Approche multitâche pour l'amélioration de la fiabilité des systèmes de résumé automatique de conversation

Eunice Akani, Benoît Favre, Frédéric Bechet, Romain Gemignani

Approche multitâche pour l'amélioration de la fiabilité des systèmes de résumé automatique de conversation

Eunice Akani[1, 2] Benoit Favre[1] Frederic Bechet[1] Romain Gemignani[2]
(1) Aix-Marseille Univ, CNRS, LIS, Marseille, France
(2) Enedis, Marseille, France
prenom.nom@lis-lab.fr, romain.gemignani@enedis.fr

RÉSUMÉ

Le résumé de dialogue consiste à générer un résumé bref et cohérent d'une conversation ou d'un dialogue entre deux ou plusieurs locuteurs. Même si les modèles de langue les plus récents ont permis des progrès remarquables dans ce domaine, générer un résumé fidèle au dialogue de départ reste un défi car cela nécessite de prendre en compte l'interaction entre les locuteurs pour conserver les informations les plus pertinentes du dialogue. Nous nous plaçons dans le cadre des dialogues humain-humain avec but. Ce cadre nous permet d'intégrer des informations relatives à la tâche dans le cadre du résumé de dialogue afin d'aider le système à générer des résumés plus fidèles sémantiquement. Nous évaluons dans cette étude des approches multitâches permettant de lier la tâche de résumé à des tâches de compréhension du langage comme la détection de motifs d'appels. Les informations liées à la tâche nous permettent également de proposer des nouvelles méthodes de sélection de résumés basées sur l'analyse sémantique du dialogue ainsi que des métriques d'évaluation basées également sur cette même analyse. Nous avons testé ces méthodes sur DECODA, un corpus français de dialogue collecté dans le centre d'appel de la RATP entre des usagers et des téléconseillers. Nous montrons que l'ajout d'informations liées à la tâche augmente la fiabilité des résumés générés.

ABSTRACT

Multitask approaches for improving reliability in goal-oriented dialogue summarization

Dialogue summarization consists of generating a brief and coherent summary of a conversation or dialogue between two or more speakers. Although the most recent language models have led to remarkable progress in this field, generating a summary faithful to the original dialogue remains a challenge because it requires taking into account the interaction between the speakers to retain the most relevant information from the dialogue. In this study we will consider human-human goal-oriented dialogues. This framework allows us to integrate task-related information into the dialogue summary framework to assist the system in generating more semantically faithful summaries. In this study, we evaluate multitask approaches that link the summary task to language comprehension tasks such as calltype classification. Task-related information also enables us to propose new methods of summary selection based on semantic analysis of the dialogue as well as semantic-based evaluation metrics. We tested these methods on DECODA, a French dialogue corpus collected in the RATP call center between users and teleadvisors. We demonstrate that the addition of task-related information increases the reliability of the generated summaries.

MOTS-CLÉS : Résumé Dialogue orienté tâche, compréhension de la parole, approches multitâches.

KEYWORDS: Task-oriented dialog summarization, spoken language understanding, multitask approaches.

1 Introduction

La tâche de résumé automatique est une sous-tâche de la génération automatique de texte qui consiste à condenser un contenu pour en avoir les informations essentielles. Il existe deux types de résumés automatiques : le résumé par extraction qui consiste à extraire les informations importantes, les agencer pour en faire un résumé et le résumé par abstraction qui consiste à résumer un texte en employant de nouveaux termes ou des paraphrases. Bien que les systèmes pour le résumé par abstraction aient une meilleure représentation syntaxique et une bonne compréhension sémantique, un problème majeur demeure. Il s'agit de la fidélité du résumé généré par rapport au document source. En effet, les systèmes de l'état-de-l'art basés sur les modèles de langue pré-entraînés générèrent des résumés qui peuvent contenir des informations erronées n'étant pas présentes dans le document original (souvent appelées *hallucinations*) (Maynez *et al.*, 2020).

Le résumé de conversation, une extension du résumé automatique de texte sur les interactions orales ou écrites, n'échappe pas à ce problème majeur. En effet, le style du discours spontané des transcriptions ne correspond pas au style attendu des résumés. Ce qui implique donc l'utilisation d'approches abstractives qui sont significativement affectées par ces hallucinations. De plus, résumer une conversation implique de *comprendre* les interactions entre les participants afin de ne pas faire de contresens sur son contenu.

Une étude menée par Cao *et al.* (2018) révèle que 30% des résumés générés par divers systèmes de résumé de texte contenaient des informations erronées par rapport au document original, qualifiées d'« hallucination » par Maynez *et al.* (2020). Plusieurs approches ont été proposées pour évaluer la fidélité des résumés générés, telles que l'implication textuelle (Falke *et al.*, 2019; Maynez *et al.*, 2020; Luo *et al.*, 2023), la vérification des questions dérivées d'un résumé (Durmus *et al.*, 2020; Fabbri *et al.*, 2022), et l'analyse des entités nommées absentes dans la source (Nan *et al.*, 2021; Ji *et al.*, 2023). Pour les résumés de dialogues, les études sur l'hallucination demeurent moins nombreuses que pour les résumés de documents textuels. Wang *et al.* (2022) a constaté que 35% des résumés du jeu de données SAMSum (Gliwa *et al.*, 2019) ne sont pas cohérents vis-à-vis des dialogues sources. Tang *et al.* (2022) ont classé huit types d'erreurs factuelles dans les résumés de dialogue, cinq étant spécifiques à ceux-ci, tandis que Wang *et al.* (2022) en ont identifié six. Ils ont proposé un schéma d'évaluation au niveau du modèle pour évaluer la fidélité. Ils ont utilisé un modèle de résumé basé sur des probabilités de génération conditionnelles pour différencier les résumés positifs des résumés négatifs.

