Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e 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

Où en sommes-nous dans la reconnaissance des entités nommées structurées à partir de la parole ?

Antoine Caubrière[1] Sophie Rosset[2] Yannick Estève[3] Antoine Laurent[1]
Emmanuel Morin[4]
(1) LIUM, Avenue Olivier Messiaen, 72085 Le Mans ; (2) LIMSI, Rue du Belvedère, 91405 Orsay
(3) LIA, 339 Chemin des Meinajaries, 84140 Avignon ; (4) LS2N, 2 Chemin de la Houssinière, 44322 Nantes
prénom.nom@[univ-lemans ; limsi ; univ-avignon ; univ-nantes].fr

RÉSUMÉ

La reconnaissance des entités nommées (REN) à partir de la parole est traditionnellement effectuée par l'intermédiaire d'une chaîne de composants, exploitant un système de reconnaissance de la parole (RAP), puis un système de REN appliqué sur les transcriptions automatiques. Les dernières données disponibles pour la REN structurées à partir de la parole en français proviennent de la campagne d'évaluation ETAPE en 2012. Depuis la publication des résultats, des améliorations majeures ont été réalisées pour les systèmes de REN et de RAP. Notamment avec le développement des systèmes neuronaux. De plus, certains travaux montrent l'intérêt des approches de bout en bout pour la tâche de REN dans la parole. Nous proposons une étude des améliorations en RAP et REN dans le cadre d'une chaîne de composants, ainsi qu'une nouvelle approche en trois étapes. Nous explorons aussi les capacités d'une approche bout en bout pour la REN structurées. Enfin, nous comparons ces deux types d'approches à l'état de l'art de la campagne ETAPE. Nos résultats montrent l'intérêt de l'approche bout en bout, qui reste toutefois en deçà d'une chaîne de composants entièrement mise à jour.

ABSTRACT

Where are we in Named Entity Recognition from speech ?

Named entity recognition (NER) from speech is usually made through a pipeline process that consists in (i) processing audio using an automatic speech recognition system (ASR) and (ii) applying a NER to the ASR outputs. The latest data available for named entity extraction from speech in French were produced during the ETAPE evaluation campaign in 2012. Since the publication of ETAPE's results, major improvements were done on NER and ASR systems, especially with the development of neural approaches for both of these components. In addition, recent studies have shown the capability of End-to-End (E2E) approach for NER / SLU tasks. In this paper, we propose a study of the improvements made in speech recognition and named entity recognition for pipeline approaches. For this type of systems, we propose an original 3-pass approach. We also explore the capability of an E2E system to do structured NER. Finally, we compare the performances of ETAPE's systems (state-of-the-art systems in 2012) with the performances obtained using current technologies. The results show the interest of the E2E approach, which however remains below an updated pipeline approach.

MOTS-CLÉS : Reconnaissance d'entités nommées structurées, Reconnaissance automatique de la parole, Chaînes de composants, bout en bout.

KEYWORDS: Named Entity Recognition, Automatic Speech Recognition, Pipeline, End-to-End.

Traduction de l'article accepté à LREC 2019 : "Where are we in Named Entity Recognition from speech ?"

Volume 1 : Journées d'Études sur la Parole, pages 64–72.
HAL : hal-02798516.

1 Introduction

La reconnaissance d'entités nommées (REN) consiste en la localisation de concepts, dans des textes non structurés, et en leurs classifications dans des catégories prédéfinies. Le projet Quaero (Grouin *et al.*, 2011) a permis la mise en place d'une définition étendue des entités nommées (EN) dans le cadre de données françaises. Cette définition possède une structure arborescente multiniveau au sein de laquelle différentes EN sont combinés pour définir les plus complexes. Dans le but de mieux décrire les entités nommées, le projet Quaero met en place la notion de composant d'EN. Ainsi, avec cette définition, la REN consiste en la localisation, la classification et la décomposition des entités. La campagne d'évaluation ETAPE (Galibert *et al.*, 2014) a utilisé cette définition étendue.

