Simulation d'erreurs de reconnaissance automatique dans un cadre de compréhension de la parole

Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, Yannick Estève

Simulation d'erreurs de reconnaissance automatique dans un cadre de compréhension de la parole

Edwin Simonnet Sahar Ghannay Nathalie Camelin Yannick Estève
LIUM, Le Mans Université, France
firstname.lastname@univ-lemans.fr

RÉSUMÉ

Cet article propose une méthode de simulation d'erreurs de systèmes de reconnaissance automatique de la parole (SRAP) à partir de transcriptions manuelles, et montre son utilité pour rendre les systèmes de compréhension automatique de la parole (SCAP) plus robustes aux erreurs de SRAP. Partant du principe que le SRAP confond les mots acoustiquement et linguistiquement proches, cette méthode s'appuie sur l'utilisation de plongements de mots acoustiques et linguistiques pour calculer une mesure de similarité entre les mots : cette mesure vise à prédire les confusions de mots faites par le SRAP. Les expériences menées sur le corpus MEDIA (réservations d'hôtel) montrent que cette approche améliore significativement les performances des SCAP avec une réduction relative de 21,2% du taux d'erreur concept/valeur, en particulier quand le SCAP est neuronal (réduction de 22,4%). Une comparaison avec une méthode de bruitage naïf montre la pertinence de l'approche de bruitage proposée.

ABSTRACT

Simulating ASR errors for training SLU systems

This paper presents an approach to simulate automatic speech recognition (ASR) errors from manual transcriptions and how it can be used to improve the performance of spoken language understanding (SLU) systems. The proposed method is based on the use of both acoustic and linguistic word embeddings in order to define a similarity measure between words. This measure is dedicated to predict ASR confusions. Actually, we assume that words acoustically and linguistically close are the ones confused by an ASR system. Experiments were carried on the French MEDIA corpus focusing on hotel reservation. They show that this approach significantly improves SLU system performance with a relative reduction of 21.2% of concept/value error rate (CVER), particularly when the SLU system is based on a neural approach (reduction of 22.4% of CVER). A comparison to a naive noising approach shows that the proposed noising approach is particularly relevant.

MOTS-CLÉS : compréhension de la parole, augmentation des données, bruitage, reconnaissance automatique de la parole, erreurs.

KEYWORDS: spoken language understanding, data augmentation, noising, automatic speech recognition, errors.

1 Introduction

Les systèmes de compréhension de la parole (SCAP) ont pour but l'extraction d'informations sémantiques dans un discours. Dans un système de dialogue, cela consiste à extraire automatiquement des concepts sémantiques sous forme de couples concept/valeur à partir de transcriptions automatiques afin d'alimenter le gestionnaire de dialogue.Nous considérons ainsi la tâche de compréhension de

la parole comme une tâche de traduction où les séquences d'hypothèses de mots issues d'un SRAP doivent être traduites en une séquence de concepts sémantiques associés à leurs valeurs. Ainsi, la bonne performance du système de compréhension est donc fortement liée à la bonne performance du système de transcription. En effet, les erreurs de transcription sont susceptibles d'affecter les mots supports d'un concept, rendant difficile à la fois la détection du concept et l'extraction de sa valeur.

Dans l'optique de rendre plus robustes les systèmes de compréhension aux erreurs de reconnaissance, il est habituel d'entrainer le modèle sur des transcriptions automatiques plutôt que sur des transcriptions manuelles. Comme les corpus requis pour l'apprentissage des systèmes de dialogue sont rares, certaines méthodes ont été proposées pour simuler les erreurs de transcription dans ce cadre (Pietquin & Beaufort, 2005; Schatzmann *et al.*, 2007). La simulation d'erreurs de transcriptions a également été utilisée pour l'entraînement de modèles de langage discriminatifs afin d'améliorer les performances des SRAP en terme de taux d'erreurs mots (Jyothi & Fosler-Lussier, 2010).

