Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Vers la production automatique de sous-titres adaptés à l'affichage

François Buet et François Yvon
Université Paris-Saclay, CNRS, LISN,
Campus universitaire bât 508, Rue John von Neumann, F - 91405 Orsay cedex
{francois.buet,francois.yvon}@limsi.fr

RÉSUMÉ

Une façon de réaliser un sous-titrage automatique monolingue est d'associer un système de reconnaissance de parole avec un modèle de traduction de la transcription vers les sous-titres. La tâche de « traduction » est délicate dans la mesure où elle doit opérer une simplification et une compression du texte, respecter des normes liées à l'affichage, tout en composant avec les erreurs issues de la reconnaissance vocale. Une difficulté supplémentaire est la relative rareté des corpus mettant en parallèle transcription automatique et sous-titres sont relativement rares. Nous décrivons ici un nouveau corpus en cours de constitution et nous expérimentons l'utilisation de méthodes de contrôle plus ou moins direct de la longueur des phrases engendrées, afin d'améliorer leur qualité du point de vue linguistique et normatif.

ABSTRACT

Towards automatic adapted monolingual captioning

A possible manner to achieve automatic monolingual closed captioning is to pair an Automatic Speech Recognition (ASR) system with a Machine Translation model turning transcripts into subtitles. The "translation" task is difficult as it must simplify and compress the text, observe norms with respect to display, as well as handle ASR errors. An added difficulty is the relative scarcity of parallel datasets pairing automatic transcripts and subtitles. We describe here the on-going process of corpus collection and we experiment the use of direct or indirect control of the output sentences, in order to improve their quality from a linguistic and normative point of view.

MOTS-CLÉS : Sous-titrage automatique, simplification de textes, traduction automatique.

KEYWORDS: Automatic subtitling, text simplification, machine translation.

1 Introduction

Du fait de l'augmentation générale des sources de contenu audio-visuel, du besoin de leur assurer une diffusion large (en d'autres langues) et des obligations légales concernant l'accessibilité de ces contenus, la production automatique de sous-titres monolingues ou traduits est aujourd'hui un champ d'application très actif du traitement automatique des langues.

La création de sous-titres monolingues nécessite en général d'effectuer une simplification du contenu, de manière à rendre le texte plus abordable pour les lecteurs potentiels. Ceux-ci sont en effet susceptibles de ne pas parfaitement maîtriser la langue écrite ; il peut s'agir par exemple de personnes

Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

ayant une autre langue maternelle, ou bien de personnes sourdes ou malentendantes locutrices de la langue des signes (pour qui l'écrit est assimilable à une langue étrangère) (Daelemans et al., 2004). De plus, les sous-titres doivent satisfaire des contraintes spatiales (les tronçons de phrases doivent rentrer dans la largeur du moniteur, sans trop obstruer le champ de vision) et temporelles (le texte doit être approximativement synchronisé avec les paroles ou l'image, et doit rester affiché suffisamment longtemps pour permettre une lecture confortable à l'écran [1]).

Récemment, les modèles neuronaux ont apporté des avancées significatives dans le domaine de la traduction automatique (Bahdanau et al., 2015; Vaswani et al., 2017), avant d'être adaptés au domaine de la « traduction monolingue », et ont notamment été utilisés pour des tâches de simplification (Zhang et al., 2017; Zhang & Lapata, 2017) et de compression de phrases (Rush et al., 2015; Takase & Okazaki, 2019). Toutefois, ces méthodes demandent de grandes quantités de données parallèles représentant la transformation attendue pour pouvoir être mises en œuvre avec succès. Pour les applications de sous-titrage, les ressources de ce type sont encore relativement lacunaires (Karakanta et al., 2020b).

Nous décrivons ici un nouveau corpus [2] associant des transcriptions automatiques et des sous-titres en français, obtenu à partir du traitement automatique de programmes télévisés contemporains. Ce corpus est utilisé pour mettre en place une chaîne de traitements capable de produire sans intervention humaine le fichier de sous-titres correspondant à une entrée vidéo. Les expériences décrites ici s'intéressent en particulier à l'usage de mécanismes pour contrôler la longueur (Kikuchi et al., 2016; Takase & Okazaki, 2019), et par extension le taux de compression et le débit des phrases engendrées. Nous comparons également différentes stratégies pour mieux contrôler la segmentation du texte en tronçons compatibles avec les normes d'affichage, en fonction du type d'émissions à sous-titrer.

Les questions que nous étudions sont les suivantes :

— Les mécanismes numériques de contrôle de la longueur sont-ils effectifs ?
— Est-il possible d'obtenir un meilleur contrôle en utilisant un marquage symbolique de l'entrée ?
— Est-il utile d'introduire des distinctions entre les émissions qui sont sous-titrées en direct et celles qui sont sous-titrées en post-production ?

Nos expériences mettent en particulier en évidence que les méthodes neuronales utilisées permettent (a) de corriger une partie des erreurs de la reconnaissance vocale ; (b) de calculer des sous-titres respectant globalement les normes d'affichage sans qu'il soit besoin d'expliciter les contraintes de longueur que ces normes imposent.

2 Corpus et métriques

2.1 Corpus

Nous avons à disposition un ensemble de vidéos, assorties de fichiers de sous-titres professionnels, correspondant à des programmes télévisés récemment diffusés en France. Le panel d'émissions qui

1. La *Charte relative à la qualité du sous-titrage à destination des personnes sourdes ou malentendantes* du CSA préconise une fréquence moyenne d'affichage des caractères aux alentours de $12 - 15$ car/s, et un écart maximum de 10 s entre le discours et le sous-titre correspondant (https://www.csa.fr/content/download/20043/334122/version/3/file/Chartesoustitrage122011.pdf, consultée le 14/01/21).

