Utilisation de mesures de confiance pour améliorer le décodage en traduction de parole

Laurent Besacier, Benjamin Lecouteux, Ngoc Luong Quang

Utilisation de mesures de confiance pour améliorer le décodage en traduction de parole

Laurent Besacier[1] Benjamin Lecouteux[1] Ngoc Luong Quang[1]
(1) LIG, Univ. Grenoble-Alpes, France
laurent.besacier@imag.fr, benjamin.lecouteux@imag.fr, quangngocluong@gmail.com

Résumé. Les mesures de confiance au niveau mot (*Word Confidence Estimation* - WCE) pour la traduction automatique (TA) ou pour la reconnaissance automatique de la parole (RAP) attribuent un score de confiance à chaque mot dans une hypothèse de transcription ou de traduction. Dans le passé, l'estimation de ces mesures a le plus souvent été traitée séparément dans des contextes RAP ou TA. Nous proposons ici une estimation conjointe de la confiance associée à un mot dans une hypothèse de traduction automatique de la parole (TAP). Cette estimation fait appel à des paramètres issus aussi bien des systèmes de transcription de la parole (RAP) que des systèmes de traduction automatique (TA). En plus de la construction de ces estimateurs de confiance robustes pour la TAP, nous utilisons les informations de confiance pour re-décoder nos graphes d'hypothèses de traduction. Les expérimentations réalisées montrent que l'utilisation de ces mesures de confiance au cours d'une seconde passe de décodage permettent d'obtenir une amélioration significative des performances de traduction (évaluées avec la métrique BLEU - gains de deux points par rapport à notre système de traduction de parole de référence). Ces expériences sont faites pour une tâche de TAP (français-anglais) pour laquelle un corpus a été spécialement conçu (ce corpus, mis à la disposition de la communauté TALN, est aussi décrit en détail dans l'article).

Abstract.
Word Confidence Estimation (WCE) for machine translation (MT) or automatic speech recognition (ASR) assigns a confidence score to each word in the MT or ASR hypothesis. In the past, this task has been treated separately in ASR or MT contexts and we propose here a joint estimation of word confidence for a spoken language translation (SLT) task involving both ASR and MT. We build robust word confidence estimators for SLT, based on joint ASR and MT features. Using these word confidence measures to re-decode the spoken language translation graph leads to a significant BLEU improvement (2 points) compared to the SLT baseline. These experiments are done for a French-English SLT task for which a corpus was specifically designed (this corpus being made available to the NLP community).

Mots-clés : Mesures de confiance, traduction automatique de la parole, paramètres joints, re-décodage de graphe.

Keywords: Word confidence estimation (WCE), spoken language translation (SLT), joint features, search graph re-decoding.

1 Introduction

L'estimation de mesures de confiance est un sujet important en reconnaissance automatique de la parole (RAP) ainsi qu'en traduction automatique (TA). En effet, tandis que ces systèmes produisent des sorties dont la qualité ne cesse de croître, une correction (ou post-édition) humaine de ces sorties est le plus souvent nécessaire pour produire des transcriptions ou des traductions parfaites. Ainsi, ces estimateurs de confiance nous permettent de répondre à des questions telles que : est-ce que ces transcriptions / traductions peuvent être publiées telles quelles ? La qualité est elle suffisante pour qu'elles soient corrigées ou mieux vaut-il les retranscrire (re-traduire) à partir de zéro ? Quelles parties de la transcription / traduction doivent être corrigées en priorité ? Pour répondre à ces questions, il est nécessaire de construire un système automatique qui détecte les erreurs sur des segments d'une sortie de TA ou de RAP ; un tel système est appelé estimateur de confiance et génère des mesures de confiance au niveau de chaque segment de texte. Lorsque les segments considérés sont simplement les mots, on parle alors de *mesures de confiance au niveau des mots* (*Word Confidence Estimation - WCE* - en anglais). En plus d'être utiles pour des scénarios interactifs où l'humain participe à la tâche, les mesures de confiance permettent

également de re-ordonner des hypothèses de type N-meilleures (Luong *et al.*, 2014b) ou de re-décoder un graphe de recherche (Luong *et al.*, 2014a) en améliorant les performances.

Un estimateur de confiance au niveau des mots (WCE) assigne donc à chaque mot de l'hypothèse un score (typiquement entre 0 et 1). Plus spécifiquement, la détection d'erreurs consiste à seuiller ce score puis à étiqueter chaque mot comme correct ou incorrect. Pour cela, un système d'étiquetage séquentiel, entrainé sur un ensemble de paramètres, apprend à prédire des labels de type correct (*Good*) ou incorrect (*Bad*). Dans le passé, l'estimation de ces mesures a le plus souvent été traitée séparément dans des contextes RAP ou TA. Nous proposons ici une estimation conjointe de la confiance associée à un mot dans une hypothèse issue d'un système de traduction automatique de la parole (TAP). Cette estimation fait appel à des paramètres issus aussi bien des modules de transcription de la parole (RAP) que des modules de traduction automatique (TA), tous deux nécessaires à une tâche de TAP.

Cet article en français présente des résultats nouveaux (re-décodage d'un graphe de traduction de parole avec mesures de confiance) mais il s'appuie néammoins sur deux publications récentes, en anglais, des mêmes auteurs qu'il convient ici de mentionner :

- une publication à IWSLT 2014 (Besacier *et al.*, 2014) qui présente en détail le corpus sur lequel s'appuie cette étude ; il nous semblait important de le présenter à la communauté TALN francophone et ce corpus est donc décrit à nouveau ici (25% de cette soumission), mais avec un peu moins de détails que dans l'article original en anglais,

- une publication à EAMT 2014 (Luong *et al.*, 2014a) qui présente un algorithme de re-décodage de graphes d'hypothèses de traduction automatique ; cet article ne concernait qu'une tâche de traduction de texte et nous reprenons ici l'algorithme (décrit de façon plus succinte - 25% de cette soumission) en l'adaptant à une tâche de traduction de parole.

Le reste de l'article décrit un résultat original et non encore publié ailleurs qui montre que l'utilisation de mesures de confiance jointes RAP+TA permet d'améliorer significativement les performances d'un système de traduction de la parole.