À notre connaissance, aucun nouveau résultat n'a été publié depuis les résultats de la campagne ETAPE concernant la REN structurées à partir de la parole sur des données françaises. En raison de leurs structures, la reconnaissance de ces types d'EN ne peut pas être abordée comme une simple tâche d'étiquetage de séquences. Lors de la campagne ETAPE, l'état de l'art était construit à l'aide de traitement en plusieurs étapes avant de reconstruire la structure arborescente des EN. Les CRF (Conditional Random Field (Lafferty *et al.*, 2001)) sont au cœur des approches d'étiquetage de séquence. Certaines approches utilisent, en plus des CRF, des grammaires probabilistes sans contexte (Johnson, 1998) (PCFG) leur permettant de mettre en œuvre un modèle en cascade. Les CRF sont appris sur les composants d'EN et les PCFG sont utilisés pour prédire l'ensemble de structure en arbre des EN. Toutefois, le système de REN ayant remporté la campagne ETAPE n'utilise que des CRF avec un modèle par concept (Raymond, 2013). La plupart des approches pour la REN à partir de la parole utilisent une chaîne de deux composants. Tout d'abord un système de reconnaissance de la parole (RAP) produisant des transcriptions automatiques, puis un système de REN est appliqué dessus. Dans cette configuration, le système de REN est appliqué sur des transcriptions automatiques et donc imparfaites. Cela signifie que la qualité des transcriptions a un impact important sur les performances finales de la chaîne (Ben Jannet *et al.*, 2015). En 2012, les systèmes basés sur les modèles de Markov cachés et les modèles à mélange de gaussienne (HMM-GMM) constituaient l'état de l'art en RAP. Depuis, les approches neuronales ont montré leur potentiel (Tomashenko *et al.*, 2016) en se basant sur une combinaison d'HMM et de réseau de neurones profonds (DNN). Les approches neuronales se sont également illustrées dans le cadre de la REN par la combinaison de CRF et de couches neuronales de type bLSTM (Lample *et al.*, 2016; Ma & Hovy, 2016). Dernièrement, une approche de bout en bout (E2E) a été proposée dans (Ghannay *et al.*, 2018) pour la REN directement à partir de la parole. Cette approche laisse un système apprendre l'alignement entre la parole et sa transcription manuelle enrichie avec des concepts d'EN non structurés. D'autres travaux utilisent des approches de bout en bout pour faire correspondre directement la parole aux concepts au lieu de la faire correspondre à des mots, puis ces mots aux concepts (Lugosch *et al.*, 2019). Ces travaux montrent l'intérêt grandissant des approches E2E pour ce type de tâche.

Dans cet article, nous proposons une étude des améliorations récentes pour la tâche de REN dans le cadre de la campagne d'évaluation ETAPE. Nous comparons une approche traditionnelle par chaîne de composants à une approche de bout en bout apprise selon deux stratégies. Notre première contribution consiste en une implémentation en trois étapes permettant d'aborder la REN structurées. Avec cette implémentation, nous séparons la structure arborescente en trois parties distinctes pour les appréhender comme différentes tâches d'étiquetage de séquence plus simple avant de reconstruire la structure arborescente. La seconde est une approche E2E pour la REN structurées. Elle consiste en l'apprentissage de l'alignement entre la parole et les transcriptions textuelles enrichies directement avec les EN structurées.

Nous commençons par une description de la tâche de REN structurées Quaero. Puis, nous décrivons notre implémentation en trois étapes, suivi de la description de nos systèmes état de l'art pour les tâches de RAP / REN dans le cadre d'une chaîne de composant sections 3.2, 3.3 et 3.4, de notre système de bout en bout section 4, de nos jeu de données sections 5 et 6, de nos expérimentation et l'analyse de nos résultats section 7, puis nous concluons.

