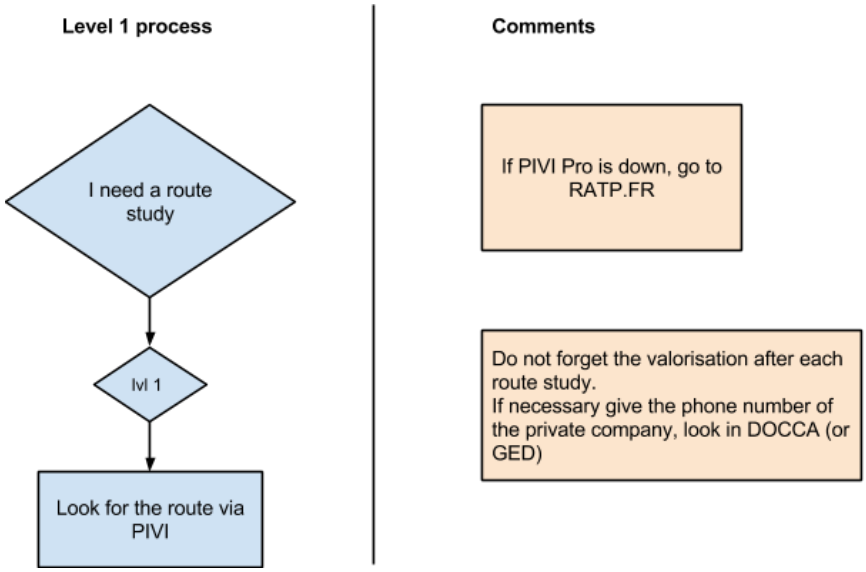
Nothing special here.

However we can notice the syntax used by annotator 1. Since the beginning he used the same method i.e. "Needs informations about". It can be a bit "word consuming" but it can also be a good way to introduce some kind of structure in every synopsis.

Call scenario

The DECODA's conversation come from a call center from the RATP. In this call center every adviser have some call scenario to help them answering the caller.

For example, the caller need to know a route, then the adviser get his route scenario and answer the question. The call scenario looks like this :



The call scenario is divided in two parts, the call process and the comments.

The comments are just some additional stuff to help the adviser during the process, it's generally very technical and then not useful for us. On the other hand the call process is very interesting.

The call process is like a state diagram. On the top on the diagram we have the main issue of the caller (here "I need a route study"). Then by following the arrow we have the way to answer depending on the other caller's issues.

Usually these call scenario draw a good base for a synopsis because if you can apply the corresponding scenario to a conversation it will give you the main theme, and the process pretty much already summed up.

How to get or generate these call scenario ?

In the DECODA corpus we already got these call scenario because of the nature of the corpus. But how could we make them if they are not available for

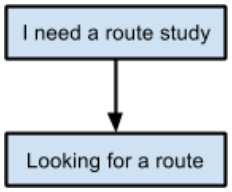
other conversations ?

- Basically the first thing to do is to establish a list with all the main theme of the conversation.

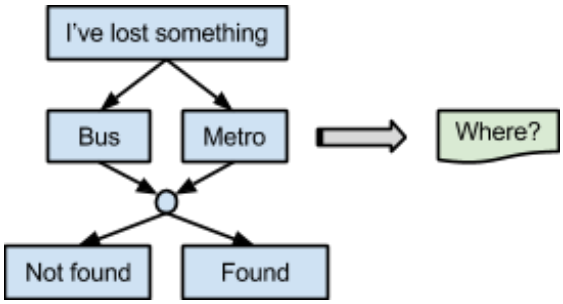
For example in DECODA we have :

1. Route
2. Loses, theft, found
3. Official report
4. Reimbursement
5. Delay, incident
6. Prices
7. Accidents

- Then for each theme you have to find if there is some redundant schema.
For instance for a route study, in most cases the caller is asking for a route and then the adviser is giving the caller the route needed. It can easily be drawn like this :

