Catégorisation multi-thématique de dialogues téléphoniques

Xavier Bost, Marc El-Bèze, Renato de Mori

▶ **To cite this version:**

Xavier Bost, Marc El-Bèze, Renato de Mori. Catégorisation multi-thématique de dialogues téléphoniques. XXXe édition des Journées d'Études sur la Parole (JEP 2014), Jun 2014, Le Mans, France. hal-01967848

HAL Id: hal-01967848
https://hal.science/hal-01967848

Submitted on 1 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Catégorisation multi-thématique de dialogues téléphoniques

Xavier Bost[1] Marc El-Bèze[1] Renato De Mori[1,2]
(1) LIA, Université d'Avignon, 339 chemin des Meinajariès, Avignon, France
(2) McGill University, School of Computer Science, Montréal, Québec, Canada
{xavier.bost, marc.elbeze, renato.demori}@univ-avignon.fr

RÉSUMÉ

Ce papier porte sur l'analyse automatique de dialogues téléphoniques entre un client et un agent d'un centre d'appel d'un service clientèle. Le but de l'analyse est d'identifier, parmi un ensemble thématique prédéfini, les thèmes des problèmes évoqués dans le dialogue. Un dialogue peut contenir des thèmes multiples mentionnés dans des segments entremêlés difficiles à délimiter. Deux méthodes sont proposées pour conjecturer les thèmes multiples. La première repose sur une mesure de similarité *cosine* appliquée au sac de mots extrait de la totalité du dialogue. La seconde méthode introduit le concept de densité thématique distribuée autour de positions spécifiques du dialogue. En plus des unigrammes, sont également pris en compte les bigrammes, avec d'éventuels trous entre les deux termes. Les résultats expérimentaux obtenus par les méthodes proposées apparaissent supérieurs à ceux obtenus par des machines à support de vecteurs appliquées aux mêmes données.

ABSTRACT

Multi-theme categorization of telephone dialogues

This paper deals with the automatic analysis of dialogues between a customer and an agent in a call centre of a customer care service. The purpose of the analysis is to identify, among a predefined thematic set, themes about problems discussed in the dialogue. A dialogue may contain multiple themes mentioned in interleaved segments that cannot be well defined. Two methods are proposed for multiple theme hypothesization. One of them is based on a cosine similarity measure using a bag of words extracted from the entire dialogue. The other method introduces the concept of thematic density distributed around specific word positions in a dialogue. In addition to unigrams, word bigrams with possible gaps between the two words are also considered. The experimental results obtained with the proposed methods turn out to outperform the results obtained with support vector machines on the same data.

MOTS-CLÉS : classification multi-thématique de documents audio, analyse de dialogues humain/humain, *speech analytics*, bigrammes à distance.

KEYWORDS: multi-topic audio document classification, human/human dialogue analysis, speech analytics, distance bigrams.

1 Introduction

Ces dernières années, l'intérêt porté aux possibilités offertes par les technologies de la parole pour le suivi des services téléphoniques est allé croissant. Dans un souci d'efficacité et de satisfaction du client, il existe actuellement un consensus sur l'importance de l'amélioration des systèmes d'analyse de dialogues entre humains afin d'obtenir des rapports sur les problèmes rencontrés par les clients et la façon dont les conseillers ont pu les résoudre. A partir de tels rapports, diverses statistiques peuvent être établies sur la typologie des problèmes, l'efficacité des solutions qui leur sont apportées, le comportement et le taux de satisfaction du client.

Le contexte applicatif pris en considération dans ce papier concerne l'analyse de dialogues entre un agent d'un centre d'appel et un client dont le comportement reste imprévisible. Le client est supposé demander des informations et/ou évoquer des difficultés en lien avec le système de transport parisien (RATP) et les services afférents. La documentation du centre d'appel décrit succinctement un nombre fini de motifs d'appels possibles, assimilables aux thèmes du dialogue.