2 Définition de la tâche

Dans ce travail, nous étudions la REN structurées suivant le formalisme d'annotation Quaero (Rosset *et al.*, 2011). Il permet une annotation suivant huit catégories principales : *amount*, *event*, *func*, *loc*, *org*, *pers*, *prod* and *time*. Ces catégories sont enrichies de sous-types dans le but de créer une hiérarchie. Elle permet de décrire davantage les concepts. Ce guide permet ainsi une annotation selon 39 concepts différents par exemple : *loc.add.phys* qui représente une adresse physique.

En complément des types d'EN, le guide Quaero met en place la notion de composants. Ils permettent de décrire davantage les EN et sont au nombre de 28. Le guide impose que chaque mot présent au sein d'une EN soit annoté en composant, sauf dans le cas de certains articles et mots de liaison. Presque tous les composants dépendent directement du type d'EN, par exemple : "day", "week" sont dépendant du type "time". Cependant certain sont transversaux, par exemple : "kind", "qualifier".

L'annotation finale en EN structurées est arborescente, ce qui signifie qu'une EN peut être composée de composants, mais aussi d'autre EN, sans limites d'imbrication. En suivant la méthode d'annotation proposée par le guide Quaero, la phrase : "la mairie de paris", serait annotée ainsi : "la <**org.adm** <**kind** mairie > de <**loc.adm.town** <**name** paris > > >".

3 Systèmes à chaîne de composants

3.1 Implémentation en trois étapes

Pour réaliser la REN structurées, les systèmes que nous mettons en place utilisent le format d'annotation BIO. Il consiste en un fichier à deux colonnes, la première pour un mot, la seconde pour le concept associé à ce mot. Le concept est préfixé par un "B", un "I" ou un "O", en fonction de la position du mot au sein du groupe de mots correspondant à ce concept. Un inconvénient de ce format est qu'il ne permet pas de représenter efficacement l'arborescence des EN structurées. En effet, cette arborescence implique qu'un mot puisse appartenir à plusieurs concepts. Le format BIO impose un unique label pour chaque mot. Pour représenter l'arborescence, il est nécessaire de concaténer les labels BIO obtenus avec les différents concepts associés à un mot. Nous illustrons cette concaténation dans la figure 1.

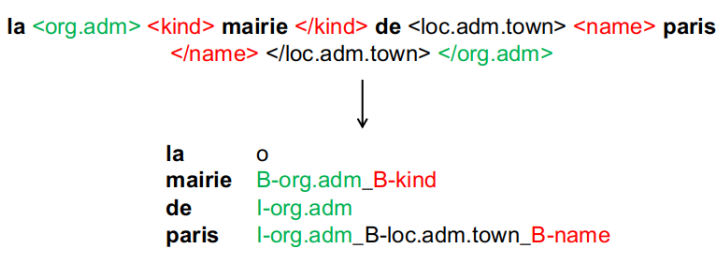

FIGURE 1 – Exemple de transformation d'une séquence d'EN arborescente au format BIO

Ainsi, le nombre final de labels possibles augmente drastiquement. Nous dénombrons désormais 1 690 labels prédictibles. À partir de la notion de composant du guide Quaero, nous pouvons déduire que la racine de l'arborescence est nécessairement un type EN. Nous pouvons également observer que ses feuilles sont en très grande majorité des composants. Enfin, les annotations entre la racine et les feuilles sont un mélange de composants et de type EN. L'augmentation drastique du nombre de labels possibles et ces observations nous motivent à mettre en place une annotation en trois niveaux :

— Le premier niveau représente la racine de l'arborescence. C'est-à-dire les concepts de plus haut niveau. Il est représenté par 96 labels distincts.
— Le troisième niveau représente les feuilles de l'arborescence. Il s'agit ainsi des concepts de plus bas niveau. Un total de 57 labels distincts sont nécessaires pour ce niveau.
— Le second niveau représente l'ensemble des concepts présents entre la racine et la feuille de l'arborescence. Pour ce niveau, nous utilisons 187 labels distincts.

