# Prédiction de l'échec d'une conversation médiée dans un contexte de dialogues à rôles asymétriques

Romain Carbou, Delphine Charlet, Géraldine Damnati, Frédéric Landragin, Jean-Léon Bouraoui

# Prédiction de l'échec d'une conversation médiée dans un contexte de dialogues à rôles asymétriques

R. Carbou[1a, 2] D. Charlet[1b] G. Damnati[1b] F. Landragin[2] J.-L. Bouraoui[1b]
(1) Orange Labs, Châtillon (a), Lannion (b), France
(2) Lattice, CNRS, ENS Paris, Université Sorbonne Nouvelle, PSL Research University, USPC, 1 rue Maurice Arnoux, 92120 Montrouge
romain.carbou@orange.com, delphine.charlet@orange.com, geraldine.damnati@orange.com, frederic.landragin@ens.fr, jeanleon.bouraoui@orange.com

## RESUME

Dans une conversation humain-humain entre un usager et un interlocuteur en centre d'assistance, on se place dans le contexte où l'issue du dialogue est caractérisée par une notion de succès ou d'échec, explicitement annotée ou extrapolée. L'étude envisage différents paramètres susceptibles d'exercer une influence sur un modèle de classification prédictive des échecs constatés. On cherchera d'une part à exploiter une modélisation de la distribution lexicale tirant parti de l'asymétrie des rôles des locuteurs. On examinera d'autre part si la partie du lexique plus étroitement liée au domaine d'assistance client abordé ici, modifie la qualité de la prédiction. On interrogera enfin les perspectives de généralisation du modèle à des corpus morphologiquement comparables.

## ABSTRACT

In a human-to-human conversation between a user and his interlocutor in an assistance center, we suppose a context where the conclusion of the dialog can characterize a notion of success or failure, explicitly annotated or deduced. The study involves different approaches expected to have an influence on predictive classification model of failures. On the one hand, we will aim at taking into account the asymmetry of the speakers' roles in the modelling of the lexical distribution. On the other hand, we will determine whether the part of the lexicon most closely relating to the domain of customer assistance studied here, modifies the quality of the prediction. We will eventually assess the perspectives of generalization to morphologically comparable corpora.

MOTS-CLES : dialogue humain-humain, échec d'un dialogue, corpus de dialogues, apprentissage artificiel, évaluation de dialogues, dialogue asymétrique.

KEYWORDS: human-to-human dialog, dialog failure, dialog corpus, artificial learning, dialogue evaluation, asymmetric dialog.

# 1 Introduction

Dans un corpus de dialogues humain-humain à rôles *asymétriques* (e.g. un utilisateur et un agent au sein d'un service client), on cherche à prédire les issues d'« échec » au regard d'une certaine définition. Pour cela on utilisera, de façon nominale, une méthode de classification supervisée utilisant les annotations de sortie disponibles, et dont on répétera les phases d'entraînement et de test en faisant varier différents paramètres. Chaque donnée d'observation est un dialogue, représenté

sous la forme d'un sac de mots dont on compte le nombre d'occurrences au sein de ce dialogue. L'objectif poursuivi est de pouvoir prédire l'échec de futurs dialogues se déroulant dans le même environnement mais aussi d'examiner si des phénomènes stables permettent d'utiliser le modèle sur des corpus de dialogues différents, selon une approche guidée par corpus [Tognini-Bonelli, 2001]. Le corpus étudié, *Datcha* [Damnati & al, 2016], rassemble des dialogues écrits (« tchats ») entre clients et téléconseillers du service technique de la société Orange.

On exploitera en particulier l'asymétrie des rôles du client (C) et du téléconseiller (T) sous l'angle lexical. Dans un cas, on considèrera dans sa globalité la distribution des mots employés, quel que soit le locuteur. Dans l'autre cas, les distributions propres à C ou à T seront considérées séparément. On parlera alors de données d'observations « différenciées » (par locuteur).

