Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ?

Mohamed Morchid[†], Richard Dufour[†], Georges Linarès[†],
Renato de Mori[†‡], *Fellow, IEEE* *
[†]Laboratoire Informatique d'Avignon (LIA), Université d'Avignon, France
[‡]McGill University, School of Computer Science, Montréal, Québec, Canada
prénom.nom@univ-avignon.fr, rdemori@cs.mcgill.ca

RÉSUMÉ

Les systèmes de reconnaissance automatique de la parole (RAP) sont désormais très performants. Néanmoins, la qualité de transcription est fortement dégradée dans des environnements très bruités, ce qui influe sur les performances des applications les utilisant, telles que les tâches de classification. Dans ce papier, nous proposons d'identifier les thèmes présent dans des services vocaux téléphoniques au moyen de l'approche classique à base de fréquences de mots (TF-IDF avec le critère de pureté Gini) et au moyen de l'approche à base d'espaces de thèmes (LDA). Ces deux représentations sont ensuite utilisées dans un processus de classification utilisant les SVM afin de retrouver le thème présent dans la conversation. Enfin, nous proposons de discuter autour de la qualité, en termes de taux d'erreur-mot, des mots identifiés comme discriminants et non-discriminants par les méthodes de représentation des dialogues étudiées dans cet article.

ABSTRACT

Classification of highly imperfect automatic transcriptions : Should we adapt the word error rate ?

Although the current transcription systems could achieve high recognition performance, they still have a lot of difficulties to transcribe speech in very noisy environments. The transcription quality has a direct impact on classification tasks using text features. In this paper, we propose to identify themes of telephone conversation services with the classical Term Frequency-Inverse Document Frequency using Gini purity criteria (TF-IDF-Gini) method and with a Latent Dirichlet Allocation (LDA) approach. These approaches are coupled with a Support Vector Machine (SVM) classification to resolve theme identification problem. Results show the effectiveness of the proposed LDA-based method compared to the classical TF-IDF-Gini approach in the context of highly imperfect automatic transcriptions. Finally, we discuss the impact of discriminative and non-discriminative words extracted by both methods in terms of transcription accuracy.

MOTS-CLÉS : classification automatique, TF-IDF, LDA, taux d'erreur-mot.

KEYWORDS: document indexing, TF-IDF, LDA, word error rate.

*. Ce travail a été réalisé dans le cadre des projets SUMACC et DECODA financés par l'Agence Nationale de la Recherche (ANR) sous les contrats ANR-10-CORD-007 et ANR-09-CORD-005.

1 Introduction

L'application étudiée dans cet article concerne l'analyse automatique de conversations téléphoniques entre un agent et un client dans le cadre du service vocal d'aide aux usagers de la RATP (*Régie Autonome des Transports Parisiens*). L'agent doit suivre un protocole particulier pendant la conversation afin de répondre aux requêtes ou plaintes des usagers. Ce papier présente un système permettant l'extraction automatique de thèmes à partir de conversations téléphoniques. Un unique thème est alors proposé pour chaque conversation. Dans cette tâche particulière, les conversations peuvent contenir des segments très bruités, ce qui a pour conséquence de fortement dégrader la qualité de leur transcription automatique au moyen d'un système de reconnaissance automatique de la parole (RAP).

Dans le domaine de la recherche d'information, la caractéristique principale utilisée est la *fréquence des mots*. Cette caractéristique spécifique permet d'obtenir un sous-ensemble de mots discriminants [1] pour une classe considérée (un *thème* dans cette étude). Au final, cet ensemble de mots discriminants doit permettre une représentation vectorielle des thèmes d'une conversation dans un espace sémantique.

Bien que la fréquence des mots soit une caractéristique performante dans le contexte de textes écrits ou transcrits manuellement, son application aux transcriptions automatiques semble être beaucoup plus difficile, et ce, à cause des erreurs de transcription. En effet, ces erreurs peuvent conduire à une représentation incorrecte des mots discriminants. Pour cette raison, la projection des mots transcrits automatiquement dans un espace de représentation plus abstrait pourrait améliorer la robustesse des applications face aux erreurs du système de RAP.