Dans ce papier, nous nous concentrons sur la détection des thèmes, éventuellement multiples, présents dans le dialogue. Différents thèmes peuvent être mentionnés dans des segments disjoints du discours. Dans certains cas, les références aux différents thèmes peuvent coexister dans de courts segments ou même dans une seule phrase. Même quand les différents thèmes sont mentionnés dans des segments distincts, les frontières des segments peuvent être difficiles à tracer, tant à cause des erreurs commises par le système de transcription automatique (ASR) que de la difficulté à appréhender convenablement les structures linguistiques mises en œuvre dans une situation réelle par des usagers occasionnels.

Ce papier est structuré comme suit. La section 2 évoque les travaux antérieurs. La section 3 introduit le cadre applicatif et les descripteurs utilisés pour la détection de thèmes. La première des deux approches utilisées pour la tâche de classification multi-thématique, fondée sur une mesure de similarité *cosine*, est introduite dans la section 4. La seconde approche, décrite dans la section 5, introduit le concept de densité thématique, propre à permettre une détection locale des segments thématiques constitutifs du dialogue. Les résultats expérimentaux sont présentés dans la section 6.

2 Travaux antérieurs

Les méthodes d'analyse de dialogues entre humains ont été récemment passées en revue dans (Tur et Hakkani-Tür, 2011). Des méthodes d'identification thématique dans les documents audio sont examinées dans (Hazen, 2011). Des solutions pour la détection de segments de dialogues porteurs de différents thèmes ont été proposées dans de nombreuses publications, récemment passées en revue dans (Purver, 2011). Des solutions intéressantes ont été proposées pour des modèles linéaires de segmentation non hiérarchique. Certaines approches proposent des méthodes d'inférence pour extraire des points de segmentation localisés aux *maxima* de fonctions de cohésion. Les fonctions utilisent des paramètres extraits de chacune des phrases du dialogue ou d'une fenêtre qui couvre quelques phrases. Certaines méthodes de recherche détectent la cohésion d'un dialogue à partir de modèles de langage et d'autres prennent en considération des thèmes cachés partagés à travers les documents.

On peut trouver une revue critique de la notion de cohésion lexicale dans (Eisenstein et Barzilay, 2008), qui propose une approche non supervisée pour conjecturer des points de segmentation en utilisant des répliques automatiquement extraites de données non annotées. On trouve une évaluation de la segmentation *à grain grossier* dans (Niekrasz et Moore, 2010).

La catégorisation multi-thématique est abordée dans (Tsoumakas et Katakis, 2007) et (de Carvalho et Freitas, 2009), qui introduit une méthode dite de *création* qui consiste dans le cas de vastes corpus à créer de nouveaux labels composites par association des thèmes multiples attribués à une instance.

Ce papier généralise certains concepts de la littérature en introduisant une version des bigrammes à distance : dans une séquence de trois unigrammes, l'unigramme intermédiaire est ignoré et le bigramme est formé par concaténation des deux termes extrêmes. Ces termes sont utilisés dans le cadre de stratégies de décision fondées sur une mesure de similarité *cosine* et sur un nouveau concept de densité thématique locale.

3 Contexte applicatif et descripteurs

Notre papier propose une nouvelle approche pour annoter automatiquement les dialogues entre un agent et un client à l'aide d'un ou plusieurs des labels thématiques définis par le cadre applicatif, éléments de l'ensemble $\mathbb{C}$ défini comme suit :

$\mathbb{C}$:= {**itinéraire**, **objets perdus**, **horaires**, **tarifs**, **carte de transport**, **état du trafic**, **pv**, **appels spécialisés**, **autres**}

Chacune des deux méthodes de catégorisation multi-thématique présentées peut être définie comme une application $\gamma_i : \mathbb{X} \rightarrow \mathscr{P}(\mathbb{C})$ $(i = 1, 2)$ qui à tout dialogue d du corpus $\mathbb{X}$ associe une partie de $\mathbb{C}$. Les méthodes sont apprises à partir d'un sous-ensemble $\mathbb{T} \subset (\mathbb{X} \times \mathscr{P}(\mathbb{C}))$ de dialogues dont les thèmes ont été manuellement annotés.