De nos jours, les SCAP sont souvent construits avec une approche guidée par les données (Sarikaya *et al.*, 2014; Mesnil *et al.*, 2015; Hakkani-Tür *et al.*, 2016). Des annotations manuelles sont habituellement produites pour étiqueter des transcriptions manuelles avec des étiquettes sémantiques afin de construire un corpus d'apprentissage. Dans l'étude présentée ici nous supposons – et le vérifions – que la construction des SCAP à partir de transcriptions automatiques est une bonne solution pour les rendre plus robustes aux erreurs de transcriptions. Or, l'obtention des transcriptions automatiques nécessite d'avoir à disposition d'une part des enregistrements audio relatifs aux annotations sémantiques et d'autre part un SRAP. Afin que ce dernier soit efficace, il nécessite lui aussi des données d'apprentissage et de validation, ces dernières étant souvent les mêmes que celles utilisées pour l'apprentissage et la validation SCAP. Il convient donc de manipuler ces données avec prudence afin d'éviter des biais et notamment celui du sur-apprentissage.

Dans le cadre de la construction d'un SCAP performant, cette étude propose une approche de simulation des erreurs de reconnaissance à partir des transcriptions manuelles afin d'une part de s'affranchir de la nécessité de données audio et d'un SRAP lors de la phase d'apprentissage et d'autre part d'avoir néanmoins à disposition un corpus proche de celui à gérer lors du déploiement. Notre approche consiste à simuler et introduire des erreurs dans les transcriptions manuelles en substituant des mots corrects par des mots similaires. Nous supposons que les mots susceptibles d'être confondus par un SRAP sont des mots acoustiquement proches. Cette hypothèse a également été retenue dans (Fosler-Lussier *et al.*, 2002; Stuttle *et al.*, 2004), où la simulation des erreurs est basée sur la similarité phonétique des mots pour évaluer leur similarité. De plus, nous considérons que ces mots confondus sont également linguistiquement proches.

Pour calculer une mesure de similarité entre les mots, nous présentons une nouvelle approche utilisant des plongements de mots acoustiques et linguistiques. Dans nos expériences, nous évaluons l'impact de cette approche en bruitant le corpus d'apprentissage de deux SCAP : un basé sur des champs aléatoires conditionnels (Lafferty *et al.*, 2001) (CRF) et l'autre sur un réseau de neurone récurent bidirectionnel encodeur-décodeur avec un mécanisme d'attention (Cho *et al.*, 2014) (RNN-EDA). Ces expériences sont menées sur le corpus français MEDIA, sur lequel les CRF fonctionnent toujours mieux que les approches neuronales (Vukotic *et al.*, 2015; Simonnet *et al.*, 2017).

2 Mesure de similarité et simulation d'erreurs de SRAP

Nous proposons une mesure de similarité qui s'appuie sur l'utilisation des plongements linguistiques et acoustiques pour prédire une liste de mots qui pourraient être substitués par un système de reconnaissance de la parole à un mot effectivement prononcé. Nous nommons cette liste une *liste de*

confusion. Elle se compose des mots les plus proches du mot analysé selon une mesure de similarité qui s'appuie sur la combinaison des similarités cosinus des plongements de types linguistique et acoustique.

Les plongements linguistiques de mots correspondent à la combinaison par analyse en composante principale de différents types de plongement de mots : *word2vecf* (Levy & Goldberg, 2014), *skip-gram* fournis par *word2vec* (Mikolov *et al.*, 2013), et *GloVe* (Pennington *et al.*, 2014), comme décrit dans (Ghannay *et al.*, 2016).

Les plongement acoustiques de mots correspondent à la projection de séquences acoustiques de longueur variable dans un espace de faible dimension de telle sorte que les mots qui se prononcent de la même manière sont projetés dans la même zone, tandis que les mots qui se prononcent différemment sont projetés dans des zones différentes. L'approche que nous avons utilisée pour construire ces représentations s'inspire de celle proposée dans (Bengio & Heigold, 2014).

2.1 Interpolation linéaire des similarités linguistique et acoustique

Dans cette étude, nous proposons d'utiliser des plongements linguistiques et acoustiques pour prédire les confusions faites par le SRAP. Pour construire une mesure de similarité combinant des plongements de mots de natures différentes, nous proposons d'utiliser l'interpolation linéaire des similarités cosinus linguistique et acoustique. La similarité résultante est appelée $LA_{SimInter}$, et est définie comme suit :

$$LA_{SimInter}(\lambda, w_1, w_2) = (1-\lambda) \times L_{Sim}(w_1, w_2) + \lambda \times A_{Sim}(w_1, w_2) \quad (1)$$

où w_1 et w_2 sont les deux mots à comparer et λ est le coefficient d'interpolation. Les similarités L_{Sim} et A_{Sim} sont calculées avec la similarité cosinus appliquée respectivement aux plongements linguistiques et acoustiques de w_1 et w_2.