En second lieu, la taille du lexique obtenu directement par segmentation des mots est supérieure à 11000 flexions. Il s'agira d'en choisir des critères de réduction pertinents – pour, d'une part, garder un lexique significatif et, d'autre part, assurer un bon comportement de l'algorithme d'apprentissage du modèle. Or, l'univers dialogique constitue ici un « domaine fermé » (l'assistance client d'un opérateur), en opposition au « domaine ouvert » (lequel met en jeu des dialogues portant sur des sujets d'une diversité arbitraire). On examinera l'influence relative de la part du lexique la plus caractéristique du domaine (ci-après désignée comme part « thématique »), en retranchant optionnellement du lexique un sous-ensemble de 815 mots usuels (articles, mots de liaison syntaxique, et mots appartenant au vocabulaire général). Les données d'observation ainsi obtenues seront ci-après qualifiées de « filtrées ».

La segmentation du texte en mots est réalisée par un outil de lemmatisation (voir § Méthode). Celui-ci permettra, en troisième lieu, de comparer la forme originale et la forme lemmatisée de la distribution lexicale. Cette dernière variante a un fondement purement syntaxique, agnostique vis-à-vis de l'asymétrie exploitée par les deux précédentes. On notera que le lemmatiseur reconnaît directement certaines expressions locutionnelles, comme par exemple « prendre en charge » (comptant alors pour une entrée dans le lexique). Enfin, on ne corrigera pas les déviations linguistiques propres à Datcha, telles que discutées dans [Damnati & al, 2016].

L'action individuelle de ces trois variantes de construction de la distribution lexicale sera comparée à la distribution nominale *non différenciée par locuteur*, *non filtrée*, et *non lemmatisée*. Un enjeu de généralisation sera de rechercher les conditions où le contraste sur les résultats est le plus marqué, à savoir dans une fenêtre où l'ensemble du lexique n'est construit ni trop finement (« surabondance » lexicale) ni trop grossièrement (« pénurie » lexicale).

La section 2 présente le corpus Datcha, ses caractéristiques globales et ses annotations. Les sections 3 et 4 décrivent respectivement la méthodologie d'apprentissage du modèle prédictif et les résultats obtenus dans chacune des variantes. La section 5 commente et interprète ces résultats. Enfin, la section 6 propose de futures extensions à l'étude menée.

# 2 Corpus

La problématique d'annotation d'un corpus de dialogue à des fins d'évaluation s'étend sur une longue période, de MADCOW [Hirschman, 1992] à ADEM [Lowe & al, 2017]. Ici, le corpus est constitué de 2775 dialogues de type tchat, issus d'un centre de contact en Assistance Technique.

Les dialogues, lors d'études précédentes sur ce même corpus, ont été annotés manuellement par un expert. Celui-ci a annoté chaque dialogue selon le type de résolution du problème client tel qu'il est observé à l'issue de la conversation :

- *échange* : T propose un remplacement matériel ;
- *résolu* : T a fourni l'information attendue ou corrigé le problème ;
- *à tester* : T fournit un scénario de résolution en cours de session, avec actions ultérieures ;
- *hors périmètre* : le problème de C n'est pas du ressort de T, dans ce cas une réorientation est le plus souvent proposée, mais il peut ne pas y avoir réorientation selon circonstances ;
- *pas de solution* : la conversation n'a abouti à aucune des résolutions précédentes.

Le succès ici n'est pas lié à l'obtention au cours du dialogue d'un ensemble d'informations caractéristiques (« slot-filling ») comme abordé par [Claveau & al, 2013] ou [Talha & al, 2014]. Le téléconseiller a pour tâche d'établir un diagnostic qui par nature induit une infinité de possibles. Aussi, on considère que toute annotation exprimant un progrès dans la résolution de la demande initiale est un succès. Ainsi, un dialogue sera considéré comme échec s'il est annoté en « *pas de solution* » ou « *hors-périmètre sans réorientation* ». Ces situations sont marginales (une dizaine d'occurrences) et la majorité des conversations considérées en échec sont celles qui n'aboutissent pas. Les raisons sont diverses : le dialogue s'est interrompu inopinément (78% des cas) ; le client raccroche de son plein gré (12% des cas) ou devient définitivement inactif (8% des cas). La tâche étudiée dans cet article est donc la prédiction de l'échec, avec une répartition des conversations en 76,6% de succès et 23,4% d'échecs.

