# Annotation en Actes de Dialogue pour les Conversations d'Assistance en Ligne

Robin Perrotin Alexis Nasr Jeremy Auguste
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
prenom.nom@lis-lab.fr

RÉSUMÉ

Les conversations techniques en ligne sont un type de productions linguistiques qui par de nombreux aspects se démarquent des objets plus usuellement étudiés en traitement automatique des langues : il s'agit de dialogues écrits entre deux locuteurs qui servent de support à la résolution coopérative des problèmes des usagers. Nous proposons de décrire ici ces conversations par un étiquetage en actes de dialogue spécifiquement conçu pour les conversations en ligne. Différents systèmes de prédictions ont été évalués ainsi qu'une méthode permettant de s'abstraire des spécificités lexicales du corpus d'apprentissage.

ABSTRACT

**Dialog Acts Annotations for Online Chats**

Technical online chats distinguish themselves from more usual natural language production. They are written dialogs between two locutors that are meant to solve the user's problems. We describe here such chats with dialog acts annotation tailored for theirs specificities. Several prediction systems are described and evaluated, as well as a method that allows to learn models that abstract away from the lexical specificities of the training corpus.

MOTS-CLÉS : TAL, Traitement Automatique des Langues, Actes de Dialogues, Conversations en Ligne, Big Data, CRF, Réseaux Neuronaux.

KEYWORDS: NLP, Natural Language Processing, Dialog Acts, Online Chats, Big Data, Conditional Random Fields, Neural Networks.

## Introduction

Ces dernières années ont vu le développement important de dispositifs de conversations instantanées en ligne, qui constituent un moyen pour de nombreuses entreprises et de services de fournir à leurs usagers un support technique. Les données générées par ces échanges constituent une trace linguistique d'une interaction entre un téléconseiller et un usager, visant à résoudre un problème rencontré par ce dernier.

Ces interactions sont intéressantes à plusieurs titres, d'un point de vue linguistique, elles permettent d'étudier sur des cas concrets la manière dont deux individus collaborent afin de résoudre un problème. Du point de vue du traitement automatique de la langues, elles constituent des productions écrites d'un genre original, souvent assez bruitées, qui donnent du fil à retordre aux outils de TAL. Du point de vue de l'entreprise, l'analyse de ces données pourrait permettre d'améliorer la qualité et l'efficacité

des échanges à la fois pour les entreprises et pour les usagers.

Nous proposons dans cet article d'annoter de telles données à l'aide d'actes de dialogues, décrivant la fonction illocutoire principale des tours dans le dialogue (Austin, 1975), notion linguistique qui a déjà été utilisée avec des jeux d'étiquettes variés pour le traitement du dialogue écrit (Core & Allen, 1997; Shriberg *et al.*, 2004; Ivanovic, 2005; Moldovan *et al.*, 2011; Salim *et al.*, 2016; Hardy *et al.*, 2003).

Une telle annotation permet de décrire le rôle que joue, dans le dialogue, chaque production des locuteurs. Elle peut participer à l'élaboration de tâches plus complexes, tel que le résumé automatique, la catégorisation automatique des conversations ou encore l'extraction d'informations. Notre étude porte plus spécifiquement sur des conversations en ligne provenant du projet ANR DATCHA [1]. Ces conversations proviennent des centres d'aide en ligne de l'entreprise Orange. Ces conversations sont de même nature que celles décrites par Damnati *et al.* (2016).

Nous avons développé un jeu d'étiquettes qui se concentre sur les aspects spécifiques des échanges où un usager et un téléconseiller coopèrent pour résoudre un enjeu dialogique [2]. Notre jeu d'étiquettes a été utilisé pour annoter un corpus, à partir duquel des outils d'annotation automatique ont été développés et évalués. Les données collectées dans le cadre du projet DATCHA, possèdent des caractéristiques générales à tout échange entre un usager et un téléconseiller ainsi que des caractéristiques spécifiques à un domaine d'activité donné, ici la téléphonie. Nous proposons d'essayer de nous abstraire du domaine d'activité en éliminant du lexique les mots spécifiques à ce dernier.