Le résumé de dialogue, une tâche récente, englobe divers types de conversations, notamment les résumés de réunions introduits en premier par des corpus comme AMI/ICSI (Carletta *et al.*, 2005; Janin *et al.*, 2003). Des approches neuronales ont été appliquées à ces corpus, mais leur efficacité est limitée par la structure des dialogues et la diversité des données d'entrée, incluant des conversations de service à la clientèle, des discussions informelles et techniques. Ainsi, des méthodes utilisant des informations auxiliaires, telles que les actes de dialogue (Goo & Chen, 2018) ou la terminologie du domaine (Koay *et al.*, 2020), ont été proposées. Des jeux de données plus récents, comme SAMSum (Gliwa *et al.*, 2019), offrent de nouveaux défis, notamment la modélisation des participants dans les dialogues.

Une tâche récente dans le domaine du résumé de conversation est le résumé de dialogue orienté tâche. Le dialogue orienté tâche fait référence à un type de conversation dont l'objectif principal est d'accomplir une tâche ou un objectif spécifique. Ces dialogues sont typiques des interactions avec un service clientèle, une assistance technique et d'autres scénarios similaires. Plusieurs corpus ont été

proposés pour la tâche entre autre TODSUM (Zhao *et al.*, 2021), DECODA (Bechet *et al.*, 2012).

Dans un dialogue orienté tâche, les participants interagissent spontanément pour résoudre un problème. Chacun joue un rôle spécifique, utilisant un langage naturel avec des hésitations. Les objectifs des participants, les procédures, entités nommées, etc. sont donc cruciaux pour caractériser l'objectif du dialogue. Ainsi, les résumés doivent refléter ces aspects pour rester fidèles à la conversation.

Dans ce papier, nous proposons d'évaluer et d'améliorer la fidélité du résumé de dialogue en utilisant des informations spécifiques à la tâche telle que le motif d'appel. Nous nous concentrons sur les centres d'appels et nous nous appuyons sur le corpus DECODA (Bechet *et al.*, 2012), qui est l'un des rares corpus de dialogue parlé humain-humain à grande échelle, enregistré dans des centres d'appels (dans des conditions réelles), avec des annotations en motif d'appel et d'entités nommées spécifiques au domaine. A notre connaissance, il n'y a pas de corpus de parole en anglais plus volumineux contenant ce type d'interactions (dialogue humain-humain avec but) annotés sémantiquement et possédant des résumés annotés.

Nos contributions sont les suivantes :

— Nous proposons d'utiliser des informations spécifiques à la tâche pour améliorer la fiabilité des modèles de résumé automatique en comparant plusieurs méthodes multitâches.
— Nous introduisons une mesure basée sur la prédiction de la distribution des motifs d'appel permettant de sélectionner des résumés à partir de critères sémantiques.
— Nous proposons une évaluation montrant l'amélioration de la qualité des résumés sur le corpus DECODA.

2 Guidage de la génération de résumés par les informations sémantiques

Dans ce chapitre, nous décrivons plusieurs méthodes permettant d'intégrer des informations spécifiques à la tâche, notamment les motifs d'appel et les entités nommées du domaine, pour la génération de résumés.

Informations sémantiques liées à la tâche Le résumé de la conversation orientée tâche implique plusieurs échanges portant sur des informations spécifiques ainsi que des instructions relatives à la tâche que les participants cherchent à accomplir. Les informations dans la conversation dépendent ainsi de l'objectif à atteindre. Le type de représentation sémantique permettant d'encoder ces informations a été particulièrement étudié dans le cadre du dialogue humain-machine avec les modèles de Compréhension Automatique de la Parole (Spoken Language Understanding) développés pour des tâches de réservation de transport (ex : corpus ATIS) ou encore de restaurants ou hotels (ex : corpus MEDIA). Dans ce type d'étude on défini généralement 3 niveaux (Lee *et al.*, 2018) sémantiques :

— **domaine** : le domaine représente le cadre sémantique dans lequel se déroule le dialogue. Par exemple la réservation de billets d'avion pour le corpus ATIS. Dans notre étude il s'agit du domaine des transports publics dans Paris pour le corpus DECODA.
— **intention** : l'intention représente le type de requête pour un système de communication humain-machine, par exemple une confirmation, une demande de renseignement, etc. Elle est le plus souvent associée à un tour de parole, mais dans le corpus DECODA les étiquettes d'intention sont mises au niveau de l'ensemble du dialogue. Elles correspondent aux motifs

d'appels tels que *demande d'itinéraire* ou *réclamation objet perdu*.

— **paires concepts/valeurs** : ces paires correspondent aux arguments des relations sémantiques exprimées dans les intentions, comme par exemple la destination pour une demande d'itinéraire. Dans le corpus DECODA les entités nommées du domaine des transports parisiens sont annotées dans le corpus.

Dans cette étude, nous explorons comment introduire ces informations dans le processus de génération de résumés, comment elles peuvent être utilisées pour créer une mesure d'évaluation de la fidélité du résumé et comment leur utilisation influence la fidélité des résumés.

Approches pipeline et multi-tâches pour le résumé de conversation avec but Les informations spécifiques à une tâche ne sont pas directement disponibles dans les enregistrements ou les transcriptions de conversations, elles doivent donc être déduites. Nous pouvons, soit exploiter un système distinct pour prédire les catégories correspondantes et les utiliser comme entrée du système de résumé automatique, soit laisser le système les apprendre dans le cadre d'une supervision multitâche. Pour le moment nous nous sommes concentré sur la prédiction du motif d'appel à travers ces deux approches.

Soit C la conversation en entrée, R le résumé généré et M le motif d'appel. Nous considérons les deux méthodes suivantes :

1. **Multitache$_X$: Motif d'appel puis synopsis** : De manière générale, il s'agit de faire générer au modèle de langue une séquence composée de l'information sémantique X suivi le résumé. $\{X\}, R =$ résumé $\circ$ sémantique$C$. Pour notre cas particulier, nous avons considéré le motif d'appel. On a donc : $\{M\}, R =$ résumé $\circ$ sémantique(C).
2. **Pipeline$_X$** : Il s'agit de conditionner la génération du résumé par par une information sémantique X. On a donc $R =$ résumé$(X, \text{information-sémantique}(M))$. Ici nous utilisons le motif d'appel comme information sémantique. Ainsi on obtient : $R =$ summary$(C, \text{information-sémantique}(M))$. Ce motif en entrée peut être issu des motifs de référence (expérience *oracle*) ou encore d'un classifieur en motif d'appel.

