Génération des prononciations de noms propres à l'aide des Champs Aléatoires Conditionnels

Irina Illina, Dominique Fohr, Denis Jouvet
Équipe Parole, INRIA-LORIA, 615, rue du Jardin Botanique, 54602 Villers-les-Nancy, France
{illina, fohr, jouvet}@loria.fr

RÉSUMÉ

Dans cet article, nous proposons une approche de conversion graphème-phonème pour les noms propres. L'approche repose sur une méthode probabiliste : les Champs Aléatoires Conditionnels (*Conditional Random Fields, CRF*). Les CRFs donnent une prévision à long terme, n'exigent pas l'indépendance des observations et permettent l'intégration de tags. Dans nos travaux antérieurs, l'approche de conversion graphème-phonème utilisant les CRFs a été proposée pour les mots communs et différents paramétrages des CRFs ont été étudiés. Dans cet article, nous étendons ce travail aux noms propres. Par ailleurs, nous proposons un algorithme pour la détection de l'origine des noms propres. Le système proposé est validé sur deux dictionnaires de prononciations. Notre approche se compare favorablement aux JMM (Joint-Multigram Model, système de l'état de l'art), et tire profit de la connaissance de la langue d'origine du nom propre.

Abstract

Pronunciation generation for proper names using Conditional Random Fields

We propose an approach to grapheme-to-phoneme conversion for proper names based on a probabilistic method: Conditional Random Fields (CRFs). CRFs give a long term prediction, assume a relafex state independence condition and allow a tag integration. In previous work, grapheme-to-phoneme conversion using CRF has been proposed for non proper names and different CRF features are studied. In this paper, we extend this work to proper names. Moreover, we propose an algorithm for origine detection of proper names of foreign origins. The proposed system is validated on two pronunciation dictionaries. Our approach compares favorably with the performance of the state-of-the-art Joint-Multigram Models and takes advantage of the knowledge of the origin of the proper name.

MOTS-CLÉS : reconnaissance de la parole, noms propres, phonétisation du lexique, champs aléatoires conditionnels (CRF)

KEYWORDS: automatic speech recognition, proper names, lexique phonetisation, conditional random field (CRF).

1 Introduction

La phonétisation d'un mot à partir de sa forme écrite consiste à trouver les variantes de prononciations de ce mot. Les principales applications sont la reconnaissance automatique de la parole, la synthèse vocale et la génération de dictionnaires de prononciations. Dans ces applications, l'utilisation de dictionnaires conçus et vérifiés manuellement est la solution qui permet la meilleure précision. Mais le coût de cette solution est souvent prohibitif. Pour les noms communs, des dictionnaires phonétiques sont parfois disponibles (comme, par exemple, le dictionnaire CMU pour l'anglais ou le Bdlex pour le français). En revanche de tels dictionnaires

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 641–648,
Grenoble, 4 au 8 juin 2012.

sont rarement disponibles et même inexistants pour les noms propres. Dans ce cas, la génération automatique d'un dictionnaire est nécessaire.

Les problèmes de phonétisation des noms propres sont nombreux et proviennent en partie de leur diversité (Bechet, 2000) : leurs différentes prononciations, leurs différentes origines, l'imprévisibilité orthographique de noms propres et leurs homographes hétérophones (formes identiques ayant des prononciations différentes), l'orthographe non complètement normalisée, les noms propres d'origine étrangère.

La tâche de phonétisation est souvent considérée comme une tâche de conversion de la suite de graphèmes vers la suite de phonèmes correspondants (*Grapheme-to-Phoneme Conversion, G2P*). Ces dernières années, différentes approches plus ou moins automatiques ont été proposées pour essayer de résoudre le problème de la phonétisation des noms propres. A ce jour, ces approches produisent entre 50 et 12% d'erreur. Ces approches se décomposent souvent en deux étapes. La première étape est la détection de l'origine du nom propre et s'appuie fréquemment sur un modèle N-gramme de graphèmes. L'étape suivante est la phonétisation dépendante de l'origine trouvée. Le système à base de règles de (Bartkova, 2003) détecte l'origine d'un nom propre à phonétiser à partir de suites de lettres caractéristiques, puis génère les variantes de prononciation en s'appuyant sur des règles propres à chaque origine détectée. Le nombre de règles, et parfois les conflits entre elles rendent cette approche assez lourde à mettre en place. (Llitjos, 2001) proposent de détecter l'origine d'un nom propre en étudiant les trigrammes de lettres, puis des arbres de décision (un pour chaque lettre) sont utilisés pour prédire la prononciation à partir des lettres et de leurs contextes. Concernant le français, la combinaison de 4 systèmes a permis de passer de 20% d'erreurs de mots à 12% (de Mareuil, 2005). Pour la détection de l'origine, (Chen, 2006) utilise *N-gram Syllable-Based Letter Clusters :* pour elargir la fenetre d'analyse, pour chaque langue, les auteurs construisent un modèle N-gramme de classes de lettres les plus fréquentes (syllabes).