Suite à ce découpage, nous proposons d'effectuer la REN structurées par l'intermédiaire de trois systèmes d'étiquetage de séquences fonctionnant de concert. Nous apprenons un système par niveau. Puis, avec la prédiction de chacun des systèmes nous pouvons reconstruire l'arborescence et ainsi obtenir la séquence finale annotée en EN structurées. Nous souhaitons aussi exploiter les liens existants entre les types EN et les composants. Ainsi, nous proposons d'injecter comme données additionnelles, les prédictions issues d'un modèle précédent dans le/les modèles suivant. C'est-à-dire, injecter les prédictions du premier niveau dans le second et le troisième et injecter les prédictions du second niveau dans le troisième. L'approche que nous proposons est représentée par la figure 2.

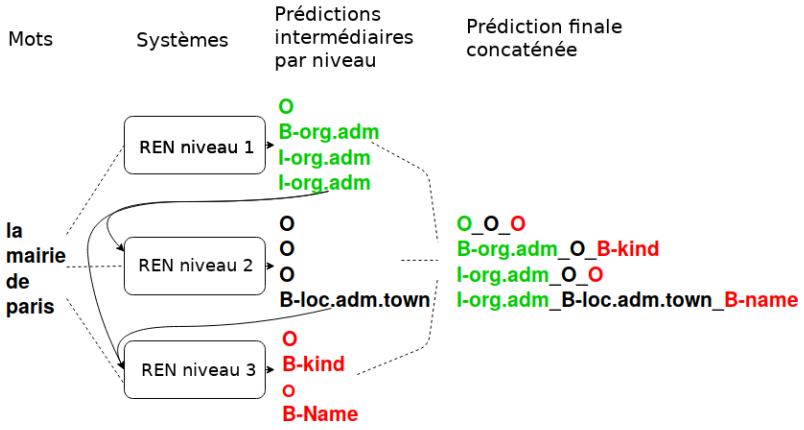

FIGURE 2 – Représentation de notre implémentation en 3 étapes

3.2 CRF

Le premier système de REN utilisé dans ces travaux est un CRF. Nous entraînons les modèles à l'aide du logiciel WAPITI (Lavergne *et al.*, 2010). Ils sont basés sur diverses caractéristiques :

— Les mots et les bigrammes des mots localisés autour des mots cibles sur une fenêtre [-2,+2].
— Les préfixes et les suffixes localisés autour des mots cibles sur une fenêtre [-2,+2].
— Plusieurs caractéristiques de types Oui / Non comme la présence de chiffre dans le mot ou la présence d'une majuscule comme première lettre.

En complément, nous utilisons des caractéristiques morphosyntaxiques extraites de la sortie de l'outil tree-tagger, ainsi que les hypothèses des modèles précédents (implémentation en 3 étapes). Tous les modèles sont entraînés à l'aide de l'algorithme rprop pour un maximum de 40 époques.

3.3 NeuroNlp2

Il s'agit d'un système d'étiquetage de séquences [1] proposé par (Ma & Hovy, 2016). Nous l'exploitons pour réaliser une tâche de REN à partir du texte. Ce système est un empilement de couches CNN, bLSTM et d'une couche CRF. La couche CNN permet l'extraction de plongements de caractères qui s'additionnent aux plongements de mots en entrée des couches bLSTM. Les vecteurs résultant des couches bLSTM sont placés en entrée du CRF finale. Dans nos expérimentations, nous conservons les paramètres par défaut, excepté le nombre de couches bLSTM, paramétré à 2, et le nombre d'unités par couches paramétré à 200.