Dans nos expériences un modèle de langue a été affiné sur une tâche de résumé automatique correspondant à chaque scénario. En complément de l'utilisation des informations sémantiques directement dans le processus de génération de résumé, nous proposons également de les utiliser en sortie du système de génération afin de sélectionner le résumé le plus fiable sémantiquement selon nos modèles.

3 Sélection de résumés sur des critères sémantiques

Il est possible de faire varier certains paramètres dans les processus de génération de texte par modèles de langue afin d'obtenir plusieurs sorties pour une même entrée. Dans cette étude, nous avons généré plusieurs résumés en utilisant 4 méthodes d'échantillonnage différentes pour sélectionner le prochain token comme dans l'article Akani *et al.*, 2023 : la recherche de type *beam-search* qui conserve les n-meilleurs chemins de probabilité les plus élevés à chaque étape ; l'échantillonnage de température qui consiste à redimensionner les logits avant d'appliquer la fonction softmax ; l'échantillonnage Top-K (Fan *et al.*, 2018) qui ne conserve que les K mots suivants les plus probables et redistribue la probabilité parmi ces K mots ; et l'échantillonnage Top-P (Holtzman *et al.*, 2020) qui consiste, étant donné une probabilité p, à prendre le plus petit ensemble possible de mots suivants dont la probabilité cumulative dépasse une masse de probabilité donnée et redistribue la probabilité parmi eux.

A partir de cet ensemble de résumés possibles, nous proposons deux méthodes de sélection basées sur les informations sémantiques liées à la tâche, l'une utilisant le motif d'appel et l'autre les entités nommées liées à la tâche.

Sélection de résumés basés sur la distribution des motifs d'appel. Le motif d'appel d'une conversation peut être vu comme une signature sémantique de l'objectif visé par la conversation. Ainsi, pour un résumé généré à partir d'une conversation, préserver le motif d'appel signifie préserver les caractéristiques sémantiques les plus importantes de cette conversation. Nous avons émis l'hypothèse qu'un résumé généré produisant un motif d'appel différent du motif d'appel de référence de la conversation contiendrait davantage d'informations incorrectes.

Pour utiliser ce critère pour choisir un résumé parmi un ensemble d'hypothèses, il est possible d'entraîner un classifieur de motifs d'appel pour prédire, à partir d'un résumé généré, son motif d'appel et de le comparer au motif d'appel qui pourrait être prédit par un autre classifieur sur la conversation entière.

Ici se pose un problème, certaines conversations peuvent avoir plusieurs motifs d'appel. En effet, un locuteur peut appeler pour une raison particulière et faire une autre demande dans la même conversation. Cela crée une frontière ambiguë entre les différents motifs d'appel. Pour traiter ce problème, nous proposons d'utiliser la divergence de Kullback-Leibler (KL) (Kullback & Leibler, 1951) sur la distribution de probabilité des motifs d'appel fournie par le classifieur, afin d'évaluer les résumés générés. La divergence de KL est une mesure de distance statistique qui quantifie la dissemblance entre deux distributions de probabilités. Elle évalue la différence entre une distribution de probabilité et une distribution de référence.

Ici, nous avons utilisé la divergence de KL pour évaluer un résumé généré sur la base de la distribution de probabilité sur les motifs d'appel. Cela est possible par l'utilisation de deux classifieurs en motifs d'appel : le premier appris sur les conversations entières et qui fournit une distribution de probabilités de motifs d'appels pour toute la conversation, et le second qui est appris directement sur les résumés de références et qui sera utilisé pour obtenir la distribution de motifs des résumés à sélectionner.

Pour $1...n$ les motifs d'appel, $G = \{g_1, ..., g_n\}$ la distribution de probabilité du résumé généré et $R = \{r_1, ..., r_n\}$ celle de la conversation entière, la KL divergence entre G et R se définit comme suit :

$$D_{\mathrm{KL}}(G \parallel R) = \sum_{x \in \{1..n\}} G(x) \log \left(\frac{G(x)}{R(x)} \right) \tag{1}$$

Il est maintenant possible de sélectionner le résumé qui minimise la distance D_{KL} parmi l'ensemble des résumés générés.

Sélection de résumés basée sur le risque d'hallucination. En complément du motif d'appel comme critère de sélection, nous utilisons le NEHR Akani *et al.*, 2023 qui correspond au pourcentage d'entités dans le résumé qui ne se trouvent pas dans le document source (la conversation dans notre cas). Il permet d'évaluer le risque d'hallucination sur les entités nommées. Lorsqu'un système de génération de résumés produit une entité nommée qui n'appartient pas au document original, cela augmente le risque d'*hallucination* de la part de ce modèle. En effet il peut s'agir d'une entité valide,

variante d'une des entités du document, mais il peut également s'agir d'une erreur de sur-génération de la part du modèle Akani, 2023.

Pour augmenter la fidélité du résumé, nous proposons de combiner au critère sur les motifs d'appels D_{KL}, la minimisation explicite du risque *NEHR*. Nous appliquons la règle de sélection suivante du résumé : $\hat{s}$. Soit H l'ensemble des résumés issus de l'échantillonnage pour la conversation C, V l'ensemble des résumés avec le NEHR minimum m, $D_{KL}(x, C)$ la divergence KL entre la résumé généré x et et la conversation C, et $\hat{s}$ la sortie finale :

$$m = \min_{x \in H} NEHR(x) \quad \text{and} \quad V = \{x \in H | NEHR(x) = m\}$$