A chaque couple (d, c) formé d'un dialogue d et d'une classe thématique c, on fait correspondre un couple de vecteurs $(\mathbf{v_d}, \mathbf{v_c})$.

Les composantes $w_c(t)$ du vecteur $\mathbf{v_c}$ sont données, pour tout terme t, par le produit $w_c(t) = df_c(t).idf(t).G(t)$, où $df_c(t)$ désigne le nombre de dialogues de la classe c contenant t, $idf(t)$ la fréquence inverse de document de t et $G(t)$ la pureté de t définie selon le critère de Gini par :

$$G(t) = \sum_{c \in \mathbb{C}} \mathbb{P}^2(c|t) = \sum_{c \in \mathbb{C}} \left(\frac{df_c(t)}{df_{\mathbb{T}}(t)} \right)^2 \quad (1)$$

avec $df_{\mathbb{T}}(t)$ correspondant au nombre de dialogues du corpus d'apprentissage qui contiennent au moins une fois t.

Les composantes $w_d(t)$ du vecteur $\mathbf{v_d}$ sont quant à elles définies, pour tout terme t, par $w_d(t) = tf_d(t).idf(t).G(t)$, où $tf_d(t)$ désigne le nombre total d'occurrences du terme t dans le dialogue d.

Afin d'enrichir les descripteurs, des bigrammes ont été ajoutés au vocabulaire initial de 7 217 mots en conjonction avec des bigrammes à distance formés par concaténation de mots distants

au plus de deux termes dans le cours du dialogue. Le nombre de descripteurs s'élève alors à 160 433 termes.

Afin d'éviter les effets induits par une telle dispersion des données, un ensemble réduit de descripteurs a été obtenu par sélection des termes sur la base de leur pureté $G(t)$, développée dans la formule (1), et de leur couverture.

4 Mesure globale de similarité *cosine*

La première méthode de classification γ_1 repose sur le calcul de l'indice de similarité *cosine* $\mathrm{sc}(d,c)$ entre les deux vecteurs $\mathbf{v_d}$ ($d \in \mathbb{X}$) et $\mathbf{v_c}$ ($c \in \mathbb{C}$), dont la formule est donnée par :

$$\mathrm{sc}(d,c) = \cos(\widehat{\mathbf{v_d}, \mathbf{v_c}}) = \frac{\sum_{t \in d \cap c} w_d(t).w_c(t)}{\sqrt{\sum_t w_d(t)^2 . \sum_t w_c(t)^2}}$$

La classe, notée $\hat{c}$, la plus similaire au dialogue traité lui est alors attribuée. D'éventuels thèmes secondaires lui sont en outre assignés si les scores qu'ils obtiennent excèdent un seuil fixé empiriquement et défini en proportion du score $\mathrm{sc}(d,\hat{c})$ obtenu par la classe la plus proche du dialogue. Toutefois, si le score maximal obtenu par la classe $\hat{c}$ reste en deçà d'un second seuil, fixé empiriquement en valeur absolue, aucun thème secondaire n'est attribué au dialogue et seul le thème dominant est retenu.

5 Mesure locale de densité thématique

Dans le cadre de la seconde approche, le dialogue n'est plus considéré comme un sac de mots dont l'ordre est indifférent mais linéairement dans l'ordre de son déroulement temporel. Un dialogue de N mots est alors vu, en incluant les espaces intermédiaires, comme une suite finie $(p_1, ..., p_n)$ de n positions (avec $n = 2N - 1$). Le $k-$ième unigramme du dialogue se situe alors à la position $2k-1$; le $k-$ième bigramme à la position $2k$ et le $k-$ième bigramme à un trou à la position $2k+1$.