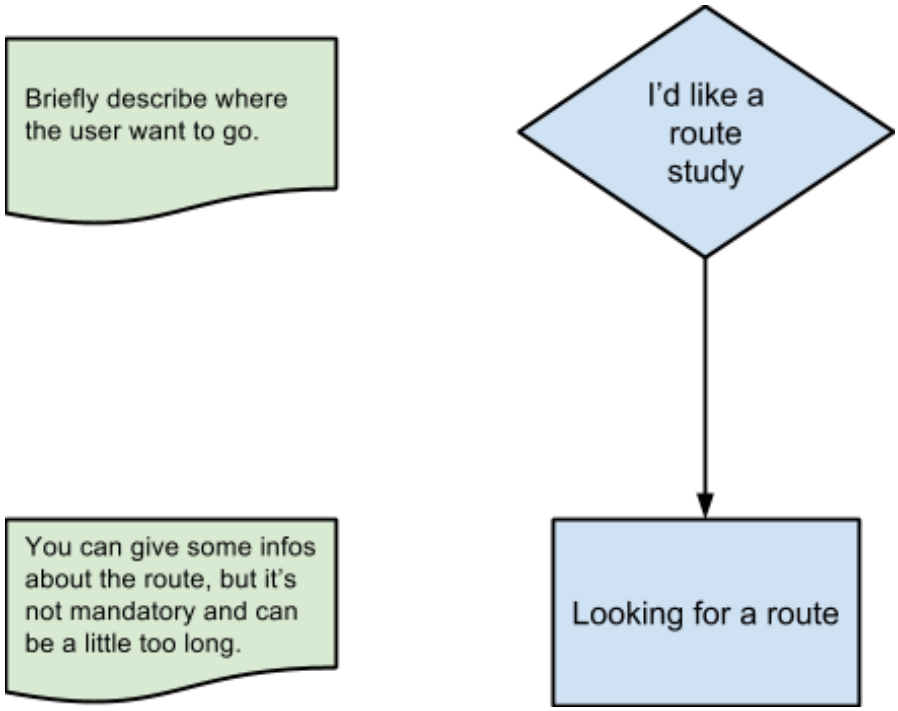


- While the state diagram can provide a lot of options, you can add some variety in it.
For instance in case of a lost, we can have several options like where the object has been lost, or is it found yet or not.



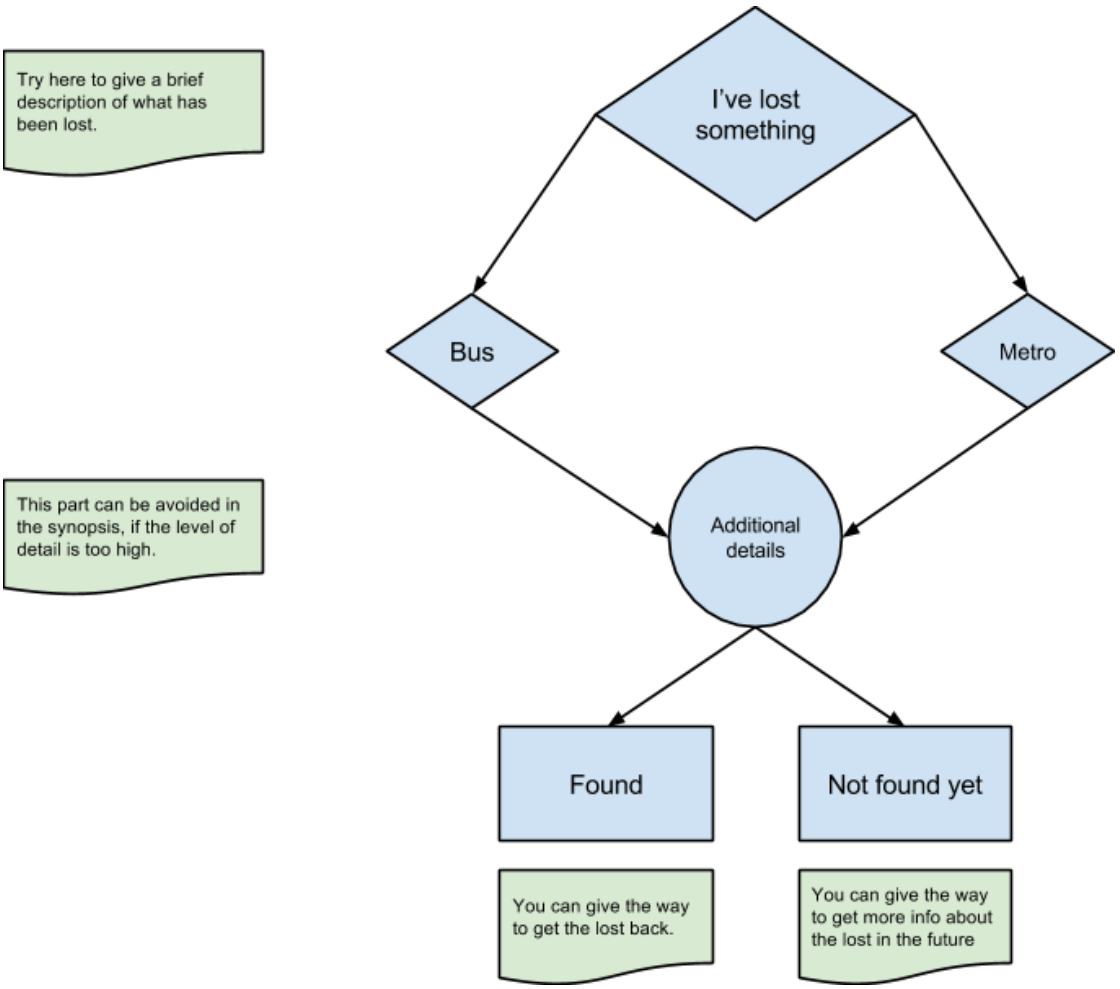
The next part gather the call scenario from DECODA listed as example above. Some comments have been added about the summarization task.

