Introduction d'informations sémantiques dans un système de reconnaissance de la parole

Stephane Level, Irina Illina, Dominique Fohr

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Introduction d'informations sémantiques dans un système de reconnaissance de la parole

Stéphane Level , Irina Illina, Dominique Fohr
Equipe MultiSpeech
Université de Lorraine, CNRS, Inria, F-54000 Nancy, France
{stephane.level,irina.illina,dominique.fohr}@loria.fr

Résumé

Malgré les avancés spectaculaires ces dernières années, les systèmes de Reconnaissance Automatique de Parole (RAP) commettent encore des erreurs, surtout dans des environnements bruités. Pour améliorer la RAP, nous proposons de se diriger vers une contextualisation d'un système RAP, car les informations sémantiques sont importantes pour la performance de la RAP. Les systèmes RAP actuels ne prennent en compte principalement que les informations lexicales et syntaxiques. Pour modéliser les informations sémantiques, nous proposons de détecter les mots de la phrase traitée qui pourraient avoir été mal reconnus et de proposer des mots correspondant mieux au contexte. Cette analyse sémantique permettra de réévaluer les *N* meilleures hypothèses de transcription (*N-best*). Nous utilisons les *embeddings* Word2Vec et BERT. Nous avons évalué notre méthodologie sur le corpus des conférences TED (TED-LIUM). Les résultats montrent une amélioration significative du taux d'erreur mots en utilisant la méthodologie proposée.

Abstract

Despite spectacular advances in recent years, the Automatic Speech Recognition (ASR) systems still make mistakes, especially in noisy environments. In order to reduce these errors, we suggest moving towards a contextualization of a ASR system, because semantic information is important for the performance of ASR. Current ASR systems mainly take into account only lexical and syntactic information. To model the semantic information, we propose to detect the words of the recognised sentence, which could have been badly recognized and to propose words corresponding better to the context. This semantic analysis will allow to re-evaluate the N-best hypotheses of recognition. We use Word2Vec embedding and Google's BERT model. We evaluated our methodology on the corpus of TED conferences (TED-LIUM). The results show a significant improvement of the word error rate using the proposed methodology.

Mots-clés : reconnaissance automatique de la parole, contexte sémantique ,*embeddings*, Word2Vec, BERT.

Keywords: automatic speech recognition, semantic context, embeddings, Word2Vec, BERT.

1 Introduction

Grace aux réseaux de neurones profonds, les systèmes de reconnaissance automatique de la parole commencent à être utilisables dans les conditions réelles de notre vie de tous les jours. D'ailleurs des nombreux industriels proposent déjà des systèmes vocaux pour nos maisons, nos voitures et nos smartphones.

Malgré des efforts constants et quelques avancées spectaculaires, la capacité d'une machine à reconnaître la parole est encore loin d'égaler celle de l'être humain. Les systèmes RAP actuels

Volume 1 : Journées d'Études sur la Parole, pages 362–369.
HAL : hal-02798559.

voient leurs performances diminuer de manière significative lorsque les conditions dans lesquelles ils ont été entraînés et celles dans lesquelles ils sont utilisés diffèrent. Les causes de variabilité existantes entre ces conditions peuvent être liées à l'environnement acoustique et/ou à l'acquisition du signal sonore. Le matériel de capture du son, le changement de microphone, l'environnement acoustique ajoutent au signal de la parole des composantes perturbatrices. L'approche classique comporte deux étapes : débruiter (rehausser) le signal puis le transmettre au système de RAP pour le décodage. Cependant, la performance d'un système de RAP sur un mot donné dépend toujours de la distorsion au moment précis où ce mot a été prononcé.

Pour résoudre ce problème, nous proposons de se diriger vers une **contextualisation** du système RAP. En effet, les informations lexicales, sémantiques et temporelles sont importantes pour qu'un système RAP soit performant. En revanche, les systèmes RAP actuels ne prennent en compte principalement que les informations lexicales et syntaxique (modèles de langage n-gramme locaux). Pour modéliser les informations sémantiques, plusieurs méthodes fondées sur des statistiques de cooccurrences, sur l'information mutuelle, sur un modèle vectoriel et sur les réseaux de neurones peuvent être utilisées (Sheihk, 2016).