Dans notre article nous nous intéressons au problème de la phonétisation des noms propres et nous proposons une nouvelle approche pour la détection de l'origine d'un nom propre. Pour ces deux problèmes, nous proposons d'utiliser les *Champs Aléatoires Conditionnels* (*Conditional Random Fields, CRFs*) car à l'issu de l'apprentissage ils permettent de trouver les coefficients optimaux même si les paramètres sont corrélés et ils permettent d'intégrer différentes sortes d'indices (Lafferty, 2001). Pour comparer notre approche à l'état de l'art, le *Modèle de Multigrammes Joints* (*Joint-Multigram Model, JMM*) (Bizani, 2008) est utilisé.

La structure de notre article est la suivante : la section 2 est consacrée à la présentation de la méthodologie proposée, la section 3 décrit les expérimentations menées et leurs résultats, et la section 4 conclut notre article.

2 Méthodologie

Dans notre travail, nous proposons d'utiliser les *Champs Aléatoires Conditionnels* (*Conditional Random Fields, CRFs*). Les CRFs sont un outil probabiliste pour l'étiquetage et la segmentation des données structurées, telles que des séquences, des arbres ou des treillis. Les CRFs donnent une prévision à long terme, contrairement aux HMM n'exigent pas l'indépendance des observations, permettent un apprentissage discriminant et finalement convergent vers un optimum global. Notre choix de CRFs est motivé par le fait que le processus d'apprentissage permet de trouver les coefficients optimaux même si les paramètres sont corrélés.

Les CRFs trouvent des applications dans le domaine de l'étiquetage et de l'analyse de données séquentielles, dans le domaine de la segmentation d'images et peuvent être utilisés comme une approche générale de combinaison de caractéristiques de différentes sources. Récemment, les CRFs ont été appliqués dans le domaine de la reconnaissance vocale : pour insérer de façon automatique les *virgules* dans les résultats de la reconnaissance (Akita, 2011), pour une mesure de confiance performante (Seigel, 2011), pour la détection des entités nommées (Mc Callum, 2003). (Lehnen, 2011) a étendu le cadre théorique des CRFs en y introduisant l'idée de « *back-off* » (similaire à l'idée de « *back-off* » dans les modèles de langage).

2.1 Phonétisation de noms propres à l'aide de CRFs

Dans (Illina, 2011), nous avons présenté notre méthodologie d'utilisation des CRFs pour la conversion G2P. Ici nous rappellerons rapidement les étapes principales :

- Comme l'apprentissage des CRFs nécessite d'avoir les associations « *un graphème – un phonème* » du corpus d'apprentissage, l'étape de pré-traitement consiste à aligner tous les graphies de mots d'apprentissage avec les phonèmes correspondants. L'obtention de ces associations s'effectue en deux sous-étapes : (1) - Tout d'abord, nous générons les associations « *un graphème – plusieurs phonèmes* » en effectuant un alignement forcé. Pour cela, nous utilisons des HMMs discrets : chaque phonème est modélisé par un HMM à un état, chaque observation de ce HMM correspond à un graphème. (2)- La deuxième sous-étape consiste, à partir de ces associations « *un graphème – plusieurs phonèmes* », à générer les associations « *un graphème – un phonème* ». Dans les cas où un phonème est aligné avec plusieurs graphèmes, nous associons ce phonème avec le graphème dont la probabilité est la plus grande. Les graphèmes restants sont alors associés avec le phonème nul.
- Durant la deuxième étape d'apprentissage, en utilisant les associations « *un graphème – un phonème* » générées, les modèles CRFs sont appris. Les CRFs apprennent les poids w en maximisant la vraisemblance de $p(\bar{y} | \bar{x}; w)$:

$$p(\bar{y} \mid \bar{x}; w) = 1/Z(\bar{x}, w) \exp \Sigma_j w_j F_j(\bar{x}, \bar{y}) \quad (1)$$

$$F_j(\bar{x}, \bar{y}) = \Sigma_i f_j(\bar{y}_{i-1}, \bar{y}_i, \bar{x}, i) \quad (2)$$

où $\bar{x}$ est la séquence de graphèmes, $\bar{y}$ est la séquence de phonèmes, w est le poids à apprendre. f_j est une fonction qui dépend de la séquence de graphèmes de mot, du phonème actuel, du phonème précédent et de sa position actuelle dans le mot. Notons, que l'équation (2) correspond aux bigrammes (le phonème courant et le phonème précédent sont pris en compte).

Lors de la phonétisation G2P, le décodage à l'aide des CRFs trouve les N-meilleures séquences de phonèmes correspondants à un mot du corpus de test.

2.2 Détection de l'origine d'un nom propre à l'aide de CRFs

Pour prédire l'origine d'un nom propre, nous avons utilisé des CRFs. Pour cette tâche, la séquence des observations (vecteur X des formules (1) et (2)) est constituée de la séquence des graphèmes du mot. A chaque graphème on associe l'étiquette $\bar{y}_i$ correspondant à l'origine du mot. Les vecteurs de caractéristiques des CRFs sont composés des graphèmes du mot dont on veut connaître l'origine. Afin d'obtenir plusieurs origines possibles pour un mot donné, un seuil de probabilité est

utilisé: les réponses fournies par les CRFs dont les probabilités sont supérieures à ce seuil sont conservées.

3 Expériences

3.1 Critères d'évaluation

Dans le cas de la génération d'une seule prononciation par mot, le critère de performance est le pourcentage de mots avec une phonétisation correcte. Ce terme est défini comme le pourcentage de mots, où tous les phonèmes de la phonétisation correspondent exactement aux phonèmes de la référence. Dans le cas où plusieurs variantes de prononciations de référence existent pour un mot, toutes les variantes sont examinées et celle qui obtient la meilleure correspondance est choisie.

Dans le cas de la génération de plusieurs variantes de prononciation par mot, le rappel et la précision sont utilisés. Le rappel est le nombre de variantes de prononciations générées qui sont correctes divisé par le nombre de total de prononciations de référence. La précision représente le nombre de variantes de prononciations correctes divisé par le nombre total de variantes de prononciations générées.

3.2 Corpora

Corpus BDLex Le corpus BDLex est une base de données lexicale développé à l'IRIT (De Calmès, 1998). Il contient des informations lexicales, phonologiques et morphologiques. BDLex est composé d'environ 440000 formes fléchies avec les attributs suivants : graphie, prononciation, traits morphosyntaxiques, forme canonique (lemme) et un indicateur de fréquence.

Nous avons divisé ce corpus de façon aléatoire en trois parties disjointes : 75% pour l'apprentissage, 5% pour le développement, 20% pour le test. Cette partition est faite selon les lemmes : toutes les formes fléchies d'un mot sont mises ensemble dans la même partie. L'ensemble de développement du corpus BDLex est utilisé pour sélectionner le paramétrage optimal. Les paramètres obtenus sont ensuite appliqués sur la partie test.

Le corpus BDLex ne contient pas d'information sur l'origine de chaque mot et ne contient pas de noms propres. Pour un taux de reconnaissance en mot de 95%, l'intervalle de confiance est de $\pm$ 0,2 avec la tolérance de 5%.

Corpus du LORIA Ce corpus de noms propres est composé de 3500 noms de personnes (noms de familles) (*NP-Lor* dans nos expériences). Pour chaque nom propre, on dispose de sa graphie, d'une ou plusieurs transcriptions phonétiques et de l'information de l'origine de ce nom propre (appelé *tag d'origine* dans la suite de l'article). Un même nom propre peut avoir plusieurs tags, par exemple, le nom « Berger » peut être un nom propre français ou allemand, avec des prononciations différentes associées à chaque tag. En tout il y a une quinzaine de tags d'origine, le tag "français" couvre environ 50% du corpus. En moyenne, il y a 1,4 prononciations par mot.