Dans ce contexte où la variable à prédire n'a pas été directement annotée, on voit que la tâche définitoire trahit des difficultés intrinsèques propres à la détermination de la « coopération » entre les acteurs [Grice, 1975].

Remarque : C'est l'échec qui est ici pris comme modalité positive. Or, les phénomènes qui orientent vers le succès sont aussi légitimes à étudier. En fait, ceux-ci semblent se concentrer sur des aspects cérémoniels (marque de politesse, satisfaction, etc.). Par exemple le test du seul mot « merci » produit pour la modalité négative, *c'est-à-dire le succès*, un triplet (rappel ; précision ; F-mesure) de (82,41% ; 94,88% ; 88,20%). Donc dès lors que le succès devient l'objet de la prédiction, *un seul mot* fait mieux que le modèle trivial attribuant à toutes les observations du corpus la classe majoritaire (100% ; 76,60% ; 86,75%). Si a fortiori l'on élargit le lexique (e.g. les lexies présentes sur 40% de dialogues), pas un modèle obtenu n'aboutit à une F-mesure inférieure à 90% (toujours pour le succès).

Il semble dès lors inutile de convoquer des métriques autres que les (rappel ; précision ; F-mesure) de la modalité d'échec, notés à présent (r ; p ; F). Il en irait autrement *si la répartition de sortie était davantage balancée*. Exactitude (« accuracy ») et Kappa sont rappelés à titre indicatif, en tant que valeurs symétrisées usuelles de la performance en classification [Japcowicz, 2014].

# 3 Méthode

L'étude envisage la prédiction de l'échec sous un angle lexical, où la démarche est d'« aplatir » chaque dialogue en un vecteur unique d'une dimension égale à la taille du lexique. Celui-ci est construit selon le mode opératoire qui suit, appelant une précision définitoire.

On appellera TAUX D'OCCURRENCE des entrées lexicales retenues, le pourcentage minimal de dialogues *distincts* où ces entrées doivent être présentes, qu'elles y apparaissent une ou plusieurs fois. Par exemple, l'adjectif « autre » fait partie de la distribution lexicale de Datcha pour un taux d'occurrence de 40%, mais pas de 80%. Il s'agit donc d'un paramètre servant à ajuster le dimensionnement de la distribution lexicale nominale. Les différentes valeurs de taux d'occurrence font l'objet d'un choix d'échelle. Pour opérer un compromis entre combinatoire et observation significative, une échelle de 5% à 75% par incrément de 5% est retenue. La valeur minimale (5%) correspond au lexique maximal, soit en l'occurrence 1141 entrées lexicales non lemmatisées.

Pour chaque entrée, la valeur de la cellule est le nombre d'occurrences dans le dialogue observé. Comme les mots-coordonnées entretiennent des relations de cooccurrence dans le champ linguistique, on choisit de ne pas effectuer de normalisation sur ces valeurs.

Pour un taux d'occurrence fixé, 3 variantes de la distribution nominale sont construites comme décrit ci-après.

- DIFFERENTIATION DES LOCUTEURS : On distingue les entrées lexicales utilisées par T et par C en les préfixant. Par exemple, « bonjour », qui est employé par les deux locuteurs, existera sous les formes « T_bonjour » et « C_bonjour » dans le lexique « différencié ». La distinction des distributions par locuteur cherche à déterminer si une même entrée, observée chez C ou T, traduit la même inclinaison de la conversation vers le succès ou l'échec.

- FILTRAGE DES « MOTS CREUX » : Un filtrage lexical selon une liste de 815 formes fléchies de « mots creux » (« stop words ») est appliqué à la distribution lexicale considérée.

- LEMMATISATION : L'entrée lexicale est le lemme de celle trouvée en corpus. L'outil utilisé est la plateforme « TiLT » (Orange Labs). Par exemple, la flexion « voudrais » trouvée en corpus est représentée par l'entrée lexicale « vouloir ».