$$\hat{s} = \min_{x \in V} D_{\text{KL}}(x \parallel C) \tag{2}$$

4 Expériences et Résultats

4.1 Contexte expérimental : Jeu de données et modèles utilisés

Le corpus DECODA Bechet *et al.*, 2012 contient pour chaque transcription de conversation un résumé entre plusieurs agents du service client de la RATP et un usager. Ce sont des résumés très courts appelés synopsis. Les synopsis présentent les principaux évènements de la conversation tels que les objectifs des participants, le processus de résolution. Ce corpus se compose de 3 parties qui représentent les années d'enregistrement des conversations et ont des particularités (Voir Annexe A.1 pour la différence entre les deux parties). Pour notre part, nous n'avons utilisé que deux (DECODA-1, DECODA-3) des trois parties qui ont des synopsis associés créés et écrits par des annotateurs humains avec deux types de consignes. Tandis que les synopsis de DECODA-3 sont plus longs, écrits avec plus de détails, ceux de DECODA-1 sont synthétiques et écris en français abrégé. Le corpus a été aussi annoté en termes d'entités nommées spécifiques au domaine, avec un étiquetage morpho-syntaxique, avec des lemmes et des dépendances syntaxiques. Pour nos expériences, nous avons fusionné les deux parties. Les conversations issues du corpus DECODA couvrent une variété de motif d'appel notamment : *Information Trafic*, *Itinéraire*, *Objets perdus/trouvés*, *Abonnement*, *Horaires*, *Tickets* (Trione, 2014). La distribution de chaque motif d'appel est donné en Annexe A.2. De plus DECODA possède des entités nommées appartenant à une ontologie du domaine. On peut notamment citer *Produit, Transport, Horaire*, etc. Ces différentes annotations ainsi que le fait qu'il soit un corpus enregistré dans des conditions réelles ont motivé notre choix.

Dans le cadre de nos expériences, nous avons divisé le jeu de données en trois partitions (train pour l'entraînement des modèles, val pour le développement et test pour l'évaluation). Afin de s'assurer que chaque motif apparaît dans chaque partitions, une stratification des données en suivant le motif d'appel a été appliquée lors de la division. Nous avons calculé quelques statistiques sur les données et les avons consignées dans le tableau 1.

Stats	Train	Val.	Test
# examples	717	99	140
# conv. mots	486.3	487.9	465.4
# synopsis mots	28.8	30.3	29.5
# conv. tokens	608.2	604.7	579.8
# synopsis tokens	35.9	37.5	36.2

TABLE 1 – Distribution du jeu de données DECODA

Résumé automatique Pour l'entraînement des systèmes pour le résumé automatique, nous avons utilisé BARThez (Kamal Eddine *et al.*, 2021), un modèle séquence-à-séquence pré-entraîné sur la partie française de CommonCrawl, Wikipédia, NewsCrawl et d'autres corpus disponibles en français. Il a été introduit pour la

Conseiller : euh NNAAMMEE bonjour *Appelant :* oui bonjour je voulais savoir de la Madeleine quel bus je dois prendre pour me rendre au Saint-Philippe+du+Roule *Conseiller :* alors Madeleine Saint-Philippe+du+Roule *Appelant :* oui *Conseiller :* je recherche un instant *Appelant :* merci *Conseiller :* le cinquante-deux en direction du parc de Saint-Cloud *Appelant :* alors le cinquante-deux je le prends où ah+ben oui *Appelant :* d'accord *Conseiller :* Madeleine+Vignon *Appelant :* d'accord *Appelant :* donc cinquante-deux direction Porte+de+Saint-Cloud *Conseiller :* parc de Saint-Cloud *Appelant :* parc de Saint-Cloud et il descend à Saint-Philippe+du+Roule *Conseiller :* tout+à+fait *Appelant :* merci *Appelant :* au monsieur merci au+revoir *Conseiller :* bonne journée au+revoir	***REFERENCE SUMMARY*** Un appelant demande quel bus prendre pour se rendre de La Madeleine à Saint-Philippe-du-Roule. Le conseiller lui indique de prendre le bus 52 en direction du parc de Saint-Cloud, qui s'arrête à Saint-Philippe-du-Roule. ***CT-syn REFERENCE SUMMARY*** [MOTIF] ITNR [SUMMARY] Un appelant demande quel bus prendre pour se rendre de La Madeleine à Saint-Philippe-du-Roule. Le conseiller lui indique de prendre le bus 52 en direction du parc de Saint-Cloud, qui s'arrête à Saint-Philippe-du-Roule.

TABLE 2 – Un exemple issu du jeu de données DECODA

tâche de résumé automatique de texte. BARThez est un modèle composé de couches de Transformers (Vaswani *et al.*, 2017) et est basé sur l'architecture de BART (Lewis *et al.*, 2020). Il existe deux versions de l'architecture BARThez, une version de base et une large mBARThez. Pour notre part, nous avons utilisé la version de base qui possède 6 couches d'encodeur et 6 de decodeur comme BART base. Pour l'entraînement des modèles, nous avons utilisé le modèle pré-entraîné fourni par les auteurs et disponible sur la librairie Transformers [1] de Hugging Face.

Dans le tableau 1, nous avons consigné le nombre moyen de tokens de chaque conversation ainsi que chaque synopsis en utilisant le tokeniseur de BARThez. En ce qui concerne les hyper-paramètres, nous avons utilisé un *learning rate* de 5×10^{-5} pour l'optimiseur AdamW et avons fixé la taille maximale pour les conversations à 1024 et ceux des synopsis à 128. Chaque modèle a été entraîné sur 12 époques en ne sauvegardant que celle qui minimise la perte sur le jeu de validation. Nous avons modifié les synopsis du jeu de données en y ajoutant les motifs pour être dans les conditions pour entraîner le modèle Multitache$_M$. Un exemple d'illustration des modifications apportées au résumé de référence est dans le Tableau 2.

Pour le modèle Pipeline$_M$, les motifs associés à chaque conversation sont nécessaires en entrée, ce qui implique de les avoir préalablement. Nous avons utilisé un classifieur de motifs d'appel entraîné via une validation croisée (k-fold) pour prédire les motifs sur 25% des données non utilisées lors de l'entraînement initial (75%). Ce processus a été répété quatre fois pour obtenir les prédictions sur l'ensemble d'entraînement, puis utilisé pour l'entraînement du système de résumé automatique. Les motifs et les conversations ont été séparés par un marqueur afin d'aider le modèle à tenir compte des motifs lors de la prédiction des résumés.