Comme la taille du corpus est faible, nous avons utilisé l'approche « *leave-one-out* ». Pour cela le corpus est divisé de façon aléatoire en 10 parties égales : 9 parties sont utilisées pour l'apprentissage et la partie restante pour le test. Cette procédure est répétée 10 fois. Pour un taux de reconnaissance en mot de 60%, l'intervalle de confiance est de $\pm$ 1,6% avec la tolérance de 5%.

3.3 Logiciels utilisés

CRF++. CRF++ est un logiciel *open source* de CRFs destiné à segmenter et étiqueter des données séquentielles. Il est écrit en C++, utilise la méthode d'apprentissage rapide fondée sur la descente de gradient et génère les N-meilleures hypothèses.

Sequitur G2P (JMM). Pour comparer notre approche à l'état de l'art, nous avons choisi d'utiliser l'approche du Modèle de Multigrammes Joints (*Joint-Multigram Model, JMM*) (Bizani, 2008) et le logiciel Sequitur correspondant. Le principe consiste à déterminer l'ensemble optimal des séquences jointes, où chaque séquence est composée d'une séquence de graphèmes et de la séquence de phonèmes associés. Un modèle de langage est appliqué aux séquences jointes. L'algorithme procède de façon incrémentale : la première passe crée un modèle simple. Puis chaque passe utilise le modèle précédemment créé pour agrandir les séquences jointes (8 passes dans nos expériences).

4 Résultats expérimentaux

Dans (Illina, 2011) nous avons étudié l'influence du POS-tag, du contexte et de l'effet de l'unigramme et du bigramme sur la performance des CRFs. Les résultats expérimentaux ont suggéré que plus le contexte de graphèmes est large, meilleurs sont les résultats. Donc, nous avons fixé le contexte de graphèmes à neuf, c'est-à-dire les quatre lettres précédentes, la lettre courante et les quatre lettres suivantes. Il est préférable d'utiliser un ensemble d'indices assez large (des indices bigrammes et unigrammes). Dans le présent travail, nous utilisons donc des indices bigrammes et des indices unigrammes avec des contextes de 1, 3, 5, 7 et 9 graphèmes.

4.1 Génération d'une seule prononciation par mot

Dans ces expériences nous générons une seule prononciation par mot. Nous effectuons les tests sur : (1) - la partie test du corpus BDLex pour mettre en évidence la différence de résultats de la génération G2P pour le corpus BDLex par rapport aux noms propres ; (2) - la partie test du corpus BDLex, en excluant les verbes car la prédiction de leur phonétisation est plus simple que pour les autres mots ; (3)- les noms propres d'origine française du corpus NP-Lor ; (4) - les noms propres d'origine non française du corpus NP-Lor.

Apprentissage de modèles sur le corpus BDLex. L'apprentissage de nos modèles est effectué sur la partie apprentissage du corpus BDLex. La figure 1 (à gauche) présente le pourcentage de mots dont les phonétisations sont correctes en fonction du corpus de test utilisé. Cette figure montre que sur le corpus BDLex de test, 97% de mots sont bien transcrits. En excluant les verbes du corpus BDLex et donc complexifiant la tâche, le taux de transcription descend à 95%. La tâche de conversion G2P pour les noms propres d'origine étrangère est la tâche la plus difficile : autour de 43% de mots sont bien transcrits. Les CRFs donnent des résultats légèrement meilleurs que ceux obtenus par les JMM. Les modèles appris sur le corpus BDLex, qui ne contient pas les noms propres, ne semblent pas être performants sur la conversion G2P de noms propres : une perte d'environ 20% de mots correctement phonétisés est observée.

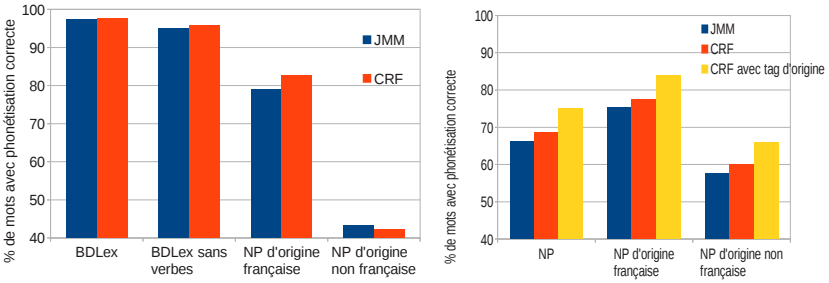

FIGURE 1 (à gauche) – Pourcentage de mots dont la phonétisation est correcte en fonction du corpus de test. Apprentissage de modèles sur le corpus BDLex. (à droite) – Pourcentage des mots dont la phonétisation est correcte en fonction du corpus de test. Apprentissage de modèles sur le corpus NP-Lor (noms propres).