Les **espaces sémantiques et thématiques** sont des espaces vectoriels utilisés pour la représentation numérique des mots, des phrases ou des documents textuels. Presque tous les modèles s'appuient sur l'hypothèse de la sémantique statistique qui stipule que: des schémas statistiques d'apparition des mots (contexte d'un mot) peuvent être utilisés pour décrire la sémantique sous-jacente (Turney & Pantel, 2010). La méthode la plus utilisée pour apprendre ces représentations est de prédire un mot en utilisant le contexte dans lequel ce mot apparaît : *embedding* (Mikolov *et al*., 2013 ; Pennington *et al*., 2014), et cela peut être réalisé avec des réseaux neuronaux. Ces représentations se sont avérées efficaces pour une série de tâches de traitement du langage naturel (Baroni *et al*., 2014). Elles sont devenues très populaires en raison de leur capacité à traiter de grandes quantités de données textuelles non structurées avec un faible coût de calcul. L'efficacité et les propriétés sémantiques de ces représentations nous motivent à explorer ces représentations sémantiques pour notre tâche de reconnaissance dans des conditions bruitées. Nous espérons que dans les parties très bruitées, le modèle de langage et le modèle sémantique peuvent permettre de lever les ambiguïtés acoustiques afin de trouver les mots prononcés par le locuteur.

Dans cet article nous proposons de compléter l'étape de RAP **par l'ajout d'informations sémantiques** afin de détecter les mots de la phrase traitée qui pourraient avoir été mal reconnus et de proposer des mots de prononciation similaire correspondant mieux au contexte. Cette analyse sémantique permet **de réévaluer** les *N* meilleures hypothèses de transcription (*N-best*) et peut être vue comme une forme d'adaptation dynamique dans le cadre spécifique des données bruitées. Les informations sémantiques sont introduites en utilisant des représentations prédictives à l'aide de vecteurs continus (*embeddings*). Toutes nos modélisations s'appuient sur les technologies performantes de DNN. Par rapport aux travaux précédents utilisant la réévaluation de la liste N-best (Song *et al.*, 2019 ; Shin *et al.*, 2019 ; Ogawa *et al.* 2018), nous nous appuyons uniquement sur des informations sémantiques. De plus, la spécificité de notre approche est l'utilisation de la partie contextuelle et des zones de possibilité de la liste des N-hypothèses: la partie contextuelle représente l'information sémantique du contexte thématique du document à reconnaître et la zone de possibilité correspond à la zone où nous voulons trouver les mots à corriger. Cela nous permet de donner moins d'importance aux mots de la zone de possibilité qui ne correspondent pas au contexte du document, et de donner un faible score sémantique à l'hypothèse correspondante.

2 Méthodologie proposée

Une façon efficace de prendre en compte les informations sémantiques est de réévaluer les meilleures hypothèses du système de reconnaissance. Le système de reconnaissance nous fournit

pour chaque mot de la phrase hypothèse un score acoustique $p_{ac}(w)$et un score linguistique $p_{lm}(w)$. La meilleure phrase est celle qui maximise la probabilité de la séquence de mots :

$$\widehat{W} = \underset{h_i \in H}{argmax} \prod_{w \in h_i} p_{ac}(w)^{\alpha} * p_{lm}(w)^{\beta} \quad (1)$$

$\widehat{W}$ est la phrase reconnue (le résultat final) ; H est l'ensemble des *N*-meilleures hypothèses de phrases ; h_i est la *i*-ème hypothèse de phrase ; *w* est un mot de l'hypothèse. α et β représentent les poids du modèle acoustique et du modèle de langage. Ils sont indispensables car les scores acoustiques et les scores linguistiques ne sont pas toujours normalisés (ce sont souvent des vraisemblances et non des probabilités). Ces poids sont ajustés sur un corpus de développement.

Nous souhaitons **ajouter de l'information sémantique** pour guider le processus de reconnaissance. L'approche la plus naturelle pour intégrer cette information consiste à modifier le calcul de la probabilité de la séquence de mots de la façon suivante :

$$\widehat{W} = \underset{h_i \in H}{argmax} \prod_{w \in h_i} p_{ac}(w)^{\alpha} * p_{lm}(w)^{\beta} * p_{sem}(w)^{\gamma} \quad (2)$$

Nous avons ajouté la probabilité sémantique de chaque mot : $p_{sem}(w)$. Pour avoir un bon équilibre entre les différents modèles, nous introduisons un troisième poids γ pour pondérer l'information sémantique. Il sera également ajusté sur un corpus de développement.