Cet article est organisé de la façon suivante : la section 2 résume rapidement les principaux travaux antérieurs concernant les estimateurs de confiance au niveau des mots (WCE). Les approches sont présentées séparément entre les tâches de RAP et de TA puisque, à notre connaissance, seule notre précédente publication (Besacier *et al.*, 2014) propose une estimation jointe. Ensuite, le corpus utilisé pour la partie expérimentale est décrit dans la section 3. Les parties 4 et 5 présentent nos systèmes de WCE pour des tâches de transcription et de traduction, respectivement. La section 6 présente, quant à elle, des résultats originaux et montre comment les estimateurs de confiance améliorent les performances sur une tâche de traduction de parole. Pour finir, nous concluons ce travail et donnons quelques perspectives dans la dernière partie.

2 Rapide aperçu des mesures de confiance pour la TA et la RAP

De nombreux travaux ont proposé d'estimer des mesures de confiance afin de détecter automatiquement les erreurs en sortie des systèmes de RAP. Dans ce domaine, ces mesures ont tout d'abord été introduites pour la détection des mots hors-vocabulaire (Asadi *et al.*, 1990). Ces travaux ont ensuité été exploités par (Young, 1994) qui a alors introduit l'utilisation des probabilités *a posteriori* comme mesures de confiance pour la RAP. Ces dernières sont estimées, le plus souvent, en utilisant le graphe (ou treillis) issu de la transcription automatique (Kemp & Schaaf, 1997). Plus récemment, d'autres paramètres sont venus enrichir les probabilités *a posteriori* (Lecouteux *et al.*, 2009) : nombre d'hypothèses concurrentes à un instant donné, paramètres linguistiques ou acoustiques (stabilité du signal, durée des phonèmes, etc.) ainsi que des paramètres sémantiques. Une liste exhaustive des différents paramètres qui peuvent être utilisés est présentée dans (Chase, 1997). Ces différents paramètres peuvent alors êtres classés selon diverses méthodes : des réseaux Bayésiens naïfs (Sanchis *et al.*, 2012), des *Support Vector Machines* (Zhang & Rudnicky, 2001), des réseaux de neurones (Weintraub *et al.*, 1997). Plus récemment, dans (Seigel *et al.*, 2011) et (Seigel & Woodland, 2012) les auteurs combinent les différents paramètres en utilisant des champs aléatoires conditionnels ((Conditionnal Random Fields (CRF) (Lafferty *et al.*, 2001)).

Par ailleurs, l'atelier WMT (*Workshop on Machine Translation*) a introduit en 2013 une nouvelle tâche d'évaluation dédiée aux mesures de confiance appliquées aux systèmes de TA. (Han *et al.*, 2013) (Luong *et al.*, 2013b) proposent d'utiliser des CRFs pour aborder le problème comme un étiquettage de séquence. En parallèle, (Bicici, 2013) a proposé un modèle permettant d'estimer la similarité sémantique entre phrases cibles et sources. Leur modèle, basé sur l'apprentissage global (*Global Learning Model*) est indépendant du moteur de traduction et utilise des paramètres liés à des informations syntaxiques, de contexte ou de forme. (Han *et al.*, 2013) proposent de s'attacher essentiellement aux combinaisons de n-grammes dans la langue cible. Finalement, dans les travaux de (Luong *et al.*, 2013b), l'ensemble des paramètres présentés précédemment sont intégrés en rajoutant des informations sur la topologie du graphe d'exploration, sur des pseudo-

références ou encore des éléments liés à la polysémie et la complexité syntaxique.

A notre connaissance, les premiers travaux proposant des mesures de confiance pour la traduction orale utilisant des paramètres joints entre RAP et TA sont ceux présentés à IWSLT 2014 (Besacier *et al.*, 2014).

3 Notre corpus pour la construction d'estimateurs de confiance en traduction de la parole

Ce travail s'appuie sur un corpus, construit par nos soins, disponible en ligne [1] et déjà décrit dans (Besacier *et al.*, 2014). Nous en présentons les principales traits ci-dessous. Le corpus présente 2643 phrases prononcées en français (3 locuteurs * 881 phrases différentes - 5h au total environ - lectures de phrases de corpus journalistiques), et traduites vers l'anglais. Plus précisément, pour chaque phrase, un quintuplet est disponible : la sortie de transcription (*src-asr*), la transcription de référence ou verbatim (*src-ref*), la traduction automatique de ce verbatim (*tgt-mt*), la traduction automatique de la transcription automatique - c'est à dire la sortie d'un système de traduction de parole (*tgt-slt*) et la post-édition de la traduction (*tgt-pe*).

Concernant les systèmes automatiques, le système de RAP est construit à partir de la boîte à outils KALDI (Povey *et al.*, 2011). Le modèle de langue de type 3-gramme est appris à partir des corpus ESTER (Galliano *et al.*, 2006) et Gigaword français (taille du vocabulaire = 55k mots). Les modèles acoustiques, de type SGMM, sont appris sur le même corpus ESTER. Par ailleurs, un post-traitement est nécessaire sur les sorties de transcription pour les rendre compatibles avec le système de traduction automatique (conversion des nombres, restauration de la casse et de la ponctuation, etc.). Le système de TA français-anglais est construit à partir de la boîte à outils MOSES (Koehn *et al.*, 2007). C'est un système statistique à base de fragments (*statistical phrase-based*) et il est décrit plus en détails dans (Potet *et al.*, 2010).

Il est important ici de souligner que les post-editions des sorties de traduction (*tgt-pe*) sont antérieures à l'enregistrement du corpus oral. Ainsi, le point de départ était un corpus de post-éditions, publié en 2012 dans (Potet *et al.*, 2012), puis les enregistrements à partir des phrases sources en français ont été réalisés. Les transcriptions automatiques issues du système de RAP ont donc été traduites par le système décrit ci-dessus pour obtenir *tgt-slt*. L'hypothèse forte, faite ici, est que nous avons supposé que les post-éditions (réellement obtenues à partir de *tgt-mt*) seraient aussi valables pour *tgt-slt*.