Nous proposons de comparer deux représentations non-supervisées des mots discriminants dans des transcriptions très imparfaites (*i.e.* taux d'erreur-mot très élevé) avec pour finalité d'identifier automatiquement les thèmes de conversations téléphoniques. La méthode classique à base de fréquences de mots (TF-IDF avec le critère de pureté Gini (Robertson, 2004)) est tout d'abord appliquée pour extraire les mots discriminants pour chaque thème à identifier. Nous proposons ensuite d'étudier une représentation des mots discriminants avec des espaces de thèmes au moyen de l'approche *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003). Chaque représentation est ensuite utilisée pour apprendre un classifieur de type *Support Vector Machine* (SVM) afin d'associer automatiquement un thème à une conversation. Enfin, nous proposons également dans cet article une discussion autour de l'impact sur les performances de classification des mots discriminants et non-discriminants choisis par les deux méthodes de représentation étudiées.

La partie suivante présente les travaux antérieurs réalisés dans ce domaine. La partie 3 introduit les méthodes de représentation des mots étudiées. Les résultats expérimentaux sont exposés dans la partie 4. Une discussion autour de la précision de la transcription des mots discriminants est proposée dans la partie 4.3, avant de conclure dans la partie 5.

2 Travaux antérieurs

Différents résumés de travaux récents ont été réalisés autour de l'analyse de conversations (Tur et De Mori, 2011), du traitement de la parole (Melamed et Gilbert, 2011), de l'identification (Hazen, 2011) et de la segmentation de thèmes (Purver, 2011). L'approche classique à base de fréquences de mots, *Term Frequency-Inverse Document Frequency* (TF-IDF) (Robertson, 2004), a

1. Le terme *discriminant* est associé à un mot s'il permet de choisir une classe par rapport à une autre.

été très largement utilisée afin d'extraire les mots discriminants contenus dans des textes. Des améliorations ont été constatées en lui associant le critère de pureté Gini (Dong *et al.*, 2011).

D'autres approches ont proposé de considérer le document comme une mixture de thèmes latents. Ces méthodes, telles que *Latent Semantic Analysis* (LSA) (Deerwester *et al.*, 1990; Bellegarda, 1997), *Probabilistic LSA* (PLSA) (Hofmann, 1999), ou *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003), permettent un niveau de représentation plus élevé des mots au moyen d'un espace de thèmes. Les documents sont alors considérés comme des sacs-de-mots (Salton, 1989) ; les performances de ces méthodes ont pu être démontrées sur de nombreuses tâches.

LDA est un modèle génératif qui considère qu'un thème est associé à chaque occurrence d'un mot composant le document, plutôt que d'associer un thème au document complet. Ainsi, un document peut changer de thèmes d'un mot à un autre. Cependant, les occurrences des mots sont connectées par une variable latente qui contrôle le respect global de la distribution des thèmes dans le document. Ces thèmes latents sont caractérisés par une distribution des probabilités des mots qui leur sont associées. Les modèles PLSA et LDA obtiennent généralement de meilleurs résultats que le modèle LSA sur les tâches de recherche d'information (Hofmann, 2001). De plus, LDA fournit une estimation directe de la pertinence d'un thème sachant un ensemble de mots.

Les machines à vecteurs de support, *Support Vector Machines* (SVM), sont un ensemble de techniques d'apprentissage supervisé. Sachant un échantillon, les SVM déterminent un plan séparateur entre les parties de l'échantillon appelé *vecteur de support*. Ensuite, un hyperplan séparateur maximisant la *marge* entre les vecteurs de support et l'hyperplan séparateur (Vapnik, 1963) est calculé. Les SVM ont été utilisés pour la première fois par (Boser *et al.*, 1992), pour des tâches de régression (Müller *et al.*, 1997) et de classification (Joachims, 1999). La popularité des SVM est due aux bons résultats atteints dans ces deux tâches et au faible nombre de paramètres nécessitant un ajustement.

Une approche à base de LDA, combinée avec une classification SVM, a été récemment étudiée dans de nombreux domaines, tels que la classification de textes (Zrigui *et al.*, 2012), la stylométrie (Arun *et al.*, 2009), la recherche d'information (Kim *et al.*, 2009), la détection d'évènements sociaux (Morchid *et al.*, 2013), ou la détection d'images (Tang *et al.*, 2009). À notre connaissance, une approche LDA-SVM n'a jamais été appliquée à la classification de thèmes au moyen de transcriptions automatiques très imparfaites, mais a été utilisée dans le contexte d'extraction de mots clés dans des transcriptions automatiques (Sheeba et Vivekanandan, 2012). La méthode TF-IDF couplée avec une classification SVM, qui constitue notre système de base, a été très largement étudiée dans le contexte de la classification de textes, comme par exemple dans (Lan *et al.*, 2005; Georgescul *et al.*, 2006).