Densité thématique : définition

La contribution $w_c(p_i)$ de la $i-$ème position du dialogue à la classe thématique $c \in \mathbb{C}$ est donnée par la somme des contributions normalisées à ce thème des termes (unigrammes et bigrammes) qui y sont localisés :

$$w_c(p_i) = \frac{1}{\| \mathbf{v_c} \|} \sum_{t \in T_{p_i}} w_c(t) \quad (i = 1, ..., n)$$

où T_{p_i} désigne l'ensemble des termes (unigrammes et bigrammes) situés à la $i-$ème position du dialogue.

A chaque position p_i du dialogue, on peut alors associer une mesure de densité thématique $d_c(p_i)$ pour chaque classe thématique c, évaluée selon la formule suivante :

$$d_c(p_i) = \frac{\sum_{j=1}^{n} \frac{w_c(p_j)}{\lambda^{d_j}}}{\sum_{j=1}^{n} \frac{1}{\lambda^{d_j}}} \; (i = 1, ..., n)$$

où $\lambda \geqslant 1$ est un paramètre de sensibilité au contexte local dont la valeur est estimée empiriquement à partir du corpus de développement ; et où d_j, la distance entre la position p_j et la position de référence p_i, est donnée par $d_j := |i - j|$.

Profil thématique d'un dialogue

La mesure de densité thématique en une position du dialogue permet d'en construire le profil thématique.

La FIGURE 1 montre le profil de la transcription automatique d'un dialogue pour $\lambda = 1.05$ (FIG. 1-a), et $\lambda = 3$ (FIG. 1-b). La densité thématique est tracée en fonction de la position dans le dialogue. Le dialogue porte sur un service de desserte aéroportuaire par bus : une mère s'enquiert pour le compte de sa propre fille, qui doit prendre dans quelques jours un avion, des horaires (notés HORR), ainsi que du tarif (noté TARF) de la desserte. Trois fonctions sont tracées : deux pour ces thèmes et une troisième pour le thème *itinéraire* (noté ITNR).

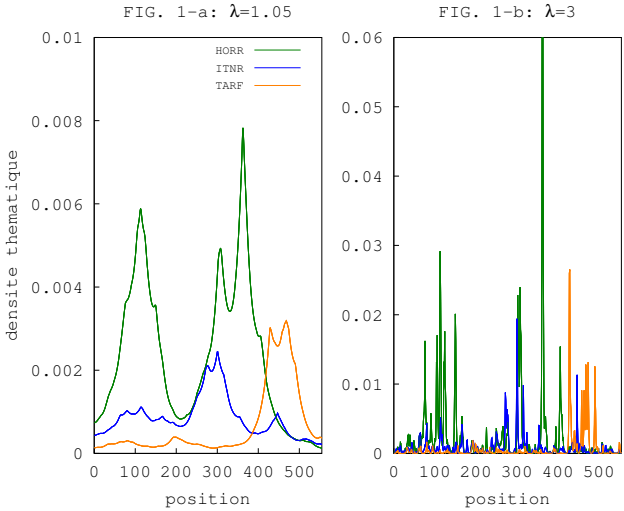

FIGURE 1 – Densités thématiques en fonction de la position dans le dialogue. Les densités sont tracées pour $\lambda = 1.05$ (FIG. 1-a) et $\lambda = 3$ (FIG. 1-b) et trois thèmes : horaires (noté HORR, tracé vert), itinéraire (noté ITNR, tracé bleu), et tarif (noté TARF, tracé orange).

On trouvera ci-après un extrait, manuellement transcrit pour plus de lisibilité, de ce dialogue. Les positions correspondantes dans la transcription automatique telles qu'elles apparaissent sur la figure 1 sont signalées entre crochets :

– *Client* [82–100] : je voudrais savoir si les (...) bus commencent tôt le matin.
– *Agent* [101–118] : cinq heures quarante-cinq le premier départ à l'opéra.