1- Route



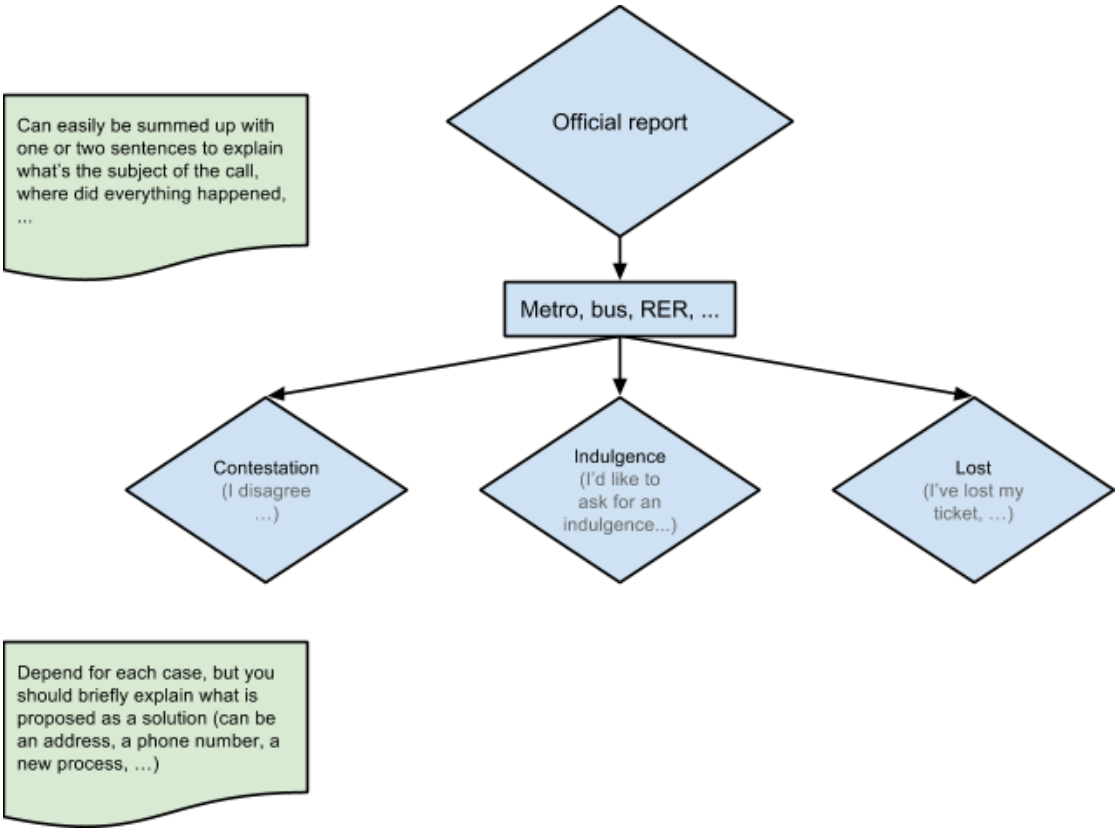
In case of several route requests in the same call, try not to focus on the destination, try to find a location where all the routes happen like for example Paris center if there are 2 or 3 route in individual different locations in Paris center.

