Vers un décodage guidé pour la traduction automatique

Benjamin Lecouteux et Laurent Besacier
Laboratoire d'Informatique de Grenoble (LIG), Université de Grenoble
benjamin.lecouteux@imag.fr, laurent.besacier@imag.fr

RÉSUMÉ

Récemment, le paradigme du décodage guidé a montré un fort potentiel dans le cadre de la reconnaissance automatique de la parole. Le principe est de guider le processus de décodage via l'utilisation de transcriptions auxiliaires. Ce paradigme appliqué à la traduction automatique permet d'envisager de nombreuses applications telles que la combinaison de systèmes, la traduction multi-sources etc. Cet article présente une approche préliminaire de l'application de ce paradigme à la traduction automatique (TA). Nous proposons d'enrichir le modèle log-linéaire d'un système primaire de TA avec des mesures de distance relatives à des systèmes de TA auxiliaires. Les premiers résultats obtenus sur la tâche de traduction Français/Anglais issue de la campagne d'évaluation WMT 2011 montrent le potentiel du décodage guidé.

ABSTRACT

Driven Decoding for machine translation

Recently, the concept of driven decoding (DD), has been sucessfully applied to the automatic speech recognition (speech-to-text) task : an auxiliary transcription guide the decoding process. There is a strong interest in applying this concept to statistical machine translation (SMT). This paper presents our approach on this topic. Our first attempt in driven decoding consists in adding several feature functions corresponding to the distance between the current hypothesis decoded and the auxiliary translations available. Experimental results done for a french-to-english machine translation task, in the framework of the WMT 2011 evaluation, show the potential of the DD approach proposed.

MOTS-CLÉS : Décodage guidé, traduction automatique, combinaison de systèmes.

KEYWORDS: Driven Decoding, machine translation, system combination.

1 Introduction

Le concept du décodage guidé (Lecouteux *et al.*, 2012, 2013) a montré un fort potentiel dans le cadre de la reconnaissance automatique de la parole. Le principe est de guider le processus de décodage via l'utilisation de transcriptions auxiliaires. Ce paradigme appliqué à la traduction automatique permet d'envisager de nombreuses applications telles que la combinaison de systèmes, la traduction multi-sources (à partir de différentes langues, ou à partir de sorties de différents systèmes de reconnaissance de la parole dans le cas de la traduction de la parole), l'utilisation de systèmes en ligne (comme *Google traduction*), le recalcul en temps réel d'hypothèses de traduction dans une interface de post-édition, etc.

Cet article présente un travail préliminaire concernant l'application du paradigme de décodage guidé à la traduction automatique (TA). Nous proposons d'utiliser les systèmes de TA Fran-

çais/Anglais de deux laboratoires (le LIA et le LIG) présentés dans (Potet *et al.*, 2011). Ces systèmes sont des systèmes de traduction statistiques à base de séquences (phrase-based (Koehn, 2010)). Dans ces approches, un score de vraisemblance est calculé pour chaque phrase candidate à la traduction, en fonction de la phrase source ; et ce score résulte de la combinaison log-linéaire d'un ensemble de paramètres.

Notre première approche introduisant le décodage guidé consiste en l'addition de paramètres, dans le modèle log-linéaire, modélisant la distance entre l'hypothèse courante (notée H) et la transcription auxiliaire (notée T) : d(T,H). Avec l'introduction de ces nouveaux paramètres, les N meilleures hypothèses sont alors réévaluées et réordonnées.

L'article s'articule ainsi : la section 2 propose un état de l'art relatif au travail présenté. La section 3 présente notre approche, les sections 4 et 5 décrivent respectivement le système de traduction étalon utilisé et nos expérimentations qui sont analysées plus finement dans la section 6. La dernière section est consacrée à nos conclusions et à quelques perspectives.

2 État de l'art

Contrairement à la reconnaissance automatique de la parole, la traduction automatique propose une grande variété de systèmes basés sur des concepts différents. Même parmi les systèmes statistiques, on trouve de nombreuses variantes telles que les systèmes à base de segments, les systèmes hiérarchiques ou les approches syntaxiques. Ceci complique la combinaison d'hypothèses en TA car on est confronté à des hypothèses potentiellement très différentes en terme de fluidité, d'ordre de mots, etc.