3.4 Système de reconnaissance de la parole

Dans cette étude, nous utilisons un système de reconnaissance de la parole traditionnel. Ce système est construit à partir de Kaldi et est composé de modèles de Markov cachés et d'un réseau de neurones de types TDNN (Time-Delay Neural Network).

4 Système de bout en bout

L'implémentation [2] que nous utilisons dans cette étude est basée sur le système de RAP DeepSpeech 2 (Amodei *et al.*, 2016). Son architecture consiste en un empilement de deux couches CNN, cinq couches bLSTM et une couche de sortie Softmax. Ce système est entraîné à l'aide de la fonction de coût Connectionist Temporal Classification (Graves *et al.*, 2006). Cette fonction de coût permet au système d'apprendre l'alignement entre un segment audio et une séquence de caractères à produire. Les caractéristiques d'entrées de ce système correspondent aux log-spectrogrammes des segments audio, calculés sur des fenêtres de 20 ms.

Nous mettons en œuvre notre système E2E tel que nous l'avons proposé dans (Ghannay *et al.*, 2018). Les séquences à produire sont des séquences de caractères constituées de mots et des frontières des EN encadrant leurs valeurs. Comme DeepSpeech 2 produit une séquence de caractères, nous représentons ces frontières par l'intermédiaire de caractères uniques. Notre système va ainsi apprendre l'alignement entre des segments audio et des séquences de caractères enrichies des frontières d'EN. L'exemple annoté de la section 2, deviendrait : "la **$ &** mairie > de **% #** paris > > >".

5 Données pour la reconnaissance d'entités nommées

Nos expérimentations sont réalisées sur le corpus français ETAPE (Gravier *et al.*, 2012). Il est composé de données issues d'émissions de radio et TV françaises enregistrées entre 2010 et 2011. Ces émissions proviennent de quatre sources différentes : France Inter, LCP, BFMTV et TV8. Il contient 36 heures de parole séparées en trois parties distinctes : Développement (7 heures), Entraînement (22 heures) et Test (7 heures). Ce corpus possède une annotation en EN selon le guide Quaero. En complément, nous augmentons nos données d'entraînement à l'aide de l'ensemble Quaero (Grouin *et al.*, 2011). Ainsi, nous ajoutons 100 heures de parole entièrement transcrites et annotées en EN manuellement. Ces données sont aussi issues d'émissions de radio et TV françaises.

1. https ://github.com/XuezheMax/NeuroNLP2
2. https ://github.com/SeanNaren/deepspeech.pytorch

6 Données pour la reconnaissance automatique de la parole

Pour effectuer la tâche de RAP, nous utilisons les ensembles ESTER 1 et 2 , REPERE et VERA. Grâce à la combinaison de ces données, nous atteignons un total de 220 heures de parole nous permettant d'apprendre le modèle acoustique du système de RAP de notre approche par chaîne de composants. Le modèle de langage est appris à l'aide des transcriptions manuelles de ces mêmes données enrichies de textes issus de journaux français. Plus de détails sont présents dans la section 4.2.3 de l'article (Deléglise *et al.*, 2009). Notre système E2E est appris sur ces données, en plus de l'ensemble d'apprentissage d'ETAPE.

7 Expérimentations

Nos expérimentations sont évaluées à l'aide de l'ensemble de test d'ETAPE et de la métrique du Slot Error Rate (SER) (Makhoul *et al.*, 1999). Le système de REN remportant la campagne d'évaluation ETAPE (Raymond, 2013) consistait en une combinaison de 68 modèles CRF binaires. Un modèle par type d'EN et par composants. Ce système couplé au meilleur système de RAP permettait d'obtenir un SER de 59,3 %. Ce qui constitue notre référentiel. Afin d'utiliser des transcriptions automatiques annotées en EN structurées, les annotations manuelles d'EN sont projetées dans les transcriptions produites par les systèmes de RAP. Aussi, comme notre système E2E produit à la fois des mots et des EN, nous supprimons les EN produites pour obtenir les transcriptions automatiques. Pour être comparable aux résultats publiés (Galibert *et al.*, 2014), nous utilisons les scripts d'évaluation et de projection de la campagne.

