# Utilisation d'annotations sémantiques pour la validation automatique d'hypothèses dans des conversations téléphoniques

Carole Lailler[1] Yannick Estève[1] Renato de Mori[2,3] Mohamed Bouallègue[1] Mohamed Morchid [2]
(1) LIUM — Université du Maine, France
(2) LIA — Université d'Avignon et des Pays de Vaucluse, France
(3) McGill University, Montréal, Canada

**Résumé.** Les travaux présentés portent sur l'extraction automatique d'unités sémantiques et l'évaluation de leur pertinence pour des conversations téléphoniques. Le corpus utilisé est le corpus français DECODA. L'objectif de la tâche est de permettre l'étiquetage automatique en thème de chaque conversation. Compte tenu du caractère spontané de ce type de conversations et de la taille du corpus, nous proposons de recourir à une stratégie semi-supervisée fondée sur la construction d'une ontologie et d'un apprentissage actif simple : un annotateur humain analyse non seulement les listes d'unités sémantiques candidates menant au thème mais étudie également une petite quantité de conversations. La pertinence de la relation unissant les unités sémantiques conservées, le sous-thème issu de l'ontologie et le thème annoté est évaluée par un DNN, prenant en compte une représentation vectorielle du document. L'intégration des unités sémantiques retenues dans le processus de classification en thème améliore les performances.

**Abstract.**
**Use of Semantic Annotations for Validating Mentions of Semantic Hypotheses in Telephone Conversations.**
The presented work focuses on the automatic extraction of semantic units and evaluation of their relevance to telephone conversations. The corpus used is DECODA corpus. The objective of the task is to enable automatic labeling theme of each conversation. Given the spontaneous nature of this type of conversations and the size of the corpus, we propose to use a semi-supervised strategy based on the construction of an ontology and a simple active learning : a human annotator analyses not only the lists of semantic units leading to the theme, but also studying a small amount of conversations. The relevance of the relationship between the conserved semantic units, sub-theme from the ontology and annotated theme is assessed by DNN, taking into account a vector representation of the document. The integration of semantic units included in the theme classification process improves performance.

**Mots-clés :** analyse de conversation humain/humain, extraction automatique d'unités sémantiques pertinentes, validation d'une ontologie.

**Keywords:** human/human conversation analysis, automatic extraction of relevant semantic units, ontology validation.

## 1 Introduction

Les travaux présentés concernent une analyse sémantique du corpus DECODA (http ://decoda.univ-avignon.fr/). Il s'agit d'un corpus issu d'une application de conseil à la clientèle par téléphone mise en place en interne par la Régie Autonome des Transports Parisiens (dorénavant RATP). L'objectif est de permettre l'extraction automatique du thème de chaque conversation en utilisant un DNN (DNN pour « *deep neural networks* »). Toutefois, deux difficultés majeures demeurent avec ce corpus : d'une part, le nombre de conversations reste faible au regard des besoins d'apprentissage d'un DNN. D'autre part, les Word Error Rate (dorénavant WER) obtenus par les systèmes de reconnaissance de la parole sur ce type d'oral demeurent élevés. Ces deux constats nous ont conduits à utiliser une méthode semi-supervisée avec annotations pour faciliter la détection en thème. Dans ce cadre, l'article propose l'utilisation de réseaux de neurones profonds pour valider l'extraction des mentions sémantiques pertinentes selon les relations sémantiques établies dans l'ontologie construite après extraction automatique de ngrams. Il s'agit de valider la relation tripartite [ngram ; sous-thème ; thème] mise au point dans l'ontologie. Les DNN utilisés ont été inspirés par des travaux récemment publiés et portent sur l'extraction de relations dans le discours (Bost *et al.*, 2015) et (Ji & Eisenstein, 2014).