Classification des motifs d'appel Nous avons entraîné un modèle CamemBERT-base (Martin *et al.*, 2020) pour la tâche de classification de séquences pour prédire les motifs d'appel. Pour ce faire, deux classifieurs ont été utilisés. L'un appliqué aux conversations *Conv-M-classif* et l'autre aux synopsis *Syn-M-classif*. Le classifieur *Conv-M-classif* prend en entrée la conversation, et prédit le motif d'appel. Ainsi, le nombre maximum de tokens pris en entrée est 512. Pour *Syn-M-classif*, on part du synopsis pour prédire le motif d'appel. La longueur maximale de l'entrée pour ce modèle est de 128. Chaque modèle a été entraîné sur 13 époques avec un batch size de 8. Le modèle ayant la loss minimale sur les données de validation a été conservé.

Reconnaissance des entités nommées du domaine DECODA possède 14 entités nommées spécifiques à un domaine, parmi lesquelles le numéro de téléphone, le prix, le type de produit, le type de transport, etc. Nous avons entraîné CamemBERT-base (Martin *et al.*, 2020) pour la tâche NER et avons obtenu un micro F1 et un macro F1 de 0,93 et 0,84 respectivement en utilisant la bibliothèque Seqeval [2]. Les entités détectées sont utilisées dans les métriques d'évaluation, telles que NEHR.

1. https://huggingface.co/moussaKam/barthez
2. https://github.com/chakki-works/seqeval

4.2 Résultats de l'évaluation

Classification des motifs d'appel Après avoir entraîné les deux modèles de classification en motifs d'appel, les résultats sont consignés dans le tableau 3. Les données étant déséquilibrées, nous avons calculé l'accuracy et le score F1 de chaque modèle. Le tableau 3 montre que la classification effectuée sur la conversation donne de meilleurs résultats que celle effectuée sur les synopsis. Ce résultat est probablement dû à la quantité d'informations supplémentaires disponibles dans la conversation qui révèlent la raison de l'appel. Le modèle Multitache$_M$ qui prédit les entités en même temps que le résumé donne des résultats significativement plus faibles que les classifieurs.

Système	Acc.	F1	W-F1
Conv-M-classif	86	65	86
Syn-M-classif	83	54	81
Multitache$_M$	79	50	76

TABLE 3 – Classification des motifs d'appel en utilisant un classifieur sur les conversations (Conv-M-classif), les synopsis (Syn-M-classif) et le modèle Multitache$_M$ générant des motifs. F1 et W-F1 sont respectivement le F1 score en macro moyenne et celui en moyenne pondérée.

Résumé automatique basé sur les motifs d'appel Dans la section 2, nous avons décrit le processus pour entraîner les différents modèles : Multitache$_M$, Pipeline$_M$. Multitache$_M$ est conçu pour une utilisation multitâche (génération du motif d'appel suivi du résumé). Il faut donc modifier les données en sortie du système pour les faire correspondre à la sortie attendue. Ainsi, nous avons ajouté le motif d'appel de référence au résumé pendant l'entraînement (voir Table 2). Pour Pipeline$_M$, nous avons construit trois différents modèles qui se distinguent par le motif d'appel donné en entrée au côté de la conversation lors de l'entraînement et de la prédiction. Le premier est un système d'oracle (Pipeline$_M$ − oracle). En effet, en entrée du modèle, nous donnons le motif d'appel de référence, que ce soit pour l'entraînement ou pour le test. Le second est Pipeline$_M$ − oracle/pred ; nous supposons avoir accès au motif d'appel de référence pendant la phase d'entraînement, mais pas pendant la phase de test.

Par conséquent, en utilisant le classifieur (Conv-M-classif), nous avons prédit les motifs d'appel pour chaque exemple du jeu de test. Ces prédictions ont ensuite été combinées à la conversation et transmises au modèle pour la génération de résumé. Pour le dernier modèle (Pipeline$_M$), nous supposons que nous n'avons pas accès aux motifs d'appel de référence de chaque conversation pendant l'entraînement. Ainsi, nous avons donc dû les prédire. Pour la phase de test, nous avons procédé de la même manière que pour le modèle (Pipeline$_M$ − oracle/pred).

Nous avons comparé ces systèmes avec un modèle baseline qui génère un synopsis à partir de la conversation uniquement en entrée.

System	R1	R2	RL
Baseline	34.04	14.91	28.83
Multitache$_M$	33.59	14.23	28.42
Pipeline$_M$ − oracle	34.65	15.10	29.33
Pipeline$_M$ − oracle/pred	34.46	14.93	29.09
Pipeline$_M$	34.09	14.79	28.99

TABLE 4 – R1, R2, RL : ROUGE-Score des différents systèmes.

Métrique	Système	1st
Fidélité	Baseline	38.9
	Multitache$_M$	16.7
	Pipeline$_M$	44.4
Informativité	Baseline	33.3
	Multitache$_M$	23.8
	Pipeline$_M$	42.9

TABLE 5 – Évaluation manuelle : 1st - Nombre de fois (en pourcentage) où chaque système a été classé premier en termes de fidélité et d'informativité.

— **Évaluation automatique** Pour évaluer les différents systèmes, nous avons calculé ROUGE (Lin, 2004)

afin de tester la capacité du modèle à générer des n-grammes proches du résumé de référence. Le tableau 4 montre les résultats obtenus. Les différents systèmes donnent des résultats similaires en termes de ROUGE. Comme Zhou *et al.* (2022) montrent que les résumés générés qu'ils ont obtenus de DECODA contiennent des hallucinations, des omissions et des erreurs grammaticales, nous évaluons manuellement le niveau de fidélité des résumés.