7.1 Expérimentations par chaîne de composants

Nous comparons les résultats obtenus avec l'utilisation du meilleur système de RAP de la campagne (nommé RAP_{2012}) et notre système de RAP état de l'art (nommé RAP_{2020}). Le système RAP_{2012} obtenait un taux d'erreur sur les mots (WER) de 21,8 %, tandis que notre système RAP_{2020} obtient 16,5 %, représentant un gain relatif de 24,3 %. Nous réalisons des expérimentations avec plusieurs combinaisons des systèmes de RAP / REN dont les résultats sont reportés dans la table 1. Il comporte également des expérimentations comparant notre implémentation en 3 étapes (3-pass) avec une approche classique en une étape (1-pass).

TABLE 1 – Résultats expérimentaux des chaînes de composants

Système	**SER**
Sys 0. Baseline ETAPE 2012	59,3
Sys A. 1-pass – CRF – RAP_{2012}	69,4
Sys B. 3-pass – CRF – RAP_{2012}	59,5
Sys C. 3-pass – CRF – RAP_{2019}	55,0
Sys D. 3-pass – bLSTM-CRF – RAP_{2012}	56,1
Sys E. 3-pass – bLSTM-CRF – RAP_{2019}	**51,1**

Le système le plus simple (A) obtient un taux d'erreur de 69,4 %. Lorsque nous employons notre approche en trois étapes dans la même configuration (B), nous atteignons un SER de 59,5 %, soit un gain relatif de 14,3 % par l'utilisation de cette approche. Ces résultats sont proches de ceux de notre référentiel en utilisant uniquement 3 modèles CRF au lieu de 68. Sans surprise, la qualité des

transcriptions automatique améliore les performances globales. Pour un système de REN à base de CRF, les résultats vont de 59,5 % de SER à 55,0 %, soit un gain relatif de 7,6 % (B et C). Avec un système bLSTM-CRF, les résultats vont de 56,1 % à 51,1 % (D et E), permettant un gain relatif de 8,9 %. La mise à jour du système de REN permet également une amélioration. Avec le système de RAP de 2012, les résultats vont de 59,5 % de SER à 55,0 %, soit un gain relatif de 7,6 % (B et D). Tandis qu'avec le système de 2020, les résultats vont de 55,0 % à 51,1 %, soit un gain relatif de 7,1 %. Avec une chaîne de composants, nous obtenons nos meilleurs résultats en mettant à jour chacun des composants et en employant notre implémentation en 3 étapes (système E).

7.2 Expérimentations de bout en bout

Afin d'apprendre notre approche E2E, nous appliquons la même stratégie que dans nos travaux précédents pour compenser le manque de données audio annotées manuellement en EN (Ghannay *et al.*, 2018). Nous effectuons un entraînement multitâche qui consiste tout d'abord à apprendre un système de RAP, puis un système de REN par transfert d'apprentissage ($RAP -> REN_{struct}$). Pour le transfert d'apprentissage, nous conservons le modèle issu de la tâche de RAP, puis nous poursuivons son entraînement en ciblant la tâche de REN. Comme les labels de sorties changent entre les tâches de RAP et de REN, nous réinitialisons la couche de sortie softmax. L'apprentissage du modèle de RAP est effectué à l'aide des données de la section 6 et le modèle de REN à l'aide de celles décrites dans la section 5. Nos travaux précédents ont montré l'intérêt d'un transfert d'apprentissage piloté par une stratégie de curriculum (Caubrière *et al.*, 2019). Cette stratégie consiste à cibler des tâches organisées de celle considérée la plus générique vers celle considérée la plus spécifique. Comme les EN structurées sont composés de type et de composant d'EN, nous proposons d'exploiter cette stratégie. Pour ce faire, nous proposons d'entraîner la tâche de REN en deux apprentissages successifs. Le premier avec des annotations de type EN uniquement, puis le second avec l'annotation complète, incluant donc les composants. Nous supposons les composants comme plus spécifiques que les types EN puisqu'ils en dépendent directement. Ainsi, nous réalisons la chaîne d'apprentissage suivante : $RAP -> REN_{struct} -> REN_{full}$.