Dans un premier temps, nous présentons le concept de décodage guidé utilisé dans les systèmes de reconnaissance automatique de la parole (SRAP). Ensuite, nous présentons les approches de combinaison de systèmes existantes dans le cadre de la TA.

2.1 Reconnaissance de la parole guidée par des transcriptions approchées

Dans (Lecouteux *et al.*, 2012, 2013), nous proposons l'utilisation de transcriptions auxiliaires pour améliorer les performances d'un SRAP. Nous montrons que même des informations bruitées peuvent apporter une aide précieuse et exploitable. Pour ce faire, deux méthodes complémentaires sont exploitées : la combinaison d'un modèle de langage générique avec un modèle estimé sur la transcription imparfaite (permettant de réduire l'espace linguistique et de le focaliser sur la tâche) et la réestimation dynamique de la fonction de coût du SRAP en fonction de la ressemblance de l'hypothèse courante avec la transcription auxiliaire. Ainsi, la probabilité de l'hypothèse courante est biaisée par la transcription auxiliaire. Différents types de transcriptions auxiliaires peuvent être utilisés, comme par exemple des transcriptions issues d'autres SRAP, aboutissant finalement à une combinaison. Ainsi, en associant une hypothèse auxiliaire et ses scores de confiance, il est possible d'influencer dynamiquement la probabilité linguistique. Cette approche a montré des gains supérieurs aux méthodes de combinaison classiques (i.e. ROVER) pour des tâches de transcription de parole.

2.2 Combinaison de systèmes de traduction automatique

2.2.1 Décodage de réseaux de confusion

De nombreux problèmes se présentent pour la fusion de réseaux de confusion (RC), dans le cadre de la TA. L'un des plus importants est relatif aux erreurs d'alignement entre hypothèses, qui génèrent des erreurs grammaticales. Le décodage de réseaux de confusion pour la TA a été proposé par (Bangalore, 2001). Les hypothèses sont alignées en utilisant une distance de Levensthein, en vue de les fusionner en RC. L'étape la plus importante consiste à sélectionner une hypothèse "patron" servant de base à l'alignement. Dans (Rosti *et al.*, 2007b), les sorties *1-best* de chaque système sont utilisées à tour de rôle comme patron et la mesure TER (Term Error Rate) entre le patron et les hypothèses concurrentes est estimée dans chaque cas. Au final, le score TER minimal permet de retenir l'hypothèse patron E_s telle que : $E_s = \arg\min_{E \in E_i} \sum_{j=1}^{N_s} TER(E_j, E_i)$ où N_s est le nombre de systèmes.

Finalement, un réseau est construit en agrégeant toutes les hypothèses. Dans cette approche les auteurs montrent que des paramètres supplémentaires peuvent être rajoutés dans le modèle log-linéaire, comme les probabilités a posteriori relatives à chaque arc du RC. Dans cette approche, l'ordre de la combinaison est fortement influencé par la qualité de l'hypothèse patron.

Dans (Rosti *et al.*, 2007a), une combinaison basée sur les scores de confiance a posteriori de différents systèmes est introduite. Dans la partie expérimentale de leurs travaux, les auteurs combinent trois systèmes à base de segments, deux systèmes hiérarchiques et un syntaxique. Tous les systèmes sont entraînés sur les mêmes données. Les poids des décodeurs sont optimisés selon TER ou BLEU en fonction du système. Les résultats de combinaison montrent une amélioration significative par rapport au meilleur système initial.

2.2.2 Réordonnancement des meilleures hypothèses.

L'article (Hildebrand et Vogel, 2009) présente une approche où les scores des N meilleures hypothèses sont réestimés. Les N meilleures hypothèses de chaque système sont combinées et des paramètres sont rajoutés au modèle log-linéaire (modèle de langage, informations lexicales, etc.). Les poids du modèle sont alors recalculés en vue de réordonner optimalement les hypothèses. Les expériences décrites dans (Hildebrand et Vogel, 2009) montrent la nécessité de sélectionner un nombre N de meilleures hypothèses optimal, 50 dans ce cas précis. Avec cette méthode, les auteurs combinent incrémentalement 4 systèmes, montrant une amélioration corrélée au nombre de systèmes introduits.