Il s'agit tout d'abord, d'obtenir une ontologie suffisamment générique et exhaustive pour faciliter l'extraction automatique d'éléments sémantiques pertinents permettant de conduire au thème de la conversation. Par ailleurs, chaque conversation doit faire l'objet d'un rapport final construit automatiquement consignant les éléments les plus informatifs des interactions Conseiller-Usager de la RATP. La documentation fournie par la société gestionnaire de l'application téléphonique contient des informations suffisantes pour en extraire une ontologie. Elle fait suite à l'extraction automatique de ngrams jugés pertinents, selon le critère de Gini (Breiman *et al.*, 1984). Elle permet de pallier le manque de données et d'organiser dans une structure suffisamment générique les ngrams extraits automatiquement. Elle étend également la recherche de ces ngrams selon des principes sémantiques et discursifs logiques. Elle est établie selon plusieurs niveaux de granularité (5 au total) en fonction du type d'informations recherchées. Elle doit permettre l'extraction d'éléments sémantiques pertinents conduisant non seulement au thème mais levant aussi les ambiguïtés de la conversation. Nous n'aborderons ici que les deux principaux niveaux sémantiques de l'ontologie, à savoir le niveau thématique et le niveau syntagmatique qui organise, selon des sous-thèmes, les ngrams les plus discriminants pour chaque thème. Le niveau dialogique fera l'objet d'une présentation ultérieure. Il s'agit pour l'instant de ne s'intéresser qu'aux éléments sémantiques permettant de conduire au thème de la conversation, indépendemment des tours de parole et de la fonction des locuteurs.

Le premier niveau est celui qui contient le thème de la conversation. Le second rassemble, en une série de vingt sous-thèmes, les éléments sémantiques les plus discriminants, ceux qui ont pour but d'inférer le thème. Ces sous-thèmes contiennent les éléments les plus signifiants au regard de l'application. Ils permettent de relier un thème aux unités sémantiques pertinentes, c'est-à-dire aux ngrams porteurs de l'information. Constitués d'un ngram, d'un sous-thème indiquant le champ sémantique abordé et d'un thème, cet ensemble tripartite, exprimé en langage naturel, est de taille variable : il va du mot au syntagme (jusqu'à six entités lexicales sans blanc) et peut intégrer des valeurs numériques (numéro de bus, de rue, etc.) et des entités nommées. Nous l'appellerons dorénavant TRSC pour « *theme-specific report component* ». Ainsi, le TRSC [mouvement social ; Grève ; ETFC] appartient au thème « ETFC » (État du trafic). Il a pour sous-thème « Grève » et pour mention sémantique pertinente « mouvement social ». La relation entre le sous-thème et la mention est le prédicat « perturber ». Ainsi, les deux arguments du prédicat sont des éléments issus de l'ontologie qui doivent être insérés dans le rapport.

Néanmoins, les ngrams porteurs de l'information conduisant au thème sont disséminés dans l'ensemble des énoncés d'une conversation. Compte tenu notamment du caractère spontané des échanges (importance des disfluences et des répétitions), il peut être difficile de les retrouver ou d'appliquer des bornes pour permettre leur identification. Pour cette raison, les méthodes généralement proposées pour l'identification en thème et en particulier pour l'identification des traits sémantiques ne sont pas appropriées. La relation exprimée par le prédicat n'est pas immédiatement appréhendable morphosyntaxiquement. Il est souvent difficile pour un agent de suivre le protocole fixé à l'avance. Le client adopte rarement un comportement prévisible et normé. Il tend à s'écarter du champ d'application se rapportant au domaine des transports. Toutes ces digressions perturbent la bonne marche de l'échange et retardent sa conclusion. Ainsi, les modèles issus de l'ontologie décrivant l'expression d'un TRSC sont prévus pour tenir compte non seulement des scenarii-protocoles mais également des aléas d'une conversation téléphonique. En se fondant sur la documentation de l'application et sur le protocole devant être suivi par l'agent, il est possible de détecter des TRSC pertinents qui conduiront au thème de la conversation et dont les sous-thèmes pourront servir de fils conducteurs dans l'élaboration du rapport.

## 2 Travaux connexes

Des résultats prometteurs ont récemment été obtenus sur l'extraction de relations unissant des entités exprimées dans une phrase (Mesquita *et al.*, 2013), selon une analyse de dépendance entre la phrase et les relations définies. Plus récemment, de nouvelles solutions ont permis l'évaluation de la cohérence dans des successions de phrases au sein d'un même texte (Li & Hovy, 2014). Ces solutions sont inspirées par la Théorie de la Structure Rhétorique (RST pour « *Rhetorical Structure Theory* ») (Mann & Thompson, 1988). La RST considère un texte comme cohérent s'il est composé d'unités de discours élémentaires (EDU), portées par des phrases et un petit ensemble de relations de discours typiques les unissant.