2 - Loses, theft, found



In this particular case we like to precise if the object has been found or not just by adding at the end “object found, go to [location] to get it back” or “object not found still [recall later]”.

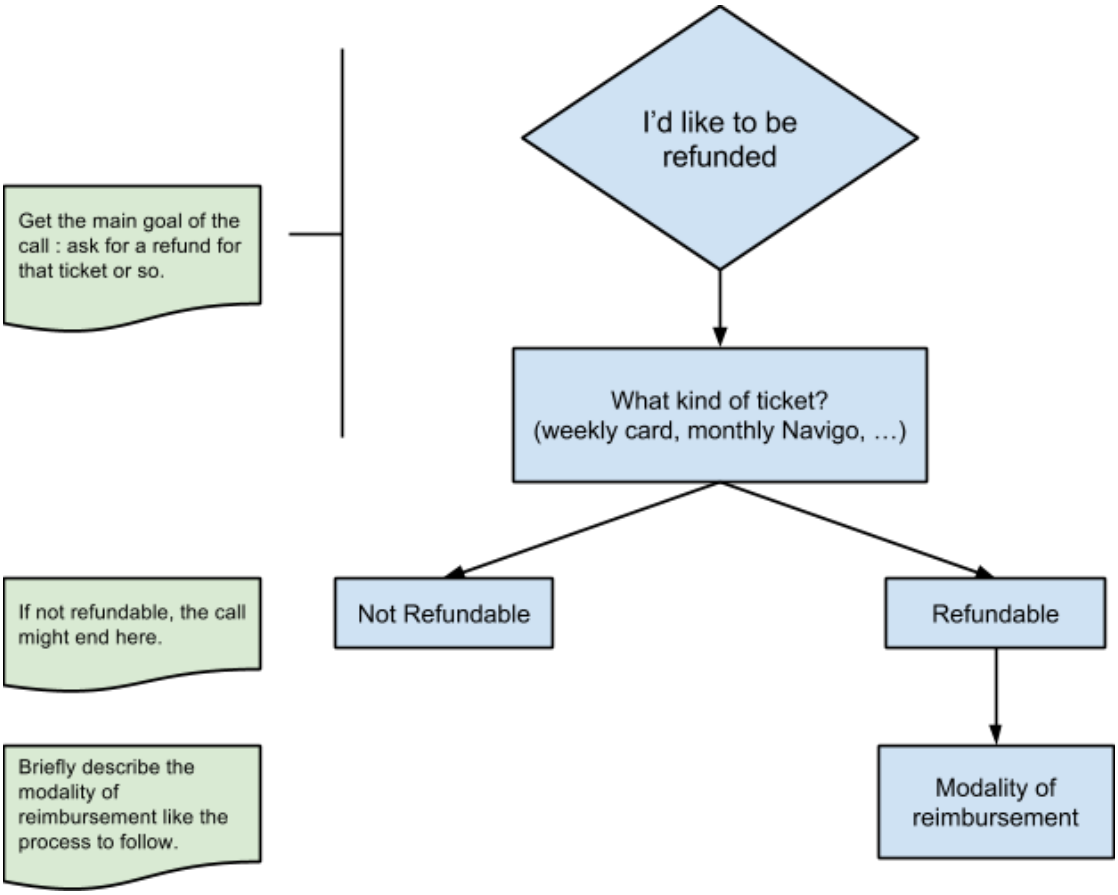
3 - Official report



Official report call are frequently pretty long (to explain everything or so).