Avec notre système E2E, nous pouvons effectuer un décodage classique de type "Greedy", mais aussi un décodage de type "Beam Search" grâce à un modèle de langage. Nous apprenons un modèle de langage 4-gramme à l'aide des données d'apprentissage ETAPE et QUAERO. Les résultats des deux types de décodages pour les deux chaînes d'apprentissages sont donnés dans la table 2.

TABLE 2 – Résultats expérimentaux de l'approche de bout en bout

Système	**ML**	**SER**
$RAP -> REN_{struct}$	X	62,9
$RAP -> REN_{types} -> REN_{full}$	X	61,9
$RAP -> REN_{struct}$	4-gramme	57,3
$RAP -> REN_{types} -> REN_{full}$	4-gramme	**56,9**

Nos résultats montrent l'intérêt de notre stratégie de curriculum pour la REN structurées, par la réduction du SER de 62,9 % à 61,9 %. Ils montrent également l'intérêt du décodage Beam Search qui nous permet d'obtenir de meilleures performances en réduisant le taux de SER de 62,9 % à 57,3 %. La stratégie de curriculum conserve son utilité et nous permet d'obtenir nos meilleurs résultats en atteignant 56,9 % de SER.

7.3 Comparaison globale

Nous reportons les résultats de notre référentiel, de notre meilleur système E2E et de notre meilleur système par chaîne de composant dans la table 3.

TABLE 3 – Résultats reportés du référentiel et de nos meilleurs systèmes

Système	**SER**
(Sys 0) Baseline ETAPE 2012	59.3
$RAP- > REN_{types}- > REN_{full}$ (4-gramme)	56.9
3-passes – bLSTM-CRF – RAP_{2019}	**51.1**

Notre approche E2E nous permet un gain relatif de 4 % par rapport aux résultats de la campagne ETAPE. Toutefois, nos résultats montrent qu'une approche classique avec notre implémentation en 3 étapes et pour laquelle chaque composants est à jour est bien meilleure. Plaçant le nouvel état de l'art à 51,1% de SER. Cette approche classique nous permet un gain relatif de 13,8 % par rapport à notre référentiel. Enfin, en comparant les résultats de nos deux meilleurs systèmes ensemble, nous pouvons observer un gain relatif de 10,2 % à l'avantage de l'approche par chaîne de composants. Pour mieux comprendre les différences entre notre approche E2E et notre chaîne de composants en 3 étapes, nous comparons leurs réponses pour chaque type d'EN. Nous pouvons noter que les cinq concepts les plus représentés (name, kind, pers.ind, name.first, name.last) bénéficient tous d'une amélioration de reconnaissance autour de 16 %. Nous pouvons cependant noter que certains concepts (year, name.nickname) sont désavantagés par notre approche état de l'art, avec respectivement -21 % et -12,3 % de taux de reconnaissance par rapport à notre approche E2E. Ils restent tout de même peu impactant pour les performances globales, car peu représentés (seulement 95 et 65 concepts de références). Il serait nécessaire, dans de futurs travaux, d'effectuer une étude plus approfondie des différences entre nos deux types d'approches.