— **Évaluation manuelle** Nous évaluons manuellement 30 résumés de systèmes en fonction de leur fidélité et de leur caractère informatif. L'idée est de choisir entre les trois systèmes les plus fidèles et les plus informatifs. Un système est dit fidèle à la conversation s'il ne contient aucune information la contredisant. Un système est informatif s'il couvre les informations les plus essentielles de la conversation. Ceci peut être mesuré en comparant le résumé d'un système avec le résumé de référence correspondant, qui devrait contenir les informations les plus essentielles de la conversation. Les résultats sont présentés dans le tableau 5. Nous observons que Pipeline$_M$ est plus souvent considéré comme le meilleur résultat, tant en termes de fidélité que d'informativité. Il est à noter que Multitache$_M$ est moins bon que la ligne de base, probablement en raison de la difficulté du système à générer des motifs d'appel.

Sélection de résumé Dans le but de maximiser la similarité en termes de motif d'appel entre le résumé prédit et la référence, nous avons généré une cohorte de résumés et effectué une sélection afin de minimiser la KL CONV. Nous n'avons utilisé que le modèle Baseline et le *Pipeline*$_M$ pour cette section. En moyenne 40 résumés ont été générés pour chaque système, et le meilleur en fonction de deux critères de sélection a été choisi. Le premier critère consiste à choisir le résumé présentant la divergence KL minimale entre la distribution de probabilité de la conversation et la distribution de probabilité du résumé généré ($\min D_{\mathrm{KL}}(G \parallel \mathrm{Conv})$). Le second critère est basé sur les équations 2 décrites dans la section 3. D'abord une sélection de l'ensemble des résumés ayant le NEHR minimum, puis, parmi eux, le résumé minimisant la $\min D_{\mathrm{KL}}(G \parallel \mathrm{Conv})$ ($\min \mathrm{NEHR} + D_{\mathrm{KL}}(G \parallel Conv)$).

Pour l'évaluation des résumés sémantiquement, nous avons introduit quatre métriques basées sur les informations sémantiques de la conversation tels que le motif d'appel et les entités nommées.

— **CT-Acc** : Il s'agit de la mesure de l'accuracy du classifieur en motif d'appel appliqué aux résumés générés par rapport aux motifs de référence. Nous considérons que plus l'accuracy est haute, plus le résumé contient d'éléments cohérents avec le motif d'appel que le classifieur utilise pour prédire la bonne étiquette.

— **NE-R/NE-P/NE-F1** : Nan *et al.*, 2021 a introduit des métriques qui comparent les entités nommées des résumés générés à ceux des résumés de référence. Basé sur leur metrique, nous mesurons la précision, le rappel et le F1 concernant les entités nommées détectées dans le résumé automatique par rapport aux entités dans le résumé de référence. La différence fondamentale avec leur proposition réside dans ce qui est considéré comme étant une entité nommée. Pour notre part, nous intégrons les quantités, les nombres, dans la définition des entités nommées et basons nos métriques sur les entités nommées du domaine à notre étude. Nous considérons que plus ces valeurs sont élevés, moins grand est le risque d'hallucination de la part du modèle.

Les métriques étant définies, nous nous attendons à ce que la minimisation de la divergence KL ($\min D_{\mathrm{KL}}(G \parallel \mathrm{Conv})$) augmente le **CT-Acc** et que le minimisation du NEHR augmente le rappel et la précision sur les entités nommées (NE-R et NE-P). Ainsi, la combinaison des deux métriques devrait nous permettre d'augmenter aussi bien les mesures sur les entités que l'accuracy sur les motifs d'appel.

Le tableau 6 présente les résultats obtenus après la sélection du résumé, tout d'abord selon le score ROUGE par rapport aux résumés de référence, puis par les mesures sémantiques définies précédemment.

Pour chaque modèle, nous indiquons également le score obtenu par le modèle sans sélection prenant juste la meilleure hypothèse (BEAM). Nous avons aussi rajouté le résumé qui permet d'avoir le NEHR minimim en suivant le papier Akani *et al.*, 2023.

Le tableau 6 montrent de faibles variations en terme de score ROUGE entre les différents modèles. Par contre les scores sémantiques évoluent fortement : la minimisation explicite de la divergence sur les motifs d'appels augmentent très significativement la mesure **CT-Acc**, particulièrement pour le modèle *Pipeline*$_M$ et en combinant

la minimisation du risque sur les entités et la divergence sur les motifs, nous montrons une augmentation significative de la précision sur les entités en conservant un gain important en accuracy sur les motifs d'appels et une dégradation légère en terme de rappel.

	R1 ↑	R2 ↑	RL ↑	CT-Acc ↑	NE-R ↑	NE-P ↑	NE-F1 ↑
Baseline model - BARThez finetuné pour générer les synopsis							
BEAM	34.04	14.91	28.83	76.4	0.46	0.57	0.51
min NEHR	34.12	14.13	27.79	73.6	**0.50**	0.58	**0.54**
min $D_{\mathrm{KL}}(G \parallel \mathrm{Conv})$	33.99	14.23	28.13	**82.9**	0.43	0.54	0.48
min NEHR + $D_{\mathrm{KL}}(G \parallel Conv)$	33.70	14.57	28.04	82.1	0.45	**0.60**	0.51
Pipeline$_M$ - baseline + motif prédit en entrée, le synopsis en sortie							
BEAM	34.09	14.79	28.99	76.4	**0.47**	0.49	0.48
min NEHR	34.18	14.69	28.52	75.7	**0.47**	0.55	**0.51**
min $D_{\mathrm{KL}}(G \parallel \mathrm{Conv})$	33.85	14.60	28.87	**85.7**	0.42	0.53	0.47
min NEHR + $D_{\mathrm{KL}}(G \parallel Conv)$	33.32	14.48	28.58	84.3	0.45	**0.57**	0.50

TABLE 6 – R1,R2,RL : ROUGE ; CT-Acc : Prédiction des motifs depuis les synopsis (ref=oracle synopsis) ; NE-R et NE-P : Le rappel et la précision des entités des résumés générés par rapport à la référence

5 Discussion et Conclusion

Nous avons examiné l'impact de deux modèles basés sur le motif d'appel pour améliorer la représentation sémantique des résumés générés. Bien que les métriques d'évaluation automatique comme ROUGE montrent des résultats similaires à la baseline, une évaluation humaine révèle que les résumés Pipeline$_M$ ont été choisi comme plus fidèle et informative en comparaison aux systèmes Baseline et Multitache$_M$.