Comme notre objectif est de prédire ou corriger les erreurs du SRAP, nous voulons optimiser la valeur λ à cette fin. Pour estimer λ, une liste connue d'erreurs de substitution générées par le SRAP est utilisée. Dans cette liste, nous définissons h comme étant l'hypothèse de mot erronée et $\overline{r}$ le mot de référence qui a été substitué par h. Pour chaque paire de mots $(h, \overline{r})$ dans la liste, nous calculons la probabilité que le mot h soit reconnu lorsque le mot de référence $\overline{r}$ est erroné : $P(h|\overline{r}) = \frac{\#(h,\overline{r})}{\#\overline{r}}$, où $\#(\mathrm{h}, \overline{r})$ est le nombre de substitutions de $\overline{r}$ par h et $\#\overline{r}$ le nombre d'erreurs sur le mot de référence $\overline{r}$.

Nous proposons alors de retenir le coefficient d'interpolation $\hat{\lambda}$ qui minimise l'erreur quadratique moyenne (MSE) entre la valeur proposée par $LA_{SimInter}(\lambda, h, \overline{r})$ et la valeur effective de $P(h|\,\overline{r})$. Nous définissons alors $\hat{\lambda}$ tel que :

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, MSE(\forall (h, \overline{r}) : P(h|\overline{r}), LA_{SimInter}(\lambda, h, \overline{r})) \quad (2)$$

où $LA_{SimInter}(\lambda, h, \overline{r})$ et $P(h|\,\overline{r})$ sont calculés sur tous les couples $(h, \overline{r})$ possibles.

En utilisant $LA_{SimInter}$ avec $\hat{\lambda}$, il est maintenant possible de proposer pour un mot donné sa liste de confusion contenant ses voisins les plus proches linguistiquement et acoustiquement. La valeur de $LA_{SimInter}(\hat{\lambda}, x, y)$ est considérée comme une mesure de similarité entre les mots x et y et nous la notons plus simplement $confus(x, y)$.

2.2 La simulation d'erreurs

Pour simuler des erreurs de reconnaissance, on applique la mesure de similarité $confus(x, y)$ afin de substituer dans la transcription manuelle des mots corrects par des mots erronés de la liste de

confusion.

En déterminant un taux d'erreur e (uniquement composés de substitutions), on modifie aléatoirement un pourcentage e des occurrences de mot. Ces substitutions sont faites après avoir défini deux seuils : le seuil c qui réfère la valeur la plus basse de $confus(\overline{r}, h)$ qui permet de substituer le mot $\overline{r}$ par le mot h, et le seuil n qui limite le nombre de substitutions possibles de $\overline{r}$ parmi les n mots h_i les plus proches (*i.e.* les mots h_i tels que la valeur $confus(\overline{r}, h_i)$ est l'une des n valeurs les plus hautes étant donné $\overline{r}$). Le mot h est choisi aléatoirement dans la liste des mots h_i qui respectent les contraintes des seuils n et c.

3 Protocole Expérimental

Cette partie décrit le protocole expérimental inspiré d'une étude précédente (Simonnet *et al.*, 2017).

3.1 Le corpus MEDIA

Le corpus utilisé est le corpus MEDIA, collecté dans le projet français Media/Evalda (Bonneau-Maynard *et al.*, 2005). Il contient trois ensembles de dialogues téléphoniques humain/ordinateur liés au tourisme, à savoir : un ensemble d'apprentissage (APP) avec environ 17,7k phrases, un ensemble de développement (DEV) avec 1,3k phrases et un ensemble d'évaluation (TEST) contenant 3,5k phrases. Le corpus a été annoté manuellement avec des concepts sémantiques caractérisés par une étiquette et sa valeur. Les évaluations sont effectuées avec les ensembles DEV et TEST et rapportent les taux d'erreur CER (concept error rate) pour les étiquettes de concepts seulement et les taux d'erreur CVER (concept-value error rate) pour les paires étiquette-valeur. Il est à noter que le nombre de concepts annotés dans une phrase a une grande variabilité et peut inclure plus de 30 concepts annotés.