On précise que, même en conservant les flexions d'origine, l'outil de segmentation et lemmatisation a une empreinte sur la distribution lexicale, e.g. pour les locutions et les transformations sur les contractions (« au » → « à » + « le »). Afin d'avoir une base homogène de comparaison, *ces transformations sont appliquées aux deux cas*. Ces effets, quoique marginaux, renvoient au débat de fond de la lemmatisation [Brunet, 2000]. Des cas de variantes flexionnelles [Valette & al, 2014] pouvant s'avérer en contexte sémantiquement opposées, étayent une invitation à la vigilance quant aux effets du remplacement par le lexème.

80% des données sont dévolues à l'entraînement avec une validation croisée sur 5 sous-groupes et 20% au test, par tirage au sort. Les modalités étant déséquilibrées mais sans excès (649 observations d'échecs) et les essais de sur-échantillonnage sans effet autre que marginal, les observations ne sont pas corrigées.

L'étude est menée selon le mode opératoire séquentiellement disposé comme suit. On y désigne par « campagne » une phase d'entraînement d'une série de modèles de classification employant différentes valeurs de paramétrage.

- NETTOYAGE DU CORPUS : Pour identifier des phénomènes liés aux entités nommées (EN), le nettoyage de Datcha effectué dans les études antérieures a été poursuivi sur les erreurs résiduelles (nom, téléphone) et sur tous les horodatages. L'exercice a dû définir des limites,

puisque peut être une EN toute « expression linguistique autonome » [Nouvel & al, 2015], pouvant « par ses seules ressources, évoquer un référent ». Ici, les EN non traitées les plus représentatives ont été les adresses postales. Leur diversité toponymique et syntaxique contribue sensiblement à la taille de la distribution lexicale. Mais leur détection et leur remplacement par un terme générique sortait des bornes de l'étude (voir § Perspectives).

- CAMPAGNE GROSSIERE : Cette série de phases d'apprentissage avait pour objet de sélectionner un algorithme de classification au comportement suffisamment et rapidement prometteur : un classifieur bayésien naïf « témoin » et des modèles usuels comme SVM, utilisés dans un contexte analogue mais multimodal par [Salim & al, 2016]. Parmi les algorithmes s'illustrant bien en classification binaire [Torlay, 2017] et [Alpaydin, 2010], le modèle XgBoost a été finalement retenu parmi 5 alternatives.

- CAMPAGNE FINE : Cette série de phases d'apprentissage a exploré de façon systématique le comportement du modèle XgBoost retenu, à chaque palier de taux d'occurrence entre 75% et 5%, pour la distribution lexicale nominale et pour les trois variantes *différenciées* par locuteur ; *filtrée* sur mots creux ; et *lemmatisée*. Soit 15 paliers de taux d'occurrence × (1 + 3) distributions = 60 entraînements-tests.

# 4 Résultats

Pour réaliser l'ajustement de la fourchette pertinente dans laquelle contraindre la largeur du lexique, on a d'abord observé la variabilité des résultats endogènes à chaque palier du taux d'occurrence.

Ainsi, dans le TABLEAU 1, chaque ligne établit les valeurs moyennes des métriques relatives aux 4 distributions lexicales construites (nominale, différenciée, filtrée, lemmatisée). La colonne 2 indique l'intervalle dans lequel varient leurs tailles respectives. Les 3 colonnes « c.v.(*) » indiquent le *coefficient de variation* (rapport de l'écart-type sur la moyenne) pour le rappel, la précision et la F-mesure, au sein de chaque groupe de 4 distributions, pour un taux d'occurrence donné.