L'utilisation de KL divergence comme critère de sélection des résumés montre des promesses, avec une augmentation d'environ 10% de l'accuracy (CT-Acc) lorsque le modèle Syn-M-classif est appliqué aux résumés générés. Cependant, certains motifs d'appel n'ont pas été suffisamment représentés dans le jeu de données, suggérant la nécessité d'une étude plus approfondie pour améliorer la prédiction du motif et, par conséquent, la qualité du résumé associé.

La combinaison de la KL divergence sur la conversation et du NEHR comme critère de sélection améliore la précision du modèle sur les entités nommées par rapport à la référence. Cependant, une évaluation manuelle plus poussée est nécessaire pour confirmer l'efficacité de ce critère combiné en termes de fidélité et d'informativité. Étant donné que toutes les méta-données du corpus n'ont pas été exploitées, nous envisageons d'intégrer d'autres aspects, notamment la structure de la conversation, pour continuer à améliorer la fidélité des résumés générés.

Les expériences ont été menées sur le corpus DECODA en utilisant le modèle BARThez, pre-entraîné sur du français. Le choix du corpus a été motivé par sa richesse en terme d'annotations mais également parce qu'il est un corpus enregistré dans un cas réel ; il a donc des conversations naturelles. Notre méthodologie a été utilisée dans ce cas précis mais nous croyons que les conclusions seront les mêmes quelques soit le corpus ou le modèle qui sera utilisé.

De plus, on peut envisager d'utiliser le corpus X-RiSaWoz (Moradshahi *et al.*, 2023) pour nos expériences. L'inconvénient de ce corpus est qu'il n'est pas issu de conversations naturelles et qu'il ne possède pas de résumé associé à chaque conversation. Néanmoins, il s'agit d'un grand corpus qui peut être utiliser suivant notre méthodologie grâce à l'usage des LLMs pour la génération de résumés. C'est une piste que nous souhaiterons explorer.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012525 attribuée par GENCI.

Références

AKANI E. (2023). Étude de la fidélité des entités dans les résumés par abstraction. In M. CANDITO, T. GERALD & J. G. MORENO, Éds., *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, p. 21–36, Paris, France : ATALA.

AKANI E., FAVRE B., BECHET F. & GEMIGNANI R. (2023). Reducing named entity hallucination risk to ensure faithful summary generation. In C. M. KEET, H.-Y. LEE & S. ZARRIESS, Éds., *Proceedings of the 16th International Natural Language Generation Conference*, p. 437–442, Prague, Czechia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.inlg-main.33.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BÈZE M., DE MORI R. & ARBILLOT E. (2012). DECODA : a call-centre human-human spoken conversation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1343–1347, Istanbul, Turkey : European Language Resources Association (ELRA).

CAO Z., WEI F., LI W. & LI S. (2018). Faithful to the original : Fact aware neural abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1). DOI : 10.1609/aaai.v32i1.11912.

CARLETTA J., ASHBY S., BOURBAN S., FLYNN M., GUILLEMOT M., HAIN T., KADLEC J., KARAISKOS V., KRAAIJ W., KRONENTHAL M., LATHOUD G., LINCOLN M., MASSON A. L., MCCOWAN I., POST W., REIDSMA D. & WELLNER P. D. (2005). The ami meeting corpus : A pre-announcement. In *Machine Learning for Multimodal Interaction*.

DURMUS E., HE H. & DIAB M. (2020). FEQA : A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 5055–5070, Online : ACL. DOI : 10.18653/v1/2020.acl-main.454.

FABBRI A., WU C.-S., LIU W. & XIONG C. (2022). QAFactEval : Improved QA-based factual consistency evaluation for summarization. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2587–2601, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.187.

FALKE T., RIBEIRO L. F. R., UTAMA P. A., DAGAN I. & GUREVYCH I. (2019). Ranking generated summaries by correctness : An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the ACL*, p. 2214–2220, Florence, Italy : ACL. DOI : 10.18653/v1/P19-1213.

FAN A., LEWIS M. & DAUPHIN Y. (2018). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the ACL (Volume 1 : Long Papers)*, p. 889–898, Melbourne, Australia : ACL. DOI : 10.18653/v1/P18-1082.

GLIWA B., MOCHOL I., BIESEK M. & WAWER A. (2019). SAMSum corpus : A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, p. 70–79, Hong Kong, China : ACL. DOI : 10.18653/v1/D19-5409.

GOO C.-W. & CHEN Y.-N. (2018). Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 735–742. DOI : 10.1109/SLT.2018.8639531.

HOLTZMAN A., BUYS J., DU L., FORBES M. & CHOI Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

JANIN A., BARON D., EDWARDS J., ELLIS D., GELBART D., MORGAN N., PESKIN B., PFAU T., SHRIBERG E., STOLCKE A. & WOOTERS C. (2003). The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, p. I–I. DOI : 10.1109/ICASSP.2003.1198793.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of hallucination in natural language generation. **55**(12). DOI : 10.1145/3571730.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9369–9390, Online and Punta Cana, Dominican Republic : ACL. DOI : 10.18653/v1/2021.emnlp-main.740.

KOAY J. J., ROUSTAI A., DAI X., BURNS D., KERRIGAN A. & LIU F. (2020). How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5689–5695, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.499.

KULLBACK S. & LEIBLER R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**(1), 79–86. Publisher : Institute of Mathematical Statistics.

LEE J., KIM D., SARIKAYA R. & KIM Y.-B. (2018). Coupled representation learning for domains, intents and slots in spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 714–719 : IEEE.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 7871–7880, Online : ACL. DOI : 10.18653/v1/2020.acl-main.703.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : ACL.

LUO Z., XIE Q. & ANANIADOU S. (2023). Chatgpt as a factual inconsistency evaluator for text summarization.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 7203–7219, Online : ACL. DOI : 10.18653/v1/2020.acl-main.645.

MAYNEZ J., NARAYAN S., BOHNET B. & MCDONALD R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the ACL*, p. 1906–1919, Online : ACL. DOI : 10.18653/v1/2020.acl-main.173.