2. La question de la diffusion de cette ressource est délicate : elle appartient au diffuseur pour la partie sous-titre, la propriété des enregistrements étant répartie sur les multiples acteurs de la chaine de production. La question de sa diffusion partielle ou complète n'a pas été décidée et ne nous appartient pas. Le corpus traité continue d'évoluer et de s'accroitre.

nous est fourni a été choisi de manière à représenter diverses catégories (dessin animé, documentaire, fiction, jeu, journal, magazine, politique, vulgarisation).

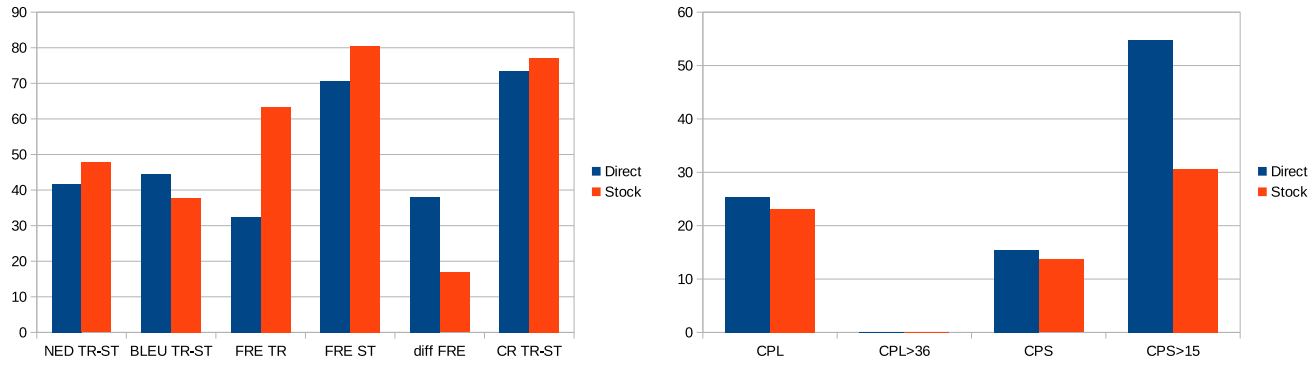
Les instances de programmes collectées, qui arrivent au fur et à mesure des diffusions, sont transcrites automatiquement (mot-pour-mot) en utilisant le système VoxSigma développé conjointement par Vocapia Research et le LIMSI [3]. Ce système délivre des performances à l'état de l'art pour la transcription du français, avec un taux d'erreur proche de 10 % pour de la parole préparée, correspondant par exemple à la transcription de journaux radio-télévisés. Il produit des transcriptions automatiques segmentées automatiquement sur la base d'indices prosodiques et acoustiques (silences, changements de locuteurs, etc) ; les transcriptions sont ponctuées automatiquement, et elles respectent principales les règles typographiques (majuscule en début de phrase, pour les noms propres, etc.). Le texte ainsi obtenu est alors aligné avec celui des sous-titres, afin de pouvoir reconstituer des paires de phrases. Nous avons décidé d'utiliser la segmentation calculée par le système de transcription automatique comme base de l'alignement ; ces segments sont assez longs (environ 40 mots en moyenne), et généralement, correspondent à plusieurs tronçons de sous-titres (voir le Tableau 1). Lors de la réalisation des expériences de la section 3, le corpus contenait environ 411 000 paires de phrases, soit environ 17 millions de mots transcrits, et près de 1600 heures de vidéo. Toutefois, après un filtrage selon la qualité de l'alignement, seulement 265 000 paires ont été utilisées pour l'apprentissage des modèles.

Une distinction notable peut être faite entre les sous-titres provenant d'émissions diffusées en *direct*, et ceux provenant d'émissions de *stock*. Dans le premier cas, les sous-titres sont produits pendant la diffusion, alors que dans le second cas, les sous-titres sont préparés en amont de la diffusion. Ces deux classes sont équitablement réparties dans le corpus (54 % direct et 46 % stock). La figure 1 met en lumière leurs différences selon plusieurs métriques (dont certaines sont détaillées dans la Section 2.2). La transcription et les sous-titres sont globalement plus simples pour les émissions de stock (le score de lisibilité FRE est plus élevé pour la transcription et les sous-titres stock) : cela correspond notamment au fait que certaines de ces émissions contiennent moins d'interventions spontanées, qui forment des énoncés moins structurés et plus longs (les pauses étant plus fréquentes et plus appuyées dans les discours préparés). La différence de FRE (nettement plus grande pour le direct) et le taux de compression (sensiblement plus faible pour le direct) entre la transcription et les sous-titres suggèrent une simplification plus importante dans le cas du direct. Cependant, la distance d'édition normalisée et le score BLEU montrent que les sous-titres sont plus proches de la transcription pour le direct que pour le stock. En fait, bien qu'opérant davantage de suppressions de mots (par exemple sur des marques de l'expression orale telles que les hésitations ou les répétitions), les sous-titres produits en direct procèdent à relativement moins de réécriture que ceux en stock. Il apparaît également qu'en dépit de la compression plus forte, les sous-titres produits en direct sont de façon générale plus denses (le nombre caractères par ligne (CPL) et le nombre de caractères par seconde (CPS) sont légèrement plus faibles pour le stock, et la recommandation de 15 car/s est bien moins souvent respectée dans les sous-titres en direct).