– (...)
– *Agent* [304–314] : à (...) huit heures quarante elle veut être **là-bas** ou neuf heures ?
– (...)
– *Agent* [442–470] : ah non tarification spécifique desserte aéroportuaire je vais vous donner le tarif (...) alors le tarif neuf euros dix
– *Client* [471–496] : neuf euros dix on vous donne (...) neuf euros dix en espèces ?

Pour $\lambda = 3$ (FIG. 1-b), le contexte thématique tend à être négligé. La décision d'affectation thématique peut alors indûment reposer sur des termes isolés fortement connotés isolément alors que le contexte thématique est tout autre. Par exemple, l'adverbe *là-bas* dans le tour de parole [304–314] (souligné dans l'extrait ci-dessus) tend à orienter la polarité thématique dans le sens d'une demande d'itinéraire avec un pic de densité pour ce thème autour de la position 300, alors qu'il est en fait utilisé dans le contexte d'une demande d'horaires.

Le contexte local est davantage pris en compte pour $\lambda = 1.05$ (FIG. 1-a) : la contribution de l'adverbe *là-bas* au thème *itinéraire* n'est pas soutenue par l'environnement thématique immédiat et voit sa valeur atténuée.

Convenablement estimé, le paramètre λ de sensibilité au contexte permet ainsi de prendre une décision fondée sur la cohérence thématique de segments entiers du dialogue en lissant les contributions thématiques locales par leur contexte.

Règles de décision

Deux règles de décision sont appliquées conjointement : un thème c est attribué à un dialogue d s'il est de densité dominante en une position quelconque du dialogue et si la somme de ses densités dans les positions où il est dominant excède un seuil fixé empiriquement.

6 Expériences et résultats

Cadre expérimental

Les expériences ont été menées en utilisant le système d'ASR décrit dans (Linarès *et al.*, 2007), à base de modèles de Markov cachés (HMM) acoustiques triphones avec des mélanges gaussiens appartenant à un ensemble de 230 000 distributions. Les paramètres du modèle ont été estimés selon une probabilité maximum *a posteriori* par adaptation des 150 heures de dialogues téléphoniques formées par les données du corpus d'apprentissage. Un corpus de 1 658 dialogues téléphoniques a été collecté au centre d'appel du service public des transports parisiens (RATP), puis décomposé en trois sous-ensembles : apprentissage (884 dialogues), développement (196) et test (578). Un modèle de langage trigramme a été obtenu en adaptant aux transcriptions de l'ensemble d'apprentissage un modèle de langage de base. Un ensemble initial d'expériences a été mené avec ce système avec un taux global d'erreurs mots (WER) sur le corpus de test de 57%. Ces taux d'erreurs élevés s'expliquent pour l'essentiel par les disfluences du discours et par des environnements acoustiques bruités pour certains dialogues quand, par exemple, des clients sont en train d'appeler depuis des gares ou des rues bruyantes avec des téléphones mobiles. Par ailleurs, le signal de certaines phrases est saturé ou de faible intensité à cause de la distance entre les locuteurs et les téléphones. Les corpus de développement et de test ont été annotés *a posteriori* par maximisation de l'accord entre trois annotateurs et le corpus d'apprentissage à la

volée par les agents eux-mêmes avec parfois pour consigne de n'indiquer que le thème principal.

Métriques d'évaluation

Les approches proposées ont été évaluées selon des procédures spécifiées dans (Tsoumakas et Katakis, 2007) en utilisant des métriques issues du domaine de la recherche d'informations, mais adaptées au cas particulier de la classification multi-thèmes.