Toutefois les relations entre les unités sémantiques et les thèmes considérés dans cette étude diffèrent de celles prises en considération dans d'autres types d'extraction de relation. En effet, les ngrams candidats à l'élection d'un thème peuvent être utilisés dans des conversations renvoyant à un autre thème. Il peut arriver, par exemple, qu'un ngram se rattachant à une notion temporelle, déclarée comme TRSC pour le thème HORR (Horaires) apparaisse dans une conversation dont le thème principal est ITNR (Itinéraire). C'est le cas notamment en fin de conversation quand le client souhaite s'assurer de son temps de trajet. Dans ce cas, la mention sémantique exprimée par le modèle n'a pas à être déclarée comme un TRSC pertinent pour le thème principal de la conversation. De plus, les ngrams pertinents sont généralement disséminés dans l'ensemble des segments de conversation. Une réflexion menée sur l'utilisation de mots avec des liens de dépendance

à longue distance qui ne peuvent donc être retrouvées avec de simples analyseurs de phrases est présentée dans (Ji & Eisenstein, 2014) et (Prasad *et al.*, 2014).

Dans notre cas, le problème relève d'une difficulté supplémentaire, puisqu'au lieu d'utiliser un document sous forme d'un texte cohérent, nous analysons des conversations téléphoniques en langue spontanée au sein desquelles les interventions du client sont souvent imprévisibles, avec une grande variabilité morphosyntaxique et de nombreux "bruits" (disfluences, répétions, etc.). Par ailleurs, il nous faut essayer de trouver une solution pour établir des relations pertinentes entre une mention sémantique locale exprimée par un ngram et un thème caractéristique qui reste commun à l'ensemble de la conversation. Les sous-thèmes ont justement pour but de permettre l'établissement de cette relation : en offrant la possibilité de diviser une conservation selon des unités sémantiques plus petites, ils soulignent la progression argumentative et thématique d'un échange en tenant compte des éléments de variabilité du discours. Les sous-thèmes permettent d'établir une connexion sémantique unique et nécessaire entre un ngram structuré et le thème de la conversation. La détection des TRSC est une étape qui va au-delà de la description d'un thème (Hazen, 2011), (Morchid *et al.*, 2014a), (Morchid *et al.*, 2014b). Le réseau de neurones, décrit dans (Estève *et al.*, 2015), est considéré ici comme le système de référence. Il sert de point de comparaison à notre étude. Nous le nommerons SystOne.

# 3 Domaine d'application

Les segments de dialogues considérés dans l'application de la RATP sont constitués d'un problème rencontré par le client, d'une phase de reformulation puis d'une réponse (ou de bribes de réponse) apportée par l'agent. Ce dernier tente de résoudre le problème posé tout en suivant un protocole prédéfini. Le problème est formulé de telle façon que les mentions sémantiques révélant le thème sont livrées tout au long de l'échange selon un principe de pertinence. Ainsi, l'énoncé du problème et sa réponse sont reliés par des relations de continuité discursive. Le discours se veut généralement collaboratif entre les deux parties. Après avoir reformulé en des éléments clairs et concis le problème de l'usager, le conseiller cherche à rapidement apporter une réponse. Suivre les protocoles imposés par la RATP induit également des types de réponse. L'ontologie créée prend en compte ces contingences discursives et les utilise. Elle se fonde sur la documentation fournie par le service et son site internet mais aussi sur les relations qui se créent au cours des échanges entre les locuteurs. Elle prend la forme d'un graphe dont les nœuds représentent les unités sémantiques les plus discriminantes et essentielles. Ces nœuds sont unis aux thèmes par des liens qui représentent les relations sémantiques prédicatives les plus couramment utilisées dans l'application. Le contenu du rapport est guidé par une stratégie de composition reposant sur une planification résultant du DNN. Cette stratégie permet alors la formulation d'hypothèses dévoilant les relations sémantiques menant au thème de l'échange.

L'ensemble des thèmes du domaine liés à l'application et leurs abréviations sont au nombre de douze : *ITNR = itinéraire, OBJT = objets trouvés et perdus, HORR = horaires, NVGO = cartes de transport et abonnement, VGC = Vente Grand Compte (il s'agit d'un thème rassemblant les ventes aux entreprises et collectivités ainsi que les cartes des ayant-droits), ETFC = état du trafic, TARF = les tarifs, PV = les infractions, OFTP = l'offre de transport palliatif (en cas de travaux de réfection par exemple), CPAG = problème avec un agent, JSTF = les justificatifs de retard, RETT = les remboursements.* Un treizième thème, AAPL, concerne les conversations qui font référence à un problème concernant non pas la RATP mais la SNCF (le réseau francilien étant partagé par ces deux compagnies). Enfin, les conversations hors domaine sont rassemblées sous le thème NULL. Des exemples sur la segmentation du dialogue et les détails de la détection des huit premiers thèmes peuvent être retrouvés dans (Morchid *et al.*, 2014a) et (Morchid *et al.*, 2014b).