Pour ces expériences, une variante du SRAP développé par le LIUM est utilisée. Elle a remporté la dernière campagne d'évaluation sur la langue française (Rousseau *et al.*, 2014). Ce système est basé sur la boîte à outils de reconnaissance vocale Kaldi (Povey *et al.*, 2011). Une description détaillée du SRAP est donnée dans (Simonnet *et al.*, 2017). Les taux d'erreur mot pour les corpus APP, DEV et TEST sont respectivement de 23,7%, 23,4% et 23,6%.

3.2 Descriptions des systèmes de compréhension

Deux systèmes de compréhension sont comparés sur le corpus MEDIA. Le premier est un RNR-EDA similaire à celui utilisé pour la traduction automatique proposé dans (Cho *et al.*, 2014). Le second est basé sur des CRF. Les deux architectures construisent leur modèle d'apprentissage sur le même ensemble de descripteurs en entrée, avec des valeurs continues pour le premier et des valeurs discrètes pour le second.

3.2.1 Descripteurs de mot

Afin d'améliorer les performances de compréhension des systèmes, un ensemble de descripteurs, inspiré de (Hahn *et al.*, 2011), représente chaque occurence de mot en entrée des SCAP. Il s'agit de : le mot, la catégorie sémantique prédéfinie qui peut être spécifique à MEDIA ou plus générale ; des caractéristiques syntaxiques et morphologiques ; et deux mesures de confiance : la probabilité *postérieure* (*pap*) et la mesure de confiance issue d'un perceptron multi-couche. Ces deux dernières caractéristiques estiment la fiabilité du mot reconnu par le SRAP. La description détaillée de ces descripteurs se trouve dans (Simonnet *et al.*, 2017).

Les deux SCAP prennent en entrée tous ces descripteurs à l'exception des mesures de confiance où seulement une est gardée dans un but de cohérence expérimentale comme cela sera décrit dans la sous-section 3.3. Ces architectures doivent également être calibrées sur leurs hyper-paramètres respectifs afin de donner les meilleurs résultats. La façon dont la meilleure configuration est choisie est décrite dans la section 4.

3.2.2 Système de compréhension de la parole basé sur les RNR-EDA

Le RNR-EDA proposé, inspiré d'une architecture de traduction automatique, a été implémenté à partir de l'outil *nmtpy* (Caglayan *et al.*, 2017). L'étiquetage de concept est considéré comme une traduction de mots (langage source) vers étiquettes sémantiques (langage cible). Une description détaillée du RNR-EDA est donnée dans (Simonnet *et al.*, 2017).

3.2.3 Système de compréhension de la parole basé sur les CRF

Les expériences passées décrites dans (Hahn *et al.*, 2011) ont montré que les meilleures performances en annotation sémantique sur les transcriptions manuelles et automatiques du corpus MEDIA ont été obtenues avec les CRF. Plus récemment, dans (Vukotic *et al.*, 2015), cette architecture a été comparée à un RNR bidirectionnel (biRNR). La conclusion fut que les CRF surpassent les biRNR sur le corpus MEDIA, alors que de meilleurs résultats ont été observés par les biRNR sur le corpus ATIS (Hemphill *et al.*, 1990). Ceci s'explique probablement par le fait que MEDIA contient des contenus sémantiques dont les mentions sont plus difficiles à désambiguïser, et les CRF exploitent plus efficacement des contextes complexes ((Vukotic *et al.*, 2015)).

Par soucis de comparaison avec le meilleur SCAP proposé dans (Hahn *et al.*, 2011), la boîte à outils Wapiti (Lavergne *et al.*, 2010) a été utilisée dans notre étude. Néanmoins, l'ensemble des descripteurs utilisés par le système proposé dans cet article est différent de celui utilisé dans (Hahn *et al.*, 2011). Parmi les nouveautés utilisées dans notre système, nous considérons des descripteurs syntaxiques et des mesures de confiance de SRAP et notre modèle de configuration est différent. Après de nombreuses expériences effectuées sur le DEV, notre modèle de descripteur final inclut les instances précédentes et suivantes pour les mots et la catégorie grammaticale dans un unigram ou un bigram afin d'associer une étiquette sémantique avec le mot en cours. De plus sont associés avec le mot courant les catégories sémantiques des deux instances précédentes et des deux suivantes. Les autres descripteurs ne sont considérés qu'à la position courante. De plus, l'outil discretize4CRF [1] est utilisé pour discrétiser les mesures de confiance de SRAP afin qu'elles soient acceptées en entrée des CRF.