Référence	The	consequence	of	the	fundamentalist	movement		also	has	its importance	.
	T	S	T	T	S	Y	I	T	D	P	
Hyp après Shift	The	result	of	the	hard-line	trend	is	also		important	.

TABLE 1 – Exemple de labels obtenus avec TERp-A.

L'étiquetage des sorties au niveau des mots (correct / incorrect) est réalisé avec l'outil TERp-A (Snover *et al.*, 2008). Comme illustré sur la table 1, chaque mot de l'hypothèse de traduction est aligné avec un mot ou un fragment de la post-édition selon différents types d'édition : "I" (insertions), "S" (substitutions), "T" (correspondance au niveau du lemme), "Y" (synonyme), "P" (substitution d'un segment) et "E" (identique). Ensuite, nous étiquetons l'hypothèse en groupant les labels E, T et Y selon la catégorie *Good* (G), tandis que les labels S, P et I correspondent à l'étiquette *Bad* (B).

Les principales statistiques sur ce corpus sont présentées dans la table 2, où nous montrons comment les étiquettes de confiance (G/B) sont obtenues. Pour l'ensemble de test, nous disposons donc de jeux de données pour construire des estimateurs de confiance pour trois tâches : RAP, TA et TAP.

– **RAP** : extraire les étiquettes G/B en calculant le taux d'erreur mots - WER - entre *src-asr* et *src-ref*,
– **TA** : extraire les étiquettes G/B en calculant le TERp-A entre *tgt-mt* et *tgt-pe*,
– **TAP** : extraire les étiquettes G/B en calculant le TERp-A entre *tgt-slt* et *tgt-pe*.

La table 3 donne un exemple de quintuplet disponible dans notre corpus. Une des transcriptions (*src-asr1*) a une erreur tandis que l'autre transcription (*src-asr2*) en a 4. Ceci donne lieu à, respectivement 2 et 4 étiquettes de type B pour (*tgt-slt1*) et (*tgt-slt2*) dans la sortie de TAP, alors que la sortie de TA *tgt-mt* ne présente qu'une seule étiquette de type B (incorrect).

Enfin, la table 4 résume les performances de nos systèmes de TA (traduction des références de transcription) et de TAP

1. https ://github.com/besacier/WCE-SLT-LIG

Jeu de données	# train	# test	Méthode pour obtenir les étiquettes G/B
src-ref	10000	881	
src-asr		881*3	wer(*src-asr*,*src-ref*)
tgt-mt	10000	881	terpa(*tgt-mt*,*tgt-pe*)
tgt-slt		881*3	terpa(*tgt-slt*,*tgt-pe*)
tgt-pe	10000	881	

TABLE 2 – Vue d'ensemble du corpus

src-ref	quand	notre	cerveau	chauffe
src-asr1	*comme*	notre	cerveau	chauffe
labels RAP	B	G	G	G
src-asr2	*qu'*	*entre*	*serbes*	*au* chauffe
labels RAP	B	B	B	B G
TA	when	our	brains	*chauffe*
labels TA	G	G	G	B
tgt-slt1	*as*	our	brains	*chauffe*
labels TAP	B	G	G	B
tgt-slt2	*between*	*serbs*	*in*	*chauffe*
labels TAP	B	B	B	B
tgt-pe	when	our	brain	heats up

TABLE 3 – Exemple de quintuplet avec étiquettes associées

(traduction des sorties de transcription) obtenues sur notre corpus et évaluées en utilisant les post-éditions comme références. Nous donnons également la distribution des étiquettes *correct* (G) et *incorrect* (B) obtenues pour les deux tâches (ces étiquettes seront considérées comme notre "vérité terrain" lorsque nous évaluerons la performance de nos estimateurs de confiance). Logiquement, le pourcentage d'étiquettes de type (B) augmente lorsqu'on passe d'une tâche de TA à une tâche de TAP. Ce corpus est téléchargeable en ligne sur *github* (le lien exact sera donné dans la version finale de cet article - si celui-ci est accepté).

4 Mesures de confiance pour la transcription de parole

Dans nos travaux, nous proposons d'extraire un certain nombre de traits issus du graphe du système de reconnaissance de la parole. Ces traits sont principalement issus des scores du modèle de langage et d'une analyse morphosyntaxique. Les traits utilisés sont les suivants (plus de détails sont donnés dans (Besacier *et al.*, 2014)) :

– Acoustiques : durée du mot (F-dur).
– Graphiques (extraits du réseau de confusion de la phrase courante) : nombre de chemins alternatifs (F-alt) entre deux noeuds et la probabilité *a posteriori* (F-post).
– Linguistiques (basés sur les probabilités du modèle de langue) : l'unigramme (F-word), probabilité du 3-gramme (F-3g) et utilisation du modèle de repli (F-back) telle que proposée dans (Fayolle *et al.*, 2010),
– Morpho-syntaxiques : les étiquettes morpho-syntaxiques (*Part-Of-Speech*) liées au mot (F-POS).

Nous utilisons un algorithme basé sur le *boosting* afin de combiner les différents traits. Le classifieur utilisé est *bonzaiboost* (Laurent *et al.*, 2014). Sa particularité est d'implémenter un algorithme de *boosting* (Adaboost.MH) sur des arbres de décision.

Pour chaque mot nous estimons les 7 traits (F-Word ; F-3g ; F-back ; F-alt ; F-post ; F-dur ; F-POS) décrits et l'apprentissage de l'estimateur est réalisé sur un corpus séparé mais de nature proche (parole lue - BREF 120 (Lamel *et al.*, 1991)). On notera que la préparation de ce corpus d'apprentissage aura nécessité le décodage de la totalité des signaux du corpus BREF 120, pour obtenir les transcriptions automatiques et leurs séquences d'étiquettes G/B associées.