3. Voir www.vocapia.com/speech-to-text-technology.html.

TR	Tout au long de la journée, des orages violents, de fortes pluies et quelles conséquences pour la population, faisons le point ce soir sur cette soudaine montée des eaux et sur les vents violents qui ont soufflé cet après-midi, dans les Bouches-du-Rhône à Marignane et je vous le disais sur la Côte-d'Azur à à Valbonne Vence ou encore à Nice, Alexandre Christophe Larocca.
ST	Des orages violents, de fortes pluies et quelles conséquences pour <p> la population ? <p> Faisons le point sur cette soudaine montée des eaux et sur les vents <p> violents qui ont soufflé cet après-midi... <p>

TABLE 1 – Exemple de segment transcrit automatiquement TR (source) et de segment sous-titre ST (cible) produit par un sous-titreur professionnel dans les conditions du direct. Les balises représentent la segmentation à l'affichage : `<br>` pour un saut de ligne au sein d'un bloc, `<p>` pour une fin de bloc (et changement d'écran).



(a) Métriques de simplification direct / stock

(b) Respect des normes de sous-titrage direct / stock

FIGURE 1 – Comparaison direct / stock au sein du corpus, selon plusieurs mesures portant sur la transcription (TR) ou les sous-titres (ST). NED et CR sont respectivement la distance d'édition normalisée et le taux de compression entre la transcription et les sous-titres (calculés au niveau caractère, et moyennés sur les segments). Les autres métriques sont décrites à la Section 2.2.

2.2 Métriques

Qualité et simplicité des phrases

BLEU (Papineni et al., 2002) est une métrique standard pour la traduction automatique. Xu et al. (2016) ont montré que dans le cas de la simplification, BLEU corrèle les jugements humains pour le sens et la grammaticalité, mais pas pour la simplicité. Nous utilisons l'implantation *SacreBLEU* de Post (2018).

SARI (Xu et al., 2016) compare les opérations d'édition (insertion, copie, suppression de n-gramme) observées entre l'entrée et la sortie, avec celles observées entre l'entrée et les références [4]. Nous utilisons l'implantation de la bibliothèque *EASSE* (Alva-Manchego et al., 2019).

Flesch Reading Ease (FRE) (Flesch, 1948) évalue la lisibilité, en se fondant sur le nombre moyen de mots par phrase et sur le nombre moyen de syllabes par mot. Nous reprenons la formule adaptée au français par Kandel & Moles (1958).

4. N'ayant qu'une seule version de sous-titres pour les émissions, nous ne mesurons SARI qu'avec une référence.

Respect des normes superficielles de sous-titrage

L'affichage de sous-titres nécessite des informations précisant certains aspects de la présentation à l'écran, tels que la segmentation du texte en blocs et en lignes, le temps d'apparition de chaque bloc, la couleur des caractères, ou encore le positionnement horizontal des lignes. Ce formatage doit se conformer à des codes et des normes qui assurent la lisibilité des sous-titres.

Le nombre de caractères par lignes (*CPL*) et le nombre de caractères par seconde (*CPS*, calculé à partir de la durée d'affichage des blocs) sont en particulier soumis à des recommandations. Pour rendre compte du respect de ces contraintes, nous calculons la proportion de lignes dont la longueur dépasse 36 car, *CPL>36*, ainsi que la proportion de blocs qui dépassent une fréquence d'affichage de 15 car/s, *CPS>15* (ces seuils correspondent à des valeurs de référence).

Qualité de la segmentation des sous-titres

Nous reprenons deux métriques proposées respectivement par Matusov et al. (2019) et Karakanta et al. (2020a) pour évaluer la segmentation des sous-titres :

— Nous calculons *BLEU* en conservant les balises de fin de ligne et de fin de bloc dans les prédictions et les références. Cette mesure, que nous notons *BLEU-br*, permet d'évaluer indirectement le positionnement des balises de sous-titrage dans les phrases.

— Nous calculons le score *TER* (Snover et al., 2006) entre la sortie du système et la référence en masquant tous les mots à l'exception des balises de segmentation `<p>` et `<br>`.

Précision du contrôle de longueur

Pour estimer la précision du contrôle de longueur (opéré par les méthodes **LRPE** et **LDPE**, voir Section 3.2), nous avons choisi de calculer l'*erreur absolue moyenne* (EAM) des taux de compression obtenus par rapport aux taux de compression visés :

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|\hat{r}_i - r_i|, \tag{1}$$

où n est la taille de l'ensemble de test, et $\hat{r}_i$ et r_i sont respectivement le taux de compression obtenu et le taux de compression visé pour la i-ème phrase.

L'*erreur absolue* (EA) $|\hat{r} - r|$ peut aussi être vue comme la différence entre la longueur produite et la longueur visée $|l_{\hat{y}} - r \times l_x|$ rapportée à la longueur source l_x. Pour compléter nos métriques, nous avons évalué la proportion d'instances pour lesquelles l'erreur absolue est inférieure à 10 %.