Initialement, seuls les thèmes ont été annotés au sein des transcriptions manuelles. Concernant les conversations qui présentent dans leur déroulé plusieurs thèmes, un unique thème majoritaire a été retenu en se référant aux règles contenues dans la documentation de service. Toutefois, ces conversations multithèmes restent un frein à l'analyse automatique. Elles déclenchent de fausses alertes : des TRSC appartenant à d'autres thèmes sont ainsi détectés et conduisent à un mauvais étiquetage. Afin de réduire l'effort d'annotation des TRSC, une procédure semi-automatique, minimisant l'effort humain, est proposée. Elle consiste à utiliser l'ensemble des conversations du corpus d'apprentissage pour trouver automatiquement une liste de ngrams selon leur indice de pureté dans chaque thème. L'indice de pureté est calculé selon le critère de Gini (Breiman *et al.*, 1984). Cette liste est ensuite analysée par un expert humain qui sélectionne et étend dans les modèles de l'ontologie les ngrams les plus pertinents en les associant à un sous-thème afin de construire un TRSC efficace et unique. Il s'agit pour l'expert de nettoyer cette première liste automatique de ses scories et d'étendre ses éléments pour la rendre plus robuste. Ainsi, l'effort humain dépend principalement de la taille de la liste initiale. L'expert se sert également de ses connaissances sur l'application et de la documentation de service pour enrichir cette première liste automatique de mentions sémantiques à conserver. Les TRSC pourront être enrichis au fur et à mesure de l'acquisition de nouvelles données

afin de ne pas voir diminuer leur pureté : les nouveaux TRSC ainsi injectés permettent d'accroître la reconnaissance en thème et de réduire les confusions inter-thèmes.

Nous considérons $\Gamma_t = \{\gamma_{t,1}, ..., \gamma_{t,i}, ..., \gamma_{t,I_t}\}$ comme un ensemble de TRSC du thème *t*. Chaque TRSC $\gamma_{t,i}$ est exprimé par un ensemble de mentions $M_{t,i} = \{m_1^{t,i}, ..., m_j^{t,i}, ..., m_{J_{t,i}}^{t,i}\}$. Une mention $m_j^{t,i}$ est exprimée par un modèle de mentions qui peut contenir des mots, des déclinaisons de syntagmes, ou encore des mots entrecoupés par des éléments spécifiques et identifiables (entités nommées ou valeurs numériques). L'objectif poursuivi est d'obtenir un nombre suffisant d'exemples positifs et négatifs pour entraîner les DNN. Pour une conversation donnée, tous les ngrams issus de TRSC conduisant à un autre thème sont considérés comme des exemples négatifs. En revanche, tous les ngrams appartenant à des TRSC du thème annotés manuellement constituent potentiellement des exemples positifs. Cette hypothèse de travail a été validée en utilisant un échantillonnage aléatoire du corpus d'apprentissage contenant une proportion d'exemples négatifs suffisants par rapport aux exemples positifs. Une vérification en OUI/NON pour chacun des TRSC est ensuite effectuée : le OUI correspond à une adéquation entre le TRSC et le thème annoté, le NON est signe de fausse alarme.

Par ailleurs, il faut noter que les conversations qui ne contenaient aucun TRSC ont été isolées. Il s'agit de constituer un sous-ensemble suffisant et cohérent permettant d'effectuer un apprentissage actif simple : en l'occurrence, une analyse minutieuse des conversations qui ne contiennent aucun TRSC pour essayer de capturer de nouveaux ngrams porteurs d'information et les insérer ensuite dans une relation tripartite [ngram ; sous-thème ; thème]. En s'assurant de la robustesse des ngrams déjà engagés dans les niveaux de l'ontologie par une vérification manuelle binaire et en y ajoutant des ngrams issus des conversations étudiées selon leur thème principal, la stratégie d'apprentissage actif mise en place n'a nécessité que peu d'effort.