5 Mesure de confiance pour la traduction automatique

Nous utilisons les CRFs (Conditional Random Fields (Lafferty *et al.*, 2001)) comme méthode d'apprentissage. En effet, la tâche est vue ici comme un étiquetage séquentiel d'une séquence de mots (avec labels G/B). Plus précisément, une implémentation du LIMSI nommée WAPITI (Lavergne *et al.*, 2010), est utilisée pour entraîner notre estimateur de confiance. Les 10000 phrases du corpus d'apprentissage présenté dans la table 2 sont utilisées pour apprendre les modèles (corpus

Tâche	RAP (WER)	TA (BLEU)	% G (correct)	% B (incorrect)
TA	0%	52.8%	82.5%	17.5%
TAP	26.6%	30.6%	65.5%	34.5%

TABLE 4 – Performances de traduction de texte (TA) et de parole (TAP) sur notre corpus (2643 phrases)

issu de (Potet *et al.*, 2012)).

La raison pour laquelle les méthodes d'apprentissage sont différentes pour les mesures de confiance en RAP (*boosting*) et en TA (*CRFs*) sont liées à la pre-existence de systèmes dans l'équipe avant ce travail. Une perspective est bien sûr d'avoir une approche unifiée pour ces estimateurs de confiance, en utilisant les CRFs par exemple.

Brièvement, un CRF permet de calculer la probabilité d'une séquence d'étiquettes $Y = (y_1, y_2, ..., y_N)$ étant donnée une séquence de mots en entrée $X = (x_1, x_2, ..., x_N)$ par :

$$p_\theta(Y|X) = \frac{1}{Z_\theta(X)} exp \left\{ \sum_{k=1}^{K} \theta_k F_k(X, Y) \right\} \quad (1)$$

où $F_k(X, Y) = \sum_{t=1}^{T} f_k(y_{t-1}, y_t, x_t)$; $\{f_k\}$ $(k = \overline{1, K})$ est un ensemble de fonctions caractéristiques ; $\{\theta_k\}$ $(k = \overline{1, K})$ sont les paramètres du modèle ; et $Z_\theta(x)$ est une fonction de normalisation.

Concernant les traits extraits pour apprendre ces modèles, de multiples sources peuvent être utilisées. Nous utilisons plusieurs dizaines de traits qui sont décrits en détail dans (Luong *et al.*, 2013a). Ils sont présentés ici seulement brièvement ; il est cependant important de préciser qu'un vecteur de traits est extrait pour chaque mot de l'hypothèse de traduction en langue cible :

– Mots ou séquences de mots en langue cible : mot cible ; séquence bigramme (et trigramme) précédent le mot cible considéré,
– Mots en langue source : mots en langue source alignés avec le mot en langue cible considéré,
– Contexte de l'alignement (Bach *et al.*, 2011) : mots en langue source qui entourent (pour une fenêtre ± 2) le mot source aligné avec le mot en langue cible considéré,
– Probabilité *a posteriori* du mot cible considéré (WPP - *Word Posterior Probability* (Ueffing *et al.*, 2003)),
– Pseudo-référence (Google Translate) : en considérant un système de TA en ligne comme une pseudo référence, nous obtenons un trait binaire indiquant si chaque mot de la phrase cible se trouve (ou pas) dans cette (pseudo-) référence,
– Topologie du graphe d'hypothèse de TA (Luong *et al.*, 2013a) : à partir d'une listes des meilleurs hypothèses (*N-best list* en anglais) rassemblée dans un réseau de confusion, les traits suivants sont extraits pour chaque mot de l'hypothèse de TA : nombre de chemins alternatifs d'un noeud à l'autre du graphe (en d'autres termes, nombre de mots qui sont des alternatives au mot considéré), valeur minimale et maximale des probabilités *a posteriori* parmi les alternatives au mot considéré,
– Traits issus du modèle de langue : nous construisons tout d'abord deux modèles de langue 4-gramme pour les deux cotés (source et cible) ; ensuite, nous comptons la longueur la plus grande possible du n-gramme couvert par le mot courant et les mots précédents dans le modèle de langue du coté cible ainsi que du coté source (en utilisant les informations d'alignement) ; par exemple, avec le mot cible w_i : si la séquence $w_{i-2}w_{i-1}w_i$ existe dans le modèle de langue cible tandis que la séquence $w_{i-3}w_{i-2}w_{i-1}w_i$ n'existe pas, la valeur du trait calculée pour w_i est 3,
– Traits lexicaux : étiquette morpho-syntaxique du mot considéré (POS) ; étiquette du mot source aligné avec le mot considéré, séquences des POS en langue cible, traits binaires indiquant si on est en présence d'une marque de ponctuation, d'un nom propre ou d'une valeur numérique,
– Traits syntaxiques : obtenus à partir de l'arbre d'analyse de la phrase cible (nous utilisons ici la sortie de l'outil *Link Grammar Parser*) et calculons pour chaque mot : le label grammatical, la distance à la racine, et un trait binaire indiquant si un mot n'a pas de dépendance (*null link* (Xiong *et al.*, 2010)),
– Traits sémantiques : nombre de sens dans *WordNet* côté cible.

Un ensemble de traits similaires a été utilisé pour constuire un estimateur de confiance pour un système de traduction anglais vers espagnol. L'estimateur, soumis à la campagne d'évaluation WMT 2013 (*Quality Estimation shared task*), s'est classé premier selon la métrique proposée par les organisateurs (Luong *et al.*, 2013a).

Tâche	Est. conf. RAP	Est. conf. TA	Est. conf. TAP	Est. conf. TAP	Est. conf. TAP
Traits issus de	RAP	TA	TA	RAP	RAP+TA
F(G)	87.85%	87.65%	77.17%	76.41%	**77.54%**
F(B)	37.28%	42.29%	39.34%	38.00%	**43.96%**

TABLE 5 – Résumé des performances de nos estimateurs de confiance (différentes tâches, différents traits utilisés)

