Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique

Irina Illina, Dominique Fohr, Georges Linarès

Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique

Irina Illina[1], Dominique Fohr[1], Georges Linarès[2]

(1) Equipe parole, LORIA-INRIA, 54602 Villers-les-Nancy, France
(2) LIA, Université d'Avignon, 84911 Avignon, France

RESUME

La reconnaissance de noms propres est une tâche difficile dans le domaine de la recherche d'information dans de grandes bases de données audio/vidéo. Les noms propres sont souvent indispensables pour comprendre l'information contenue dans un document. Notre travail se concentre sur l'augmentation du vocabulaire d'un système de transcription automatique de la parole. L'idée est de récupérer automatiquement des noms propres à partir de documents diachroniques. Nous avons proposé des méthodes qui augmentent de façon dynamique le vocabulaire du système de reconnaissance en utilisant des informations lexicales et temporelles. Nous faisons l'hypothèse que les mêmes noms propres apparaissent fréquemment dans des documents relatifs à la même période temporelle. Nous avons étudié une méthode fondée sur l'information mutuelle et nous avons proposé une nouvelle méthode utilisant la similarité cosinus. Dans cette nouvelle méthode, le contexte d'un nom propre est représenté par un modèle vectoriel (sac de mots). Nous avons également étudié différents paramètres de sélection de noms propres afin de limiter l'augmentation du vocabulaire et donc l'impact sur les performances de l'ASR. Les résultats de reconnaissance montrent une réduction significative du taux d'erreur de mots en utilisant un vocabulaire augmenté.

ABSTRACT

Proper names are usually keys to understand the information contained in a document. Our work focuses on increasing the vocabulary size of a speech transcription system by automatically retrieving proper names from contemporary diachronic text documents. We proposed methods that dynamically augment the automatic speech recognition system vocabulary, using lexical and temporal features. We assume that the same proper names frequently appear in documents relating to the same time period. We studied a method based on Mutual Information and we proposed a new method based on cosine similarity to retrieve new proper names. In this new method, proper name context is represented by vector space model (Bag of Words). We also studied different metrics for proper name selection in order to limit the vocabulary augmentation and therefore the impact on the ASR performances. Recognition results show a significant reduction of the word error rate using augmented vocabulary with retrieved proper names.

MOTS-CLES : reconnaissance de la parole, mots hors vocabulaire, noms propres, augmentation du vocabulaire.

KEYWORDS: speech recognition, out-of-vocabulary words, proper names, vocabulary augmentation.

1 Introduction

Dans notre travail, nous nous intéressons à la reconnaissance automatique de la parole (RAP) et plus particulièrement à la transcription de documents audio. Même en utilisant un très grand vocabulaire, les systèmes RAP sont confrontés au problème des *mots hors vocabulaire* (*Out Of Vocabulary*, OOV). Ces mots OOV sont des mots qui se trouvent dans le signal de parole, mais pas dans le vocabulaire du système RAP. Dans ce cas, le système RAP ne pourra pas les transcrire correctement et les remplacera par un ou plusieurs mots du vocabulaire, impactant négativement l'intelligibilité de la transcription.

La reconnaissance de *noms propres* (NP) est une tâche complexe, car les noms propres sont en constante évolution et aucun vocabulaire statique ne pourra contenir tous les noms propres existants : par exemple, les NP représentent environ 10 % des mots des articles de journaux en anglais ou en français et ils sont vitaux pour caractériser le contenu d'un texte (Friburger, 2002). Bechet et Yvon (Bechet, 2000) ont montré que 72% des mots OOV sont des NP dans le cas d'un vocabulaire de 265K mots.

Notre travail utilise la modélisation du contexte temporel afin de capturer l'information lexicale de manière à récupérer des noms propres OOV et augmenter la taille du vocabulaire du système de reconnaissance. Nous nous concentrons sur l'exploitation du contexte lexical grâce à des informations temporelles de documents *diachroniques* (documents qui évoluent dans le temps) (Allauzen, 2005). Notre hypothèse est que l'information temporelle est un élément important pour capturer des dépendances NP-contexte (Kobayashi, 1998). Notre approche a été inspirée par (Bigot, 2013) et (Oger, 2008) : nous utilisons également la notion du contexte pour les noms propres. Cependant, nos approches se concentrent sur l'exploitation de la temporalité des documents en utilisant des documents diachroniques. Nous supposons que les NP sont souvent liés à un événement qui se produit dans une période de temps spécifique dans des documents diachroniques. Nous émettons l'hypothèse que les NP évoluent dans le temps, et que pour une date donnée, les mêmes NP vont apparaître dans des documents qui appartiennent à la même période. Les contextes temporels ont été proposés auparavant par Federico et Bertoldi (Bertoldi, 2001) pour augmenter le vocabulaire, et par (Prada, 2010) pour la prédiction des OOV dans les sorties du système de reconnaissance. Contrairement à ces travaux, notre travail étend le vocabulaire du RAP en utilisant des périodes temporelles plus courtes et plus précises pour éviter l'augmentation excessive du vocabulaire. Nous recherchons un bon compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire. Dans le cas contraire, cela pourrait conduire à augmenter considérablement les ressources requises pour un système ASR.