Cet article se décompose en trois parties. Il commence par une présentation de notre jeu d'étiquettes. S'ensuit une présentation de notre corpus de travail et de l'abstraction du lexique spécialisé en vue d'une généralisation des outils de prédiction. Dans un troisième temps, nous présentons les résultats obtenus à l'issue de la prédiction automatique de l'annotation en actes de dialogues.

# 1 Annotation en Actes de Dialogues

Le jeu d'étiquettes présenté ici a été conçu pour annoter des interactions langagières entre deux locuteurs distants lors d'une première interaction où l'un des locuteur (le Client C) présente un problème à l'interlocuteur (le Télé-Conseiller TC). Une spécificité importante de ces interactions est la collaboration des locuteurs en vue de la résolution d'un enjeu commun (le problème du client).

L'annotation décrit chacun des tours de parole (en respectant la segmentation des locuteurs) de la conversation par l'une des dix étiquettes présentées dans le tableau 1.

Cet ensemble d'étiquettes permet de décrire sommairement les tours, établissant leur fonction illocutoire principale [3] au sein du dialogue. L'annotation est volontairement simple et ne décrit pas tous les phénomènes dialogiques pouvant survenir. Un guide technique d'annotation (Asher *et al.*, 2017) décrit plus en détails ces étiquettes.

Dans l'exemple 1, un dialogue annoté extrait de notre corpus est présenté. Structurellement, le dialogue se construit par une courte séquence d'ouverture suivie de la description du problème à

1. http ://datcha.lif.univ-mrs.fr

2. Un enjeu dialogique étant une raison pour l'usager de contacter le téléconseiller, par exemple la résolution d'un problème ou une demande d'information.

3. Pour être plus précis, l'annotation permet le multi-étiquetage des tours, mais pour tout ce qui est présenté ici, seule la fonction principe a été conservée.

| Étiquette | Signification | Description |
|---|---|---|
| OPE | Opening | Tours d'ouverture du dialogue |
| PRO | Problem Description | Description du problème du client |
| INQ | Information Question | Demande d'information de la part d'un des locuteurs |
| CLQ | Clarification Question | Demande de clarification |
| STA | Statement | Apport de nouvelles informations à l'interlocuteur |
| TMP | Temporisation | Mise en pause temporaire du dialogue |
| PPR | Plan Proposal | Proposition de résolution du problème. |
| ACK | Acknowledgment | Acquiescement des propos de l'interlocuteur |
| CLO | Closing | Tour de fermeture du dialogue |
| OTH | Other | Tour n'étant pas décrit par les autres étiquettes. |

Tableau 1: Jeu d'étiquettes utilisé pour l'annotation en actes de dialogue.

résoudre (tour 2). Après plusieurs questions destinées à préciser divers aspects du problèmes (tours 4 et 7), le téléconseiller donne un début de réponse informel. Le tour 10 est une relance du problème de la part du client, finalement résolu au tour 12. La fin du dialogue est une suite d'échanges assez protocolaires visant à évaluer la qualité du service produit puis à clore le dialogue. L'annotation en actes permet de décrire toutes ces informations.