6 Re-décodage d'un graphe de traduction de parole à l'aide de mesures de confiance

6.1 Estimateurs de confiance fusionnés RAP+TA

La table 5 présente tout d'abord les résultats obtenus par nos estimateurs de confiance sur des tâches de RAP et de TA (2 premières colonnes). Pour la traduction automatique (TA), l'estimateur peut être considéré comme "à l'état de l'art" puisqu'il présente des performances similaires à celle obtenues (même si c'est pour un autre couple de langues) au cours des campagnes d'évaluation WMT 2013 et 2014 auxquelles nous avons participé. Concernant la transcription de parole (RAP), il est plus difficile de se comparer car il n'existe pas, à notre connaissance, de corpus *étalon* sur lequel il est possible d'évaluer notre approche. Cependant, nous atteignons des résultats acceptables qui sont bien au delà d'une décision "au hasard". Les trois dernières colonnes de la table sont quant à elles directement comparables entre elles puisqu'elles rapportent des performances évaluées selon une même vérité terrain correspondant à des étiquettes de confiance pour une tâche de traduction de parole (TAP - 65.5% d'étiquettes G et 34.5% d'étiquettes B). Plus précisément, ces trois colonnes correspondent aux résultats des estimateurs suivants :

– Le premier système (estimateur de confiance TAP / traits issus de la TA) est celui décrit dans la section 5 et n'utilise que des traits issus du système de traduction automatique (TA). La seule différence est que, en entrée, la phrase source provient de la sortie de RAP (*src-asr*) plutôt que d'être une phrase texte sans erreur (*src-ref*).
– Le second système (estimateur de confiance TAP / traits issus de la RAP) est celui décrit dans la section 4 et n'utilise que des traits issus du système de transcription automatique (RAP). Ceci revient donc à prédire la confiance d'une sortie de traduction de parole en n'utilisant que des informations issues du module de transcription. L'alignement en mots, obtenu grâce au décodeur *moses* entre *src-asr* et *tgt-slt* est utilisé pour projeter les scores de confiance issus du système de RAP - qui sont donc sur la langue source - vers la sortie de traduction en langue cible.
– Le troisième système (estimateur de confiance TAP / traits issus de RAP+TA) combine les informations issues des deux estimateurs de confiance utilisés précédemment. Dans cette expérimentation, le score issu de l'estimateur RAP est projeté sur chaque mot cible comme pour le second système (présenté dans l'item précédent) puis combiné linéairement (simple moyenne pour cet expérimentation) avec le score issu de l'estimateur TA (*0.5score(MT) + 0.5score(ASR)*). Il est important de noter que les estimateurs de confiance ne sont pas re-entraînés ici, puisque nous réalisons une "fusion tardive" des scores de confiance issus des systèmes de RAP et de TA. Une perspective de ce travail consistera à entraîner un nouvel estimateur à partir de traits joints ASR+MT (en concaténant simplement les vecteurs de traits, par exemple).

Les résultats de ces trois systèmes sont donnés dans les trois dernières colonnes de la table 5. On voit clairement que la fusion d'informations RAP+TA permet d'améliorer les performances de l'estimateur de confiance pour une tâche de TAP[2]. En effet, la performance (F-mesure) pour l'étiquette *"B"* passe de 39.34% (traits TA seulement) et 38% (traits RAP seulement) à 43.96% (traits RAP+TA fusionnées), tout en conservant un score similaire pour l'étiquette *"G"*. Il est aussi intéressant de remarquer que l'utilisation des traits issus de la transcription seule donnent des performances très intéressantes. On peut d'ailleurs s'interroger sur les gains finalement limités obtenus en combinant des traits qui, séparément, donnent toutes les deux des résultats honorables. Une explication possible est que la fusion tardive proposée ici n'est sans doute pas la meilleure solution car l'observation des scores *"G"* et *"B"* obtenus par chacune des méthodes fait apparaître des distributions biaisées vers le label *"G"* ; une normalisation de ces scores avant combinaison serait nécessaire, ainsi que des stratégies de fusion plus avancées qu'une simple moyenne des scores (arbres de décision par exemple). Il semble aussi que l'estimateur de confiance utilisant des traits uniquement TA (et appris sur des données dont la répartition des labels G/B est plutôt 80%/20%) n'est pas bien adapté aux données issues de TAP dont la répartition de labels G/B est plus équilibrée. Un re-entraînement et une optimisation de l'estimateur de confiance RAP+TA seraient donc une tâche à court terme importante.

2. Tous les résultats sont donnés avec un seuil de décision sur les scores p(G) et p(B) fixé à 0.7 - c'est à dire que l'étiquette est fixée à G si p(G)>0.7 et l'étiquette est B sinon - ce seuil de 0.7 est fixé empiriquement et permet de favoriser la détection d'erreurs

6.2 Re-décodage d'un graphe de traduction de parole

6.2.1 Quelques travaux antérieurs sur le sujet

Plusieurs travaux ont proposé une "seconde passe" de post-édition (Parton *et al.*, 2012) ou de re-ordonnancement des N-meilleures hypothèses (Duh & Kirchhoff, 2008; Bach *et al.*, 2011; Zhang *et al.*, 2006).

Concernant le re-décodage de graphes, (Zens & Ney, 2006) proposent un système de traduction en 2 passes qui utilise, au cours de la seconde passe, des paramètres de longueur de phrase et de probabilité *a posteriori* de séquences de mots. Les expérimentations sur un système de traduction Mandarin-Anglais (tâche NIST) montrent une amélioration significative des performances. Par ailleurs, (Tromble *et al.*, 2008) propose de re-décoder le graphe de traduction en trouvant l'hypothèse candidate qui correspond à la minimisation du risque de Bayes (décodage MBR - *Minimum Bayes Risk*). Les résultats expérimentaux sur des tâches de traduction Arabe-Anglais, Mandarin-Anglais et Anglais-Mandarin montrent que l'approche par décodage MBR de graphes surpasse le ré-ordonnancement d'hypothèses (fondé sur le même critère MBR). De son côté, (Venugopal *et al.*, 2007) utilise un modèle de langue (utilisant un historique de taille importante) pour re-décoder le graphe d'hypothèses de traduction d'un système fondé sur une grammaire probabiliste hors contexte.

Notre approche, qui consiste à utiliser des informations externes, rassemblées via un estimateur de confiance, au cours d'une seconde passe de traduction, est présentée dans la section suivante. Elle peut être comparée à celle de (Zens & Ney, 2006) mais avec un nombre de traits (ayant conduit à l'estimation de confiance) beaucoup plus important.