3 Système d'identification de thèmes

Cette partie présente le système de classification de thèmes proposé utilisant les mots discriminants extraits à partir de transcriptions très imparfaites. Le système est composé de deux parties principales. La première crée une représentation vectorielle des mots au moyen de deux approches non-supervisées : un vecteur de fréquences de mots Okapi/BM25 (Robertson, 2004) avec la méthode TF-IDF-Gini (Dong *et al.*, 2011), et une représentation par espace de thèmes avec l'approche LDA (Blei *et al.*, 2003). La seconde partie utilise les vecteurs extraits afin d'apprendre des classifieurs SVM. La figure 1 présente l'architecture globale du système de classification proposé utilisant des transcriptions manuelles (MAN) et automatiques (RAP).

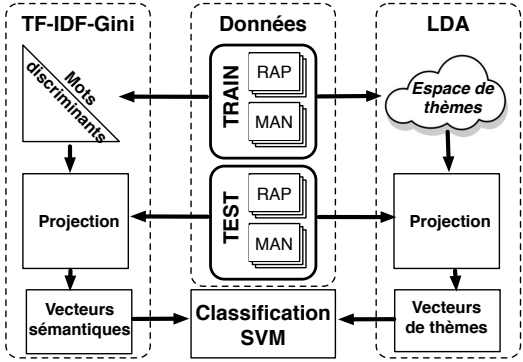


FIGURE 1 – Architecture globale du système de classification.

3.1 Représentation des dialogues

3.1.1 Représentation par fréquence des mots

Considérons un corpus D de dialogues d ayant un vocabulaire $\mathbf{V} = \{w_1, \ldots, w_N\}$ de taille N où d est vu comme un sac-de-mots (Salton, 1989). Un mot de $\mathbf{V}$ est choisi en fonction de son importance δ_t dans le thème t en calculant sa fréquence (TF), sa fréquence inverse (IDF) (Robertson, 2004), et le critère de pureté Gini (Dong *et al.*, 2011) commun pour tous les thèmes. Cet ensemble de scores δ compose le modèle de fréquence f :

$$\delta_t^w = tf_t(w) \times idf_t(w) \times gini_t(w) \tag{1}$$

Ensuite, les mots ayant les scores les plus élevés Δ pour tous les thèmes $\mathbf{T}$ sont extraits et constituent le sous-ensemble de mots discriminants $\mathbf{V}_\Delta$ (chaque thème $t \in \mathbf{T}$ possède son propre score δ_t) et sa propre fréquence γ dans le modèle f :

$$\gamma_f^t = \frac{\#d \in t}{\#d \in D} \tag{2}$$

Notons qu'un même mot w peut être présent dans différents thèmes, mais avec des scores différents (TF-IDF-Gini) en fonction de sa pertinence dans le thème :

$$\begin{aligned}
&\Delta(w) = P(w|f) = \textstyle\int_t P(w|t)P(t|f)\,\mathrm{dt} \\
&= \sum_{t \in \mathbf{T}} P(w|t)P(t|f) \\
&= \sum_{t \in \mathbf{T}} \delta_t^w \times \gamma_f^t \\
&= \overrightarrow{\delta^w}\overrightarrow{\gamma^f}
\end{aligned}$$

Pour chaque dialogue $d \in D$, un vecteur de caractéristiques sémantiques V_d^s est déterminé. La n^{ieme} $(1 \leq n \leq |\mathbf{V}_\Delta|)$ caractéristique $V_d^s[n]$ contient le nombre d'occurrences du mot w_n $(|w_n|)$ dans d, et le score Δ de w_n (voir équation 3.1.1) dans l'ensemble des mots discriminants $\mathbf{V}_\Delta$:

$$V_d^s[n] = |w_n| \times \Delta(w_n) \tag{3}$$

3.1.2 Représentation par espace de thèmes

La représentation par espace de thèmes est réalisée au moyen de l'approche LDA (voir partie 2). Un espace thématique m de n thèmes est alors obtenu avec, pour chaque thème z, la probabilité de chaque mot w de $\mathbf{V}$ sachant z ($P(w|z) = V_z^w$), et pour le modèle complet m, la probabilité de chaque thème z sachant le modèle m ($P(z|m) = V_m^z$).