3 Méthodes

Les systèmes de production de sous-titres que nous évaluons ont pour entrée des phrases issues de la transcription automatique des paroles prononcées dans une émission. Nous prenons comme référentiel un système qui conserve telle quelle la transcription mot-à-mot réalisée par l'outil de reconnaissance de parole (ce système est noté **Source** dans le Tableau 2). Nos modèles réalisent

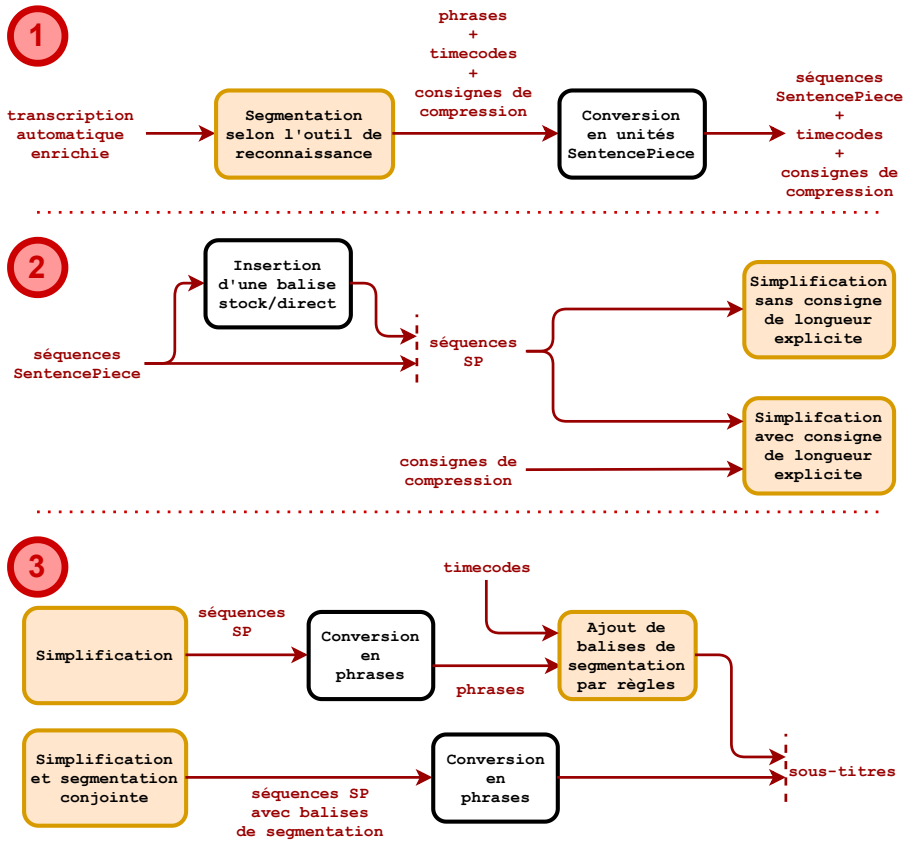


FIGURE 2 – Architecture pour le sous-titrage monolingue

une simplification via l'architecture *Transformer* (Vaswani et al., 2017). Nous expérimentons en plus l'emploi de mécanismes de contrôle de longueur, et l'intégration dans les données de balises pour la segmentation à l'affichage des sous-titres, ou la qualification du type d'émission. Nous pré-traitons les données en utilisant *Sentencepiece* (Kudo & Richardson, 2018), avec un vocabulaire de 16 000 unités.

L'architecture globale des modèles est représentée sur la figure 2.

3.1 Modèles *Transformer* pour la simplification

Pour réaliser la simplification de la transcription nous avons utilisé des modèles à base de *Transformer*, en ré-implémentant l'architecture de Vaswani et al. (2017) (dimension de plongement $d_{\text{model}} = 256$, dimension du perceptron multicouche $d_{\text{ff}} = 1024$, nombre de couches pour l'encodeur / le décodeur $N = 6$, nombre de têtes d'attention $h = 8$). L'optimisation a été faite avec *Adam* (Kingma & Ba, 2015) (avec $\beta_1 = 0,9$, $\beta_2 = 0,98$, $eps = 10^{-9}$). Nous avons également repris la méthode de variation du taux d'apprentissage proposée par Vaswani et al. (2017), en fixant le nombre d'étapes d'échauffement à 4000. Les modèles ont été entraînés sur 265 000 paires de phrases jusqu'à ce que la fonction de perte n'ait pas augmenté pendant 5 époques.

3.2 Contrôle de la longueur

Vaswani et al. (2017) ont défini un encodage positionnel, qui dans le *Transformer* est combiné avec le plongement de chaque mot de la phrase d'entrée (dans la partie encodeur) ou de l'amorce de phrase produite (dans la partie décodeur) :

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right), \tag{2}$$

où pos est la position du mot dans la phrase, et $2i$ (resp. $2i+1$) correspond aux dimensions paires (resp. impaires) de l'encodage. Les modèles avec encodage classique sont notés $\mathbf{Transf}$ (Tableau 2).

Pour contrôler la longueur de la sortie de certains de nos modèles, nous avons ré-implémenté les variantes $\mathbf{LRPE}$ et $\mathbf{LDPE}$ proposées par (Takase & Okazaki, 2019). Ces encodages intègrent une consigne sur la longueur à atteindre l, par ratio ($\mathbf{LRPE}$) ou différence ($\mathbf{LDPE}$) avec la position pos :

$$LRPE_{(pos,l,2i)} = \sin\left(\frac{pos}{l^{2i/d_{model}}}\right), \quad LRPE_{(pos,l,2i+1)} = \cos\left(\frac{pos}{l^{2i/d_{model}}}\right), \tag{3}$$

$$LDPE_{(pos,l,2i)} = \sin\left(\frac{l-pos}{10000^{2i/d_{model}}}\right), \quad LDPE_{(pos,l,2i+1)} = \cos\left(\frac{l-pos}{10000^{2i/d_{model}}}\right). \tag{4}$$

l est égal à la longueur de la séquence cible de référence pendant la période d'entraînement, mais est fixé par l'utilisateur pendant la période de test. $\mathbf{LRPE}$ caractérise à la fois la position courante pos et la longueur totale voulue l, tandis que $\mathbf{LDPE}$ exprime une distance à l'objectif de longueur.