6.2.2 Notre approche

Maintenant que nous avons des estimateurs de confiance au niveau mot, pour une tâche de TAP, nous allons les intégrer dans une seconde passe de décodage pour la traduction. La technique proposée peut se résumer comme suit : les chemins dans le graphe de recherche passant par des mots étiquetés comme incorrects (B) seront pénalisés, tandis que des chemins passant par des mots étiquetés comme corrects (G) seront récompensés. Une fois ce principe énoncé, il convient cependant de préciser que nos estimateurs de confiance ne sont pas capables d'étiqueter directement un graphe (dans leur état actuel, ils doivent être appliqués sur des chaînes de mots). Ainsi, pour couvrir un maximum de mots présents dans le graphe de recherche, nos estimateurs de confiance sont appliqués sur une liste de N-meilleures hypothèses de traduction afin d'étiqueter un maximum de mots différents. Ensuite, pour chaque mot différent rencontré dans la liste des N-meilleures hypothèses, nous mettons à jour les scores des hypothèses du graphe de recherche qui contiennent ces mots. Enfin, nous recherchons à nouveau le meilleur chemin dans le graphe de recherche pour trouver la nouvelle traduction considérée comme "la meilleure". Cette approche est décrite en détail dans (Luong *et al.*, 2014a) où elle est appliquée sur une tâche de traduction de texte uniquement. Dans ce même article, plusieurs façons de mettre à jour les scores dans le graphe de recherche sont aussi présentées. Nous ne décrivons ici que la méthode utilisée dans les expérimentations de cette présente soumission, pour une tâche de TAP.

Si on formalise, notre décodeur génère N-meilleures hypothèses de traduction $T = \{T_1, T_2, ..., T_N\}$ à la fin de la "première passe". Toutes ces hypothèses sont ensuite étiquetées par notre estimateur de confiance et nous obtenons alors, pour le *j-ème* mot dans l'hypothèse T_i, noté t_{ij}, une étiquette de qualité c_{ij} (e.g. *"G"* (correct, pas d'erreur), *"B"* (incorrect, doit être édité)). On remarquera que les scores p(G) ou p(B) auraient pu être utilisés plutôt que les étiquettes, mais des expériences conduites sur une tâche de TA (reportées dans (Luong *et al.*, 2014a)) montrent que ceci fait peu de différence au final. Ensuite, la seconde passe considère chaque mot t_{ij} et son étiquette c_{ij}. Si $c_{ij} = "G"$ (les mots sont pris en compte séquentiellement en parcourant la liste des N-best, de la meilleure à la moins bonne, et un mot et son étiquette sont "ignorés" si le mot a déjà été rencontré dans une hypothèse de meilleur rang - ainsi l'étiquette considérée pour un mot est celle correspondant à l'hypothèse placée la plus haut dans la liste N-Best), toutes les hypothèses H_k dans le graphe de recherche contenant ce mot t_{ij} vont être récompensées. En revanche, si $c_{ij} = "B"$, toutes les hypothèses H_k contenant ce mot seront pénalisées. Les autres hypothèses (ne contenant pas t_{ij}) ne seront, quant à elles, pas modifiées. Si on définit $reward(t_{ij})$ et $penalty(t_{ij})$ comme les récompenses (ou pénalités) pour t_{ij}, alors le nouveau score (de transition) de H_k, après mise à jour, sera défini par :

$$transition'(H_k) = transition(H_k) + \begin{cases} reward(t_{ij}) & \text{si } c_{ij} = G \\ penalty(t_{ij}) & \text{sinon} \end{cases} \quad (2)$$

La mise à jour des scores étant faite de la façon suivante :

$$penalty(t_{ij}) = -reward(t_{ij}) = \alpha * \frac{score(H^*)}{\#mots(H^*)} \quad (3)$$

Où $\#mots(H^*)$ est le nombre de mots cible dans H^*, le coefficient α (>0) pondère l'impact de la pénalité (ou de la récompense) sur le score final de l'hypothèse. Ce paramètre doit être optimisé : dans ce travail, en raison de la taille du corpus disponible, nous effectuons une validation croisée avec optimisation sur une moitié du corpus de test et validation sur l'autre moitié (et vice versa en inversant les données d'optimisation et de validation). Ainsi, $penalty(t_{ij})$ (négatif car $score(H^*) < 0$) sera ajouté au score de toutes les hypothèses contenant t_{ij} lorsque ce mot est étiqueté *"B"* ; tandis que $reward(t_{ij})$ (même valeur absolue mais signe opposé) est utilisé dans le cas contraire. Finalement, la mise à jour des scores s'arrête quand tous les mots différents de la liste des N-meilleures hypothèses ont été traités. Les scores des hypothèses complètes sont alors recalculés à partir du graphe de recherche modifié.

6.3 Résultats

Le graphe d'hypothèse de TAP est généré de la façon suivante : l'hypothèse de RAP (*src-asr*) est traduite par le système de TA (fondé sur l'outil moses) qui génère un graphe d'hypothèses de traduction. Nous appelons ce graphe : *graphe de traduction de parole* ; cependant, une perspective de ces travaux consistera à construire un graphe plus riche en informations, où le système de RAP fournit lui-même un treillis d'hypothèses en entrée de la TA (ce qui n'est pas le cas ici).