Des approches basées sur le réordonnancement d'hypothèses sont également présentées dans (Li *et al.*, 2009) et (Hildebrand et Vogel, 2008) où les auteurs sélectionnent les hypothèses faisant consensus avec différents systèmes : pour cela ils introduisent dans le modèle log-linéaire des paramètres de consensus. À la différence de notre approche, les systèmes auxiliaires ne sont pas considérés comme des boîtes noires.

La prochaine section présente le paradigme du décodage guidé où seules les meilleures hypothèses (1-best) issues des systèmes auxiliaires sont exploitées en vue d'améliorer un système primaire. Il est donc important de mentionner que notre approche considère les systèmes auxiliaires comme étant des "boites noires".

3 Décodage guidé pour la traduction automatique

3.1 Principe général

Dans un premier temps, notre implémentation consiste en l'ajout de plusieurs paramètres dans le modèle log-linéaire, afin de réordonner les hypothèses. D'un point de vue pratique, ces scores sont rajoutés aux N-meilleures hypothèses directements issues du décodeur. Les scores additionnels correspondent à la distance entre l'hypothèse courante (notée H) et la traduction auxiliaire (notée T) : d(T,H). Nous utilisons dans notre cas les hypothèses fournies par le système du LIA et utilisons deux transcriptions auxiliaires (LIG et Google). Dans cette situation, deux scores de distance sont rajoutés au modèle log-linéaire. La distance utilisée est décrite dans la section suivante.

3.2 Mesure de distance utilisée

Nous proposons d'utiliser le BLEU comme distance entre les systèmes. Le score BLEU correspond à la moyenne géométrique de la précision n-gramme. Un score BLEU élevé suggère donc une traduction de meilleure qualité, d'où son utilisation comme métrique d'évaluation de similarité entre différents systèmes. Pour le décodage guidé, nous utilisons une distance BLEU lissée au niveau de la phrase comme présenté dans (Lin et Och, 2004). Évidemment, nous souhaitons introduire des mesures de distance supplémentaires dans des travaux futurs, mais seul BLEU est utilisé dans cet article qui peut être vu comme une "preuve de concept".

3.3 Réordonnancement des hypothèses et combinaison

La combinaison est appliquée sur les 500 meilleures hypothèses extraites du système primaire (LIA) en utilisant l'option *distinct* de Moses (ceci élimine les doublons). Chaque hypothèse comporte un ensemble de 14 scores : 1 pour le modèle de langage, 5 pour le modèle de traduction, 1 score de distorsion, 7 scores de réordonnancement et un score de pénalité. A ces scores, nous ajoutons donc une mesure de similarité pour chaque système auxiliaire.

Les poids de combinaison sont optimisés en maximisant le score BLEU au niveau de la phrase en utilisant l'algorithme MIRA (Margin Infused Relaxed Algorithm) (Hasler *et al.*, 2011). Le choix de MIRA est motivé par une meilleure stabilité observée dans le cas d'optimisation de nombreux paramètres. Nous effectuons une centaine d'itérations et le paramètre C est fixé à 0.001.

En ce qui concerne le décodage, un score est calculé pour chaque phrase (via la combinaison log-linéaire) et les phrases sont réordonnées en fonction des nouveaux scores calculés.

4 Système de référence

4.1 Données

Les systèmes LIG et LIA ont été entraînés à partir des données fournies lors de la campagne d'évaluation WMT 2011 et sur le corpus Gigaword fourni par le LDC. La Table 4.1 récapitule l'ensemble des données utilisées et introduit les notations pour les corpus qui seront utilisées dans la suite de l'article. Quatre corpus ont été utilisés pour construire le modèle de traduction : *news-c*, *euro*, *UN* et *giga* et trois corpus sont utilisés pour apprendre le modèle de langage. Enfin, deux corpus parallèles ont servi à optimiser les paramètres : *tuning-mt-LIG-LIA* a été utilisé pour

