# Segmentation de phases de dialogue dans des retranscriptions de conversations de centres d'appels

Guillaume Dubuisson Duplessis, Manon Richard, Anne-Laure Guénet

EDF Commerce, Direction des Systèmes d'Information et du Numérique (DSIN),
CSC datascience & IA, 420 rue Estienne d'Orves, 92700 Colombes

{guillaume.dubuisson-duplessis, anne-laure.guenet}@edf.fr

## Résumé

*La segmentation de phases de dialogue dans les retranscriptions de conversations de centres d'appels est cruciale pour leur exploitation opérationnelle. Cet article présente une approche d'apprentissage automatique supervisée qui nécessite une charge d'annotation manuelle raisonnable pour la segmentation. Cette approche est basée sur une ingénierie des caractéristiques qui prend en compte la nature conversationnelle des données. Les résultats de l'étude fournissent des enseignements clés et amènent une discussion sur la définition même des phases de dialogue.*

## Mots-clés

*TALN, segmentation de phase de dialogue, retranscription*

## Abstract

*The task of dialogue phase segmentation in call center conversation transcriptions is a crucial step for operational purposes. This research presents a supervised machine learning approach that requires a reasonable amount of manual annotation for segmentation and is based on feature engineering that takes into account the conversational nature of the data. The study describes key insights, including a discussion on the very definition of dialogue phases.*

## Keywords

*NLP, dialogue segmentation, retranscription*

## 1 Introduction

Les appels téléphoniques représentent la majeure partie des contacts de la relation client sur le marché des clients particuliers d'un acteur comme « Électricité de France » (EDF). Les retranscriptions automatiques peuvent nourrir de nombreux cas d'usage [2] comme, par exemple, l'optimisation des appels (aide à la professionnalisation, optimisation du discours des conseillers, optimisation des temps de traitements) et la vérification de la qualité des conversations en fonction de standards établis.

L'exploitation des retranscriptions de conversations de centres d'appels à des fins opérationnelles se fonde généralement sur une tâche clé : la segmentation en phases de dialogue. Le Tableau 1 présente un exemple d'enchaînement de deux phases de dialogue dans un appel de mise en service. La tâche de segmentation vise à reconstruire automatiquement les phases de dialogue à partir des retranscriptions. Un dialogue est un objet structuré [4] impliquant des structures locales comme la paire adjacente [11] et des structures plus étendues comme la structure dite « intentionnelle » liée à la tâche sous-jacente [6]. L'expression « phase de dialogue » peut référer à ces structures plus ou moins étendues. Dans ces travaux, nous nous focalisons sur des dialogues orientés par une tâche sous-jacente de mise en service d'un contrat de fourniture d'énergie. Par phase de dialogue, nous référons aux étapes suivies par le conseiller pour mener à bien une mise en service. Schématiquement, les appels suivent une structure enchaînant les étapes suivantes : salutations, description de la demande, description du logement, localisation du logement, choix des offres et services, mise en place du paiement, signature électronique des contrats, rendez-vous technicien et clôture de l'appel. Ces travaux se focalisent sur l'extraction de la phase de signature électronique des contrats à des fins analytiques en vue de mieux connaître celle-ci (temps passé, sujets abordés par le conseiller et le client) et d'identifier des leviers d'op-

| | | |
|---|---|---|
| CC : | 1 | oui d'accord vous pouvez vous connecter sur votre boîte mail on va créer ensemble votre espace client |
| | 2 | on va faire une signature dématérialiser de vos contrats donc pas besoin d'imprimer et de scanner d'accord |
| C : | 3 | oui |
| CC : | 4 | quand tout sera signé je pourrai vous envoyer les documents pour votre bailleur d'accord |
| C : | 5 | d'accord ok |
| ... | | ... |
| CC : | 6 | ensuite pour la date de mise en service |
| C : | 7 | oui j'aurais voulu savoir si c'est possible d'avoir l'électricité dès jeudi |