Les performances de traduction de parole (TAP) obtenues avec ou sans re-décodage de graphe, et utilisant nos estimateurs de confiance, sont présentées dans la table 6. Il est important de noter ici qu'une seconde passe qui n'utiliserait aucun estimateur de confiance donnerait lieu à une hypothèse équivalente au système en une passe. La première colonne montre notre *baseline* de TAP (système à une seule passe) dont les résultats ont déjà été donnés dans la table 4. Les seconde, troisième et quatrième colonnes montrent l'amélioration de notre système en prenant en compte des mesures de confiance pour re-décoder le graphe de recherche. Si on compare la première colonne avec la dernière (système 1-passe *vs* système 2-passes avec le meilleur estimateur), les gains observés sont significatifs (p-valeur dans l'intervalle [0.00 ; 0.01]). On remarque également que l'estimateur de confiance obtenu à partir des traits joints RAP+TA donne une amélioration légèrement plus importante que l'estimateur obtenu à partir de traits TA uniquement (BLEU de 32.82% au lieu de 31.89%) ce qui peut paraître étonnant alors que l'amélioration de la qualité de l'estimateur RAP+TA donnée dans la table 5 était faible par rapport à un estimateur TA seul. Une première explication peut être liée au fait qu'ici, l'estimateur de confiance est appliqué non pas sur 2643 phrases, mais sur 2643*N phrases (avec dans ce cas N=100 meilleures hypothèses) et les différences de performances entre TA et RAP+TA sont peut être plus importantes dans ce cas. Une autre possibilité peut être aussi que même une faible amélioration de la détection d'erreurs (mots dont l'étiquette est B) peut conduire à un gain non négligeable du score BLEU (qui est, rappelons le, évalué ici avec comme référence la post-édition ayant servi également à générer la vérité terrain de nos étiquettes G/B). Enfin, l'estimateur de confiance obtenu à partir de traits RAP seuls permet d'améliorer le décodage par rapport à un système à une seule passe (31.12% au lieu de 30.60%) ; cependant, cette configuration est la plus faible en terme d'amélioration.

système	**TAP baseline (BLEU)**	**TAP rédécodage (BLEU)**	**TAP rédécodage (BLEU)**	**TAP rédécodage (BLEU)**
estimateurs conf.	aucun	RAP	TA	RAP+TA
Perf.	30.60%	31.12%	31.89%	**32.82%**

TABLE 6 – Performance de TAP (BLEU) *avec* ou *sans* re-décodage de graphes - 2643 phrases

Des exemples de traduction de parole (TAP) obtenues avec ou sans re-décodage de graphe sont donnés dans la table 7 (sans chercher ici à analyser les différences fines entre les estimateurs TA et TA+RAP - ainsi, la ligne *avec re-décodage* indique l'un ou l'autre des estimateurs selon les cas). L'exemple 1 illustre un premier cas où le re-décodage du graphe de traduction permet une légère amélioration de l'hypothèse de traduction. L'analyse des labels issus de l'estimateur de confiance indique ici que les mots *a* (en début de phrase) et *penalty* étaient étiquetés comme incorrects ici ; le redécodage a permis de faire émerger une hypothèse très légèrement meilleure, même si l'erreur de reconnaissance n'a pas pu être rattrapée (puisque de toute façon, seule la meilleure hypothèse de RAP est traduite ici - et pas le graphe d'hypothèse de RAP complet). Pour l'exemple 2, l'estimateur de confiance a étiqueté comme incorrects les séquences *it has*, *speech that was* et *post route* ; à nouveau, une meilleure hypothèse de traduction a été trouvée via re-décodage (pronom correct, fin de phrase de meilleure qualité). Pour finir, l'exemple 3 indique un cas où cette fois-ci, la traduction issue de la première passe

s'est dégradée après redécodage ; l'analyse des sorties de l'estimateur de confiance montre que dans ce cas, la chaîne *to open* était bien étiquetée comme incorrecte, mais le re-décodage a fait émerger une hypothèse encore plus mauvaise. Cet exemple illustre, entre autres choses, les limites de notre approche actuelle qui, dans ce cas précis, aurait de toute façon été incapable de retrouver l'entité nommée *opel* puisque celle-ci n'était pas présente dans le graphe de traduction de parole. Nos travaux à venir, consistant à exploiter aussi le graphe issu de la transcription, nous laissent espérer qu'un tel cas aura une chance d'être résolu dans le futur.

src-ref1	une démobilisation des employés peut déboucher sur une démoralisation **mortifère**
src-asr1	une démobilisation des employés peut déboucher sur une démoralisation **mort y faire**
tgt-slt1 baseline	a **demobilisation employees** can lead to a **penalty demoralisation**
tgt-slt1 avec redécodage	a **demobilisation of employees** can lead to a **demoralization death**
tgt-pe1	**demobilization of employees** can lead to a **deadly demoralization**
src-ref2	celui-ci a indiqué que l'intervention **s'**était parfaitement bien **déroulée** et que les examens post-**opératoires** étaient normaux
src-asr2	celui-ci a indiqué que l' intervention **c'**était parfaitement bien **déroulés** , et que les examens post **opératoire** étaient normaux .
tgt-slt2 baseline	**it** has indicated that the speech **that was well** conducted , and that the tests were **normal post route**
tgt-slt2 avec redécodage	**he** indicated that the intervention **is very well done** , and that the tests **after operating were normal**
tgt-pe2	**he** indicated that the operation **went perfectly well** and the **post-operative tests were normal**
src-ref3	general motors repousse jusqu'en janvier le plan pour **opel**
src-asr3	general motors repousse jusqu' en janvier le plan pour **open**
tgt-slt3 baseline	general motors postponed until january **the plan to open**
tgt-slt3 avec redécodage	general motors puts until january **terms to open**
tgt-pe3	general motors postponed until january **the plan for opel**

TABLE 7 – Exemples de sortie des systèmes avec et sans redécodage de graphes

7 Conclusion

Cet article démontre que des mesures de confiance, issues d'estimateurs automatiques performants, sont utiles pour re-décoder des graphes d'hypothèses de traduction du langage parlé. Par ailleurs, les estimateurs de confiances obtenus à partir de traits multiples (RAP et TA) sont plus performants que des estimateurs fondés sur l'une ou l'autre de ces modalités. Cette bonne intégration des informations des modules de transcription et traduction, conduit à des gains de performance (mesurés avec BLEU) encore plus importants. Enfin, ce travail a été possible en raison de la constitution d'un corpus spécifique (oral/écrit avec quintuplets *transcription/référence/TA(transcription)/TA(référence)/post-édition* pour 2643 phrases), également présenté ici, mis à disposition de la communauté TALN. Les perspectives de ce travail sont les suivantes : proposer une tâche d'estimation de confiance pour la TAP pour un workshop international tel que IWSLT, entraîner un estimateur de confiance à partir de véritables traits joints RAP+TA (au lieu de fusionner les sorties de deux estimateurs différents), utiliser les mesures de confiance dans un scénario interactif de traduction de parole (ce qui nécessiterait une étape indispensable d'optimisation de nos méthodes qui, à ce jour, ne sont pas en état de fonctionner "en ligne") ou dans un scénario interactif de transcription de discours (où une hypothèse de TAP serait recalculée "à la volée" en fonction des éditions de l'utilisateur).