| taux occurrence (%) | amplitude Ω(lexique) | rappel moyen | **c.v. (*) rappel** | précision moyenne | **c.v. (*) précision** | F-mesure moyenne | **c.v. (*) F-mesure** |
|---|---|---|---|---|---|---|---|
| 75 | 4 – 64 | 69,19 | **17,31** | 55,58 | **18,93** | 61,62 | **18,15** |
| 70 | 6 – 74 | 72,78 | **14,59** | 64,81 | **13,99** | 68,53 | **14,02** |
| 65 | 7 – 84 | 72,42 | **18,34** | 65,54 | **17,83** | 68,70 | **17,55** |
| 60 | 12 – 104 | 75,08 | **8,74** | 72,58 | **8,65** | 73,81 | **8,65** |
| 55 | 15 – 118 | 77,44 | **9,54** | 71,88 | **13,39** | 74,53 | **11,52** |
| 50 | 18 – 128 | 76,73 | **7,23** | 72,67 | **10,44** | 74,62 | **8,83** |
| 45 | 21 – 152 | 77,93 | **6,52** | 75,19 | **8,43** | 76,53 | **7,50** |
| 40 | 26 – 182 | 80,03 | **5,33** | 77,75 | **7,33** | 78,86 | **6,35** |
| 35 | 31 – 202 | 80,97 | **2,70** | 78,24 | **7,14** | 79,55 | **4,99** |
| 30 | 37 – 230 | 83,30 | **2,06** | 80,28 | **7,38** | 81,68 | **4,09** |
| 25 | 52 – 274 | 82,86 | **1,80** | 80,43 | **3,25** | 81,60 | **1,91** |
| 20 | 82 – 358 | 83,71 | **1,74** | 82,03 | **2,29** | 82,86 | **1,73** |
| 15 | 117 – 465 | 82,36 | **2,83** | 83,48 | **0,95** | 82,90 | **1,04** |
| 10 | 184 – 645 | 83,25 | **2,17** | 84,75 | **2,15** | 83,99 | **1,95** |
| 5 | 390 – 1141 | 82,60 | **3,50** | 83,79 | **3,71** | 83,18 | **3,52** |

**(*)** coefficient de variation

TABLEAU 1 : Variabilités (rappel ; précision ; F-mesure) par palier de taux d'occurrence

L'intervalle [30% ; 50%] apparaît comme zone centrale où la variabilité de la F-mesure est décelable mais comprise entre 4% et 9%, traduisant des résultats bons à médiocres selon le cas (différentiation par locuteur ; filtrage de mots creux ; lemmatisation).

Au-delà, l'amplitude de la variabilité des paliers a tendance à augmenter et le modèle peut rester acceptable comme devenir pire qu'un modèle aléatoire (« zone carencée »). En-deçà [5% ; 25%], on reconnaît un symptôme « d'école » en ce que le modèle n'apprend plus que marginalement de la prise en compte d'un lexique plus large (les mots trop peu fréquents n'améliorent pas le modèle). Mais surtout, les 4 distributions tendent à voir leurs effets respectifs disparaître (« zone saturée »).

Le TABLEAU 2 indique que toutes valeurs paramétriques et tous paliers confondus, le meilleur modèle est trouvé pour un taux d'occurrence de 10% dans le mode de génération du lexique sous forme *lemmatisée*, *indifférenciée par locuteur* et *sans filtrage des mots creux*. Les (r ; p ; F) en sont (85,82% ; 87,05% ; 86,43%). Ancré dans la zone saturée, le modèle ne se situe pas dans un intervalle de confiance permettant de conclure *dans cette zone*, à une influence du paramétrage. L'exactitude (« accuracy ») et le Kappa sont respectivement supérieurs à 93% et 80%, sachant qu'un Kappa supérieur à 70% est usuellement considéré excellent :

| Modèle | Ω(lexique) | exactitude | Kappa | rappel | précision | F-mesure |
|---|---|---|---|---|---|---|
| Nominal | 324 | 91,53 | 77,28 | 83,21 | 82,61 | 82,91 |
| Filtré | 184 | 91,53 | 77,71 | 81,94 | 84,89 | 83,39 |
| Différencié | 645 | 91,71 | 77,71 | 82,01 | 84,44 | 83,21 |
| Lemmatisé | 295 | 93,15 | 81,85 | 85,82 | 87,05 | 86,43 |

TABLEAU 2 : Détail du palier à 10%, hébergeant le meilleur modèle parmi 60

Hors zone saturée, observons dans le TABLEAU 3 la différence de F-mesure que les modèles différencié, filtré sur mots creux, et lemmatisé, entretiennent respectivement au modèle nominal.