TABLE 1 – Un exemple de segmentation de phases de dialogue dans des retranscriptions d'appel téléphonique illustrant l'enchaînement de la phase de signature électronique et la phase de date de mise en service. Il contient sept unités inter-pausales (UIP). CC=« conseiller client », C=« client ».

timisation via l'analyse des phases anormalement longues. Les contributions de ces travaux sont multiples. Tout d'abord, nous nous intéressons à la tâche de segmentation de retranscriptions d'appels, peu traitée par manque de corpus disponible mais pourtant clé pour l'exploitation de ces données riches [2, 12]. Ensuite, nous présentons une approche d'apprentissage automatique supervisée opérationnelle nécessitant une charge d'annotation manuelle raisonnable. Elle est fondée sur une ingénierie des caractéristiques prenant en compte la nature conversationnelle des données. Enfin, nous présentons les principaux enseignements parmi lesquels une discussion sur la définition même des phases de dialogue.
La suite de l'article se découpe en trois parties. La Section 2 pointe les travaux connexes les plus saillants. La Section 3 forme le cœur de l'article. Elle présente l'approche proposée pour la segmentation de phases de dialogue, les données d'apprentissage utilisées, les principales expérimentations réalisées et discute les résultats obtenus. Enfin, la Section 4 clôt cet article en soulignant les principales conclusions et en identifiant quelques perspectives prometteuses.

# 2 Travaux connexes

La segmentation en phases de dialogue de retranscriptions de conversations de centres d'appels est une tâche clé permettant l'exploitation des retranscriptions à de nombreuses fins opérationnelles [2]. Compte-tenu du coût important en termes d'annotation de données et de l'indisponibilité de données visant cette tâche, la plupart des approches de segmentation de phases de dialogue se base sur une approche non-supervisée fondée sur une hypothèse de cohérence lexicale et sémantique [7, 1]. Ces approches ont des performances plutôt modestes et limitées pour un usage opérationnel visant à calculer des indicateurs précis sur des phases. Des travaux récents visent à améliorer les performances de ces approches non-supervisées en apprenant automatiquement des modèles évaluant la cohérence d'une paire d'énoncés [12]. Ces travaux combinent les apports de l'apprentissage par transfert via un modèle BERT [3] à la relative facilité de construction d'un corpus d'apprentissage pour entraîner un modèle évaluant la cohérence d'une paire d'énoncés. Nos travaux se positionnent dans un contexte opérationnel de minimisation de l'utilisation de la donnée. Ils se fondent sur une tâche de classification de texte employant des caractéristiques conversationnelles hétérogènes permettant un niveau de performance satisfaisant.

# 3 Approche proposée

## 3.1 Vision globale

Afin de détecter une phase dans un appel, la retranscription n'est pas traitée comme un bloc, mais est découpée en unité inter-pausales (UIP). Une UIP est une unité de parole ne contenant pas de pause et provenant d'un seul interlocuteur (cf. Tableau 1). Une retranscription est donc découpée en une multitude d'UIP. La tâche revient alors à une classification binaire : une UIP appartient-elle ou non à la phase d'intérêt ? (cf. Figure 1) Ce découpage permet d'avoir une plus

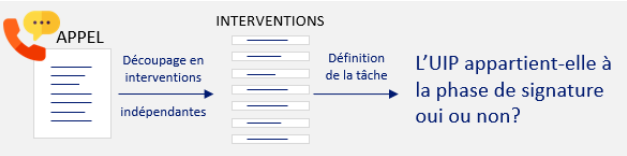


FIGURE 1 – Reformulation du problème de segmentation de phases de dialogue en une tâche de classification au niveau des unités inter-pausales (UIP).

grande volumétrie pour la modélisation. Une fois la classification réalisée, chaque UIP a une prédiction – dans ou hors de la phase – mais il peut parfois manquer de continuité dans la prédiction. Il est alors nécessaire de reconstruire la phase, par exemple en exploitant la densité de prédiction.