Références

ASADI A., SCHWARTZ R. & MAKHOUL J. (1990). Automatic detection of new words in a large vocabulary continuous speech recognition system. *Proc. of International Conference on Acoustics, Speech and Signal Processing*.

BACH N., HUANG F. & AL-ONAIZAN Y. (2011). Goodness : A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, p. 211–219, Portland, Oregon.

BESACIER L., LECOUTEUX B., LUONG N. Q., HOUR K. & HADJSALAH M. (2014). Word confidence estimation for speech translation. In *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA.

BICICI E. (2013). Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, p. 343–351, Sofia, Bulgaria : Association for Computational Linguistics.

CHASE L. (1997). *Error-responsive feedback mechanisms for speech recognizers*. PhD thesis, Carnegie Mellon University.

DUH K. & KIRCHHOFF K. (2008). Beyond log-linear models : Boosted minimum error rate training for n-best reranking. In *Proc. of ACL, Short Papers*.

FAYOLLE J., MOREAU F., RAYMOND C., GRAVIER G. & GROS P. (2010). Crf-based combination of contextual features to improve a posteriori word-level confidence measures. In *Interspeech*.

GALLIANO S., GEOFFROIS E., GRAVIER G., BONASTRE J.-F., MOSTEFA D. & CHOUKRI K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, p. 315–320.

HAN A. L.-F., LU Y., WONG D. F., CHAO L. S., HE L. & XING J. (2013). Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, p. 365–372, Sofia, Bulgaria : Association for Computational Linguistics.

KEMP T. & SCHAAF T. (1997). Estimating confidence using word lattices. *Proc. of European Conference on Speech Communication Technology*, p. 827–830.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, p. 177–180, Prague, Czech Republic.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting et labeling sequence data. In *Proceedings of ICML-01*, p. 282–289.

LAMEL L. F., GAUVAIN J.-L., ESKÉNAZI M. *et al.* (1991). Bref, a large vocabulary spoken corpus for french1. *training*, **22**(28), 50.

LAURENT A., CAMELIN N. & RAYMOND C. (2014). Boosting bonsai trees for efficient features combination : application to speaker role identification. In *Interspeech*.

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

LECOUTEUX B., LINARÈS G. & FAVRE B. (2009). Combined low level and high level features for out-of-vocabulary word detection. *INTERSPEECH*.

LUONG N. Q., BESACIER L. & LECOUTEUX B. (2013a). Word confidence estimation and its integration in sentence quality estimation for machine translation. In *Proceedings of The Fifth International Conference on Knowledge and Systems Engineering (KSE 2013)*, Hanoi, Vietnam.

LUONG N. Q., BESACIER L. & LECOUTEUX B. (2014a). An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation. In *European Association for Machine Translation (EAMT)*, Dubrovnik, Croatie.

LUONG N.-Q., BESACIER L. & LECOUTEUX B. (2014b). Word Confidence Estimation for SMT N-best List Reranking. In *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT) during EACL*, Gothenburg, Suède.

LUONG N. Q., LECOUTEUX B. & BESACIER L. (2013b). LIG system for WMT13 QE task : Investigating the usefulness of features in word confidence estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, p. 396–391, Sofia, Bulgaria : Association for Computational Linguistics.

PARTON K., HABASH N., MCKEOWN K., IGLESIAS G. & DE GISPERT A. (2012). Can automatic post-editing make mt more meaningful ? In *Proceedings of the 16th EAMT*, p. 111–118, Trento, Italy.

POTET M., BESACIER L. & BLANCHON H. (2010). The lig machine translation system for wmt 2010. In A. WORKSHOP, Ed., *Proceedings of the joint fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT2010)*, Uppsala, Sweden.

POTET M., EMMANUELLE E R., BESACIER L. & BLANCHON H. (2012). Collection of a large database of french-english smt output corrections. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

SANCHIS A., JUAN A. & VIDAL E. (2012). A word-based naïve Bayes classifier for confidence estimation in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(2), 565–574.

SEIGEL M. S. & WOODLAND P. C. (2012). Using sub-word-level information for confidence estimation with conditional random field models. In *INTERSPEECH*.

SEIGEL M. S., WOODLAND P. C. *et al.* (2011). Combining information sources for confidence estimation with crf models. In *INTERSPEECH*, p. 905–908.

SNOVER M., MADNANI N., DORR B. & SCHWARTZ R. (2008). Terp system description. In *MetricsMATR workshop at AMTA*.

TROMBLE R., KUMAR S., OCH F. J. & MACHEREY W. (2008). Lattice minimum bayes risk decoding for statistical machine translation. In *Lattice minimum bayesrisk Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 620–629.

UEFFING N., MACHEREY K. & NEY H. (2003). Confidence measures for statistical machine translation. In *Proceedings of the MT Summit IX*, p. 394–401, New Orleans, LA.

VENUGOPAL A., ZOLLMANN A. & VOGEL S. (2007). An efficient two-pass approach to synchronous-cfg driven statistical mt. In *Proceedings of Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics*.

WEINTRAUB M., BEAUFAYS F., RIVLIN Z., KONIG Y. & STOLCKE A. (1997). Neural-network based measures of confidence for word recognition. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, p. 887–890.

XIONG D., ZHANG M. & LI H. (2010). Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, p. 604–611, Uppsala, Sweden.

YOUNG S. R. (1994). Recognition confidence measures : Detection of misrecognitions and out-of-vocabulary words. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, p. 21–24.

ZENS R. & NEY H. (2006). N-gram posterior probabilities for statistical machine translation. In *Workshop on Statistical Machine Translation - StatMT*, Stroudsburg, PA, USA.

ZHANG R. & RUDNICKY A. I. (2001). Word level confidence annotation using combinations of features.

ZHANG Y., HILDEBRAND A. S. & VOGEL S. (2006). Distributed language modeling for n-best list re-ranking. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, p. 216–223, Sydney.