2.1 Définition de contexte et des zones des possibilités

Pour estimer cette probabilité sémantique, nous proposons d'introduire les notions de ***contexte et* des *zone des possibilités. Un contexte*** est constituée des mots qui sont communs à toutes les *N*-meilleures hypothèses générées par le SRAP. Ce contexte nous permet d'extraire des informations sémantiques sur le contexte thématique du document ou de la partie courante du document à reconnaître. Pour obtenir un contexte sémantique plus significatif, il peut être intéressant d'ajouter à ce contexte les mots des phrases précédemment reconnues. ***Une zone des possibilités*** est une zone située entre les parties contexte. C'est dans cette zone que nous souhaitons retrouver les mots pour corriger la phrase reconnue. À partir des *N*-meilleures hypothèses d'une phrase, nous allons extraire un contexte et une ou plusieurs zones des possibilités. Chaque zone peut être constituée de plusieurs mots. La Figure 1 illustre ces notions sur un exemple. Ici les *3*-meilleurs hypothèses sont les suivantes :

H1: le chat mange la souris grise
H2: le chat ange la souris grise
H3: le chat mange la sous rit grise

Dans ce cas le contexte est composée des mots *le, chat, la* et *grise*. Ce sont les mots qui sont communs aux trois hypothèses. Entre ces mots, nous définissons deux zones des possibilités : la première est constituée de deux possibilités *mange et ange*. La deuxième est aussi constituée de deux possibilités *souris* et *sous rit*. Nous supposons que les zones de possibilités correspondent aux zones où le système de reconnaissance hésite entre différentes solutions.

Pour obtenir le contexte, nous utilisons un algorithme de programmation dynamique qui va permettre d'apparier les hypothèses deux à deux afin de déterminer les mots communs à toutes les hypothèses.

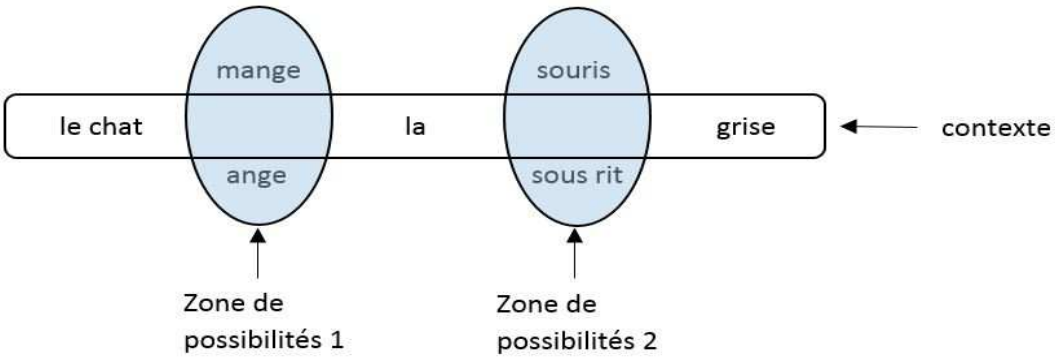

Figure 1. Illustration du contexte et des zones de possibilités pour un exemple.

2.2 Représentation sémantique du contexte et des zones des possibilités

Pour prendre en compte la sémantique du document à reconnaître, nous proposons de représenter chaque mot des N-meilleures hypothèses par un vecteur d'*embedding*. Dans notre approche, nous avons utilisé *word2vec (*Mikolov *et al*., 2013) et *BERT* (Devlin, 2018). ll est important de noter, que dans les représentations word2vec, l'*embedding* d'un mot est statique, c'est-à-dire un mot donné a un seul *embedding* quel que soit la phrase dans laquelle il apparaît. Dans le cas de représentations de mots à l'aide de BERT, l'*embedding* d'un mot dépend des autres mots de la phrase dans laquelle il apparaît et donc un mot peut avoir plusieurs *embeddings* en fonction du contexte de la phrase. Pour prendre en compte cette aptitude de BERT, lors de génération d'*embeddings* avec BERT pour le **contexte**, nous remplaçons chaque zone de possibilités par un masque spécial [*mask*] (comme dans le processus d'apprentisssage de BERT) et nous gardons inchangé le contexte. À partir de cette phrase avec les masques, BERT génère un *embedding* pour le contexte. Pour obtenir l'*embedding* pour une **zone de possibilités,** nous utilisons l'hypothèse de reconnaissance correspondante. Dans le cas où il y plusieurs zones de possibilités, toutes les zones de possibilités sont remplacées avec [*mask*] sauf la zone pour laquelle nous calculons l'*embedding*.

