Sujet :	**Extraction de contenus thématiques dans des dialogues**
Stagiaire :	(étudiant de M2)
Co-encadrement :	Renato de Mori (Pr. émérite), Marc El-Bèze (Pr.).

L'objectif de ce stage est l'exploration de plusieurs méthodes permettant l'extraction de contenus sémantiques présents dans un dialogue et au-delà dans un ensemble de dialogues. Le problème peut être considéré comme difficile pour les raisons suivantes :

- le nombre de contenus à extraire est *a priori* inconnu ;
- il n'est pas toujours évident de rattacher de expressions de contenus à des classes conceptuelles, (leur nombre et leur taille restent à définir) ;
- même quand la taille du jeu de ces classes n'est pas très grande, il arrive que certaines d'entre elles se recouvrent. Pour un dialogue n'abordant éventuellement qu'un seul thème, un être humain peut hésiter entre l'affectation à l'une ou l'autre des classes.
- les langues naturelles sont par essence ambiguës et la complexité qui résulte de cette ambiguïté est accrue quand il s'agit de dialogues oraux ;
- les conversations ayant été enregistrées en conditions réelles, en environnement bruité, donnent lieu à beaucoup d'erreurs dans les transcriptions obtenues par le biais d'un système de reconnaissance de la parole.

Application :

Des tests seront faits sur le corpus fourni par la RATP dans le cadre du projet ANR DECODA. Ce corpus a déjà été l'objet d'expérimentations dans un contexte mono label (Maza *et al*., 2011) ou dans le cas multi labels (Bost *et al.*, 2013). Pour chaque méthode envisagée, on essaiera de minimiser la perte entre les performances obtenues sur les transcriptions manuelles et celles forcément moins bonnes obtenues sur les sorties d'un système de reconnaissance de la parole (en l'occurrence le système Speeral du LIA). Il est à noter que la qualité de la reconnaissance peut être fortement perturbée par de mauvaises conditions acoustiques, et que le système reconnaît en général mieux ce que dit le conseiller que ce que dit l'usager appelant.

Caractéristiques propres à l'application :

- le conseiller a pour consigne de répéter à sa façon ce que dit l'appelant pour s'assurer qu'il a bien compris le motif de l'appel.
- Il existe des liens logiques entre les thèmes et les contenus sémantiques qui leur sont associés. Par exemple, il est naturel qu'après s'être renseigné sur un itinéraire, un appelant pose des questions sur les tarifs et/ou les horaires.

Une attention particulière sera portée sur 2 points précis

1. l'étude de stratégies de décision permettant de faire des hypothèses sur l'identification de passages contenant des contenus sémantiques, (selon que les segments correspondants aient été affectés ou pas par des erreurs de reconnaissance). La recherche des *features* appropriés est un des points clés de ce stage de recherche.
2. si on est intéressé par le fait de faire un résumé dialogue par dialogue de leur contenu sémantique, l'optique qui consiste à aboutir à une vision statistique globale des contenus d'un ensemble de dialogues (par exemple sur une période donnée) ne doit pas être négligée pour autant. De ce fait, les contenus extraits doivent être suffisamment génériques pour pouvoir être agrégés dans un second temps.

Méthodes :

Au moins trois méthodes seront envisagées. En premier lieu, une méthode qualifiée de basique sera implémentée.

Une seconde méthode visera à faire l'hypothèse de la co présence de représentants de 2 classes conceptuelles. En conséquence, on pourra recourir à des approches de type information mutuelle, ou du critère *log likelihood* employé pour détecter les collocations.

Une troisième approche devra s'appuyer sur l'état de l'art dressé par (G Tur et R. De Mori, 2005).

Les trois méthodes envisagées ici pourront être comparées avec une quatrième méthode tentant de dépasser cet état de l'art et devront être combinées avec comme objectif principal, l'amélioration quantitative et qualitative des résultats.

Notes :

Dans le cadre du projet Diamans soumis à l'ANR, un des lots est centré en partie sur les problèmes qui seront abordés dans le cadre de ce stage. En cas d'acceptation du projet Diamans, un financement permettrait une prolongation en thèse.

Références Bibliographiques :

Bechet F., Maza B, Bigouroux N, Bazillon T, El-Beze M, De Mori R, Arbillot E, « *DECODA: a call-centre human-human spoken conversation corpus* », Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istambul.

Bost X, El-Beze M, De Mori R : *Multiple topic identification in telephone conversations*; InterSpeech 25-29 August 2013, Lyon (France).

de Carvalho A, Freitas A : *A tutorial on multi-label classification techniques*, volume Foundations of Computational Intelligence Vol. 5 of ***Studies in Computational Intelligence 205***, pages 177-195 Springer, September 2009.

Maza B., El-Bèze M, Linares G, de Mori R : *On the Use of Linguistic Features in an Automatic System for Speech Analytics of Telephone Conversations.* Interspeech 2011: 2049-2052

Tsoumakas G, Katakis I, *Multi-Label Classification: An Overview*, International Journal of Data Warehousing and Mining, 3(3):1-13, 2007.

Tur G,. De Mori R, Eds, "*Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*", J. Wiley, 2011.