| | | | |
|---|---|---|---|
| OPE | 1 | TC | Bonjour, je suis _TC1_, que puis-je pour vous ? |
| PRO | 2 | C | impossible pendant la lecture d'avancer la lecture |
| STA | 3 | C | _NUMTEL_ |
| CLQ | 4 | TC | Si je comprends bien, le problème concerne la vidéo à la demande ? |
| STA | 5 | C | mais aussi l'**enregistreur** et la tv à la demande |
| INQ | 6 | C | pouvez vous m'appeler sur le portable ? |
| INQ | 7 | TC | Est ce que vous avez un message d'erreur ? |
| STA | 8 | C | non |
| STA | 9 | TC | Si vous avez **débuté** le **visionnage**, mais que le téléchargement n'est pas terminé, l'avance et le retour **rapides** sont **indisponibles**. Vous pouvez uniquement **stopper** ou reprendre le **visionnage** au début de votre vidéo. |
| PRO | 10 | C | seulement l'**enregistreur** avait terminé l'enregistrement et au cours de la lecture je n'arrive pas à avancer |
| CLQ | 11 | TC | Donc le téléchargement a terminé mais vous n'y **arrivez** pas à l'avancer ? |
| STA | 12 | C | après l'avoir **débranché** puis **rebrancher** ça **refonctionne** merci |
| INQ | 13 | TC | Ca fonctionne maintenant ? |
| STA | 14 | C | oui |
| ACK | 15 | TC | Parfait. |
| INQ | 16 | TC | Puis-je faire autre chose pour vous ? |
| STA | 17 | C | non merci |
| CLO | 18 | TC | Je vous en **prie** Mr _CLIENT_. |
| CLO | 19 | TC | Orange vous remercie de votre confiance. Je vous souhaite une bonne journée. |

Exemple 1: Dialogue Annoté

# 2 Corpus DATCHA

Notre corpus de travail provient du projet ANR DATCHA. L'objectif du projet est de réaliser des tâches d'extraction d'information sur une grande base de données de conversations en ligne. Le corpus est constitué de conversations instantanées des différents services techniques et commerciaux d'Orange. Dans cette partie sont présentés d'une part les résultat statistiques de l'annotation manuelle partielle de ce corpus, et, d'autre part, la méthode que nous avons utilisée pour nous abstraire du corpus en retirant des informations relatives au lexique métier.

## 2.1 Annotation Manuelle du Corpus

Nous avons réalisé l'annotation manuelle de 2990 conversations dont la longueur moyenne est de 31,5 tours. Le tableau 2 montre la répartition des étiquettes dans le corpus. STA est l'étiquette la plus fréquente avec presque 40% des tours, tandis que TMP, CLQ et OTH représentent ensemble moins de 5% du corpus. Nous pouvons sommairement comparer la répartition de nos étiquettes avec d'autres travaux sur l'annotation des actes de dialogue, même s'il est difficile d'aligner parfaitement des jeux d'étiquettes différents.

Notamment, l'étiquette *Statement* existe et est utilisée à une échelle similaire dans Ivanovic (2005), Moldovan *et al.* (2011), avec respectivement 36%, 34.5% d'occurrences. Salim *et al.* (2016) a 57.1% de tours avec comme fonction Inform ou Answer.

En regroupant les différents types de questions (INQ et CLQ), notre corpus en contient 21% contre respectivement 19.2% pour Ivanovic, 10.6% pour Moldovan et 17.2% pour Salim pour les équivalents les plus directs dans leurs jeux d'annotation respectifs.

Les différences importantes s'expliquent par les différences de nature des corpus : les conversations de support clients étudiés par Ivanovic sont plus proches de notre corpus que les conversations spontanées de Moldovan ou les conversations de support entre clients (et non pas client/conseiller) pour Salim.

| étiquette | occurrences | pourcentage |
|---|---|---|
| STA | 36825 | 39.16% |
| INQ | 18064 | 19.21% |
| PPR | 14276 | 15.18% |
| ACK | 6225 | 6.62% |
| PRO | 5388 | 5.73% |

| étiquette | occurrences | pourcentage |
|---|---|---|
| CLO | 5156 | 5.48% |
| OPE | 3831 | 4.07% |
| TMP | 2286 | 2.43% |
| CLQ | 1638 | 1.74% |
| OTH | 353 | 0.38% |

Tableau 2: Statistiques par Étiquette

3027 tours de parole ont été annotés par deux annotateurs afin d'évaluer la précision du guide d'annotation et évaluer la tâche d'annotation. Avec un Kappa de Cohen de $\kappa = 0.67$, l'annotation manuelle peut paraître fortement dépendante de l'annotateur. Cependant, une analyse plus fine des divergences a montré qu'environ la moitié des divergences entre annotateurs proviennent d'omissions du guide, imprécis dans certaines situations fréquentes dans notre corpus. Toutefois de réelles ambiguïtés existent, notamment entre les deux types de questions (CLQ pour Clarification et INQ pour Information), ou encore pour distinguer Statement d'autres étiquettes, notamment PRO, PPR et ACK.