Cet article est organisé de la façon suivante : la section suivante donne la méthodologie proposée pour la sélection des nouveaux NP à partir des documents diachroniques. La section 3 décrit les expériences et les résultats.

2 Méthodologie

Notre idée consiste à extraire des noms propres OOV automatiquement à partir des documents diachroniques, en utilisant le contexte lexical et temporel. Nos méthodes d'extraction des noms OOV sont fondées sur l'idée que les noms propres manquants se retrouvent probablement dans des documents contemporains, c'est-à-dire correspondant à la même période de temps que le document que nous voulons transcrire. Nous émettons

l'hypothèse que les NP évoluent dans le temps, et que pour une date donnée, les mêmes noms propres apparaitront dans les documents qui appartiennent à la même période. Par exemple, dans un document de mars 2011 contenant les NP *Japon* et *Fukushima*, il y a de fortes chances que les NP *TEPCO*, *Daiichi* et *Naoto Kan* apparaissent.

Nous proposons d'utiliser les documents du corpus diachronique qui sont contemporains de chaque document de test, et de construire un vocabulaire augmenté pour chaque document audio de test. En résumé, nous avons un document audio de test (à transcrire) qui contient des mots OOV, et nous avons un corpus de textes diachroniques, utilisé pour rechercher de nouveaux NP. Un vocabulaire augmenté est construit pour chaque document de test.

Nous supposons que, pour une date donnée, un nom propre du corpus de test co-occurrera avec d'autres NP des documents diachroniques correspondant à la même période de temps. Parmi ces NP, nous faisons l'hypothèse qu'un certain nombre seront présents dans le document de test, c'est-à-dire seront des NP OOV. L'idée est d'exploiter la relation entre les NP pour un enrichissement performant du vocabulaire.

Dans cet article, différentes stratégies de sélection de NP seront proposées pour construire ce vocabulaire augmenté :

- Méthode de référence : Sélection des documents diachronique en utilisant uniquement une période de temps correspondant au document de test.

- Méthode fondée sur l'information mutuelle : même stratégie que la méthode de référence, mais l'information mutuelle est utilisée pour mieux choisir les noms propres OOV.

- Méthode fondée sur la similarité cosinus : même stratégie que la méthode de référence mais les documents sont représentés par le modèle vectoriel (Chingal, 2011).

Dans une étude précédente (Nkairi, 2013), nous avons présenté les résultats des méthodes de référence et celle fondée sur l'information mutuelle. Ici, une nouvelle méthode, la méthode de similarité cosinus, est proposée et comparée à la méthode fondée sur l'information mutuelle.

2.1 Méthode de référence

Cette méthode consiste à extraire une liste de tous les noms propres OOV qui apparaissent dans les documents du corpus diachronique, qui correspondent à la même période que le document de test. Ensuite, notre vocabulaire est augmenté avec tous les NP extraits de ces documents. Nous considérons cette méthode comme méthode de référence. Le problème de cette approche est que si le corpus diachronique est de grande taille, nous risquons d'augmenter la taille du vocabulaire de façon démesurée.

2.2 Méthode fondée sur l'information mutuelle

Pour obtenir un meilleur compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire, nous allons filtrer les NP à l'aide de l'information mutuelle :

A) *Extraction des NP de chaque document de test* : Pour chaque document de test, nous extrayons la liste des NP. L'objectif est d'utiliser ceux-ci comme points d'ancrage pour chercher de nouveaux noms propres dans le corpus diachronique.