| zone | zone de pénurie | | | | | zone centrale | | | | | zone saturée | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| taux occur. | 75 | 70 | 65 | 60 | 55 | 50 | 45 | 40 | 35 | 30 | 25 | 20 | 15 | 10 | 5 |
| diff vs nom. | 6,1 | 6,0 | 10,4 | 3,3 | 6,5 | 1,8 | 3,2 | 2,2 | 2,1 | 0,9 | 3,7 | -0,8 | -0,7 | 0,3 | 2,0 |
| filt. vs nom. | -17,9 | -16,2 | -16,7 | -10,6 | -13,9 | -12,9 | -9,8 | -9,0 | -7,0 | -6,5 | 1,5 | -3,3 | -2,0 | 0,5 | -3,4 |
| lem. vs nom. | 5,3 | 0,8 | 6,9 | 2,2 | -0,3 | -2,6 | 0,9 | -4,7 | -0,4 | -0,5 | 2,4 | -1,8 | -0,5 | 3,5 | 3,3 |

TABLEAU 3 : Différence de F-mesure des 3 variantes de distribution, à la distribution nominale

Le filtrage se traduisant par une réduction de la dimension, il appelle une comparaison spécifique (TABLEAU 4), mesure par mesure, avec le modèle nominal de dimension la plus proche, par équité informationnelle (on prend la plus proche inférieure, qui accentue les conclusions). Les mesures sont comparées en demeurant dans la zone centrale du modèle nominal.

| modèle filtré par mots creux | | | | | modèle nominal de plus proche dimension inférieure | | | | |
|---|---|---|---|---|---|---|---|---|---|
| taux | Ω(lexique) | rappel | précision | F-mesure | taux | Ω(lexique) | rappel | précision | F-mesure |
| 5 | 390 | 78,72 | 79,86 | 79,29 | 10 | 324 | 83,21 | 82,61 | 82,91 |
| 10 | 184 | 81,94 | 84,89 | 83,39 | 20 | 179 | 84,67 | 84,06 | 84,36 |
| 15 | 117 | 79,45 | 84,06 | 81,69 | 30 | 115 | 83,82 | 82,61 | 83,21 |

TABLEAU 4 : Comparaison en F-mesure entre modèle filtré et nominal de dimension inférieure

On discute à présent le comportement des paramètres dans les trois zones instable, centrale et saturée – qui constitue l'éclairage privilégié par la présente étude.

# 5 Discussion

La différentiation de la distribution améliore les résultats du modèle, en zones d'expression. C'est en zone carencée que son effet est le plus prononcé. La F-mesure y est en moyenne de 6,45% supérieure au modèle nominal (sa meilleure observation y est de 82,96% au taux de 55%). Son effet reste positif en zone centrale mais chute à 2% en moyenne. Il devient indiscernable en zone saturée où l'information purement lexicale se suffit à elle-même. La dynamique à l'œuvre semble indiquer que plus le lexique est large, *moins la connaissance du locuteur apporte à la prédiction*.

L'effet du filtrage des mots creux engage plusieurs lectures. Hors zone saturée, il est en soi naturel que le filtrage de *quelque mot que ce soit* dégrade les performances prédictives (destruction informationnelle). Mais là, le phénomène survient dès la zone centrale (9% de dégradation de F-mesure en moyenne) pour s'accentuer en zone carencée (-15%). La zone saturée présente le même symptôme d'annulation d'effet que précédemment.

En second lieu, il s'agit de comparer la performance du lexique filtré avec un lexique nominal de taille voisine. C'est ce dernier qui l'emporte dans sa propre zone centrale, même en prenant, par convention, la taille inférieure la plus proche. À taille comparable, la part « thématique » du lexique est ainsi *moins indicatrice de l'échec* que les mots réputés creux, ce qui peut constituer le socle légitimant une étude des conditions de généralisation du modèle à d'autres corpus.