2.3 Calcul de la probabilité sémantique

A partir des représentations sémantiques du contexte et des zones de possibilités, nous pouvons calculer **une probabilité sémantique d'une hypothèse *h***. Cette probabilité sera utilisée dans la formule (2).

Pour prendre en compte la sémantique du document, nous représentons chaque mot des N-meilleures hypothèses par un vecteur d'*embedding,* comme décrit précédemment. Nous calculons un embedding moyen E_{cont} pour la partie contexte qui est égale à la moyenne des vecteurs d'*embedding* de tous les mots de la partie contexte. De la même manière, nous calculons un *embedding* moyen E_{pos} *(i, a)* pour la *i*-ème zone de possibilité de l'alternative a_h de l'hypothèse *h* comme la moyenne des vecteurs d'*embedding* de tous les mots dans cette alternative de la zone de possibilité. Une alternative correspond à un choix dans la zone de possibilité. Nous utilisons la similitude angulaire pour estimer un score sémantique entre chaque zone de possibilité et la partie contextuelle:

$$S_{sem}\left(E_{cont},, E_{pos}\left(i, a_h\right)\right) = 1 - \frac{\cos^{-1}\left(\cos\left(E_{cont}, E_{pos(i,\, a_h)}\right)\right)}{\pi} \qquad (3)$$

A partir des représentations sémantiques de la partie contexte et des zones de possibilité, nous calculons une probabilité sémantique $P_{sem}(h)$ d'une hypothèse h :

$$P_{sem}(h) = \prod_{i=1}^{N_p} S_{sem}\left(E_{cont}, E_{pos}(i, a_h)\right) \qquad (4)$$

où N_p est le nombre de zones de possibilité. Nous supposons que l'équation (2) peut être approximée comme suit:

$$\widehat{H} = \underset{h \in H}{argmax} \; P_{ac}(h)^{\alpha} P_{lm}(h)^{\beta} P_{sem}(h)^{\gamma} \qquad (5)$$

L'équation (5) est utilisée pour réévaluer la liste des N meilleures hypothèses. Pour chaque hypothèse, nous calculons le score sémantique et l'associons aux scores acoustiques et linguistiques selon (5). L'hypothèse qui obtient du meilleur score est considérée comme la phrase reconnue.

3 Expérimentations

3.1 Corpus utilisé

Nous avons utilisé le corpus TED-LIUM, la distribution standard (Hernandez *et al.,* 2018). Ce corpus contient les enregistrements des conférences TED. Le corpus est bien adapté à notre étude car chaque conférence est centrée sur un sujet particulier. L'ajout d'information sémantique avec un large contexte devrait permettre d'améliorer les performances de notre système de reconnaissance.

Le découpage du corpus TED en corpus d'apprentissage, développement et de test est proposé dans la distribution TED-LIUM et correspond à 452 heures d'apprentissage, 8 conférences pour le développement et 11 conférences pour le test. La Table 1 donne quelques statistiques sur le corpus de développement et de test car ce sont ces deux corpus qui nous intéressent pour l'introduction de l'information sémantique.

	Nbr. de documents audio	*Nbr. de phrases*	*Nbr. de mots*
Développement	8	500	17926
Test	11	1091	27021

Table 1 : Corpus de développement et de test du TED-LIUM.

3.2 Système de reconnaissance

Notre système de reconnaissance est fondé sur la boîte à outils de reconnaissance vocale *Kaldi* (Povey *et al*., 2011). Nous avons utilisé des modèles acoustiques triphones de type TDNN. L'apprentissage des modèles acoustiques TDNN a été réalisé en utilisant les 452 heures du corpus d'apprentissage de TED. Le lexique est composé de 150k mots et le modèle de langage contient 2 millions 4-grams, appris sur un corpus textuel de 250 millions de mots.

De façon classique, nous avons utilisé le corpus de développement pour choisir la meilleure configuration et ajuster les paramètres. Le corpus de test permet d'évaluer la méthode proposée avec les meilleurs paramètres obtenus sur le corpus de développement.