Pour un corpus $\mathbb{X}$, un ensemble de classes $\mathbb{C}$ et une méthode de classification multi-thématique $\gamma : \mathbb{X} \to \mathscr{P}(\mathbb{C})$, ces métriques d'évaluation, en notant $L(d)$ l'ensemble des thèmes manuellement attribués au dialogue d, sont données par :

$$\textbf{Rappel :}\ R(\gamma,\mathbb{X}) = \frac{1}{|\mathbb{X}|}\sum_{d\in\mathbb{X}}\frac{|\gamma(d)\cap L(d)|}{|L(d)|} \quad - \quad \textbf{Précision :}\ P(\gamma,\mathbb{X}) = \frac{1}{|\mathbb{X}|}\sum_{d\in\mathbb{X}}\frac{|\gamma(d)\cap L(d)|}{|\gamma(d)|}$$

$$\textbf{F-score :}\ F(\gamma,\mathbb{X}) = \frac{2P(\gamma,\mathbb{X})R(\gamma,\mathbb{X})}{P(\gamma,\mathbb{X})+R(\gamma,\mathbb{X})} \quad - \quad \textbf{Accuracy :}\ A(\gamma,\mathbb{X}) = \frac{1}{|\mathbb{X}|}\sum_{d\in\mathbb{X}}\frac{|\gamma(d)\cap L(d)|}{|\gamma(d)\cup L(d)|}$$

Résultats

Aux deux méthodes de catégorisation γ_1 et γ_2 respectivement introduites dans les sections 4 et 5, un troisième classifieur γ_3 fondé sur des SVM a été adjoint à titre de référence. Pour l'entraînement et l'application des SVM, un noyau linéaire a été utilisé avec le même lexique de référence que les méthodes γ_i $(i = 1,2)$. Pour chaque classe $c_j \in \mathbb{C}$, un classifieur binaire $\gamma_j : \mathbb{X} \to \{c_j, \overline{c_j}\}$ est défini. A tout couple $(d, c_j) \in \mathbb{X} \times \mathbb{C}$, on peut ainsi associer un score issu du classifieur binaire correspondant. Les thèmes conjecturés pour le dialogue courant d sont alors ceux dont le score représente une certaine proportion du score maximal. On impose de plus, comme condition d'attribution d'un thème secondaire, que le score de la classe dominante excède un certain seuil empiriquement fixé.

Les résultats obtenus par les SVM, la mesure de similarité *cosine* et la mesure de densité thématique sont reportés dans le tableau 1 pour le corpus de développement et dans le tableau 2 pour le corpus de test.

DEV	MAN			ASR		
	SVM	**cos.**	**dens.**	**SVM**	**cos.**	**dens.**
Accuracy	0.76	0.86	0.85	0.74	0.82	0.80
Précision	0.84	0.95	0.94	0.82	0.92	0.92
Rappel	0.87	0.89	0.88	0.88	0.84	0.85
F-score	0.86	**0.92**	0.91	0.85	**0.88**	**0.88**

TABLE 1 – Résultats obtenus par les SVM, la mesure de similarité *cosine* et la densité thématique sur le corpus de développement.

Analyse des résultats

Le corpus de développement a été collecté pendant la même période que le corpus d'apprentissage (automne) alors qu'une partie du corpus de test a été collectée pendant l'été. Les différences

TEST	MAN			ASR		
	SVM	**cos.**	**dens.**	**SVM**	**cos.**	**dens.**
Accuracy	0.73	0.81	0.79	0.69	0.75	0.75
Précision	0.81	0.90	0.88	0.76	0.84	0.84
Rappel	0.86	0.85	0.84	0.85	0.79	0.79
F-score	0.84	**0.88**	0.86	0.80	**0.81**	**0.81**

TABLE 2 – Résultats obtenus par les SVM, la mesure de similarité *cosine* et la densité thématique sur le corpus de test. Pour la méthode fondée sur la similarité *cosine*, l'intervalle de confiance estimé est de ± 2.74% pour les transcriptions manuelles et de ± 3% pour les sorties de l'ASR.

entre les résultats obtenus sur les ensembles de test et de développement peuvent donc en partie s'expliquer par l'hétérogénéité des motifs d'appel entre ces deux périodes (par exemple, les appels relatifs aux grèves, plus fréquents à l'automne, deviennent plus rares l'été et inversement, les appels relatifs aux travaux de maintenance sont caractéristiques de la période estivale).