La lemmatisation, enfin, a un caractère oscillatoire de moyenne amplitude qui rend a priori caduque toute perspective d'interprétation. Dans la présente structure d'apprentissage de type sacs de mots, elle peut à des paliers adjacents (cf. 75%, 70%, 65%) améliorer sensiblement la performance comme apparaître transparente. En zone centrale, elle peut *dégrader* la performance, suggérant des ambiguïtés de même nature que [Valette & al, 2014] (par exemple, il est plausible qu'un diagnostic de panne occasionnant « des redémarrages » ne soit pas corrélé au succès comme le serait celui occasionnant « un redémarrage » : la confusion flexionnelle serait sur cet exemple génératrice de faux négatifs, si les redémarrages multiples ont une corrélation positive à l'échec).

# 6 Perspectives

Une combinatoire complète des cas étudiés aurait à court terme vocation à examiner les effets *cumulés* de plusieurs des critères d'apprentissage (e.g. filtrage + différenciation par locuteur). La séquentialité de la dynamique dialogique inspire en fait la mise en œuvre d'approches dédiées. La plus immédiate consisterait à observer la différence de distribution lexicale des locuteurs, de part et d'autre d'un *point d'avancement* du dialogue défini par un *taux arrondi de nombre de tours de parole sur le nombre total de tours de parole du dialogue considéré*. Il constituerait un nouveau paramètre variable du processus d'apprentissage. Une caractérisation fine des comportements séquentiels passe par un changement de la technique d'apprentissage, e.g. au profit de réseaux de neurones récurrents adaptés au texte (LSTM). L'observation unitaire y devient l'entrée lexicale, ou une transformée de type word2vec ou GloVe [Pennington & al, 2014]. Elle est affublée de traits B, I, O délimitant les dialogues et tours de parole. Enfin, dans chaque scénario, une révision de l'état de l'art en reconnaissance d'entités nommées permettrait d'envisager une préparation plus poussée du corpus, réduisant l'inflation lexicale.

**Remerciements** : Ce travail a été financé partiellement par l'Agence Nationale de la Recherche : ANR-15-CE23-0003 (DATCHA).

# Références


ALPAYDIN E. (2010). *Introduction to Machine Learning*. Cambridge, MA: The MIT Press. 431.

BRUNET É. (2000). Qui lemmatise dilemme attise. *Lexicometrica, n°2*.

CLAVEAU V., NCIBI A. (2013). Découverte de connaissances dans les séquences par CRF non-supervisées. *TALN 2013.*

DAMNATI G., GUERRAZ A., CHARLET D. (2016). Web Chat Conversations from Contact Centers: a Descriptive Study, *LREC 2016*.

GRICE H. PAUL (1975). Logic and Conversation. *Syntax and Semantics, Vol. 3. Speech Acts*.

HIRSCHMAN L. (1992). MADCOW: Multi-Site Data Collection for a Spoken Language Corpus.

JAPKOWICZ N., SHAH M. (2014). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, MA: The MIT Press. 93.

LOWE R., NOSEWORTHY M., SERBAN I., ANGELARD-GONTIER N., BENGIO Y., PINEAU J. (2017). Towards an Automatic Turing Test: Learning to evaluate dialogue responses. *ICLR 2017*.

NOUVEL D., EHRMANN M., ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. Londres : ISTE Editions.

PENNINGTON J., SOCHER R., MANNING C. D. (2014). GloVe: Global Vectors for Word Representation. *Computer Science Department, Stanford University*.

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. *TALN 2016*.

TALHA M., BOULAKNADEL S., ABOUTAJDINE D. (2014). RENAM: Système de Reconnaissance des Entités Nommées Amazighes. *TALN 2014.*

TOGNINI-BONELLI, E. (2001). *Corpus linguistics at work*. Amsterdam : John Benjamins.

TORLAY L., PERRONE-BERTOLOTTI M., THOMAS E., BACIU M. (2017). Machine learning – XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics Vol. 4, Issue 3*. New York: Springer. 159–169.

VALETTE M., GRABAR N. (2004). Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP. *Le poids des mots, Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. 1111.

YANG Z., LEVOW G.-A., MENG H. (2012). Predicting User Satisfaction in Spoken Dialog System Evaluation With Collaborative Filtering. *IEEE Journal of Selected Topics in Signal Processing*.