## 3.2 Description du corpus d'apprentissage

Le corpus est composé de 700 retranscriptions d'appels téléphoniques récoltées dans le cadre d'appels de mise en service. Dans ces données, un appel a une durée moyenne de 23min. Les retranscriptions sont donc de taille conséquente : en moyenne un appel comporte 960 UIP. En vertu du règlement général sur la protection des données (RGPD), les retranscriptions ont été désidentifiées et ne comportent aucune donnée à caractère personnel telles que des noms, prénoms, adresses, numéros [5]. Les données à caractère personnel sont substituées par le type de l'entité (e.g., "je suis monsieur *[person]* je vous appelle pour mon logement situé à *[localisation]*"). Les retranscriptions ne font apparaître aucune ponctuation et peuvent comporter des erreurs de retranscription.
Afin de détecter les phases de signature électronique dans les appels, une annotation des retranscriptions a été réalisée sur les 700 retranscriptions par deux personnes, ce qui représente une charge opérationnelle acceptable dans le cadre du projet. L'annotation a consisté à noter le début et la fin de la phase. Le corpus a ensuite été découpé en 3 sous-échantillons : un échantillon *train* composé de 420 appels (257K UIP) qui représente 60% des annotations, un échantillon *dev* (20%) composé de 140 appels (85K UIP) et un échantillon *test* (20%) de 140 appels (90K UIP).

## 3.3 Expérimentations autour de la classification des unités inter-pausales

### 3.3.1 Métriques d'évaluation considérées

La phase de signature électronique ne représente qu'une phase mineure dans une retranscription d'appel. L'échantillon est déséquilibré : 15% des UIP font partie de cette phase. Afin de mesurer les performances du modèle de classification, nous avons utilisé la métrique MCC (*Matthews Correlation Coefficient*) afin de prendre en compte ce déséquilibre. D'autres métriques ont également été visualisées à titre indicatif : la précision, le rappel et l'*accuracy*.

### 3.3.2 Choix de modélisation

**Caractéristiques considérées** Afin de prédire si une UIP fait partie de la phase considérée ou non, plusieurs variables ont été intégrées dans le modèle. Certaines décrivent direc-

tement l'UIP comme le texte de l'UIP, l'interlocuteur de l'UIP (soit « client » soit « conseiller »), la durée de l'UIP, la position de l'UIP dans l'appel (entre 0 – début – et 1 – fin – de la conversation). D'autres concernent le contexte dans lequel l'UIP a été mentionnée. Dans une fenêtre de $T$ UIP avant et après – taille à définir – la durée des blancs (qui peut marquer un changement de phase dans le dialogue) ainsi que les UIP avant et après (ce qui a été dit par le client et conseiller dans une fenêtre de $T$ UIP avant ou après) sont intégrées en variables supplémentaires dans le modèle.Ainsi, six caractéristiques structurées (interlocuteur, durée, position dans l'appel, durée des blancs avant et après l'UIP) et deux caractéristiques non-structurées (texte de l'UIP et contexte avant/après) sont intégrées en variables explicatives du modèle.

**Vectorisations considérées pour les données texte** À partir de l'ensemble de ces variables explicatives, nous avons construit un modèle de classification binaire. Afin de traiter les données textuelles, nous avons testé différentes vectorisations. Dans un premier temps, nous avons privilégié un TF-IDF, puis nous avons réduit le nombre de variables du TF-IDF en le combinant à une SVD (*Singular Value Decomposition*). Enfin, nous avons testé GloVe [10] en dimension 100, modèle pré-entraîné sur des UIP de retranscriptions. La représentation au niveau du document est obtenue par *mean* ou *max pooling*. Les données ont certaines spécificités : 50% des UIP ont entre 1-2 *tokens* uniquement et le mot le plus utilisé apparaît uniquement dans 19% des UIP contrairement à un corpus commun où le mot le plus utilisé comme un mot vide apparait généralement dans plus de 95% des documents. Les hyperparamètres du TF-IDF sont à adapter à cette particularité. En outre, nous n'avons pas utilisé de modèles comme CamemBERT [9] ou FlauBERT [8] pour des raisons de tractabilité sur des conversations qui sont longues. De plus, les résultats rapportés pour le modèle CamemBERT indiquent des performances moindres sur des données de retranscription [9].