Pour chaque dialogue d du corpus D, un premier paramètre θ est défini en fonction d'une loi de Dirichlet de paramètre α. Un second paramètre ϕ est défini en fonction de la même loi de Dirichlet de paramètre β. Ensuite, pour générer tous les mots w du document d, un thème latent z est défini à partir d'une distribution multinomiale sur θ. Sachant ce thème z, la distribution des mots est une multinomiale de paramètre ϕ. Le paramètre θ est défini pour tous les documents à partir du même paramètre initial α. Cela permet d'obtenir un paramètre reliant tous les documents ensemble (Blei *et al.*, 2003).

Projection des conversations/espace de thèmes : L'algorithme de Gibbs sampling (Griffiths et Steyvers, 2002) est utilisé pour inférer un dialogue d avec les n thèmes de l'espace thématique m. Cet algorithme s'appuie sur la méthode *Markov Chain Monte Carlo* (MCMC). Ainsi, le Gibbs sampling permet d'obtenir des échantillons des paramètres de distribution θ sachant un mot w d'un document de test et un thème donné z. Un vecteur de caractéristiques V_z^d de la représentation du thème de d est alors obtenu. La k^{ieme} caractéristique (où $1 \leq k \leq n$) est la probabilité du thème z_k sachant le dialogue d :

$$V_z^d[k] = P(z_k|d) \tag{4}$$

3.2 Classification à base de SVM

Durant cette étape, les classifieurs sont entrainés à partir de la représentation vectorielle des mots afin d'attribuer automatiquement le thème le plus pertinent à chaque conversation. Ce processus de classification nécessite un classifieur multi-classes. Pour ce problème multi-thème, T représente le nombre de thèmes et $t_i, i = 1, \ldots, T$ représente les thèmes. Un classifieur binaire (méthode *un-contre-un*) est utilisé avec un noyau linéaire (Yuan *et al.*, 2012) pour chaque paire de thèmes distinct : tous les classifieurs binaires $T(T-1)/2$ sont ensuite construits ensemble. Le classifieur binaire $C_{i,j}$ est entrainé, t_i étant une classe positive et t_j une classe négative ($i \neq j$). Une fois le vote de tous les classifieurs achevé, du thème ayant le plus grand nombre de votes est attribué au dialogue d.

4 Expériences

4.1 Protocole expérimental

Les expériences sur l'identification du thème d'une conversation sont menées sur le corpus du projet DECODA (Bechet *et al.*, 2012). Ce corpus est composé de 1 076 conversations téléphoniques découpées en un corpus d'apprentissage (740 dialogues) et un corpus de test (327 dialogues). Ces dialogues ont été manuellement annotés selon 8 thèmes : *problème d'itinéraire*, *objet perdu et trouvé*, *horaire*, *carte de transport*, *état du traffic*, *prix du ticket*, *infraction* et *offre spéciale*.

L'ensemble d'apprentissage est utilisé pour définir un sous-ensemble de mots discriminants (voir partie 3.1). Ce sous-ensemble permet d'élaborer un espace sémantique pour chaque conversation du corpus de test au moyen de la méthode basique TF-IDF-Gini. Dans ces expériences, le nombre de mots discriminants a été varié de 800 à 7 920 mots (nombre total de mots de l'apprentissage). Le corpus de test contient 3 806 mots (70,8 % étant contenu dans le corpus d'apprentissage).

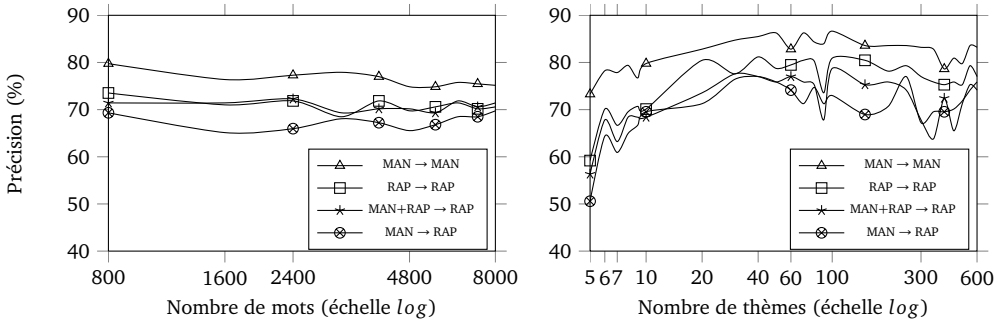
De la même manière, un vecteur de thèmes est calculé en projetant chaque dialogue du corpus de test dans chaque espace de thèmes. Un ensemble de 25 espaces de thèmes associé à un nombre de thèmes différents ($\{5, \ldots, 600\}$) est élaboré au moyen d'un modèle LDA construit à partir du corpus d'apprentissage avec l'outil Mallet (McCallum, 2002).