3.3 Simulation d'erreurs de transcriptions

La méthode présentée dans la sous-section 2 est appliquée afin de simuler des erreurs de SRAP. À partir des annotations manuelles du corpus MEDIA, nous construisons différents ensembles de données. Dans ces simulations, nous avons fixé la valeur de e à 20%, ce qui représente le taux de mots que nous corrompons au hasard dans les transcriptions manuelles.

Deux simulations différentes ont été testées, en choisissant différentes valeurs de seuil n et c;

— **corpus B.7** : $n = 7$ et $c = 0.4$;

— **corpus B.10** : $n = 10$ et $c = 0.5$.

Un autre ensemble de données artificiel a été créé, appelé **corpus B.n** : ce corpus ne prend pas en compte la mesure de similarité. Dans cet ensemble de données, le même pourcentage de mots $e = 20\%$

1. https ://gforge.inria.fr/projects/discretize4crf/

issus des transcriptions manuelles est substitué de manière aléatoire, en choisissant simplement un mot au hasard dans l'ensemble du vocabulaire MEDIA. Quand un mot correct est remplacé par un mot confondu, nous utilisons la mesure de similarité comme mesure de confiance de SRAP.

Dans un but de cohérence expérimentale, lorsque nous travaillons sur des sorties de SRAP, nous donnons seulement une mesure de confiance parmi les deux disponibles afin d'avoir toujours le même nombre de mesures de confiance dans tous les cas.

4 Résultats Expérimentaux

Pour les deux SCAP, l'apprentissage est fait sur l'APP et les meilleures configurations sont choisies pour optimiser le CVER sur le DEV. Les résultats sur le TEST en CER et CVER sont reportés dans les tables 1 et 2, où **M** fait référence au corpus manuel, **A** à un corpus composé de transcriptions automatiques, et **B** à un corpus bruité. Le TEST est constitué uniquement de transcriptions automatiques, alors que la nature des corpus APP ou DEV varie dans nos expériences.

4.1 Analyse de l'apport des transcriptions bruitées à l'apprentissage

Puisque l'évaluation sur le TEST est faite sur des transcriptions automatiques, nous considérons dans un premier temps qu'un corpus DEV composé de transcriptions automatiques est également disponible. Ce corpus est moins difficile à collecter qu'un corpus d'entraînement (1.3k phrases *vs.* 17.7k) et sa manipulation n'entraine ni biais, ni sur-apprentissage. Les résultats expérimentaux de cette configuration sont visibles dans la table 1.

	APP	M	A	B.7	B.7 x2	M +B.7	M +B.10	M +B.n	M +A	M +B.7+A
	DEV	A	A	A	A	A	A	A	A	A
RNR EDA	*CER*	31,6	22,5	23,8	23,2	22,7	23,3	23,7	20,7	20,2
	CVER	36,2	28,3	29	28,8	28,1	28,5	28,8	25,8	26
CRF	*CER*	27,5	19,9	22,6	26,3	22,6	23,2	25	20,2	29,1
	CVER	31,6	25,1	27,7	31,3	27,7	28,3	30,3	25,3	33

TABLE 1 – Comparaison de différents APP en CER et CVER sur un TEST et un DEV automatique.

Nous pouvons d'abord noter que notre hypothèse sur l'importance d'apprendre sur des données proches des données de test (avec des transcriptions automatiques ou contenant des simulations d'erreurs) est vérifiée : avec l'APP **A**, les résultats des RNR-EDA et des CRF sont significativement meilleurs que ceux fait avec un APP **M**. On voit également que les CRF surpassent significativement les RNR-EDA sus les corpus d'entraînement **M** et **A**. Il est également clair que l'entraînement d'un SCAP sur des transcriptions manuelles est largement insuffisant pour gérer les transcriptions automatiques. Le système doit être préparé aux erreurs de transcriptions.