## 2.2 Abstraction du Lexique Spécifique

Notre corpus de travail n'est pas publiquement disponible, ce qui est souvent le cas pour ce type de données. Toutefois, l'annotation en actes de dialogue que nous proposons, ainsi que les modèles de prédiction appris sur nos données pourraient être utilisés sur une catégorie de corpus qui est bien plus large que les thématiques et spécificités lexicales propres à notre corpus. Pour essayer de s'abstraire de ces spécificités, nous avons tenté d'éliminer automatiquement du corpus la partie du lexique qui est trop liée à Orange ou aux tâches spécifiques à la résolution des pannes par son service technique. L'objectif est de produire des outils d'annotation automatiques utilisables dans pour des corpus différents.

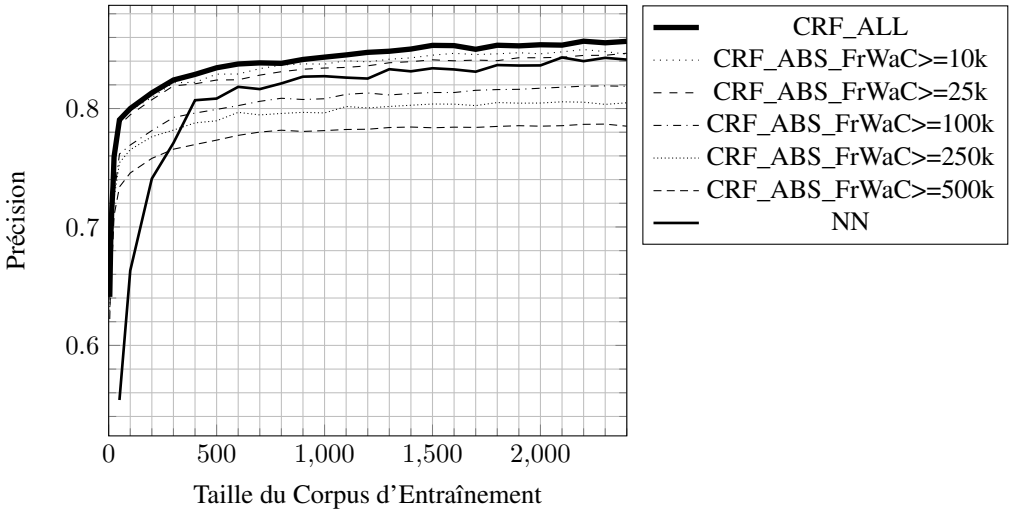
Pour construire ce lexique, nous avons extraits de frWaC (Baroni *et al.*, 2009), un grand corpus regroupant des productions de diverses sources sur le web, des lexiques de formes apparaissant plus d'un nombre fixé de fois. Par exemple, frWaC>=25k est le lexique des 5097 formes apparaissant plus de 25 000 fois dans frWaC. frWac>=500k est un lexique très restreint ne contenant que 284 formes différentes. À partir de chacun de ces lexiques, il est possible de filtrer les phrases de notre corpus de travail en considérant toutes les formes absentes du lexique comme inconnues. Dans l'exemple 1, les formes en gras sont celles absentes du lexique frWaC>=25k. Ce dernier conserve 45,5% des formes du lexique de notre corpus, soit 81,5% des occurrences produites par les locuteurs.

Parmi les 20 formes les plus fréquentes dans le corpus DATCHA qui ne sont pas dans frWaC>=25k et sont donc abstraites du corpus se trouvent 'livebox' 'décodeur' 'echat' 'sosh' 'box' 'wifi' 'câble' 'télécommande' 'chaînes' et 'redémarrer'. Ces mots sont de bons candidats pour l'abstraction. En revanche, des mots comme 'prie' 'confirmer' 'patienter' 'plait' ou 'svp' que l'on aurait souhaité conserver sont abstraits du corpus, n'apparaissant pas suffisamment dans frWaC. Il est difficile de mesurer à quel point cette méthode permettrait d'annoter des corpus proches dans la forme mais distincts dans les thématiques. Toutefois, nous présentons dans la section suivante des mesures expérimentales de l'impact de cet appauvrissement sur les performance de la prédiction automatique.