Basically we try here to sum up the report in one or two sentences and then briefly give the solution given by the adviser (like “communication of the right service”)

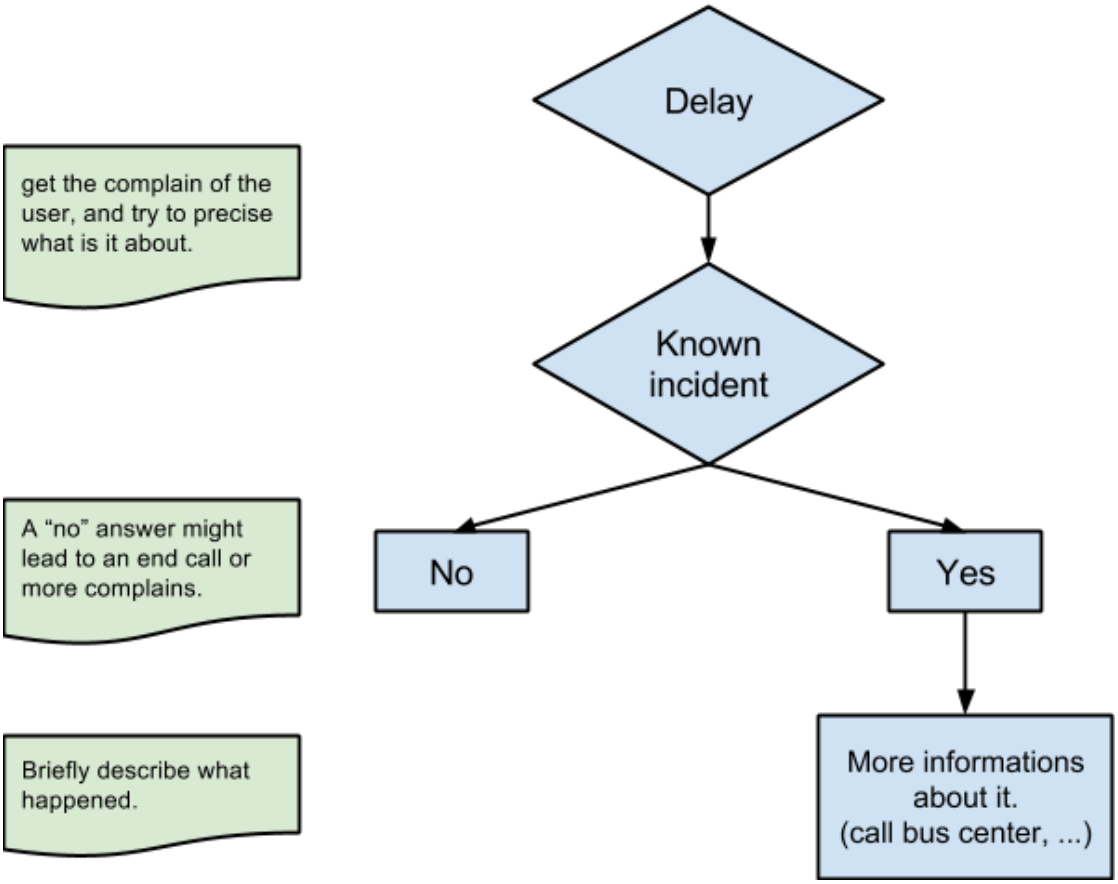
4 - Reimbursement



There is a lot of misunderstanding in these kind of call, depending on the knowledge of the caller, but because of the variety of the cards/contracts/errors/... the exchange between the caller and the adviser are pretty numerous.

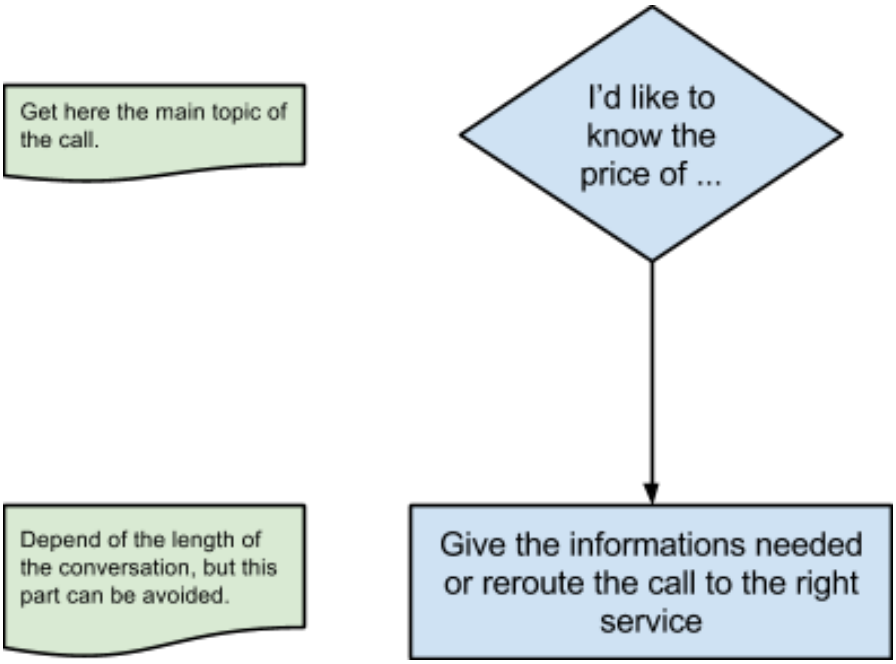
So just get the main reimbursement issue summed up and then as for the other scenario get the solution if it's relevant.