B) *Extraction du contexte à partir des documents diachroniques* : Après extraction de la liste des noms propres du document de test, nous pouvons commencer à extraire leurs «contextes» dans le corpus diachronique. Seuls les documents qui correspondent à la même période de temps que le document de test sont pris en compte. A partir de l'étiquetage

morphosyntaxique de ces documents diachroniques (*TreeTagger*), nous extrayons les mots qui ont été étiquetés comme « nom propre ».

Afin de limiter l'augmentation excessive du vocabulaire, nous proposons d'utiliser l'information mutuelle. Nous calculons cette information entre les NP du document de test (qui appartiennent au vocabulaire du système de reconnaissance) et les NP des documents contemporains du corpus diachronique. Si deux NP ont une information mutuelle élevée, cela augmente la probabilité qu'ils apparaissent tous les deux dans le document de test.

Dans la théorie des probabilités et de la théorie de l'information, l'information mutuelle de deux variables aléatoires est une grandeur qui mesure la dépendance mutuelle des deux variables aléatoires. Formellement, l'information mutuelle de deux variables aléatoires discrètes X et Y est défini comme :

$$I(X;Y) = \sum\sum p(x,y)\, log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Dans notre cas, X et Y représentent des noms propres et $x = 1$, si x est présent dans le document, et $x = 0$ sinon. Plus la probabilité de la cooccurrence de deux noms propres dans le corpus diachronique est grande, plus la probabilité de leur cooccurrence dans un document de test est élevée.

Enfin, on calcule l'information mutuelle entre toutes les combinaisons de la variable X (X est un NP du vocabulaire apparaissant dans le document de test) et la variable Y (Y est un NP OOV extrait des documents contemporains diachroniques).

C) *Augmentation du vocabulaire* : Les nouveaux NP dont l'information mutuelle est supérieure à un seuil sont retenus. Afin de sélectionner les NP les plus pertinents, nous éliminons ceux qui ont une faible fréquence d'apparition. Les prononciations de ces NP sont générées en utilisant un dictionnaire phonétique ou un outil graphème-phonème (Illina, 2011). En utilisant cette méthodologie, nous nous attendons à en extraire une liste réduite (par rapport à la méthode de référence) de tous les NP potentiellement manquants.

2.3 Méthode fondée sur la similarité cosinus

Par rapport à la méthode MI, seule l'étape B est modifiée. Nous représentons chaque document du corpus diachronique comme un *vecteur de mots* (sac de mots, BOW). Seuls les mots significatifs sont conservés : les verbes, les adjectifs, les noms et les NP. Pour chaque NP OOV présent dans les documents diachroniques contemporains du document de test, un *vecteur NP* est calculé comme étant la somme de *vecteurs de mots* des documents dans lesquels ce NP apparait. Le document de test est représenté par son *vecteur de mots*. Puis, la similarité cosinus entre ce vecteur de mots et chaque *vecteur NP* est calculée. Les NP dont la similarité cosinus est supérieure à un seuil sont sélectionnés et ajoutés au vocabulaire.

3 Expériences

3.1 Corpus de test

Doc1	Doc2	Doc3	Doc4	Doc5
2007/12/20	2007/12/21	2008/01/17	2008/01/18	2008/01/24

TABLE 1 – Date des documents de test

Pour valider la méthodologie proposée, nous avons utilisé comme corpus de test 5 documents audio du corpus ESTER2 (voir la Table 1). L'objectif de cette campagne était

d'évaluer la transcription automatique d'émissions de radio en français (Galliano, 2009). La Table 2 présente les occurrences de tous les NP (appartenant au vocabulaire et OOV) dans chaque document de test par rapport au vocabulaire de 97k mots de notre système de RAP. Pour augmenter le taux d'OOV, nous avons choisi au hasard 75 NP, qui apparaissent dans les corpus de test et nous les avons retirés de notre vocabulaire 97k mots. Cela conduit à un taux de NP OOV d'environ 1% (404/38525). Au total, le corpus de test contient 148 NP OOV différents.

	Nombre de mots différents	**Nombre d'occurrences**	**NP appartenant au vocabulaire**	**NP OOV**	**Occ. des NP OOV**
Doc1	1350	4099	86	44	93
Doc2	1446	4604	89	39	70
Doc3	1958	11803	43	25	63
Doc4	2107	10152	90	39	71
Doc5	1432	7867	48	27	107
All	-	38525	-	-	404

TABLE 2 – Couverture des noms propres du corpus de test.