	CORPUS	DÉSIGNATION	NB PHRASES
Apprentissage bilingue Anglais/Français	News Commentary v6	*news-c*	116 k
	Europarl v6	*euro*	1.8 M
	UN corpus	*UN*	12 M
	10^9 corpus	*giga*	23 M
Apprentissage monolingue Anglais	News Commentary v6	*mono-news-c*	181 k
	Shuffled News Crawl (2007 à 2011)	*news-s*	25 M
	Europarl v6	*mono-euro*	1.8 M
Développement	newstest2008 + newssyscomb2009	*dev*	2553
	newstest2009	*optimisation-LIG-LIA*	2525
Test	newstest2010	*test10*	2489
	newstest2011	*test11*	3005

TABLE 1 – Corpus utilisés pour construire les systèmes LIG et LIA (dans la campagne d'évaluation WMT 2011).

le développement des deux systèmes LIG et LIA (via MERT (Och, 2003)) tandis que le corpus *dev* a été utilisé pour estimer les poids dédiés au décodage guidé. Les corpus *test10* et *test11* ont quant à eux servi pour l'évaluation du décodage guidé.

4.2 Caractéristiques du système primaire utilisé (LIA)

Le système LIA est un système à base de segments (phrase-based). L'ensemble des données utilisées provient de la campagne d'évaluation WMT 2011 et les données sont tokenisées avec les outils fournis lors de la campagne. Le modèle de langage 4-gramme a été appris à l'aide de la boîte à outils SRILM (Stolcke, 2002) avec un modèle de repli Kneyser-Ney modifié. Le corpus parallèle a été aligné au niveau des mots en utilisant Giza++ (Och et Ney, 2003) et MGiza++ (Gao et Vogel, 2008) pour les corpus très volumineux. La table de phrases et les modèles de réordonnancement ont été appris en utilisant les outils d'apprentissage de la suite Moses (Koehn *et al.*, 2007). Au final, un ensemble de 14 paramètres a été utilisé dans le système (cf 3.3). Ces scores ont été optimisés sur le corpus newstest2009 comprenant 2525 phrases en utilisant l'algorithme MERT. Plus de détails se trouvent dans (Potet *et al.*, 2011).

4.3 Performances du système primaire et des systèmes auxiliaires

La Table 2 résume les scores BLEU obtenus par le système LIA sans la casse (tous les résultats de l'article sont donnés sans la casse). L'évaluation des performances est effectuée sur 3 corpus : dev qui correspond aux corpus newstest2008 + newssyscomb2009 de WMT (2553 phrases) ; tst10 qui correspond à newstest2010 (2489 phrases) et tst11 qui correspond à newstest2011 (3005 phrases). Nous présentons également les scores obtenus par les systèmes auxiliaires LIG (non décrit ici faute de place mais présenté dans (Potet *et al.*, 2011)) et Google (système en ligne dans sa version de Février 2012). Nous sommes conscients du risque que le système Google utilisé en 2012 puisse contenir des données issues de WMT 2011, mais à la vue des performances, ça ne semble pas être le cas. C'est principalement pour cette raison que nous avons également introduit le système du LIG dont nous contrôlons parfaitement les données d'apprentissage.

Système	dev	tst10	tst11
LIA (1)	25.45	29.30	29.30
LIG (2)	24.38	27.64	28.54
Google (3)	24.62	28.38	29.83
MANY 1,2,3	26.3	30.46	30.6
ORACLE 1,2,3	*29.5*	*34.0*	*34.63*

Système	dev	tst10	tst11
DDA Google	26.37	30.16	30.52
DDA LIG	25.71	29.57	29.51
DDA LIG+G (4)	**26.41**	**30.44**	**30.91**
ORACLE 2,3,4	*29.16*	*33.8*	*34.35*
ORACLE 1,2,3,4	*30.0*	*34.7*	*35.2*

TABLE 2 – Performances des systèmes LIA, LIG et Google , d'une combinaison état de l'art via MANY et performances du décodage guidé du système LIA par les systèmes LIG et/ou Google

Afin de nous comparer à un système de combinaison de référence, nous proposons aussi une combinaison utilisant MANY (Barrault, 2010). MANY utilise un modèle de langage (dans notre cas, celui du LIA) afin de décoder un réseau de confusion constitué de l'ensemble des meilleures hypothèses des différents systèmes.

