Décodeur Dynamique Basé sur des WFSTs pour la Reconaissance Automatique de la Parole

Étudiant : Danilo de Oliveira Ribeiro e Silva
Tuteur en entreprise : Marc Ferras-Font (SONY)

Les systèmes de reconnaissance automatique de la parole (ASR) convertissent la parole d'un signal audio enregistré en texte. Ce type de système vise à déduire les mots d'origine d'un signal observé, le plus souvent selon une approche probabiliste. On appelle décodage le processus de prédiction de la séquence de mots qui correspond le plus au signal acoustique [1].

Une manière très répandue de rechercher la meilleure séquence de mots comprend l'utilisation de transducteurs pondérés à états finis (WFST), des graphes statiques contenant l'ensemble de séquences de mots possibles. Ils contiennent de nombreux niveaux d'informations et déterminent ce qui est autorisé (et plus probable) dans la langue. Lors du décodage, le décodeur les parcourt afin de prédire la ou les meilleures séquences, c'est-à-dire les chemins plus probables (les moins coûteux).

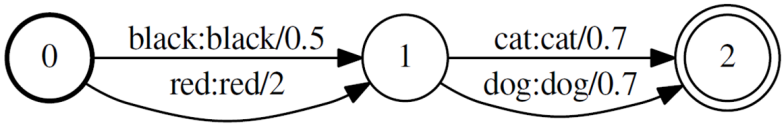

Figure 1 – Exemple de grammaire WFST. Les transitions contiennent des étiquettes d'entrée et de sortie et une probabilité (ou un coût) d'effectuer une telle transition. Un chemin d'un état initial à un état final *transduit* une séquence d'étiquettes d'entrée à une séquence de sortie

Le stage a consisté à étudier et à mettre en place des moyens de modifier dynamiquement les graphiques de décodage et/ou le processus de décodage d'un décodeur basé sur des WFSTs. Les objectifs principaux étaient de créer des décodeurs basés sur des classes de mots, ainsi que l'utilisation d'approches dynamiques pour la construction de cascades de WFST.

Nous avons réussi à implémenter des décodeurs fonctionnels basés sur des classes; les chiffres étaient conservés dans une grammaire séparée, et ce graphe a été remplacé dans un graphique principal chaque fois qu'une quantité était attendue. Cela facilite les modifications, car les OOV peuvent être plus facilement insérés dans leurs graphes de classes respectifs en raison de la taille réduite de ces derniers. Cette technique peut être étendue à d'autres classes, telles que les noms de personnes et de villes, qui sont souvent inconnus au système.

Le WFST examiné lors du décodage est une combinaison (composition) de différents graphes avec plusieurs niveaux d'informations. Ces graphes sont généralement combinés avant l'opération de décodage, mais cela augmente considérablement la taille du WFST. Nous avons réussi à implémenter cette combinaison lors du décodage, afin que l'espace de recherche soit construit *on-the-fly*. Cela réduit considérablement l'utilisation de la mémoire à la fois dans le stockage sur disque dur et dans les ressources RAM lors du décodage.

Ces deux caractéristiques ont été combinées dans un seul décodeur incluant le remplacement de classe *on-the-fly* et la composition *on-the-fly*. Le Tableau 1 présente les performances des différents décodeurs mis en œuvre au cours de ce stage. Le taux d'erreur sur les mots (WER) exprime l'erreur des prédictions (moins c'est mieux) et le facteur temps réel (RTF) indique la vitesse (moins c'est mieux, 1.0 signifie temps réel). *Static* est le cas original; *Static Replace* ajoute le remplacement des graphes de classe dans le transducteur principal; *Lazy* met en œuvre la combinaison *on-the-fly* des différents niveaux de représentation; et *Lazy Replace* utilise les deux techniques dans le même décodeur.

La performance du décodeur dynamique est moins bonne que celle du graphe statique, non remplacé, tant en termes de précision que de rapidité. Néanmoins, cet impact sur la performance

Type	WER	RTF
Static	10.0%	0.37
Static Replace	11.2%	0.39
Lazy	13.4%	1.28
Lazy Replace	14.4%	1.48

Table 1 – Scores obtenus pour des expériences de composition dynamique.

était attendu; Le graphe original, bien que peu pratique en termes de taille, est davantage optimisé, car il s'agit d'une pièce unique. Notre graphe implémenté peut également être optimisé pour obtenir une performance aussi proche que possible du graphe original.

References

[1] R.E. Gruhn, W. Minker, and S. Nakamura. *Statistical Pronunciation Modeling for Non-Native Speech Processing*. Signals and Communication Technology. Springer Berlin Heidelberg, 2011.