Une mention $m_j^{t,i}$ est une expression de $\gamma_{t,i}$ quand elle est pertinente avec le thème *t* dans une conversation annotée avec *t*. Sinon, il ne s'agit pas d'une expression de $\gamma_{t,i}$ et cela entraîne une ambiguïté dans les hypothèses formulées concernant le thème *t*. Comme cela a été observé empiriquement, la cooccurrence d'une mention $m_j^{t,i}$ et du thème correspondant est fréquente, une mesure d'ambiguïté peut être estimée en suivant l'entropie conditionnelle suivante :

$$H[t|m_j^{t,i}] = -P[t|m_j^{t,i}] * logP[t|m_j^{t,i}] \quad (1)$$

Une mention $m_j^{t,i}$ provoque peu d'ambiguïté si elle a un haut degré de pureté $P[t|m_j^{t,i}]$. Les mentions sémantiques pertinentes sont sélectionnées manuellement au sein d'une liste $L_t$ de candidats issus d'une procédure de sélection automatique, qui classe des ngrams de mots en fonction de leur pureté par rapport au thème *t*. La sélection est effectuée et complétée par un expert humain selon une méthode semi-supervisée simple, en utilisant les connaissances recueillies auprès de la documentation de service. L'expert humain introduit également des déclinaisons de syntagmes, des mots sélectionnés pour leur unicité voire des séquences de mots exprimant des domaines sémantiques identifiés, par exemple, l'heure, la date, la localisation, le type d'objets perdus, etc. Un ngram $m_j^{t,i}$ est exprimé dans une conversation *d* par une instance $m_j^{t,i}(n, d)$ débutant avec le ngram qui doit conduire à un unique thème de la transcription de *d*.

# 4 Détection des relations de discours

Récemment, des architectures de réseaux de neurones ont été utilisées pour extraire des relations de discours en ayant recours à la distribution du texte et à son évolution paragraphe après paragraphe (Li & Hovy, 2014). Par ailleurs, (Ji & Eisenstein, 2014) et (Prasad *et al.*, 2014) ont démontré que les caractéristiques syntaxiques ne peuvent suffire à elles-seules pour capturer le contenu sémantique d'un document, notamment en raison de "réalisations lexicales alternatives" (Prasad *et al.*, 2014). C'est la raison pour laquelle il a été proposé d'utiliser une représentation des documents par sacs de mots, censée permettre de révéler des caractéristiques cachées idoines qui identifient les relations de discours. Compte-tenu de l'application utilisée et des conversations humain/humain obtenues, l'objectif est ici de détecter la relation de pertinence entre un TRSC issu de la liste obtenue de manière semi-supervisée et le thème annoté en amont.

Un DNN est proposé pour évaluer la relation de pertinence

$\mathcal{R}_{pertinence}[m_j^{t,i}(n, d), t]$ entre une mention $m_j^{t,i}(n, d)$ et un thème *t*. Une telle relation de discours valide la relation de pertinence, représentée telle que $m_j^{t,i}(n, d) \rightarrow \gamma_{t,i}$, conduisant de la mention au thème. Ceci valide alors l'inférence suivante :

$$\mathcal{R}_{pertinence}[m_j^{t,i}(n, d), t] \Rightarrow [m_j^{t,i}(n, d) \rightarrow \gamma_{t,i}] \quad (2)$$

Les entrées du DNN sont des représentations continues de TRSC de $m_j^{t,i}(n, d)$ obtenues en sommant les vecteurs de mots

les composant, un vecteur, appelé C-vector, représentant la totalité de la conversation (dont le calcul est décrit dans (Morchid *et al.*, 2014a)) et les scores de pureté de $m_j^{t,i}(n,d)$. Ces scores sont obtenus sur les données du corpus d'apprentissage pour chaque thème en ajoutant le vecteur des scores calculés avec *d* pour chaque thème (en utilisant le système SystOne). Le vecteur présenté en entrée du réseau résulte d'une concaténation de ces éléments. Pour cette étude, des représentations continues de mots sur 100 dimensions ont été calculées à partir d'un corpus conséquent, d'environ 2 milliards de mots. Ce corpus a été élaboré à partir de l'outil word2vec, décrit dans (Mikolov *et al.*, 2013), en utilisant les articles du quotidien français "Le Monde", le corpus Gigaword en français, des articles de Google News et des transcriptions manuelles de journaux télévisuels français, à hauteur de 400 heures d'enregistrement.