MORADSHAHI M., SHEN T., BALI K., CHOUDHURY M., DE CHALENDAR G., GOEL A., KIM S., KODALI P., KUMARAGURU P., SEMMAR N., SEMNANI S., SEO J., SESHADRI V., SHRIVASTAVA M., SUN M., YADAVALLI A., YOU C., XIONG D. & LAM M. (2023). X-RiSAWOZ : High-quality end-to-end multilingual dialogue datasets and few-shot agents. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 2773–2794, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.174.

NAN F., NALLAPATI R., WANG Z., NOGUEIRA DOS SANTOS C., ZHU H., ZHANG D., MCKEOWN K. & XIANG B. (2021). Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the ACL : Main Volume*, p. 2727–2733, Online : ACL. DOI : 10.18653/v1/2021.eacl-main.235.

TANG X., NAIR A., WANG B., WANG B., DESAI J., WADE A., LI H., CELIKYILMAZ A., MEHDAD Y. & RADEV D. (2022). CONFIT : Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the ACL : Human Language Technologies* : ACL. DOI : 10.18653/v1/2022.naacl-main.415.

TRIONE J. (2014). Extraction methods for automatic summarization of spoken conversations from call centers (méthodes par extraction pour le résumé automatique de conversations parlées provenant de centres d'appels) [in French]. In *Proceedings of TALN 2014 (Volume 4 : RECITAL - Student Research Workshop)*, p. 104–111, Marseille, France : Association pour le Traitement Automatique des Langues.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG B., ZHANG C., ZHANG Y., CHEN Y. & LI H. (2022). Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4897–4908, Abu Dhabi, United Arab Emirates : ACL. DOI : 10.18653/v1/2022.emnlp-main.325.

ZHAO L., ZHENG F., HE K., ZENG W., LEI Y., JIANG H., WU W., XU W., GUO J. & MENG F. (2021). Todsum : Task-oriented dialogue summarization with state tracking.

ZHOU Y., PORTET F. & RINGEVAL F. (2022). Effectiveness of French language models on abstractive dialogue summarization task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3571–3581, Marseille, France : European Language Resources Association.

A Corpus DECODA

A.1 Example du corpus DECODA

DIALOGUE : *autre :* NNAAMMEE va vous répondre *autre :* NNAAMMEE bonjour *appelant :* bonjour monsieur *appelant :* monsieur j' ai un renseignement à vous demander *autre :* oui *appelant :* j' ai grand garçon handicapé mental qui a une un coupon Améthyste *autre :* hm *appelant :* bon évidemment ça ne dure pas un an il se démagnétise et avant j' allais euh trente rue Championnet *autre :* oui *appelant :* le faire changer *autre :* hm *appelant :* mais leur numéro de téléphone a changé auriez vous l' amabilité de lu de me donner le numéro le nouveau numéro *autre :* alors VGC à Championnet c' est le zéro un cinquante-huit *appelant :* oui *autre :* soixante-dix-sept trois fois *appelant :* soixante-dix-sept trois fois *autre :* hm *appelant :* très bien je vous en remercie *appelant :* monsieur *autre :* bonne *autre :* journée madame au revoir *appelant :* au revoir **RÉSUMÉ :** numéro VGC pour faire changer carte Améthyste démagnétisée
autre : va vous répondre *conseiller :* bonjour *appelant :* oui allô bonjour *appelant :* je vous appelle car euh j'ai perdu mon porte-monnaie dans le... dans le train, et ça serait pour savoir quelles sont, euh comment on fait en fait ? *conseiller :* oui quel euh, quel train ? *appelant :* euh c'était le train pour aller à Montreau *conseiller :* oui il faut voir avec la SNCF madame *appelant :* d'accord *appelant :* je vous remercie *conseiller :* mais je vous en prie *appelant :* au revoir, bonne journée *conseiller :* bonne journée, au revoir **RÉSUMÉ :** Une appelante a perdu son porte-monnaie en Gare pour aller à Montreau. Le conseiller lui rappelle donc qu'il s'agit d'une requête à adresser à la SNCF et non à la RATP.

TABLE 7 – Différence entre les parties 1 (Premier exemple : 20091112_RATP_SCD_0152) et 3 (deuxième exemple : 20110704_RATP_SCD_0018) de DECODA.

A.2 Distribution des motifs d'appel de DECODA

Cette annexe présente la distribution des motifs d'appel dans notre jeu de données. On distingue 15 motifs d'appel selon la demande de l'appelant. On peut en citer : ITNR (demande d'itinéraire), OBJT (Objet trouvé/perdu) NVGO (Pass navigo), etc.

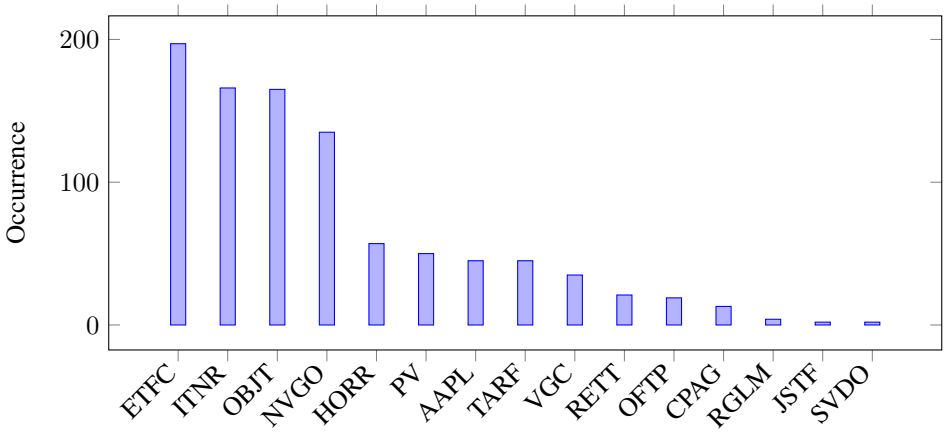

FIGURE 1 – Distribution du motif d'appel de DECODA pour le jeu de données présenté au tableau 1