**Classifieurs considérés** Nous avons considéré plusieurs classifieurs : une régression logistique, un SVM (*Support Vector Machine*) avec kernel linéaire et un perceptron multi-couche (MLP) non-linéaire avec une couche cachée de taille 100. Les conversations étant des documents de taille très grande, les classifieurs de type réseau de neurones récurrents n'ont pas été testés pour des raisons de temps d'entraînement. Au cours de nos essais de modélisation, le perceptron multi-couche a produit de meilleures performances que la régression logistique et le SVM avec kernel linéaire. Dans la suite, nous ne reportons que les résultats basés sur les MLP.

## 3.4 Résultats

Un premier modèle *baseline* a été construit sans prise en compte du contexte de l'UIP. Il est représenté en première ligne du Tableau 2 et est le modèle qui a donné les meilleures performances sans le contexte. Plusieurs essais de modélisation successifs ont permis d'itérer et de constater quelle variable, quelle vectorisation, quel type de modèle impactent les résultats de la classification. Il en ressort tout d'abord que la prise en compte des variables structurées en plus du texte – telles que la position de l'UIP dans la conversation – augmentent sensiblement la performance du modèle. La vectorisation de l'UIP basée sur le modèle GloVe préentraîné fonctionne mieux qu'une vectorisation basée sur TF-IDF et d'autant plus avec du *mean pooling*.

| Vectorisation | $T$ | $N$ | MCC CV5 | MCC dev |
|---|---|---|---|---|
| – | – | 106 | 0.454 | 0.461 |
| TF-IDF + SVD | 3 | 306 | 0.631 | 0.656 |
| GloVe | 3 | 206 | 0.667 | 0.704 |
| TF-IDF + SVD | 6 | 306 | 0.7 | **0.736** |
| GloVe | 6 | 206 | **0.719** | 0.732 |

TABLE 2 – Performances des modèles en fonction de plusieurs approches de vectorisation du contexte et de la taille du contexte considérée en nombre d'UIP avant et après. $T$ = taille du contexte. $N$ = nombre de caractéristiques.

Une fois la *baseline* fixée, le contexte a été rajouté dans la modélisation. Les résultats sont retranscrits dans le tableau 2 et permettent de dégager des enseignements clés. Tout d'abord, le contexte apporte énormément d'information au modèle, en particulier une fenêtre $T$ de taille 6. A noter que différencier le contexte avant du contexte après ne nous a pas apporté de gains de performance. De nouveau, la vectorisation du contexte basée sur notre modèle GloVe préentraîné fonctionne mieux qu'une vectorisation basée sur TF-IDF et a l'avantage de limiter le nombre de caractéristiques nécessaires pour une bonne classification. Nous avons également constaté que les UIP sont moins bien prédites lorsqu'elles sont au bord de la phase de signature électronique (mais la vérité terrain contient aussi des erreurs sur les bords).

Pour les besoins opérationnels de notre projet, nous avons retenu le modèle basé sur une vectorisation GloVe du contexte de taille 6 (dernière ligne du Tableau 2). Nous avons vérifié la bonne généralisation de ce modèle sur la partie "test" de notre corpus (MCC=0.736 / accuracy=0.94).

## 3.5 Reconstruction des phases de dialogue

La reconstruction de la phase de dialogue se fonde sur les prédictions d'appartenance à la phase au niveau des UIP. L'algorithme de reconstruction utilise une fenêtre glissante qui considère le nombre d'UIP prédites comme appartenant à la phase dans la fenêtre pour déterminer la phase de dialogue. Cela permet d'éliminer des UIP isolées faussement détectées comme faisant partie de la phase de signature électronique. Les paramètres qui nous ont donné les résultats les plus satisfaisants sont une taille de fenêtre de 10 UIP et une nombre minimum de voisins prédits dans la phase à 4. Nous avons mesuré la performance de la reconstruction de phase via l'indice de Jaccard (correspondant au calcul de l'intersection entre la phase prédite divisée par l'union de la phase de référence avec celle prédite). L'indice de Jaccard médian est de 0.73. 75% des appels ont un indice de Jaccard entre 0.5 et 0.87.