Pour se rapprocher des conditions réelles d'utilisation, nous avons décidé d'ajouter du bruit aux corpus de développement et de test. Nous avons ajouté un bruit additif de 10 dB et de 5dB (bruit de

F16 de la base NOISEX-92 (Varga, 1993)). La performance de notre système sur TED-LIUM sans ajout de bruit est autour de 8 % de taux d'erreur mots.

Les *embeddings word2vec* générés sont de taille 300 et modélisent 700 000 mots. Nous avons utilisé le modèle pré-entraîné *BERT Base* fourni par Google. Il est composé de 12 couches de *transformer,* chacun composés de 12 têtes d'attention. La taille de l'espace est 768. La métrique que nous avons utilisée est le taux d'erreur mots (WER).

4 Résultats expérimentaux

Le système de reconnaissance de base donne le taux d'erreur de 15.9 % WER pour le corpus de développement bruitée au Rapport Signal Bruit (RSB) de 10dB. Pour connaître le taux d'erreur minimal que nous pouvons obtenir en utilisant les *N* meilleures hypothèses, nous avons évalué le taux d'erreur *oracle* : 9.8 %. Ce taux est obtenu en sélectionnant l'hypothèse de la liste de *N*-meilleures hypothèses qui minimise le WER pour chaque phrase. Pour le corpus de développement bruité à 5dB nous avons obtenu un WER de 32,2 % et taux d'erreur mots *oracle* de 25,2 %.

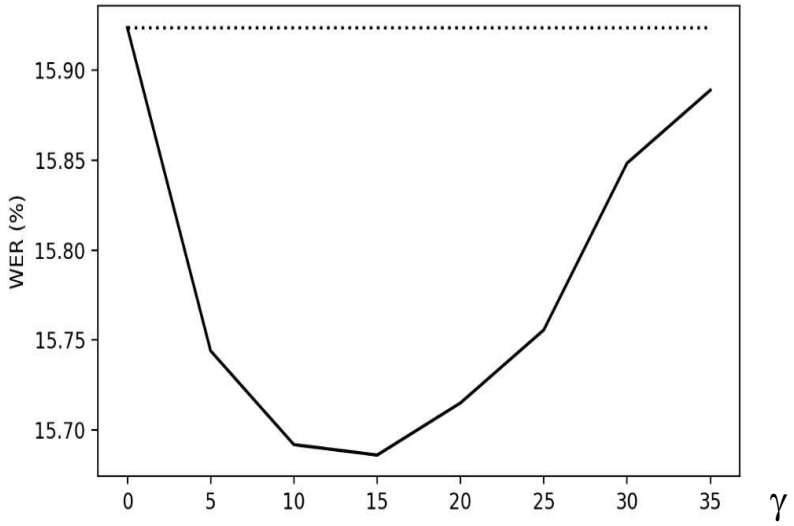

Figure 2. WER en fonction du coefficient sémantique γ. Méthode sémantique avec l'*embedding* Word2Vec. Corpus de développement TED-LIUM.

Nous avons évalué l'influence du paramètre γ (cf. formule 2) utilisé pour équilibrer les scores **acoustiques**, **langagiers** et **sémantiques**. La Figure 3 représente la courbe de ce paramètre pour l'*embedding* de Word2Vec en fonction de taux d'erreur obtenu. Nous observons que ce paramètre est important.

La Table 2 présente les résultats de reconnaissance sur le corpus TED-LIUM pour la partie développement et la partie test, ainsi que dans deux conditions de bruits : 10dB et 5dB. La première ligne de la table correspond au système sans le module sémantique, la dernière ligne à la performance maximale qu'on peut obtenir en recherchant dans N-meilleures phrase (oracle). Les lignes 2 et 3 correspondent aux approches proposées. Sur le corpus de test, nous obtenons une amélioration absolue de 0,4 % pour un RSB de 10dB (21,8 % versus 22,2 %) et de 0,8 % pour un RSB de 5dB (38,2 % versus 39 %) pour les parties test. Nous observons que les approches sémantiques proposées permettent de réduire le taux d'erreur mots. La meilleure performance est obtenue en utilisant l'*embedding* de Word2Vec. Dans les conditions plus bruitées (5dB) l'amélioration est un peu plus importante. Toutes les améliorations sont significatives par rapport au système de base.