8 Conclusion

Dans ces travaux, nous présentons les performances désormais atteignables pour la campagne d'évaluation française ETAPE s'étant déroulée en 2012. Nos expérimentations comparent des approches traditionnelles par chaîne de composants et des approches de bout en bout plus récente. Nous proposons dans ce papier une nouvelle implémentation en trois étapes pour la reconnaissance d'entités nommées structurées, dans le cadre des approches par chaîne de composants. En séparant l'arborescence de ces EN en trois parties distinctes, nous sommes capables de réaliser trois tâches plus simples d'étiquetage de séquences. Cette implémentation nous permet d'obtenir des performances similaires au meilleur système de REN en 2012, avec trois CRF classiques au lieu de 68 CRF binaires. Basé sur nos précédents travaux sur les entités nommées non-structurées avec une approche de bout en bout, nous proposons ce type d'approche pour la reconnaissance des entités nommées structurées. Nous exploitons aussi nos précédents travaux sur le transfert d'apprentissage piloté par une stratégie de curriculum pour obtenir nos meilleurs résultats avec une approche de bout en bout. Nous obtenons un gain relatif de 4 % avec ce type d'approche par rapport aux résultats de la campagne ETAPE. Toutefois, nous obtenons nos meilleurs résultats avec une approche par chaîne de composants entièrement mise à jour et exploitant notre implémentation en 3 étapes. Les résultats expérimentaux montrent un gain relatif de 13,8 % entre les résultats de 2012 et le nouvel état de l'art.

Références

AMODEI D., ANANTHANARAYANAN S., ANUBHAI R., BAI J., BATTENBERG E., CASE C., CASPER J., CATANZARO B., CHENG Q., CHEN G. *et al.* (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of ICML'16*, p. 173–182.

BEN JANNET M. A., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015). How to evaluate asr output for named entity recognition ? In *Interspeech*, Dresden, Germany.

CAUBRIÈRE A., TOMASHENKO N., LAURENT A., MORIN E., CAMELIN N. & ESTÈVE Y. (2019). Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *Interspeech*, Graz, Austria.

DELÉGLISE P., ESTEVE Y., MEIGNIER S. & MERLIN T. (2009). Improvements to the lium french asr system based on cmu sphinx : what helps to significantly reduce the word error rate ? In *Interspeech*, Brighton, United Kingdom.

GALIBERT O., LEIXA J., ADDA G., CHOUKRI K. & GRAVIER G. (2014). The ETAPE speech processing evaluation. In *Language Resources Evaluation Conference (LREC)*, Reykjavik, Iceland.

GHANNAY S., CAUBRIÈRE A., ESTÈVE Y., CAMELIN N., SIMONNET E., LAURENT A. & MORIN E. (2018). End-to-end named entity and semantic concept extraction from speech. In *IEEE SLT*.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *ICML*.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC*, Istanbul, Turkey.

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In *Linguistic Annotation Workshop*, p. 92–100, Portland, OR : ACL.

JOHNSON M. (1998). Pcfg models of linguistic tree representations. *Computational Linguistics*.

LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML*.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv :1603.01360*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical Very Large Scale CRFs. In *Annual Meeting of the Association for Computational Linguistics*, p. 504–513.

LUGOSCH L., RAVANELLI M., IGNOTO P., TOMAR V. S. & BENGIO Y. (2019). Speech model pre-training for end-to-end spoken language understanding. In *Interspeech*, Graz, Austria.

MA X. & HOVY E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv :1603.01354*.

MAKHOUL J., KUBALA F., SCHWARTZ R. & WEISCHEDEL R. (1999). Performance measures for information extraction. In *DARPA Broadcast News Workshop*, p. 249–252, Herndon, United States.

RAYMOND C. (2013). Robust tree-structured named entities recognition from speech. In *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, Vancouver, Canada.

ROSSET S., GROUIN C. & ZWEIGENBAUM P. (2011). Entités nommées structurées : guide d'annotation quaero. limsi–cnrs, orsay, france.

TOMASHENKO N., VYTHELINGUM K., ROUSSEAU A. & ESTÈVE Y. (2016). Lium asr systems for the 2016 multi-genre broadcast arabic challenge. In *IEEE Spoken Language Technology Workshop*.