Pour finir, l'algorithme MIRA (Hasler *et al.*, 2011) est utilisé pour recalculer tous les poids relatifs au décodage guidé (entraînés sur le corpus *dev*).

5 Expériences et résultats

Le décodage guidé (Driven Decoding Algorithm, DDA) a été utilisé avec le LIA comme système primaire dont les 500 meilleures hypothèses ont été extraites. Les transcriptions auxiliaires sont ici les transcriptions issues des systèmes LIG et Google dont les performances sont données dans la Table 2. Cette table présente également les résultats du décodage guidé.

Nous constatons que le système LIA est meilleur que le système LIG. Cependant, la transcription auxiliaire du LIG permet tout de même d'améliorer ses performances par décodage guidé. Nous observons également une amélioration qui se cumule lorsqu'on ajoute le système de Google. Le décodage guidé du système LIA améliore son score BLEU d'environ 1 point en comparaison du meilleur système individuel. De plus, les scores Oracle des combinaisons entre différents systèmes sont donnés à titre d'information. Il est intéressant de noter qu'en substituant le système LIA au système DDA, le score Oracle baisse mécaniquement puisque le décodage guidé dégage un consensus.

Les résultats sont également très légèrement supérieurs à ceux obtenus avec MANY, qui est un système de combinaison état de l'art. Cependant, tandis que MANY nécessite un redécodage à l'aide d'un modèle de langage cible, le décodage guidé permet une combinaison différente et peu coûteuse à mettre en oeuvre.

6 Analyse plus fine du décodage guidé

La Table 3 et la Figure associée montrent les distances BLEU entre le décodage guidé par LIG+Google, LIG, Google et l'ensemble des systèmes utilisés seuls. Les similarités présentées ont été calculées sur le corpus test11 (elles sont similaires sur les autres ensembles). Nous observons que le décodage guidé LIG+Google se rapproche à la fois des systèmes Google et LIG. Lorsqu'il utilise uniquement le système auxiliaire Google, au contraire il s'éloigne du LIG. En revanche, en utilisant uniquement le système auxiliaire LIG, l'hypothèse obtenue ne s'éloigne pas de Google. Ceci s'explique sans doute que le LIG et le LIA sont entraînés sur des données similaires, et les systèmes sont du même type, tandis que les hypothèses de Google diffèrent un peu plus.

Système	LIA	DDA tout	DDA LIG	DDA Google
LIG	63.13	66.14	72.8	61.02
LIA	100	77.2	83.6	77.19
Google	51.01	66.29	51.76	65.93
DDA tout	77.2	100	79.68	90.96

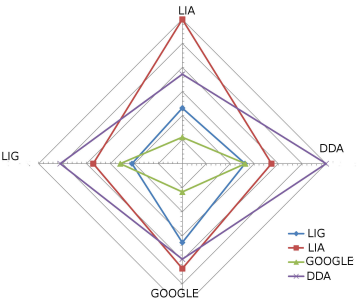

TABLE 3 – Similarité entre les systèmes. La métrique utilisée est le BLEU : Chaque sommet du graphe correspond à un système qui est considéré comme référence par rapport aux autres. DDA tout correspond au système DDA guidé à la fois par les systèmes Google et LIA

La Figure montre le comportement induit par le décodage guidé : les hypothèses se rapprochent ou s'éloignent des systèmes auxiliaires. Il est intéressant de noter que le système LIG, *a priori* moins performant que le système LIA a finalement une similarité très élevée avec ce dernier. L'utilisation d'une similarité BLEU entre les systèmes permet donc de trouver un consensus inter-hypothèses.