En conclusion, pour la phonétisation de noms propres, il est difficile d'apprendre un modèle performant à partir du corpus d'apprentissage qui ne contient que les noms communs (comme BDLex). Dans la section suivante, nous présenterons quelques résultats en utilisant les modèles appris sur le corpus de noms propres.

Apprentissage de modèles sur le corpus NP-Lor. Nous avons exploré également l'influence de l'apprentissage des modèles sur le corpus NP-Lor de noms propres. Les tests sont effectués sur le corpus NP-Lor. Deux configurations de modèles CRFs sont utilisées. La première configuration prend en compte le tag d'origine de mot. Pendant le test, le mot et son origine sont fournis aux CRFs pour effectuer la conversion G2P. Ce genre de test n'était pas possible pour la configuration de la section précédente (apprentissage de modèles sur le corpus BDLex) car le corpus BDLex ne contient pas d'information sur l'origine de mots. Dans la deuxième configuration les modèles n'utilisent pas le tag d'origine de mots.

La figure 1 (à droite) permet de tirer les conclusions suivantes. Comme précédemment, les CRFs donnent de meilleurs résultats que les JMMs. Comme attendu, l'ajout du tag d'origine permet d'améliorer de façon significative les résultats.

En comparant les figures (à gauche) et (à droite) nous observons que les noms propres d'origine française sont transcrits presque aussi bien en utilisant les modèles appris sur le corpus BDLex que les modèles appris sur les noms propres (82% versus 84%). En revanche pour la phonétisation de noms propres d'origine étrangère, l'apprentissage sur le corpus de noms propres permet d'améliorer les résultats d'environ 22% absolu (43% versus 65%). Il est probable qu'en utilisant le corpus d'apprentissage de noms propres plus large, le résultat pourrait être sensiblement meilleur. Une autre possibilité est d'ajouter une partie du corpus BDLex dans le corpus d'apprentissage tout en maintenant un bon équilibre entre les données de différentes origines.

4.2 Génération de plusieurs prononciations par mots

Un système efficace « de conversion G2P devrait générer toutes les variantes de prononciation possibles pour un mot donné. Dans les expériences précédentes, une seule prononciation par mot

a été générée. Dans cette section, nous générons plusieurs prononciations pour chaque mot et étudions leur qualité par rapport aux références multiples de prononciations dans le corpus. En variant un seuil de décision, nous générons une ou plusieurs prononciations par mot.

Nous avons utilisé le critère suivant : les variantes de prononciation générées ne sont conservées que si leur probabilité est supérieure à un seuil S. Dans nos expériences, nous avons

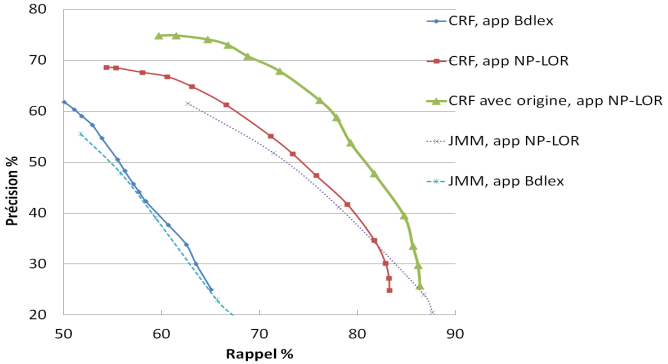

FIGURE 2 – Précision et rappel en fonction du corpus d'apprentissage et l'approche G2P utilisé (CRFs et JMM).

fait varier ce seuil S entre 0,0 et 0,4. Les résultats sont présentés sur la figure 2 en terme de précision et de rappel en faisant varier le seuil S. Les modèles sont appris sur le corpus BDLex et sur le corpus NP-Lor. Les tests sont effectués sur le corpus NP-Lor. Cette figure monte que dans le cas de la génération plusieurs variantes de prononciations de mots, comme dans le cas de génération d'une seule prononciation par mot, l'apprentissage sur le corpus BDLex donne les moins bons résultats quelle que soit l'approche G2P utilisée. Le meilleur résultat est obtenu en apprenant nos modèles sur le corpus de noms propres et en prenant en compte le tag d'origine du mot. Les CRFs surpassent légèrement les JMMs. Rappelons qu'il n'est pas possible de prendre en compte le tag d'origine pour les JMM.