Ensuite, pour ces deux configurations, un classifieur SVM est entrainé au moyen de la librairie LIBSVM (Chang et Lin, 2011). Les paramètres sont optimisés par validation croisée sur le corpus d'apprentissage.

Le système de RAP Speeral (Linarès *et al.*, 2007) a été utilisé. Les paramètres des modèles acoustiques (230 000 gaussiennes / modélisation triphone) sont estimés au moyen d'une adaptation par maximum *a-posteriori* (MAP) à partir de 150 heures de parole (conditions téléphoniques). Un modèle de language tri-gramme a été obtenu en adaptant un modèle de language basique avec les transcriptions du corpus d'apprentissage de DECODA. Le vocabulaire contient 5 782 mots. Le taux d'erreur-mot (TEM) initial atteint 45,8 % sur le corpus d'apprentissage et 58,0 % sur le corpus de test. Ces TEM élevés sont principalement dus à la présence de nombreuses disfluences, et à des conditions acoustiques bruitées, quand, par exemple, les utilisateurs appellent à partir de gares avec un téléphone portable. Une liste de rejet de 126 mots [2] a été utilisée, ce qui donne au final un TEM de 33,8 % (apprentissage), et 49,5 % sur l'ensemble (test).

Les expériences sont menées au moyen des deux méthodes non-supervisées proposées (TF-IDF-Gini / LDA) sur les transcriptions manuelles (MAN) et les transcriptions automatiques (RAP) seules. Nous proposons également d'étudier la combinaison des transcriptions manuelles et automatiques (MAN+RAP) afin de voir si les erreurs de RAP peuvent être compensées par les mots corrects (*i.e.* issus de la référence).

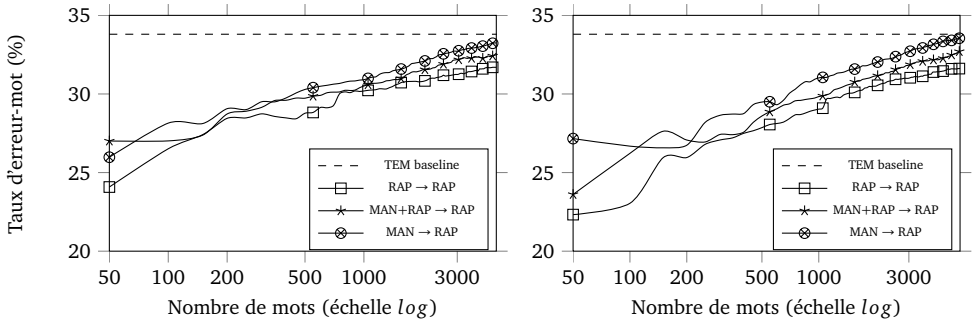
FIGURE 2 – Performance de la classification de thèmes en faisant varier le nombre de mots discriminants (a) et le nombre d'espaces de thèmes (b).



(a) **TF-IDF-Gini** (b) **LDA**

2. http ://code.google.com/p/stop-words/

FIGURE 3 – Taux d'erreur-mot des n mots discriminants extraits avec TF-IDF-Gini (a) et LDA (b).



(a) **TF-IDF-Gini** (b) **LDA**

4.2 Performance de l'identification de thèmes

La figure 2 présente les précisions de la classification de thèmes obtenues par les approches TF-IDF-Gini et LDA sur le corpus de test pour les différentes configurations étudiées (MAN / RAP) en faisant varier les conditions d'extraction des mots (nombre de mots discriminants et nombre de thèmes). Nous pouvons noter que la méthode LDA surpasse tous les résultats obtenus par l'approche TF-IDF-Gini (voir tableau 1).

TABLE 1 – Précision de la classification de thèmes.