5 - Delay, incident



In these calls the adviser will frequently call the bus center or another service to ask for more information. Generally we don’t really consider that call in call, but we prefer to sum up the whole thing just with the final answer/solution to give to the caller.

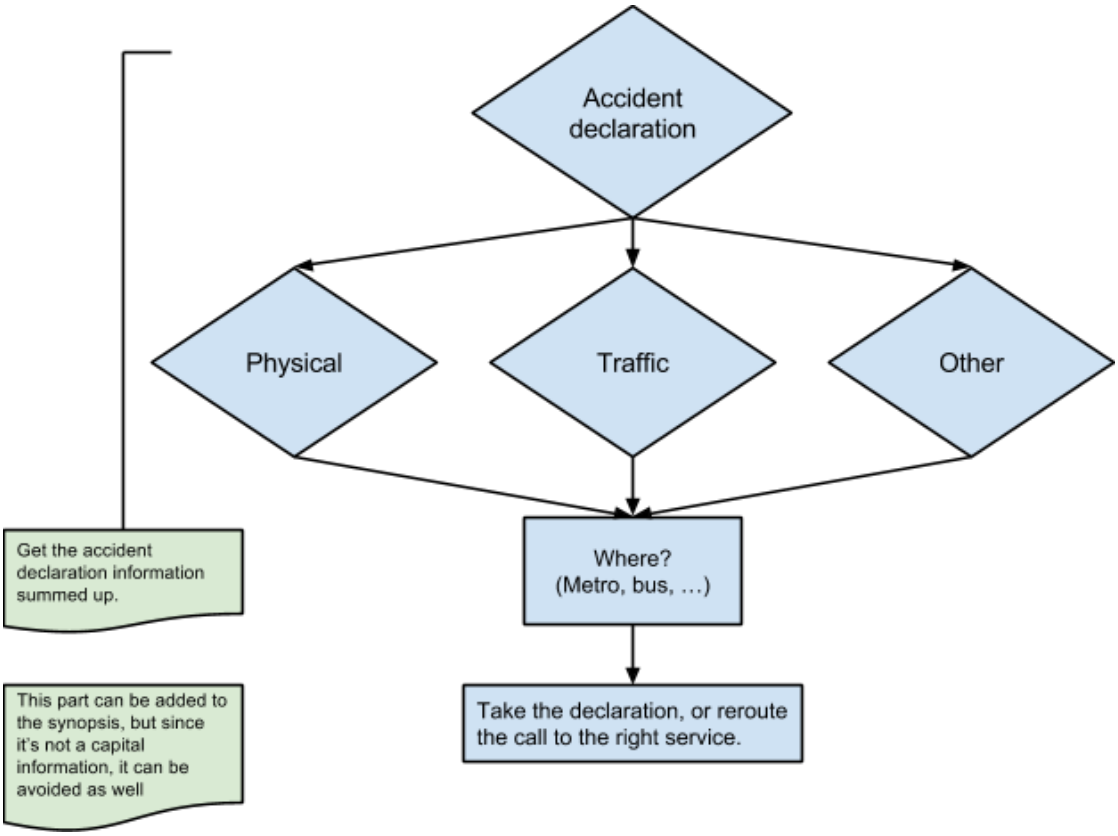
6 - Prices



Pretty much the same cases that in the previous scenario (5 - Delay, incident).

in some cases (the easiest and the shortest) we like to just sum up the whole conversation by the description of the issue (e.g. “asking for prices about an orange card”)

7 - Accidents



Pretty much the same as 5 - Delay, incident. We sum up the declaration, and then get the solution/answer.