Méthode	*RSB 10dB*		*RSB 5dB*	
	dev	*test*	*dev*	*test*
Système de base	15,9	22,2	32,2	39,0
Word2Vec *embedding*	15,7	**21,8**	**31,5**	**38,2**
BERT *embedding*	**15,5**	22,0	31,6	38,5
Oracle	9,8	14,0	25,2	29,8

Table 2 : Résultats de reconnaissance en terme de taux d'erreur de mots (WER %). Corpus de développement et de test TED-LIUM**,** RSB 10dB et 5 dB.

5 Conclusion et discussion

Dans cet article, nous voulions améliorer les performances d'un système de reconnaissance automatique de la parole en ajoutant des **informations sémantiques**. Nous proposons une méthodologie novatrice de la prise en compte de la sémantique à travers les représentations prédictives qui capturent les caractéristiques sémantiques des mots et de leur contexte. L'efficacité et les propriétés sémantiques de ces représentations récentes de type *embeddings* nous motivent à explorer ces représentations pour notre tâche de reconnaissance de la parole. Nous avons exploré les modèles **Word2Vec et BERT.** Les informations sémantiques sont prises en compte à travers le module de **réévaluation des N-meilleures hypothèses du système de reconnaissance.** Nous avons évalué notre méthodologie sur le corpus des conférences TED-LIUM. Les résultats montrent une amélioration significative du taux d'erreur mots en utilisant la méthodologie proposée.

Il existe de nombreuses extensions possibles de ce travail. Par exemple, il peut être possible d'améliorer les performances en explorant d'autres façons de calculer un *embedding* pour une zone. Il serait également intéressant d'étudier les différentes possibilités pour calculer le score d'une hypothèse.

Remerciements
Les auteurs remercient la DGA (Direction Générale de l'Armement), Thalès AVS et Dassault Aviation qui soutiennent le financement de cette étude et du programme scientifique «*Man-Machine Teaming*» dans lequel se déroule ce projet de recherche.

6 Références

BARONI M., DINU G., KRUSZEWSKI G. (2014) Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *In proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.*
CORONA R., THOMASON J., MOONEY R.J. (2017). Improving Black-box Speech Recognition using Semantic Parsing. *In proceedings of the The 8th International Joint Conference on Natural Language Processing.*
DEVLIN J., CHANG M.-W., LEE K., TOUTANOVA K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805.*

Fernandez H., Nguyen H., Ghannay S., Tomashenko N., and Estève Y. (2018) TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *In proceedings of SPECOM.*
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. Dean, J. (2013) Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems 26, 3111–3119.*
Ogawa A., Delcroix M., Karita S., Nakatani t (2018) Rescoring N-best speech recognition list based on one-on-one hypothesis comparaison using encoder-classifier model. *In: Proceedings of the ICASSP.*
Pennington J., Socher R., Manning C.D. (2014) Glove: Global vectors for word representation. *In the Proceedings of the 2014 conference on empirical methods in natural langage.*
Povey D., Ghoshal A., Boulianne G.i, Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K. (2011). The Kaldi Speech Recognition Toolkit. *In proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.*
Sheikh, I. (2016) Exploitation du contexte sémantique pour améliorer la reconnaissance des noms propres dans les documents audio diachroniques*, These de doctorat en Informatique, Université de Lorraine.*
Shin J., Lee Y., Jung K. (2019) Effective Sentence Scoring Method Using BERT for Speech Recognition. *In: Proceedings of Machine Learning Research*, pp.1081-1093.
Song Y., Jiangy D., Zhao X., Xuy Q., Wong R., Fany L., Yang Q. (2019) L2RS: a learning-to-rescore mechanism for automatic speech recognition. *arXiv:1910.11496v1.*
Turney P., Pantel P. (2010) From Frequency to Meaning: Vector Space Models of Semantics. *In Journal of Artificial Intelligence Research,* 37, pp.141-188.
Varga A., Steeneken H. (1993) Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, *Speech Communication,* Volume 12, Issue 3, pp. 247-251
Velikovich L., Williams I., Scheiner J., Aleksic P., Moreno P., Riley M. (2018) Semantic Lattice Processing in Contextual Automatic Speech Recognition for Google Assistant, *In the Proceedings of Intersppech.*