Données		*Meilleure Précision (%)*			
Apprentissage	**Test**	**#mots**	**TF-IDF-Gini**	**#thèmes**	**LDA**
MAN	MAN	*800*	79,7	*100*	**86,6**
MAN	RAP	*8 000*	69,7	*40*	**77,0**
RAP	RAP	*800*	73,5	*60*	**81,4**
RAP+MAN	RAP	*2 400*	72,2	*100*	**78,7**

Comme attendu, la configuration *MAN → MAN* donne les meilleurs résultats de classification avec un gain de 6,9 points avec la méthode LDA. Si nous comparons les configurations du corpus d'apprentissage, nous notons également que les meilleurs performances sur le corpus de test RAP sont obtenues avec le corpus d'apprentissage RAP. Un gain de 10,9 points est constaté avec la méthode LDA en comparaison de l'approche TF-IDF-Gini sur les transcriptions automatiques. Il semble évident qu'utiliser des configurations d'apprentissage et de test comparables permet d'atteindre les meilleurs résultats de classification, et ce, peu importe que l'on traite des transcriptions manuelles ou automatiques.

Nous pouvons enfin noter que les performances obtenues avec l'approche LDA ont tendance à fluctuer lorsque le nombre de thèmes varie. Ceci peut s'expliquer par les taux d'erreur-mot (TEM) du corpus traité : en effet, les mots choisis comme *discriminants* dans des conditions particulières de l'espace de thèmes peuvent être mal transcrits dans des proportions élevées. Nous pouvons étayer cette remarque en analysant les résultats obtenus avec 90 thèmes (voir figure 2). Une baisse importante des performances est observée pour la condition d'apprentissage RAP (RAP → RAP et MAN → RAP) alors qu'une faible baisse est constatée lorsque les transcriptions de références sont ajoutées dans le processus d'apprentissage (RAP+MAN → RAP et MAN → MAN).

4.3 Précision de la transcription des mots discriminants

Alors que l'approche TF-IDF-Gini est clairement meilleure sur les transcriptions manuelles (tableau 1), les performances sont quasiment identiques sur les transcriptions manuelles et automatiques avec la méthode LDA (avec une précision respective de 86,6 % et 81,4 %). Nous pensons que l'approche LDA doit mieux gérer les erreurs contenues dans les transcriptions automatiques en choisissant les mots discriminants les mieux transcrits (*i.e.* ayant le plus faible TEM). La figure 3 permet de comparer les taux d'erreur-mot (TEM) des n mots discriminants extraits au moyen des méthodes TF-IDF-Gini et LDA sur toutes les configurations (MAN / RAP). Le score $s(w)$ est utilisé pour trouver les mots les plus pertinents de l'approche LDA selon la formule :

$$\begin{aligned} \mathrm{s(w)} &= \mathrm{P(w|m)} = \int_z P(w|z)P(z|m)\,\mathrm{dz} \\ &= \sum_{z \in m} P(w|z)P(z|m) \\ &= \sum_{z \in m} V_z^w \times V_m^z \\ &= \overrightarrow{V^w} \overrightarrow{V^m} \end{aligned}$$

où $\overrightarrow{V^w}$ est la représentation vectorielle d'un mot w dans tous les thèmes z de l'espace de thèmes m, $\overrightarrow{V^m}$ est la représentation vectorielle de tous les thèmes z dans m et $\cdot\cdot$ est le produit scalaire. Le TEM est ensuite calculé sur les n mots les plus discriminants (poids de 1 pour chaque mot).

Si nous comparons tout d'abord les différentes configurations (MAN / RAP), nous pouvons noter que plus la précision de la classification est élevée (tableau 1), moins le TEM l'est. Ce constat est observé pour les deux méthodes. De plus, nous pouvons voir que le TEM obtenu avec l'approche LDA est légèrement plus bas que celui obtenu avec la méthode TF-IDF-Gini, peu importe la configuration considérée. Cela signifie que les mots discriminants extraits avec l'approche LDA sont mieux transcrits en comparaison de ceux obtenus avec la méthode TF-IDF-Gini, ce qui peut expliquer les meilleures performances de classification obtenues avec l'approche LDA.

5 Conclusions

Dans cet article, nous avons présenté une architecture permettant d'identifier le thème d'une conversation en utilisant des transcriptions très imparfaites. Deux méthodes non-supervisées de représentation des conversations (TF-IDF-Gini et LDA) ont été comparées. Nous avons montré que la représentation par espace de thèmes obtenue avec la méthode LDA surpasse les résultats de classification obtenus avec la représentation classique TF-IDF-Gini. La précision de la classification atteint 86,6 % sur les transcriptions manuelles et 81,4 % sur les transcriptions automatiques, avec un gain respectif de 6,9 et 10,9 points.