La sortie du DNN est un score mesurant la validité de l'inférence $\mathcal{R}_{pertinence}[m_j^{t,i}(n,d),t] \Rightarrow [m_j^{t,i}(n,d) \rightarrow \gamma_{t,i}]$ Le DNN est entraîné avec les données du corpus d'apprentissage en réglant la sortie pour s'approcher au plus près de l'annotation de référence de la conversation *d* sur le thème *t* , pour lequel $[m_j^{t,i}(n,d) \rightarrow \gamma_{t,i}]$. La sortie doit s'approcher de zéro quand $[m_j^{t,i}(n,d) \rightarrow \gamma_{\bar{t},i}]$ où $\bar{t}$ ne constitue pas le thème avec lequel la conversation *d* a été annotée. L'évaluation sur les transcriptions manuelles du corpus de développement a montré que 91,3% des décisions prises par le DNN sont pertinentes. Ce pourcentage descend à 88,2% sur les sorties issues d'un système de reconnaissance de la parole. Compte tenu du taux de WER relativement élevé pour ce type de conversations, l'écart entre les résultats obtenus sur des transcriptions manuelles ou des sorties de système est faible : les décisions du DNN restent fiables. Ainsi, sur les 636 occurrences de ngrams appartenant à la première liste de 365 éléments détectés dans le corpus de développement, 389 sont conservées et utilisées par le DNN. Une autre évaluation a été effectuée pour établir la pertinence de la conversation *d* dans sa totalité en prenant en compte les mentions annotées avec le même TRSC $\gamma_{t,i}$ pour la liste de thème *t*. Ce type d'évaluation peut s'avérer utile pour lever les erreurs de thème et les hypothèses de TRSC erronées, ainsi que pour révéler la présence éventuelle de plusieurs thèmes dans une conversation. Plusieurs TRSC conduisant à des thèmes différents peuvent en effet entrer en conflit au sein d'une même conversation.

Considérons

$$S_{t,i}^d(n_1, q_{t,i}) = [m_1^{t,i}(n_1,d), ..., m_j^{t,i}(n_j,d), ..., m_{Q_{t,i}}^{t,i}(n_{Q_{t,i}},d)] \tag{3}$$

comme étant la séquence pleine et entière de toutes les instances des mentions de $\gamma_{t,i}$ détectées dans une conversation *d*, en commençant avec une mention à la position $n_1$ et contenant $q_{t,i}$ mentions. La pureté de $S_{t,i}^d(n_1, q_{t,i})$ correspond à la probabilité de $P[t \| S_{t,i}^d(n_1, q_{t,i})|]$. L'approximation suivante est utilisée pour son calcul :

$$P[t|S_{t,i}^d(n_1, q_{t,i})] \approx P[t \| S_{t,i}^d(n_1, q_{t,i})|] \tag{4}$$

La stratégie conduit à construire une séquence de mentions $S_{t,i}^d(n_1, q_{t,i})$ si au moins un ngram possède un haut degré de pureté avec *t*. Une heuristique destinée à opérer des corrections est utilisée. Elle se fonde sur l'inférence logique suivante : SI {la pureté de $S_{t,i}^d(n_1, q_{t,i})$ est élevée et que *t* n'est pas annoté automatiquement pour *d* ET qu'il n'existe aucun ngram pertinent appartenant à un TRSC pour le thème annoté automatiquement pour $d$} ALORS {$t$ remplace le thème annoté automatiquement pour la conversation $d$}. Il s'agit, ce faisant, de lever les erreurs d'annotations en thème en utilisant un haut degré de précision dans la détection des ngrams pertinents.

# 5 Expériences

L'objectif de ces expériences est d'évaluer la robustesse de la méthode semi-supervisée proposée. Le corpus utilisé pour les expériences comprend 2109 conversations.