# 3 Prédiction des actes de dialogue

À partir de notre corpus annoté manuellement (respectivement 2390, 300 et 300 dialogues pour Train/Dev/Test), nous avons entraîné des modèles pour étiqueter automatiquement en actes de dialogue. Les résultats présentés dans cette section sont issus de trois modèles. [CRF_ALL], un modèle basé sur des Conditional Randm Fields (CRF) (Lafferty *et al.*, 2001) utilisant la totalité du lexique d'entraînement, [CRF_ABS_FrWaC>=N], une classe de modèles également basé sur des CRF mais dont le lexique est restreint par FrWaC, et [NN], un modèle basé sur des réseaux de neurones.

Les features utilisées pour les modèles CRF pour décrire un tour de parole sont : la longueur du tour, la position relative du tour dans le dialogue, quel locuteur a produit le tour et quels sont les mots apparaissant dans le tour (après restriction pour les CRF_ABS_FrWaC>=N). La longueur du tour est un ensemble de features binaires d'appartenance à une catégorie parmi dix, lesquelles ont été prédéterminées pour être équilibrées suivant le corpus Dev (ici les tours de 1 mots sont une catégorie, les tours de 2 mots une autre et les tours de 3 ou 4 mots sont regroupés ensembles, etc...). La position relative du tour suit la même idée de catégorisation mais pour des catégorie du type "le tour est dans le i-éme décile du dialogue".




Courbe 1: Courbe d'apprentissage des modèles suivant la quantité de données d'entraînement

Complexifier davantage le modèle n'a pas produit de meilleurs résultats dans nos tests. Cela est probablement dû à la nature de l'annotation, qui ne nécessite pas une analyse fine de la sémantique ou de la syntaxe des tours de parole. L'implémentation a été effectué par l'utilisation de CRFSuite (Okazaki, 2007). Pour catégoriser un tour par sa position relative dans le dialogue, l'algorithme sépare chaque dialogue en dix sections de taille égales et chaque tour est décrit par une unique feature précisant son segment. Cela a pour objectif de permettre à l'algorithme de pouvoir s'adapter à des conversations de taille très variable (d'une dizaine à quelques centaines de tours), tout en détaillant une information utile à l'analyseur (notamment, les tours de clôture sont souvent concentrés dans le dernier segment, peu importe la taille du dialogue).

Inspiré par Yang *et al.* (2016), NN est un réseau de neurones récurrent hiérarchique à deux niveaux. Le premier niveau s'intéresse aux tours de parole individuellement à partir de la séquence de mots de chaque tour, alors que le second niveau s'intéresse à la conversation dans son ensemble à partir des états cachés représentant les tours de paroles provenant du premier niveau. Les deux niveaux sont des réseaux récurrents bidirectionnels de type Long-Short Term Memory (LSTM). Pour un tour de parole $i$, la couche de décision utilise l'état caché $i$ du dernier LSTM pour obtenir une prédiction de l'acte de dialogue du tour de parole. En entrée, le système utilise les mots, qui sont transformés par des embeddings de mots au premier niveau du LSTM, ainsi qu'une identification du locuteur.

La courbe d'apprentissage 1 présente l'évolution de la précision des classifieurs en fonction de la quantité de données d'apprentissage. Comme souvent, l'algorithme utilisant des réseaux de neurone a besoin d'une quantité plus importante de données pour arriver à des résultats comparables (ici au moins 500 conversations). Notre procédure d'abstraction du lexique affecte légèrement (de l'ordre de 1%) la précision du modèle, mais reste plus performante en tout point que NN pour FrWaC>=10k et FrWaC>=25k. Avec l'apprentissage sur tout le corpus d'entraînement, les résultats sont de 0.857, 0.847 et 0.841 pour respectivement CRF_ALL, CRF_ABS_FrWaC>=10k, et NN. CRF_ABS_FrWaC>=25k apparait comme un bon compromis entre taille du lexique et performances.