Nous avons également discuté du lien possible entre performance de classification et précision de la transcription. L'analyse proposée a montré que les meilleurs résultats de classification sont obtenus avec des configurations extrayant les mots discriminants ayant les taux d'erreur-mot les plus faibles. Ces résultats prometteurs conduiront à une analyse qualitative plus détaillée dans des travaux futurs. En effet, cette étude préliminaire pourrait être fortement étendue avec de nouvelles analyses, en prenant par exemple en compte le poids des mots discriminants dans l'évaluation de la précision des transcriptions. Une perspective générale serait de proposer une solution pour estimer les performances de classification selon la qualité des transcriptions. Dans un contexte lié aux métriques d'évaluation, il serait intéressant de trouver une façon d'estimer la précision des transcriptions automatiques sur des tâches spécifiques, le taux d'erreur-mot n'étant pas un bon indicateur de la qualité d'une transcription dans un cadre applicatif.

Références

Arun, R., Saradha, R., Suresh, V., Narasimha Murty, M. et Veni Madhavan, C. E. (2009). Stopwords and Stylometry : A Latent Dirichlet Allocation Approach. *In NIPS Workshop on Applications for Topic Models : Text and Beyond*, Canada.

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Beze, M., De Mori, R. et Arbillot, E. (2012). Decoda : a call-centre human-human spoken conversation corpus. LREC'12.

Bellegarda, J. (1997). A latent semantic analysis framework for large-span language modeling. *In Fifth European Conference on Speech Communication and Technology*.

Blei, D., Ng, A. et Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Boser, B., Guyon, I. et Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *In 5th annual workshop on Computational learning theory*, pages 144–152.

Chang, C.-C. et Lin, C.-J. (2011). Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Dong, T., Shang, W. et Zhu, H. (2011). An improved algorithm of bayesian text categorization. *Journal of Software*, 6(9):1837–1843.

Georgescul, M., Clark, A. et Armstrong, S. (2006). Word distributions for thematic segmentation in a support vector machine approach. *In Conference on Computational Natural Language Learning*.

Griffiths, T. et Steyvers, M. (2002). A probabilistic approach to semantic representation. *In 24th annual conference of the cognitive science society*, pages 381–386. Citeseer.

Hazen, T. (2011). Topic identification. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pages 319–356.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *In Proc. of Uncertainty in Artificial Intelligence*, page 21.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *In Machine learning-international workshop then conference*, pages 200–209.

Kim, S., Narayanan, S. et Sundaram, S. (2009). Acoustic topic model for audio information retrieval. *In Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 37–40.

Lan, M., Tan, C.-L., Low, H.-B. et Sung, S.-Y. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. *In International Conference on World Wide Web*, pages 1032–1033.

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.

McCallum, A. (2002). Mallet : A machine learning for language toolkit.

Melamed, I. et Gilbert, M. (2011). Speech analytics. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pages 397–416.

Morchid, M., Dufour, R. et Linarès, G. (2013). Event detection from image hosting services by slightly-supervised multi-span context models. *In CBMI'13*.

Müller, K., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J. et Vapnik, V. (1997). Predicting time series with support vector machines. *ICANN'97*, pages 999–1004.

Purver, M. (2011). Topic segmentation. *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pages 291–317.

Robertson, S. (2004). Understanding inverse document frequency : on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520.

Salton, G. (1989). Automatic text processing : the transformation. *Analysis and Retrieval of Information by Computer*.

Sheeba, J. I. et Vivekanandan, K. (2012). Article : Improved keyword and keyphrase extraction from meeting transcripts. *International Journal of Computer Applications*, 52(13):11–15.

Tang, S., Li, J., Zhang, Y., Xie, C., Li, M., Liu, Y., Hua, X., Zheng, Y.-T., Tang, J. et Chua, T.-S. (2009). Pornprobe : an lda-svm based pornography detection system. *In International Conference on Multimedia*.

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

Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.

Yuan, G.-X., Ho, C.-H. et Lin, C.-J. (2012). Recent advances of large-scale linear classification. 100(9):2584–2603.

Zrigui, M., Ayadi, R., Mars, M. et Maraoui, M. (2012). Arabic text classification framework based on latent dirichlet allocation. *CIT*, 20(2):125–140.