Un sous-ensemble du corpus d'apprentissage contenant 600 conversations sur les 1 489 disponibles a été annoté automatiquement en utilisant 345 ngrams sélectionnés avec la procédure semi-automatique décrite en section 3. Ces ngrams sont inclus dans un TRSC, selon un jeu de 20 sous-thèmes possibles connectés à 13 thèmes appartenant à l'application, auxquels s'ajoute le thème NULL. Chaque ensemble tripartite est unique et ne permet de conduire qu'à un seul thème. Les sous-thèmes dévoilent les arguments les plus couramment utilisés dans une conversation. Toutes les conversations qui ne contiennent pas de TRSC sont étiquetées NULL. Sur ces 600 conversations, plus de 60% lèvent une alarme pour au moins un TRSC appartenant au thème annoté, 11% constituent des exemples négatifs sur lesquels entraîner un DNN. Les 29% restants sont des conversations qui ne contiennent aucune mention pertinente pour le thème annoté. Ces conversations ont donc formé un sous-ensemble destiné à être analysé par un expert humain selon un processus d'apprentissage actif, qui a permis d'ajouter 435 TRSC et d'obtenir une liste finale de 780 occurrences tripartites. Cette méthode présente l'intérêt d'être économe dans la mesure où, à partir d'un nombre relativement restreint de données et d'un apprentissage

semi-supervisé simple, on aboutit à un résultat. Toutefois, l'évaluation de ce résultat nécessiterait des éléments de comparaison voire une métrique ajustée. Cette dernière permettrait d'associer au volume de données, le temps de construction de l'ontologie et celui d'annotation pour mieux mesurer le gain par rapport à une seule extraction automatique de ngrams selon des critères de pureté. Ainsi, la portabilité de ce type de méthode pourrait être évaluée.

Toutes les conversations ont été transcrites manuellement et annotées en thème. Parmi elles, 322 conversations ont été utilisées pour constituer le corpus de TEST et 298 pour le corpus de développement. Les ngrams composant un TRSC tripartite des corpus de test et de développement pour lesquels $m_j^{t,i}(n,d) \rightarrow \gamma_{t,i}$ ont été vérifiés manuellement. Toutes les conversations ont été transcrites automatiquement avec un système de reconnaissance automatique de la parole (ASR). Ce système, décrit dans (Rousseau *et al.*, 2014), génère, pour chaque conversation, les hypothèses de séquence de mots les plus probables. Le modèle de langage du système a été adapté au domaine et un procédé du Leave-One-Out a été utilisé. Les taux de « *Word Error Rate* » (WER) du sytème sont respectivement de 33,8% pour le DEV et de 34,5% pour le TEST. Le tableau 1 présente les résultats de l'annotation automatique en thème (en utilisant les 13 thèmes à notre disposition plus le thème NULL) avant et après l'utilisation d'heuristiques de correction décrites à la fin de la section précédente. La dernière ligne du tableau est consacrée aux résultats en termes de pertinence de détection des ngrams en utilisant le DNN décrit plus haut.

| | DEV | TEST |
|---|---|---|
| Classification en thèmes | 83.2% | 81.4% |
| Classification en thèmes après correction | 84.6% | 83.2% |
| Pertinence de la détection en mentions | 89.4% | 87.0% |

TABLE 1 – Résultats obtenus en termes de détection d'un thème

Le haut degré de précision quant à la classification en thème, observé avec les résultats de sorties du système ASR, est encourageant. Il est de 96% sur le Dev et de 89% sur Test. Le rappel est, quant à lui, de 87% sur le DEV et de 85% sur le Test. En se fondant sur un effort de modélisation des TRSC, on peut thématiser des conversations téléphoniques mais aussi concevoir une stratégie de correction d'erreurs. La présence de ngrams inclus dans des TRSC non compatibles avec le thème annoté est souvent dû au fait que certaines conversations sont multithématiques. Or, seul le thème dominant a été annoté et évalué.

Enfin, on doit noter que si l'on ne tenait compte que de la détection des TRSC dans les conversations sans appliquer la méthode de correction décrite plus haut (fondée sur l'utilisation d'un DNN), aucune amélioration ne serait introduite. Au contraire, nos expérimentations montrent que, sur le corpus de Test, le taux de bonne classification en thème passerait de 81,4% à 79,9%. Dans certaines applications de dialogue sur des tâches courantes, comme c'est le cas ici avec le corpus DECODA, on s'aperçoit que l'utilisation d'un DNN accompagnée d'une approche semi-supervisée simple permet de mieux percevoir l'objet de l'échange et la détection en thème alors même que le WER montre que la tâche de reconnaissance de la parole reste délicate.