Une analyse des erreurs des meilleurs classifieurs nous a montré que l'étiquette la plus problématique,

est, sans surprise, l'étiquette STA. Sur notre corpus de test, 75% des erreurs d'annotation de nos meilleurs classifieurs sont en effet soit des annotations erronées de tours en STA ($\sim$45% des erreurs), soit annoté avec une autre étiquette alors que STA correspond à notre annotation manuelle ($\sim$30% des erreurs). Parmi ces erreurs, STA/PPR, STA/PRO STA/INQ sont les distinctions les plus difficiles pour nos modèles.

La ponctuation est souvent en cause pour les erreurs de type STA/INQ, en particulier lorsque le point d'interrogation est omis dans une question. Des tours comme *et que doit il se passer* ou encore *ok et le montant* sont incorrectement classifiés en STA, dans leurs contextes respectifs.

Pour STA/PRO, il est plus difficile de décerner une raison commune aux erreurs des classifieurs. Toutefois, il y a une fréquence plus élevé d'erreurs de classification au milieu des dialogues. Des tours comme *ça ne fait rien d autre que afficher perte de synchronisation* qui auraient été des PRO en début de dialogue sont au milieu de dialogue des STA, et les classifieurs étiquettent ceux-ci incorrectement.

# Conclusion

Dans cet article, nous avons présenté et détaillé un jeu d'étiquettes pour décrire les actes de dialogues de conversations en ligne. Au-delà des résultats issus d'une annotation manuelle et de la classification automatique supervisée en utilisant soit des CRF soit des réseaux de neurones, nous avons proposé une méthode simple permettant de ne pas prendre en compte une partie du lexique spécialisé du corpus. Une poursuite naturelle de ces travaux est d'utiliser ces actes de dialogue dans une analyse de plus haut niveau, comme la catégorisation des conversations. Une autre piste de travail est l'élaboration d'une structure dialogique détaillant davantage l'interaction entre locuteurs, en s'appuyant sur les actes comme base.

# Remerciements

Ces travaux ont été en partie réalisés grâce au soutien financier apporté par l'Agence Nationale pour la Recherche par le biais du projet DATCHA (ANR-15-CE23-0003).

# Références

ASHER N., NASR A. & PERROTIN R. (2017). *Manuel d'annotation en actes de dialogue pour le corpus Datcha*. Rapport interne. `http://datcha.lif.univ-mrs.fr/files/dialogue-acts-manual-french.pdf` .

AUSTIN J. L. (1975). *How to do things with words*. Oxford university press.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, **43**(3), 209–226.

CORE M. G. & ALLEN J. (1997). Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56 : Boston, MA.

DAMNATI G., GUERRAZ A. & CHARLET D. (2016). Web chat conversations from contact centers : a descriptive study. In *LREC*.

HARDY H., BAKER K., BONNEAU-MAYNARD H., DEVILLERS L., ROSSET S. & STRZALKOWSKI T. (2003). Semantic and dialogic annotation for automated multilingual customer service. In *Eighth European Conference on Speech Communication and Technology*.

IVANOVIC E. (2005). Dialogue act tagging for instant messaging chat sessions. In *Proceedings of the ACL Student Research Workshop*, p. 79–84 : Association for Computational Linguistics.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data.

MOLDOVAN C., RUS V. & GRAESSER A. C. (2011). Automated speech act classification for online chat. *MAICS*, **710**, 23–29.

OKAZAKI N. (2007). Crfsuite : a fast implementation of conditional random fields.

SALIM S., HERNANDEZ N. & MORIN E. (2016). 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*.

SHRIBERG E., DHILLON R., BHAGAT S., ANG J. & CARVEY H. (2004). *The ICSI meeting recorder dialog act (MRDA) corpus*. Rapport interne, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

YANG Z., YANG D., DYER C., HE X., SMOLA A. & HOVY E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1480–1489.