Dans nos expériences, nous avons modulé les objectifs de longueur afin de contraindre les modèles $\mathbf{LRPE}$ et $\mathbf{LDPE}$ à générer des phrases respectant soit un taux de compression constant r (auquel cas l est égale à la longueur de la phrase d'entrée multipliée par r), soit une fréquence d'affichage des caractères constante f (auquel cas l est égale à la durée allouée à l'affichage des tronçons de la phrase multipliée par f).

3.3 Intégration de balises

Dans notre processus de production de sous-titres, le découpage temporel est effectué à partir des périodes de parole identifiées par l'outil de reconnaissance de parole (en permettant à l'affichage de durer quelques secondes supplémentaires pendant les éventuels silences).

Concernant le découpage spatial, notre solution de base est un système à règles implémentant une heuristique simple, qui produit des tronçons de phrases dont la longueur appartient à un intervalle jugé acceptable, et qui favorise la segmentation au niveau des ponctuations. Nous avons appliqué cette méthode avec le référentiel conservant la transcription, et avec la sortie d'un modèle *Transformer* sans contrôle de longueur. Les systèmes résultants sont respectivement notés $\mathbf{Source_R}$ et $\mathbf{Transf_R}$.

L'autre méthode mise en place consiste à intégrer des balises aux emplacements des coupures dans les sous-titres utilisés pour l'apprentissage, comme dans l'exemple du tableau 1. Les systèmes utilisant cette méthode réalisent conjointement la simplification et la segmentation, et sont notés $\mathbf{Transf_B}$, $\mathbf{LRPE_B}$ et $\mathbf{LDPE_B}$.

Enfin, en suivant la littérature sur la génération de phrases contrôlée (Sennrich et al., 2016; Kobus

et al., 2017; Martin et al., 2020), nous avons entraîné un modèle *Transformer* ($\mathbf{Transf_{BT}}$ [5]) en ajoutant au début des phrases sources une balise spécifique qui indique si la phrase cible attendue est un sous-titre de stock ou de direct.

4 Résultats

Le tableau 2 présente les résultats de l'évaluation que nous avons réalisée sur un ensemble de 10 vidéos d'émissions représentatives des programmes traités, pour une durée cumulée d'environ 10 h. Les segments sous-titres de références ont été constitués par alignement automatique avec les phrases de la transcription automatique [6].

Il apparaît que les mécanismes de contrôle de longueur n'améliorent pas la qualité de la simplification, les modèles *Transformer* à encodage positionnel classique obtenant des scores BLEU et SARI meilleurs. Les scores TER-br légèrement inférieurs semblent néanmoins indiquer que **LRPE** et **LDPE** permettent un meilleur positionnement des coupures dans les phrases. Ces modèles respectent aussi de façon plus régulière la norme sur la fréquence d'affichage des caractères (en particulier lorsque l'objectif de longueur est modulé pour suivre une fréquence constante).

L'ajout d'une balise pour spécifier le type d'émission (stock ou direct) semble être bénéfique pour la segmentation, dans la mesure où $\mathbf{Transf_{BT}}$ est meilleur que $\mathbf{Transf_{B}}$ pour TER-br, CPL>36 et CPS>15, tout en étant comparable par ailleurs.

La précision du contrôle de longueur est relative, puisque la différence entre la consigne de longueur et sa réalisation représente en moyenne entre 16 et 20 % de la longueur source (EAM). **LRPE** et **LDPE** sont ici comparables du point de vue de l'effectivité de ce contrôle. Concernant la qualité des phrases produites (SARI, BLEU, BLEU-br), **LRPE** est supérieur à **LDPE**, et la poursuite d'une fréquence de caractères constante semble préférable à l'application d'un unique taux de compression (ce qui paraît effectivement plus proche de ce que ferait un sous-titreur humain).

Afin d'estimer l'importance dans les résultats des erreurs liées à la reconnaissance automatique de parole, nous avons fait réaliser une transcription manuelle (professionnelle) des émissions de notre ensemble de test, et avons évalué certains de nos modèles à partir de cette transcription considérée comme une version « idéale » de la transcription automatique. Nous observons dans ces cas un gain substantiel pour le score BLEU (entre 1,5 et 2 points), mais une baisse de SARI (d'entre 2 et 2,5 points) : les défauts dans la transcription automatique pourraient pousser les systèmes à réaliser certaines simplifications pour produire des phrases plausibles.

À titre de comparaison, Gangi et al. (2019) obtiennent des scores BLEU de l'ordre de 30 sur une traduction anglais-italien à partir d'une transcription automatique, et Matusov et al. (2019) donnent des scores BLEU-br de l'ordre de 40 pour un documentaire et 30 pour une sitcom, pour un sous-titrage multilingue de l'anglais en espagnol (ces tâches sont néanmoins plus difficiles que la nôtre).

Une première observation qui se dégage du Tableau 3 est la grande variabilité des scores selon les émissions. Les variations sont similaires pour les différents modèles : les meilleurs résultats (pour la qualité des phrases au moins) sont obtenus sur la catégorie *journal*, et les moins bons sur la catégorie *jeu*, la catégorie *magazine* étant intermédiaire. Ces écarts s'expliquent en partie par la qualité de la

5. Les balises de segmentation étaient également utilisées avec ce système.

6. Une partie des phrases transcrites (représentant dans l'ensemble $\sim$ 6 % des mots) n'ont pas pu être alignées avec les phrases des sous-titres ; nous avons décidé de les écarter pour l'évaluation.