7 Conclusion et perspectives

Nous avons présenté une adaptation préliminaire du décodage guidé à la traduction automatique. Ce paradigme permet une combinaison efficace de systèmes de traduction automatique, en réévaluant le modèle log-linéaire au niveau des N meilleures hypothèses, en utilisant des systèmes auxiliaires. Le principe est de guider le processus de recherche en utilisant des sorties existantes. Nous avons évalué différentes configurations sur le corpus WMT 2011. Les résultats montrent que l'approche est efficace et obtient des gains significatifs en terme de score BLEU. Par ailleurs, ces résultats préliminaires sont équivalents (voire légèrement meilleurs) à ceux obtenus en utilisant des méthodes de combinaison état de l'art. Enfin, cette méthode a été récemment utilisée avec succès lors de deux campagnes d'évaluation :

– Une campagne d'évaluation arabe/français (TRAD) où nous avons utilisé Google comme système auxiliaire et le système du LIG comme système primaire.
– La campagne d'évaluation IWSLT 2012 (anglais/français) où nous avons utilisé le même système en ligne pour améliorer les performances du système primaire LIG. Les résultats de cette campagne se trouvent dans (Besacier *et al.*, 2012).

Nos futurs travaux vont se concentrer sur l'intégration du décodage guidé au sein du décodeur Moses, au niveau de la fonction objective. Le second axe envisagé est l'utilisation de mesures de confiance associées aux transcriptions auxiliaires, afin de les exploiter plus finement.

Références

BANGALORE, S. (2001). Computing consensus translation from multiple machine translation systems. *In In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001*, pages 351–354.

BARRAULT, L. (2010). Many : Open source machine translation system combination. *In In Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation(93), p.145-155*.

BESACIER, L., LECOUTEUX, B., AZOUZI, M. et LUONG NGOC, Q. (2012). The LIG English to French Machine Translation System for IWSLT 2012. *In In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*.

GAO, Q. et VOGEL, S. (2008). Parallel implementations of word alignment tool. *In Proceedings of the ACL Workshop : Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, OH, USA.

HASLER, E., HADDOW, B. et KOEHN, P. (2011). Margin infused relaxed algorithm for moses. *In The Prague Bulletin of Mathematical Linguistics*, pages 96 :69–78.

HILDEBRAND, A. S. et VOGEL, S. (2008). Combination of machine translation systems via hypothesis selection from combined n-best lists. *In AMTA conference*.

HILDEBRAND, A. S. et VOGEL, S. (2009). Combination of machine translation systems via hypothesis selection from combined n-best lists. *In Proceedings of Association for Machine Translation in the Americas (AMTA)*, Hawaï, USA.

KOEHN, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York.

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. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume*, pages 177–180, Prague, Czech Republic.

LECOUTEUX, B., LINARES, G., ESTÈVE, Y. et GRAVIER, G. (2013). Dynamic combination of automatic speech recognition systems by driven decoding. *IEEE Transactions on Audio, Speech and Signal Processing*, 21, issue 6:1251 – 1260.

LECOUTEUX, B., LINARES, G. et OGER, S. (2012). Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. *Computer Speech and Language*, 26(2):67 – 89.

LI, M., DUAN, N., ZHANG, D., LI, C.-H. et ZHOU, M. (2009). Collaborative decoding : Partial hypothesis re-ranking using translation consensus between decoders. *In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*.

LIN, C.-Y. et OCH, F. J. (2004). Orange : a method for evaluating automatic evaluation metrics for machine translation. *In In COLING '04 : Proceedings of the 20th international conference on Computational Linguistics*.

OCH, F. J. (2003). Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, Sapporo, Japan.

OCH, F. J. et NEY, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

POTET, M., RUBINO, R., LECOUTEUX, B., HUET, S., BESACIER, L., BLANCHON, H. et LEFEVRE, F. (2011). The LIGA machine translation system for WMT 2011. *In Proceedings EMNLP and ACL Workshop on Machine Translation (WMT)*, Edinburgh (Scotland).

ROSTI, A.-v., AYAN, N.-F., XIANG, B., MATSOUKAS, S., SCHWARTZ, R. et DORR, B. (2007a). Combining outputs from multiple machine translation systems. *In In Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pages 228–235.

ROSTI, A.-v., MATSOUKAS, S. et SCHWARTZ, R. (2007b). Improved word-level system combination for machine translation. *In In Proceedings of ACL*.

STOLCKE, A. (2002). SRILM — an extensible language modeling toolkit. *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, USA.