4.3 La détection de l'origine d'un nom propre

Le but de cette étude préliminaire est de déterminer l'origine d'un nom propre à partir de sa graphie. Dans ces expériences la phonétisation de mots n'est pas effectuée. Cette information servira pour phonétiser les noms propres dont l'origine est inconnue. Nous avons effectué quelques expériences préliminaires en utilisant des CRFs : l'apprentissage et le test sont faits sur le corpus NP-Lor avec 7 origines (l Français, Anglais, Allemand, Italien, Slave, Espagnol, autre). Pendant le test, à partir de la graphie, les CRFs déterminent l'origine du mot de test. Le premier résultat (65.7% de détection correcte) est encourageant, néanmoins il nous montre le manque de données d'apprentissage. Nous collectons actuellement un corpus de noms propres et leurs origines à partir de pages Web (listes de sportifs, de joueurs d'échecs ou de Go, etc.).

5 Conclusion

Nous avons exploré dans cet article la problématique de la phonétisation de noms propres en vue d'améliorer la taille et la qualité d'un lexique. Les champs Aléatoires Conditionnels sont utilisés pour effectuer la conversion phonème-graphème et pour la détection de l'origine d'un nom propre. Les résultats montrent que les CRFs sont plus performants que le JMM dans le cas de la génération d'une seule ou de plusieurs variantes de prononciation. Nos futurs recherches porteront sur la détection de l'origine d'un mot et son intégration dans le processus de la phonétisation de noms propres.

6 Références

AKITA, Y., KAWAHARA, T. (2011). Automatic comma insertion of lecture transcripts based on multiple annotations. In INTERSPEECH.

ALLAUZEN, A. GAUVAIN, J.-L. (2004). Construction automatique du vocabulaire d'un système de transcription *IN JEP*.

BARTKOVA, K. (2003). Generating proper name pronunciation variants for automatic speech recognition. I*n 15th ICPhS,* pages 1321-1324.

BECHET, F., YVON, F. (2000). Les noms propres en traitement automatique de la parole. *In* Revue Traitement Automatique des Langues – TAL, pages 672-708, vol. 41/3.

BISANI, M. NEY, H., Joint-Sequence (2008). Models for grapheme-to-phoneme conversion, *In Speech Communication Journal*, 50: 434-451, *Elsevier.*

CHEN, Y., YOU, J., CHU, M., ZHAO, Y., WANG, J. (2006). Identifying language origin of person names with N-grams of different units. *In ICASSP*, pages 729-731.

DE CALMES, M., PERENNOU, G. (1998). BDLex: a lexicon for spoken and written French. *In LREC.*

ILLINA, I. FOHR, D., JOUVET, D. (2011). Grapheme-to-phoneme conversion using Conditional Random Fields. *In INTERSPEECH.*

LAFFERTY, J. MCCALLUM, A. PEREIRA, F. (2001). Conditional Random Fields: Modèles probabilistes pour la segmentation et l'étiquetage des données de séquence", *In Proc. Conférence internationale sur l'apprentissage automatique,* 282-289.

LEHNEN, P., NEY H. (2011). N-grams for CRF or a Failure-transitional posterior for acyclic FSTs. *In INTERSPEECH.*

LLITJOS, A.F., Black, A.W. (2001). Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names. *In INTERSPEECH*, pages 1919-1922.

DE MAREUIL, P., ALESSANDRO, C., BAILLY, G., BECHET, F., GARCIA, M.-N., MOREL, M., PRUDON, R., VERONIS, J. (2005). Evaluating the pronunciation of proper names by four French grapheme-to-phoneme converters. *In INTERSPEECH.*

MC CALLUM, A., LI, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons *In CONLL '03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*

SEIGEL, M.S., WOODLAND, P.C. (2011). Combining Information sources for confidence estimation with CRF models. *In INTERSPEECH.*