Nous extrayons les NP de la transcription automatique générée par notre système ANTS (*Automatic News Transcription System*). Pour cela, le vocabulaire utilisé pour la reconnaissance a été étiqueté en termes de NP et non-NP. Comme nous utilisons des documents diachroniques, nous construisons un vocabulaire spécifique pour chaque document de test en fonction de la période choisie.

Les résultats sont présentés en termes de rappel (%): nombre de NP OOV retrouvés divisé par le nombre total de NP OOV du document de test. Pour les expériences de reconnaissance, le taux d'erreur de mots (*Word Error Rate,* WER) est calculé.

3.2 Corpus diachronique

Comme corpus diachronique, nous avons utilisé le corpus *GigaWord* : Agence France Presse (AFP) et *Associated Press Worldstream* (APW). Le corpus français *GigaWord* est une archive d'articles d'agences de presse : pour l'AFP de mai 1994 à décembre 2008, pour l'APW de novembre 1994 à décembre 2008. Le choix de *GigaWord* a été motivé par le fait qu'il est *contemporain* du corpus de test Ester, qu'il est rédigé dans le *même style* (journalistique) et qu'il traite de domaines similaire (politique, sports, etc.). De plus, sa granularité temporelle est fine, journalière. En revanche, *GigaWord* est centré sur l'information instantanée.

3.3 Système de transcription

Le système ANTS (Illina, 2004) utilisé pour ces expériences est fondé sur des modèles HMM dépendants du contexte appris sur un corpus audio de 200 heures d'émissions de radio. Le moteur de reconnaissance est Julius (Lee, 2009). Le modèle de langage est estimé en utilisant la boîte à outils SRILM (Stolcke, 2002) sur des corpus de textes d'environ 1.8 milliard de mots. Le modèle de langage est ré-estimé pour chaque vocabulaire augmenté. Le lexique phonétique de base contient 218K prononciations pour les 97K mots.

4 Résultats expérimentaux

4.1 Méthode de référence

En utilisant l'outil d'étiquetage morphosyntaxique *TreeTagger*, nous avons extrait 40982 NP différents à partir de 2 mois du corpus diachronique. Parmi ces 40982 NP, 23210 ne sont pas dans notre vocabulaire. Parmi ces 23210, seulement 102 NP sont présents dans le corpus de

test et qui sont donc OOV. Cela montre qu'il est nécessaire de filtrer la liste des NP pour avoir un meilleur compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire. Afin d'étudier si la période temporelle joue un rôle important, nous avons étudié trois intervalles de temps dans les documents diachronique : le même jour que le document de test (1 jour); 3 jours avant et 3 jours après la date du document de test (1 semaine); le mois courant du document de test (1 mois).

Période temporelle	Nombre moyen de NP sélectionnés par fichier de test	Nombre moyen de NP OOV retrouvés par fichier de test	Rappel (%)
1 jour	925	16	44.0
1 semaine	4305	21	58.6
1 mois	13069	**24**	**67.6**

TABLE 3 –Résultats pour la méthode de référence.

Nous appelons ***NP sélectionnés*** les noms propres que nous avons récupérés à partir des documents diachroniques en utilisant notre méthode et qui ne sont pas dans notre vocabulaire.

Nous appelons ***NP OOV retrouvés*** les noms propres OOV de la liste *NP sélectionnés* qui sont présents dans les documents de test.

La Table 3 montre que l'utilisation des documents diachroniques d'une journée nous permet de sélectionner en moyenne 925 nouveaux NP par fichier de test. Parmi ces NP, en moyenne, 16 NP sont présents dans un fichier de test (OOV). Cela correspond à un rappel de 44%.

Nous pouvons remarquer qu'en limitant l'intervalle de temps des documents diachroniques utilisés à une semaine réduit l'ensemble de candidats NP à 4305 (Table 3), tout en récupérant 58.6% des OOV manquants. Ce résultat confirme l'idée que l'utilisation de l'information temporelle peut aider à réduire la liste des nouveaux candidats NP pour l'enrichissement du vocabulaire tout en conservant un bon taux de rappel.

4.2 Méthode fondée sur l'information mutuelle

La Table 4 montre les résultats de la méthode fondée sur l'information mutuelle en utilisant différentes périodes de temps et différentes valeurs de seuil.