Par ailleurs, les différentes stratégies développées pour conjecturer un ou plusieurs thèmes secondaires en plus du thème dominant permettent toutes d'améliorer les performances par rapport à une stratégie qui affecterait au dialogue le seul thème dominant. Étendue à la recherche d'éventuels thèmes secondaires dans les dialogues du corpus de développement, la méthode de densité thématique permet par exemple d'améliorer le F-score de 0.89 à 0.91 pour les transcriptions manuelles et de 0.84 à 0.88 pour les transcriptions automatiques. Les améliorations obtenues sur le corpus de test sont du même ordre.

Enfin, en utilisant le corpus de développement pour inférer une règle de rejet d'un dialogue si les scores obtenus selon les deux classes dominantes sont trop proches, le F-score obtenu sur le corpus de test se trouve rehaussé à 0.83 si l'on applique un taux de rejet de près de 10%, soit le taux de désaccord entre des annotateurs humains.

7 Conclusion et perspectives

Pour conjecturer un ou plusieurs des thèmes évoqués dans un dialogue entre un agent de centre d'appel et un client, on a proposé d'utiliser un lexique d'unigrammes, de bigrammes et de bigrammes à distance automatiquement extraits du domaine de l'application.

Deux approches ont été introduites pour conjecturer des thèmes multiples. La première repose sur une mesure globale de similarité *cosine* appliquée aux termes extraits du dialogue ; la seconde se fonde sur une nouvelle notion de densité thématique sensible au profil thématique du dialogue. Chacune de ces deux approches s'est montrée plus performante qu'un classifieur à base de SVM fondé sur le même vocabulaire et appliqué aux mêmes données.

Les recherches à venir se concentreront sur la question des mesures de confiance appropriées à la tâche de catégorisation multi-thématique et sur l'extraction automatique de brefs rapports des dialogues à partir de leur profil thématique. Sur la base de tels rapports, des statistiques utiles à l'amélioration du service rendu à l'usager pourront être déduites, par exemple sur la fréquence des problèmes rencontrés par les clients et l'efficacité des solutions qui leur sont apportées par les conseillers.

Références

de Carvalho, A. C. et Freitas, A. A. (2009). A tutorial on multi-label classification techniques. *In Found. of Computational Intelligence Volume 5*, pages 177–195. Springer.

Eisenstein, J. et Barzilay, R. (2008). Bayesian unsupervised topic segmentation. *In Proceedings of the 2008 Conference on Emp. Methods in Nat. Lang. Processing*, pages 334–343. Association for Computational Linguistics.

Hazen, T. J. (2011). Topic identification. *In Spok. Lang. Underst.*, pages 319–356. John Wiley & Sons, Ltd.

Linarès, G., Nocéra, P., Massonie, D. et Matrouf, D. (2007). The lia speech recognition system : from 10xrt to 1xrt. *In Text, Speech and Dialogue*, pages 302–308. Springer.

Niekrasz, J. et Moore, J. D. (2010). Unbiased discourse segmentation evaluation. *In Spok. Lang. Technology Workshop (SLT)*, pages 43–48. IEEE.

Purver, M. (2011). Topic segmentation. *Spok. Lang. Underst. : Syst. for Extracting Semantic Inf. from Speech*, pages 291–317.

Tsoumakas, G. et Katakis, I. (2007). Multi-label classification : An overview. *Int. J. of Data Warehous. and Min. (IJDWM)*, 3(3):1–13.

Tur, G. et De Mori, R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.

Tur, G. et Hakkani-Tür, D. (2011). Human/human conversation understanding. *Spok. Lang. Underst. : Syst. for Extracting Semantic Inf. from Speech*, pages 225–255.