Modèle	BLEU-br	BLEU	SARI	TER-br	CPL>36	CPS>15	EAM	EA<10 %
Source_R	27,5	34,3	18,1	0,608	0 %	83,1 %	-	-
Cible	100	100	100	0	0 %	46,2 %	-	-
Transf_R	38,1	43,3	52,2	0,390	0 %	63,5 %	-	-
Transf_B	41,5	43,8	52,9	0,381	6,1 %	63,8 %	-	-
Transf^*_B	43,2	46,0	50,5	0,360	6,1 %	53,7 %	-	-
$\text{LRPE}_{B;r=0,75}$	35,6	35,9	49,8	0,351	5,5 %	16,8 %	15,7 %	18,9 %
$\text{LRPE}_{B;f=14,5}$	38,7	39,5	51,2	0,313	5,3 %	1,2 %	20,2 %	28,5 %
$\text{LRPE}^*_{B;f=14,5}$	39,9	41,1	49,3	0,317	5,3 %	1,3 %	21,1 %	25,7 %
$\text{LDPE}_{B;r=0,75}$	34,5	35,2	49,5	0,351	7,0 %	16,3 %	16,5 %	16,3 %
$\text{LDPE}_{B;f=14,5}$	37,3	38,7	50,6	0,316	7,6 %	0,8 %	20,4 %	27,4 %
Transf_{BT}	41,6	43,9	53,1	0,375	4,1 %	62,2 %	-	-
Transf^*_{BT}	42,6	45,3	50,6	0,364	4,1 %	57,6 %	-	-

TABLE 2 – Résultats de l'évaluation des modèles sur un groupe d'émissions de test. Les métriques EAM et EA<10 % ne sont testées que pour les systèmes qui intègrent des objectifs de longueur. Les évaluations de modèles utilisant une version de référence des transcriptions réalisée manuellement sont notées par (*).

transcription automatique : les taux d'erreur par mot (*Word Error Rate*) par rapport à la transcription manuelle ont été estimés à entre 10 et 40 % en fonction des émissions (les journaux obtenant les taux les plus bas, et les jeux les taux les plus hauts). La reconnaissance vocale est notamment affectée par le débit de parole, la clarté de la prononciation, les dialogues avec recouvrement, et généralement la présence de bruits parasites. Par exemple, le jeu télévisé choisi dans notre test contient beaucoup de séquences avec de la musique ou des rires, et des échanges rapides. De plus, la nature du programme fait qu'une partie des phrases ont une structure assez spécifique (énoncé d'une question de culture générale, ou réponse très courte d'un candidat). Nous remarquons enfin que le respect de la norme CPS (fortement lié au débit de parole initial) change significativement d'une émission à l'autre, notamment dans les sous-titres de référence.

Le tableau 4 présente des exemples de transformations réalisées par les systèmes $\mathbf{Transf_B}$ et $\mathbf{LRPE_{B;f=14,5}}$. Nous observons que les modèles ont appris à re-segmenter les phrases (après « émission » dans le premier exemple, « mâts » dans le deuxième, et « majoritaires » dans le troisième), et à reconnaître la forme interrogative (premier exemple). Occasionnellement, ils peuvent également corriger des erreurs de grammaire (« avait », premier exemple). Les conventions orthographiques et typographiques, telles que l'écriture de « % » (troisième exemple) sont globalement gérées efficacement. Le deuxième exemple montre que le modèle $\mathbf{Transf_B}$ élague trop la phrase initiale par rapport à la référence. $\mathbf{LRPE_{B;f=14,5}}$ fait mieux dans ce cas, bénéficiant de l'information de longueur en rapport avec le temps d'affichage disponible. Toutefois, la contrainte de longueur induit souvent des suppressions abruptes, provoquant la perte du sens original (troisième exemple) ou de la cohérence syntaxique. Outre l'abandon d'éléments importants, les erreurs fréquemment commises par les modèles comptent de mauvaises segmentations, et la conservation d'artefacts de la transcription automatique.

En alignant les phrases produites par $\mathbf{Transf_B}$ avec les références de notre ensemble de test, nous avons noté que les opérations d'édition les plus fréquentes étaient les suppressions de mots connecteurs ou introductifs : « et », « que », « c'est », « ça », « donc », « qui »...

Modèle	BLEU-br	BLEU	SARI	FRE	TER-br	CPL>36	CPS>15
Jeu (stock)							
Source_R	24,1	24,0	14,6	68,6	0,68	0 %	74,4 %
Cible	100	100	100	94,2	0	0 %	28,9 %
Transf_B	37,1	31,9	48,7	94,5	0,41	4,7 %	46,9 %
$\text{LRPE}_{B;f=14,5}$	37,5	31,1	48,0	95,1	0,33	4,0 %	4,3 %
Journal (direct)							
Source_R	36,4	49,0	23,0	59,3	0,36	0 %	81,1 %
Cible	100	100	100	82,1	0	0 %	62,3 %
Transf_B	45,6	56,9	57,5	83,2	0,28	5,9 %	72,6 %
$\text{LRPE}_{B;f=14,5}$	40,1	48,4	54,0	84,1	0,29	7,9 %	0,8 %
Magazine (stock)							
Source_R	29,0	33,4	18,3	68,9	0,42	0 %	64,7 %
Cible	100	100	100	94,3	0	0 %	21,0 %
Transf_B	45,8	45,0	52,9	93,3	0,29	5,7 %	43,7 %
$\text{LRPE}_{B;f=14,5}$	41,5	40,4	51,3	94,1	0,25	5,5 %	0,4 %

TABLE 3 – Résultats par émission de l'évaluation des modèles.