### 3.6 Discussion

Cette approche opérationnelle a la particularité de découper le problème en deux tâches (classification des UIP, reconstruction de la phase). Avec un nombre restreint d'annotations, la modélisation a donné des résultats stables sur les échantillons de *dev* et *test* via la prise en compte de caractéristiques conversationnelles hétérogènes.

L'une des difficultés dans la segmentation de phase de dialogue et qui a pu complexifier la reconstruction des phases réside dans deux particularités inhérentes aux phases de dialogue. La première est la présence de sous-dialogues incidents qui peuvent s'immiscer au sein de la phase [4] (e.g., « ok attendez je voulais savoir si je suis sur le principe des heures creuses heures pleines »). La seconde est la présence de phases mêlées (e.g., « donc pendant que vous êtes en train de valider je vais en profiter pour vous donner les conseils en matière d'économie d'énergie »). Dans ces deux cas, la phase segmentée n'est plus unifiée, mais séparée en deux ou plusieurs sous-phases. Cela présente un défi aussi bien de formalisation de la tâche de segmentation que technique pour de futurs travaux.

## 4 Conclusion et perspectives

Cet article a présenté une approche par apprentissage automatique pour la segmentation d'une phase de dialogue dans de réelles retranscriptions d'appels en centres d'appels ; approche plus générique, rigoureuse, réplicable et maintenable qu'une conception de règles spécifiques. Cette approche supervisée se base sur un corpus d'apprentissage de petite taille et des caractéristiques conversationnelles adaptées à cette tâche. Ces travaux se sont limités à une méthode de reconstruction d'une phase très simple à partir des prédictions au niveau des UIP. Une suite directe vise à améliorer cette méthode de reconstruction de phases. Une autre direction est d'étudier l'adaptation des modèles d'apprentissage profond pour segmenter des phases dans des séquences d'UIP pouvant être très longues. Enfin, une direction intéressante est d'explorer l'opérationnalité d'approches nécessitant pas ou peu de supervision manuelle [12].

## Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues de près ou de loin sur ce projet : Gilles Pouëssel, Mélanie Cazes, Laura Rouhier, Sonia Audheon, Marie Hervé, François Raynaud.

## Références

[1] Laurent Bozzi, Philippe Suignard, and Claire Waast-Richard. Segmentation et classification non supervisée de conversations téléphoniques automatiquement retranscrites. In *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 331–336, 2009.

[2] Chloé Clavel, Gilles Adda, Frederik Cailliau, Martine Garnier-Rizet, Ariane Cavet, Géraldine Chapuis, Sandrine Courcinous, Charlotte Danesi, Anne-Laure Daquo, Myrtille Deldossi, et al. Spontaneous speech and opinion detection : mining call-centre transcripts. *Language resources and evaluation*, 47(4) :1089–1125, 2013.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4171–4186, 2019.

[4] Guillaume Dubuisson Duplessis. *Modèle de comportement communicatif conventionnel pour un agent en interaction avec des humains : Approche par jeux de dialogue*. PhD thesis, INSA de Rouen, 2014.

[5] Guillaume Dubuisson Duplessis, Elliot Bartholme, Sofiane Kerroua, Mathilde Poulain, Ahès Roulier, and Anne-Laure Guénet. Désidentification de données texte produites dans un cadre de relation client. In *Actes de la 27eme conférence Traitement Automatique des Langues Naturelles (TALN) – démonstrations*, pages 10–13, 2020.

[6] Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3) :175–204, 1986.

[7] Marti A. Hearst. Text tiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1) :33–64, 1997.

[8] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert : Unsupervised language model pre-training for french. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 2020.

[9] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[11] Emanuel A Schegloff and Harvey Sacks. Opening up closings. 1973.

[12] Linzi Xing and Giuseppe Carenini. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177, 2021.