Période temporelle	Seuil	Nb moyen de NP sélectionnés par fichier de test	Nb moyen de NP OOV retrouvés par fichier de test	Rappel (%)
1 jour	0.05	10	5	13.8
	0.01	295	13	36.8
	0.005	421	14	40.8
	0.001	532	15	44.2
1 semaine	0.05	4	3	8.1
	0.01	51	9	25.3
	0.005	229	12	34.5
	0.001	1749	**19**	**55.2**
1 mois (occ>2)	0.05	3	2	4.6
	0.01	19	7	19.5
	0.005	41	9	25.8
	0.001	222	14	40.2

TABLE 4 – Résultats pour la méthode fondée sur l'information mutuelle.

Comme nous construisons un vocabulaire augmenté pour chaque fichier de test, les résultats présentés dans la Table 4 sont donnés en termes de moyenne des valeurs calculées sur les 5

fichiers de test. Pour les résultats correspondant à une période d'un mois, nous avons introduit un seuil de fréquence (*occ* dans la Table 4) afin de limiter les efforts de calcul. Utiliser uniquement les documents diachroniques d'un seul jour s'avère être insuffisant pour extraire un grand nombre de NP OOV. Cela peut être dû au fait que nous sommes dans le cadre des agences de presse et qu'un nouvel événement fait l'objet d'articles pendant plusieurs jours. Avec le seuil de 0,05 très peu de NP sont sélectionnés, mais une grande partie de ces NP sont des NP OOV. Par exemple, pour une période d'une semaine, parmi les 4 NP sélectionnés, 3 sont de NP OOV. Le meilleur rappel est obtenu en utilisant une période de temps d'une semaine (55,2%).

4.3 Méthode fondée sur la similarité cosinus

Les résultats pour la méthode fondée sur la similarité cosinus sont présentés dans la Table 5. Afin de réduire encore le nombre de NP sélectionnés, nous ne gardons que les NP ayant un nombre d'occurrences supérieur à un seuil dépendant de la période temporelle.

Comme dans le cas de la méthode MI, ne considérer qu'une seule journée pour récupérer des NP semble insuffisant. Le meilleur compromis entre le rappel et le nombre de NP sélectionnés est obtenue pour la période d'un mois et un seuil de 0,05 (59,8% de rappel).

Période temporelle	Seuil	Nb moyen de NP sélectionnés par fichier de test	Nb moyen de NP OOV retrouvés par fichier de test	Rappel (%)
1 jour (occ>0)	0.025	813	15	44.3
	0.05	438	14	41.4
	0.075	131	11	32.2
	0.1	52	8	24.1
1 semaine (occ>1)	0.025	1880	19	55.8
	0.05	1128	19	54.0
	0.075	432	17	48.9
	0.1	152	13	38.5
1 mois (occ>2)	0.025	3796	**21**	**61.5**
	0.05	2474	**21**	59.8
	0.075	1010	19	55.8
	0.1	334	17	48.9

TABLE 5 – Résultats pour la méthode fondée sur la similarité cosinus.

4.4 Résultats de reconnaissance

Pour valider les approches proposées, nous avons réalisé la transcription automatique des 5 documents de test. Pour chaque document de test et pour chaque période temporelle, un vocabulaire augmenté est généré.

	Vocabulaire standard	Vocabulaire augmenté		
		1 jour	1 semaine	1 mois
Doc 1	19.7	17.9	18.0	18.4
Doc 2	20.9	20.3	19.7	20.0
Doc 3	28.3	28.2	28.1	28.2
Doc 4	24.5	24.0	24.2	24.2
Doc 5	36.5	36.1	36.1	36.0
All	27.1	**26.6**	**26.6**	26.7

TABLE 6 – WER (%) pour la méthode MI, seuil de 0.001.

La Table 6 donne les résultats obtenus pour la méthode fondée sur l'information mutuelle. Il montre que, en moyenne, le vocabulaire augmenté en utilisant des données diachroniques réduit de manière significative le WER. Nous n'observons pas de différence significative entre les WER correspondant aux trois périodes temporelles évaluées.

Les résultats pour la méthode fondée sur la similarité cosinus sont donnés dans la Table 7. Comme pour la méthode MI, une amélioration significative est obtenue en utilisant les vocabulaires augmentés. Les résultats obtenus avec les périodes temporelles d'une semaine et d'un mois sont légèrement meilleurs que ceux d'une journée.

	Vocabulaire standard	**Vocabulaire augmenté**		
		1 jour seuil: 0.025 occ>0	**1 semaine seuil: 0.025 occ>1**	**1 mois seuil: 0.05 occ>2**
Doc 1	19.7	18.1	18.0	18.2
Doc 2	20.9	20.1	19.5	19.4
Doc 3	28.3	28.2	27.9	28.0
Doc 4	24.5	24.0	23.8	23.7
Doc 5	36.5	36.1	35.8	35.8
All	27.1	26.7	**26.4**	**26.4**

TABLE 7 – WER (%) pour la méthode cosinus.