L'entraînement sur un corpus bruité (colonne B.7) obtient des résultats intéressants. On obtient une nette amélioration par rapport aux mauvais résultats obtenus sur les transcriptions manuelles seulement. Il se rapproche des résultats utilisant les transcriptions automatiques pures et confirme ainsi que notre approche pour simuler des erreurs de transcription est adaptée à cette tâche. Entraîner sur un corpus bruité doublé (colonne double B.7, dans laquelle deux simulations d'erreurs de SRAP successives sur l'APP ont été utilisées) permet d'améliorer un peu les résultats sur le RNR-EDA tout en aggravant fortement ceux des CRF.

De meilleurs résultats peuvent être obtenus en combinant des corpus manuels et bruités. En utilisant l'ensemble de données B.7 combiné au manuel, les résultats sont tout aussi bons que des transcriptions automatiques pures pour le RNR-EDA. Les CRF obtiennent les mêmes résultats que pour B.7 seulement.

Nous pouvons également comparer les différents types de bruit. Le B.7 obtient de meilleurs résultats que le B.10, ce qui montre qu'en substituant des mots corrects à des mots globalement moins semblables, les résultats diminuent. De plus, même si l'application de bruit naïf (B.n) obtient de meilleurs résultats que l'utilisation de transcriptions manuelles (APP M), nous obtenons les plus mauvais scores parmi les approches bruitées. Ceci montre l'importance d'un bruit généré intelligemment, et valide implicitement notre approche de simulation d'erreurs de transcription.

Finalement, les meilleurs résultats obtenus qui surpassent les transcriptions automatiques pures (A) sont obtenus en entraînant les SCAP sur une combinaison de sorties automatiques et manuelles (M+A). Les deux SCAP trouvent leur meilleure performance dans cette configuration et l'écart entre CRF et RNR-EDA a été fortement réduit par rapport aux expériences sur A ou M seulement. L'entraînement sur une triple combinaison de corpus manuel, automatique et bruité n'augmente pas davantage ces résultats.

En général, les CRF surpassent significativement les RNR-EDA lorsque ces systèmes sont entraînés sur un corpus manuel ou automatique. Mais les RNR-EDA tirent meilleur parti de la simulation d'erreurs, ou de la combinaison manuelle et automatique par rapport aux CRF. Au final, les meilleurs résultats des RNR-EDA et CRF sont très proches, montrant un potentiel des réseaux de neurones, non partagé par les CRF, à apprendre des informations pertinentes à partir de données bruitées.

4.2 Apprentissage sans transcriptions automatiques

Dans cette section, nous explorons le scénario dans lequel aucune donnée issue d'un SRAP n'est disponible pour entraîner le système de compréhension (DEV inclus). Cela peut devenir problématique lorsque le système de compréhension doit effectuer des phases de validation durant le processus d'apprentissage, ce qui est le cas des RNR-EDA. Les CRF pour leur part n'utilisent pas le DEV pendant l'entraînement (la configuration optimale n'est pas modifiée et les scores des CRF restent inchangés). Ainsi, les résultats visibles dans la table 2 ne concernent que les RNR-EDA.

	APP	M	B.7	M+B.7
	DEV	M	B.7	B.7
RNR	*CER*	33,9	23,5	23,1
EDA	*CVER*	38,2	28,6	28,5

TABLE 2 – Comparaison en CER et CVER obtenus sur un TEST automatique mais sans données automatiques pour l'APP ou le DEV.

En général, sauf pour l'APP bruité seul, de meilleurs résultats sont atteints en validant sur un DEV automatique, plus proche des données de TEST. Néanmoins, même si ces résultats sont un peu moins bons que ceux obtenus en validant sur un DEV automatique, on peut remarquer qu'il est possible d'améliorer très significativement les performances des SCAP en appliquant notre approche de simulation d'erreurs pour enrichir ou bruiter les données d'apprentissage et de développement des SCAP ne disposant que de transcriptions manuelles.