5 Travaux connexes

Les besoins d'accessibilité audio-visuelle ainsi que l'exportation de programmes vidéos ont depuis un certain temps suscité un intérêt pour l'automatisation du sous-titrage (Daelemans et al., 2004; Koponen et al., 2020). La procédure fondée sur la mise en cascade d'un système de reconnaissance de parole et d'un système de simplification / compression a longtemps été privilégiée pour le sous-titrage intralingue (Gangi et al., 2019). Récemment, les progrès dans la direction de la traduction de parole sans transcription intermédiaire (Bérard et al., 2016) ont permis l'émergence d'approches bout-en-bout pouvant prendre en compte les indices prosodiques (Karakanta et al., 2020a).

La simplification et la compression de phrases ont abondamment été étudiées pour leur multiples applications, parmi lesquelles la production de sous-titres. Ces tâches ont particulièrement été abordées en reprenant les méthodes de la traduction automatique, des systèmes à règles (Cohn & Lapata, 2008) aux modèles neuronaux (Zhang & Lapata, 2017; Dong et al., 2019).

Pour le sous-titrage automatique, la question du contrôle de la longueur des phrases engendrées est essentielle du fait des contraintes spatiales et temporelles (Angerbauer et al., 2019). Kikuchi et al. (2016) ont mis en place un tel contrôle en introduisant des consignes de longueur dans les états cachés d'un RNN. Takase & Okazaki (2019) ont adapté cette idée à l'architecture *Transformer*, en tirant parti de l'encodage positionnel pour intégrer l'objectif de longueur.

L'ajout d'une étiquette spécifique en début de séquence afin de qualifier un attribut attendu dans la phrase de sortie rejoint une riche littérature de production contrôlée de texte, cette stratégie ayant notamment mise en œuvre pour la formalité (Sennrich et al., 2016), le domaine (Kobus et al., 2017) ou encore la longueur (Lakew et al., 2019; Martin et al., 2020).

TR	**Suite à** votre passage dans **l'émission** comment les élèves à l'époque **avait réagi.**
Transf_B	Suite à votre passage dans **l'émission,** comment les élèves <p> **avaient réagi ?** <p>
$\text{LRPE}_{B;f}$	Suite à votre passage dans **l'émission.** <p> Comment ? <p>
ST	**A** votre passage dans **l'émission,** comment les élèves **avaient réagi ?** <p>
TR	**Donc nous** on monte les mâts **oui mais bon, on va laisser finir**.
Transf_B	On monte les mâts. <p>
$\text{LRPE}_{B;f}$	On monte les mâts. <p> On va laisser finir. <p>
ST	**Nous,** on monte les mâts. <p> **Ah oui.** <p> **Stop.** On va **les** laisser finir. <p>
TR	En Alabama, les anti-avortement sont **majoritaires, 70 pourcent** des habitants se déclarent pour **une interdiction de l'IVG**.
Transf_B	En Alabama, les anti-avortement sont **majoritaires.** <p> **70 %** des habitants se déclarent pour **une interdiction de l'IVG**. <p>
$\text{LRPE}_{B;f}$	En Alabama, les anti-avortement sont **majoritaires.** <p> **70 %** des habitants se déclarent pour **une IVG**. <p>
ST	En Alabama, les anti-avortement sont **majoritaires.** <p> **70%** des habitants se déclarent pour **une interdiction de l'IVG**. <p>

TABLE 4 – Exemples de phrase engendrées par les modèles $\mathbf{Transf_B}$ et $\mathbf{LRPE_{B;f=14,5}}$, comparées à la transcription initiale (TR) et au sous-titre de référence (ST).

6 Conclusion

Nous avons présenté un nouveau corpus pour le sous-titrage automatique, et en avons fait une première utilisation en comparant différentes stratégies pour la production de sous-titres segmentés en vue de l'affichage sur un écran. Dans nos premiers essais, l'implémentation d'un contrôle de longueur améliore le respect de certaines normes superficielles, mais diminue la qualité de la simplification. L'ajout de balises caractérisant le type d'émission pour lequel sont engendrés les sous-titres semble être modérément bénéfique. Compte tenu des variations de résultats observées entre émissions, il pourrait être intéressant à l'avenir d'utiliser des balises pour des catégories d'émissions plus spécifiques. L'apprentissage pourrait être renforcé en intégrant davantage de données, provenant de ressources accessibles publiquement (MOOC ou TedTalk par exemple), ou créées artificiellement à partir de sous-titres sans transcription parallèle (une pseudo-transcription peut être engendrée automatiquement, par rétro-traduction notamment). Enfin il est envisageable de tester des architectures alternatives, adaptées à la traduction monolingue, telles que *Levenshtein Transformer* (Gu et al., 2019), ou encore *Pointer Networks* (Vinyals et al., 2015).

Remerciements

Nous remercions J.-L. Gauvain (LISN, CNRS) pour son aide dans la mise en œuvre des systèmes de transcription automatique, et E. Florence (france.tv access) pour l'accès aux données des émissions. Ce travail a bénéficié de calculs réalisés sur la plateforme LabIA. Ces travaux ont été menés dans le cadre du projet "Rosetta - RObot de Sous-titrage Et Toute Traduction Adaptés", financé par le Programme d'Investissements d'Avenir "Grands défis du numérique" de la Banque Publique d'Investissement (BPI).