Pour les deux méthodes, les performances dépendent du document de test. Pour les documents 1 et 2, quelle que soit la période de temps utilisée pour créer le vocabulaire augmenté, l'amélioration est significative. En revanche, l'amélioration pour le document 3 est plus faible. Enfin, les deux méthodes proposées donnent une amélioration significative par rapport au vocabulaire standard, mais la différence de WER entre les deux méthodes n'est pas significative.

5 Conclusion

Dans le cadre de la reconnaissance automatique de la parole, notre étude a porté sur le problème de l'augmentation du vocabulaire à l'aide de documents diachroniques. Nous avons étudié des méthodes qui augmentent le vocabulaire avec des noms propres en utilisant la notion de contexte et des informations temporelles. L'idée est d'utiliser les noms propres reconnus du document de test comme ancres pour rechercher de nouveaux NP dans le corpus diachronique. Notre modèle de contexte est fondé sur l'information mutuelle ou le modèle vectoriel (sac de mots). Ces nouveaux NP sont ajoutés au vocabulaire du système.

Des expériences ont été menées sur des émissions de radio en utilisant des données textuelles d'agences de presse comme corpus diachronique. Les résultats valident l'hypothèse que la période temporelle et le contexte lexical permettent de récupérer des noms propres manquants et d'obtenir une réduction significative du taux d'erreur de mots en ajoutant ces NP dans le vocabulaire. Une perspective intéressante pourrait être d'exploiter des informations "sémantiques" contenues dans le document de test : quand une date précise est identifiée, les documents diachroniques temporellement proches de cette date pourraient être utilisés pour extraire de nouveaux noms propres.

Remerciements

Les auteurs tiennent à remercier l'ANR *ContNomina* SIMI-2 de l'Agence Nationale de la Recherche française (ANR) pour son soutien.

Références

Allauzen, A. and Gauvain, J.-L. (2005). Diachronic vocabulary adaptation for broadcast news transcription, *Proc. of Interspeech*.

Bechet, F. and Yvon, F. (2000). Les Noms Propres en Traitement Automatique de la Parole, *Revue Traitement Automatique des Langues*, vol. 41, num. 3 pp. 672-708.

Bertoldi, N. and Federico, M. (2001). Lexicon adaptation for broadcast news transcription*, In Adaptation-2001*, pp. 187- 190.

Bigot, B., Senay, G., Linares, G., Fredouille, C., and Dufour, R. (2013). Person name recognition in ASR outputs using continous context models, *Proc. of ICASSP*.

Friburger, N. and Maurel, D. (2002). Textual Similarity Based on Proper Names, *Proc. of the workshop Mathematical/Formal Methods in Information Retrieval*, pp. 155-167.

Galliano, S., Gravier, G., Chaubard, L. (2009). The ESTER2 Evaluation Campaign for the Rich Transcription of French Radio Broadcast, *Proc. of Interspeech*.

Illina I., Fohr D., Mella O., Cerisara C. (2004). The Automatic News Transcription System: ANTS, some Real Time experiments, *Proc. ICSLP*.

Illina I., Fohr D., Jouvet D. (2011). Grapheme-to-Phoneme Conversion using Conditional Random Fields, *Proc. Interspeech*

Kobayashi, A., Onoe, K., Imai, T., Ando, A. (1998). Time dependent language model for broadcast news transcription and its post-correction, *Proc ICSLP*

Lee, A. and Kawahara, T. (2009). Recent Development of Open-Source Speech Recognition Engine Julius, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).*

Nkairi, I., Illina, I., Linarès, G., Fohr, D. (2013). Exploring temporal context in diachronic text documents for automatic OOV proper name retrieval, *Proc. of LTC*.

Oger, S., Linarès, G. and Béchet, F. (2008). Local methods for on-demand out-of-vocabulary word retrieval, *Proc. of the Language Resources and Evaluation Conference (LREC)*.

Parada, C., Dredze, M., Filimonov, F. and Jelinek, F. (2010). Contextual Information Improves OOV Detection in Speech, *Proc. of NAACL*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees*, Proc. of ICNMLP*.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit, *Proc. of ICSLP.*