5 Conclusion

Deux architectures de compréhension de la parole basées sur des RNR-EDA et des CRF ont été comparées dans cette étude. Une simulation d'erreur de transcription basée sur une mesure de similarité construite à partir de plongements de mots acoustiques et linguistiques a été proposée et utilisée pour bruiter un corpus manuel annoté. Les expériences montrent que ce bruitage est pertinent pour enrichir et préparer un corpus d'entraînement de SCAP. Si aucun SRAP n'est disponible pour préparer ces données, notre proposition offre une amélioration très significative des performances des SCAP, de 36,2% de CVER avec seulement des annotations manuelles dans le corpus d'entraînement, contre 28,5% de CVER en appliquant notre approche : ceci représente une réduction relative de 21,2% des erreurs en concept-valeur. Un autre résultat intéressant dans cette étude est la diminution des écarts, en terme de CER ou de CVER, entre CRF et RNR-EDA sur le corpus MEDIA. Aucun changement n'a été fait sur ce corpus depuis 2011 (Hahn *et al.*, 2011) et les CRF sont toujours dominants. Nos résultats montrent qu'il est maintenant possible d'obtenir des résultats similaires avec une architecture neuronale. Nous nous attendons à proposer de nouvelles contributions pour rendre les réseaux de neurones plus efficaces que les CRF, qui ont atteint un plateau il y a plusieurs années sur cette tâche. Dans un avenir proche, nous considérerons également d'autres approches de simulation d'erreurs de SRAP pour comparer leur impact aux nôtres afin de préparer et d'enrichir le corpus d'entraînement des SCAP. Nous expérimenterons également l'utilisation de notre simulation de SRAP sur d'autres tâches, comme la détection d'erreurs de SRAP par exemple.

Références

BENGIO S. & HEIGOLD G. (2014). Word embeddings for speech recognition. In *INTERSPEECH*, p. 1053–1057.

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *Ninth European Conference on Speech Communication and Technology*.

CAGLAYAN O., GARCÍA-MARTÍNEZ M., BARDET A., ARANSA W., BOUGARES F. & BARRAULT L. (2017). Nmtpy : A flexible toolkit for advanced neural machine translation systems. *arXiv preprint arXiv :1706.00457*.

CHO K., VAN MERRIENBOER B., GULCEHRE C., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

FOSLER-LUSSIER E., AMDAL I. & KUO H.-K. J. (2002). On the road to improved lexical confusability metrics. In *ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*.

GHANNAY S., FAVRE B., ESTEVE Y. & CAMELIN N. (2016). Word embedding evaluation and combination. In *of the Language Resources and Evaluation Conference (LREC 2016), Portoroz (Slovenia)*, p. 23–28.

HAHN S., DINARELLI M., RAYMOND C., LEFEVRE F., LEHNEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2011). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(6), 1569–1583.

HAKKANI-TÜR D., TUR G., CELIKYILMAZ A., CHEN Y.-N., GAO J., DENG L. & WANG Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association*.

HEMPHILL C. T., GODFREY J. J., DODDINGTON G. R. *et al.* (1990). The atis spoken language systems pilot corpus. In *Proceedings of the DARPA speech and natural language workshop*, p. 96–101.

JYOTHI P. & FOSLER-LUSSIER E. (2010). Discriminative language modeling using simulated asr errors. In *Eleventh Annual Conference of the International Speech Communication Association*.

LAFFERTY J., MCCALLUM A., PEREIRA F. *et al.* (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, p. 282–289.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

LEVY O. & GOLDBERG Y. (2014). Dependency based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, p. 302–308.

MESNIL G., DAUPHIN Y., YAO K., BENGIO Y., DENG L., HAKKANI-TUR D., HE X., HECK L., TUR G., YU D. *et al.* (2015). Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, **23**(3), 530–539.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12.

PIETQUIN O. & BEAUFORT R. (2005). Comparing asr modeling methods for spoken dialogue simulation and optimal strategy learning. In *Ninth European Conference on Speech Communication and Technology*.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584 : IEEE Signal Processing Society.

ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). LIUM and CRIM ASR system combination for the REPERE evaluation campaign. In *International Conference on Text, Speech, and Dialogue*, p. 441–448 : Springer.

SARIKAYA R., HINTON G. E. & DEORAS A. (2014). Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, **22**(4), 778–784.

SCHATZMANN J., THOMSON B. & YOUNG S. (2007). Error simulation for training statistical dialogue systems. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, p. 526–531 : IEEE.

SIMONNET E., GHANNAY S., CAMELIN N., ESTEVE Y. & RENATO D. M. (2017). ASR error management for improving spoken language understanding. In *INTERSPEECH*.

STUTTLE M., WILLIAMS J. & YOUNG S. (2004). A framework for dialog systems data collection using a simulated asr channel. In *ICSLP 2004*.

VUKOTIC V., RAYMOND C. & GRAVIER G. (2015). Is it time to switch to word embedding and recurrent neural networks for spoken language understanding ? In *InterSpeech*.