# 6 Conclusion et Perspectives

Nous avons proposé une méthode d'extraction automatique d'hypothèses pour la détection de thèmes dans des conversations téléphoniques. Le but est de permettre l'élaboration automatique de rapports indiquant, entre autres, le thème des échanges. Les TRSC, qui constituent autant de déclencheurs permettant de remonter directement au thème, sont annotés à partir du corpus d'apprentissage en utilisant une approche semi-supervisée et une stratégie d'apprentissage actif simple, ne nécessitant qu'un effort humain limité. Les résultats expérimentaux montrent une bonne capacité de détection en thème lorsque seul le thème principal de la conversation est annoté. Certes, les conversations multi-thématiques viennent brouiller les pistes et constituent une difficulté majeure dans ce type de travail. Toutefois, la mesure de précision indique de fortes valeurs pour les mentions validées par le DNN. Les taux de rappel sont eux-aussi corrects. Actuellement, de nouvelles annotations sont menées pour faciliter la formulation d'hypothèses y compris pour les thèmes abordés dans un second temps au sein d'une conversation. Les bons résultats obtenus par l'annotation manuelle sur l'ensemble du TEST indiquent qu'une approche fondée sur un effort humain limité mais guidée par une ontologie respectant l'application peut être efficace. En sélectionnant et en généralisant le contenus des listes de ngrams générées et évaluées automatiquement, il est possible d'entraîner des DNN et de valider les hypothèses ainsi émises. Parallèlement, des travaux de recherche sont actuellement en cours pour étendre cette approche à la détection des actes de dialogue.

## Remerciements

Ce travail a été partiellement financé par la Commission Européenne à travers le projet EUMSSI, sous le numéro de contrat 611057, selon le numéro d'identification FP7-ICT-2013-10. Ce travail a également fait l'objet d'un financement de la part de l'Agence Nationale pour la Recherche (ANR) à travers le projet VERA qui porte le numéro ANR-12-BS02-006-01.

## Références

BOST X., DENAY G., EL-BEZE M. & MORI R. D. (2015). Multiple topic identification in human/human conversations. In *Actes de In Computer Speech and Language Journal2015(ICSLJ2015)*.

BREIMAN L., FRIEDMAN J. H., OLSHEN R. & STONE C. (1984). Classification and regression trees. In *Technical report, Wadsworth international*, Monterey, CA.

ESTÈVE Y., BOUALLEGUE M., LAILLER C., MORCHID M., DUFOUR R., LINARÈS G. & MORI R. D. (2015). Integration of word and smeantic features for theme identification in telephone conversations. In *Actes de International Worshop on Spoken Dialogue System2015(IWSDS2015)*, Buzan, South Korea.

HAZEN T. J. (2011). Topic identification. In G. TUR & J. R. DE MORI, Eds., *Spoken language understanding : Systems for extracting semantic information from speech. John Wiley and Sons*, p. 319–356.

JI Y. & EISENSTEIN J. (2014). Representation learning for text-level discourse parsing. In *Actes de the 23rd International Conference on Computational Linguistics2014*, p. 595–603.

LI J. & HOVY E. (2014). A model of coherence based on distributed sentence representation. In *Actes de Conference on Empirical Methods in Natural Language Processing*, p. 2039–2048, Doha, Qatar.

MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. In *Text, 8(3)*, p. 243–281.

MESQUITA F., SCHMIDEK J. & BARBOSSA D. (2013). Effectiveness an efficiency of open relation extraction. In *Actes de Conference on Empirical Methods in Natural Language Processing*, p. 447–457, Seattle, Washingthon, USA.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Actes de The Second International Conference on Learning Representations*.

MORCHID M., DUFOUR R., BOUALLEGUE M., LINARÈS G., MATROUF D. & MORI R. D. (2014a). An i-vector based approach to compact multi-granularity topic spaces representation of textual documents. In *Actes de Conference on Empirical Methods in Natural Language Processing*, p. 443–454, Doha, Qatar.

MORCHID M., DUFOUR R., BOUSQUET P.-M., LINARÈS G. & MORI R. D. (2014b). Improving dialogue classification using a topic space representation and a gaussian classifer based on the decision rule. In *Actes de the IEEE International Conference on Acoustic, Speech and Signal Processing*, p. 126–130, Florence, Italy.

PRASAD R., JOSHI A. & WEBBER B. (2014). Realization of discourse relations by other means : alternative lexicalizations. In *Actes de the 23rd International Conference on Computational Linguistics*, p. 1023–1031 : Association for Computational Linguistics.

ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). Lium and crim asr system combination for the repere evaluation campaign. In *Actes de 17th Internation Conference on Text, Speech and Dialogue*, Brno, Czech republic.