Références

ALVA-MANCHEGO F., MARTIN L., SCARTON C. & SPECIA L. (2019). EASSE : easier automatic sentence simplification evaluation. CoRR, **abs/1908.04567**.

ANGERBAUER K., ADEL H. & VU N. T. (2019). Automatic compression of subtitles with neural networks and its effect on user experience. In G. KUBIN & Z. KACIC, Éds., Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, p. 594–598 : ISCA. DOI : 10.21437/Interspeech.2019-1750.

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. BENGIO & Y. LECUN, Éds., 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

BÉRARD A., PIETQUIN O., BESACIER L. & SERVAN C. (2016). Listen and Translate : A Proof of Concept for End-to-End Speech-to-Text Translation. In NIPS Workshop on end-to-end learning for speech and audio processing, Barcelona, Spain. HAL : hal-01408086.

COHN T. & LAPATA M. (2008). Sentence compression beyond word deletion. In Proceedings of the 22nd International Conference on Computational Linguistics, (COLING 2008), p. 137–144, Manchester, UK : Coling 2008 Organizing Committee.

DAELEMANS W., HÖTHKER A. & TJONG KIM SANG E. (2004). Automatic sentence simplification for subtitling in Dutch and English. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal : European Language Resources Association (ELRA).

DONG Y., LI Z., REZAGHOLIZADEH M. & CHEUNG J. C. K. (2019). EditNTS : An neural programmer-interpreter model for sentence simplification through explicit editing. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, p. 3393–3402, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1331.

FLESCH R. (1948). A new readability yardstick. Journal of applied psychology, **32**(3), 221.

GANGI M. A. D., ENYEDI R., BRUSADIN A. & FEDERICO M. (2019). Robust neural machine translation for clean and noisy speech transcripts. CoRR, **abs/1910.10238**.

GU J., WANG C. & ZHAO J. (2019). Levenshtein transformer. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX & R. GARNETT, Éds., Advances in Neural Information Processing Systems, volume 32, p. 11181–11191 : Curran Associates, Inc.

KANDEL L. & MOLES A. (1958). Application de l'indice de Flesch à la langue française. Cahiers Etudes de Radio-Télévision, **19**(1958), 253–274.

KARAKANTA A., NEGRI M. & TURCHI M. (2020a). Is 42 the answer to everything in subtitling-oriented speech translation ? In Proceedings of the 17th International Conference on Spoken Language Translation, p. 209–219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.iwslt-1.26.

KARAKANTA A., NEGRI M. & TURCHI M. (2020b). MuST-cinema : a speech-to-subtitles corpus. In Proceedings of the 12th Language Resources and Evaluation Conference, p. 3727–3734, Marseille, France : European Language Resources Association.

KIKUCHI Y., NEUBIG G., SASANO R., TAKAMURA H. & OKUMURA M. (2016). Controlling output length in neural encoder-decoders. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, p. 1328–1338 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1140.

KINGMA D. P. & BA J. (2015). Adam : A method for stochastic optimization. In Y. BENGIO & Y. LECUN, Éds., 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

KOBUS C., CREGO J. & SENELLART J. (2017). Domain control for neural machine translation. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, p. 372–378, Varna, Bulgaria. DOI : 10.26615/978-954-452-049-6_049.

KOPONEN M., SULUBACAK U., VITIKAINEN K. & TIEDEMANN J. (2020). MT for subtitling : Investigating professional translators' user experience and feedback. In Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation, p. 79–92, Virtual : Association for Machine Translation in the Americas.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

LAKEW S. M., GANGI M. D. & FEDERICO M. (2019). Controlling the output length of neural machine translation. In Proceedings of IWSLT'2019.

MARTIN L., DE LA CLERGERIE É., SAGOT B. & BORDES A. (2020). Controllable sentence simplification. In Proceedings of the 12th Language Resources and Evaluation Conference, p. 4689–4698, Marseille, France : European Language Resources Association.

MATUSOV E., WILKEN P. & GEORGAKOPOULOU Y. (2019). Customizing neural machine translation for subtitling. In Proceedings of the Fourth Conference on Machine Translation (Volume 1 : Research Papers), p. 82–93, Florence, Italy. DOI : 10.18653/v1/W19-5209.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, p. 311–318, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : http ://dx.doi.org/10.3115/1073083.1073135.

POST M. (2018). A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation : Research Papers, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

RUSH A. M., CHOPRA S. & WESTON J. (2015). A neural attention model for abstractive sentence summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, p. 379–389 : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1044.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Controlling politeness in neural machine translation via side constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, p. 35–40, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1005.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas, volume 200 : Cambridge, MA.

TAKASE S. & OKAZAKI N. (2019). Positional encoding to control output sequence length. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 3999–4004, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1401.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. VON LUXBURG, S. BENGIO, H. M. WALLACH, R. FERGUS, S. V. N. VISHWANATHAN & R. GARNETT, Éds., Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, p. 6000–6010.

VINYALS O., FORTUNATO M. & JAITLY N. (2015). Pointer networks. In C. CORTES, N. LAWRENCE, D. LEE, M. SUGIYAMA & R. GARNETT, Éds., Advances in Neural Information Processing Systems, volume 28, p. 2692–2700 : Curran Associates, Inc.

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016). Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics, **4**, 401–415.

ZHANG X. & LAPATA M. (2017). Sentence simplification with deep reinforcement learning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, p. 584–594, Copenhagen, Denmark : Association for Computational Linguistics.

ZHANG Y., YE Z., FENG Y., ZHAO D. & YAN R. (2017). A constrained sequence-to-sequence neural model for sentence simplification. CoRR, **abs/1704.02312**.