Use of auxiliary translation for improving decoding in statistical machine translation

Benjamin Lecouteux, Laurent Besacier

Use of auxiliary translation for improving decoding in statistical machine translation

Benjamin Lecouteux and Laurent Besacier

Laboratoire d'Informatique de Grenoble (LIG), University of Grenoble (France)

Abstract. 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.

Keywords: Driven Decoding Algorithm, System Combination, Machine Translation

1 Introduction

The concept of Driven Decoding Algorithm (DDA), introduced by (Lecouteux, Linares, & Oger, 2012) has been successfully applied to the automatic speech recognition (speech-to-text) task. This idea is to use an auxiliary transcription (manual or automatic) to guide the decoding process. There is a strong interest in applying this concept to Statistical Machine Translation (SMT). The potential applications are: system combination, multi-source translation (from several languages, from several ASR outputs in the case of speech translation), use of an online system (like Google translate) as auxiliary translation, on-line hypothesis re-calculation in a post-edition interface, etc.

This paper presents some explorations on this topic and must be seen as a proof-of-concept, illustrated by a few experiments. We used machine translation systems (LIA and LIG french-to-english systems - presented in (Potet et al., 2011)) based on the standard phrase-based approach (Koehn, 2010). In such an approach, the likelihood score of a candidate translation in target language, given a source sentence, is evaluated as a log-linear combination of several feature functions.

Our first attempt in DDA consists in adding several feature functions corresponding to the distance between the current hypothesis decoded (called H) and the auxiliary translation available (T) : d(T,H). We started to experiment in a re-scoring framework for which N-Best hypotheses from the baseline MT system are re-ordered after adding the new distance functions.

The outline of this paper is as follows: Section 2 is dedicated to the works that can be considered as related to this paper proposal. Section 3 presents our approach and describes the baseline SMT systems used in our experiments further presented in

Section 4. Section 6 presents a deeper analysis of the DD behavior while Section 7 proposes a validation of the DD concept for a speech translation task. Finally, last Section concludes this work and gives some perspectives.

2 Related Work

Unlike speech recognition, system combination in SMT involves systems based on potentially different standards such as phrasal, hierarchical and syntax based. This introduces new issues such as breaking up of phrases and alterations of word order. We first propose a description of the application of DDA in ASR systems. Then, various system combination attempts in MT are presented.

2.1 Imperfect transcript driven speech recognition

In the paper introduced by (Lecouteux et al., 2012), the authors use auxiliary transcripts associated with speech signals to improve ASR performance. It is demonstrated that those imperfect transcripts could actually be taken advantage of. Two methods were proposed. First method involved the combination of generic language model (LM) and a LM estimated on the imperfect transcript resulting in cutting down the linguistic space. Second method involved modifying the decoding algorithm by rescoring the estimate function. The probability of the current hypothesis which results from partial exploration of the search graph is dynamically rescored based on the alignment (with imperfect transcript) scores (done using DTW). The experimental results which used both dynamic synchronization and linguistic rescoring displayed interesting gains. Another kind of imperfect transcript that can be used is the output hypothesis of another system, leading to an integrated approach for system combination (Lecouteux, Linares, Estève, & Gravier, 2013). The principle proposed is that firstly, one-best hypothesis is generated from the auxiliary system and a confidence score is evaluated for each word. Then these informations are used to dynamically rescore the linguistic probabilities. The method allows substantial gains compared to ROVER approaches. More details on this approach applied to ASR can be found in (Lecouteux et al., 2012, 2013).

2.2 System Combination for Machine Translation

Confusion Network (CN) Decoding. There are important issues to address for MT system combination using confusion network (CN) decoding. An important one is the presence of errors in the alignment of hypotheses which lead to ungrammatical combination outputs. (Rosti, Matsoukas, & Schwartz, 2007) proposed arbitrary features that can be added log-linearly into the objective function in this method. This adding of new features is the core idea we followed in our proposal.

CN decoding for MT system combination has been proposed in (Bangalore, 2001). The hypotheses have to be aligned using Levenshtein alignment to generate the CN. One hypothesis is chosen as skeletal hypothesis and others are aligned against it. In (Rosti, Matsoukas, & Schwartz, 2007), 1-best output from each system is used as the skeleton to develop the CN and the average of the TER scores between the skeleton

and other hypotheses were used to evaluate the prior probability. Finally a joint lattice is generated by aggregating all the CN parallely. Through this work it is shown that arbitrary features could be added log-linearly by evaluating log-posterior probabilities for each CN arc. In CN decoding, the word order of the combination is affected by the skeletal hypothesis. Hence the quality of the output from the combination also depends on the skeletal hypothesis. The hypothesis with the minimum average TER-score on aligning with all other hypothesis is proposed as an improved skeletal hypothesis: $E_s = \arg\min_{E \in E_i} \sum_{j=1}^{N_s} TER(E_j, E_i)$ where N_s is the number of systems and E_s is the skeletal hypothesis.

In (Rosti, Ayan, et al., 2007) system specific confidence scores are also introduced. The better the confidence score the higher the impact of that system. In the experimental part of this same work, three phrase-based (A,C,E), two hierarchical (B,D) and one syntax based (F) systems are combined. All of them are trained on the same data. The decoder weights are tuned to optimize TER for systems A and B and BLEU for the remaining systems. Decoder weight tuning is done on the NIST MT02 task. The results of the combination system were better than single system on all the metrics but for only TER and BLEU tuning. The experiments were performed on Arabic and Chinese NIST MT tasks.

N-Best Re-ranking. Another paper (Hildebrand & Vogel, 2009) presents a slightly different method where N-Best hypotheses are re-scored instead of building a synthesis CN of the MT outputs (as described in previous sub-section). The N-Best lists from all input systems are combined and then the best hypothesis is selected according to feature scores. Three types of features are: LM features, lexical features, N-Best list based features. The feature weights are modified using Minimum Error Rate Training (MERT). Experiments are performed to find the optimal size for N-Best list combination. Four systems are used and analysed on combination of two best systems and all the systems. 50-best list was found to be optimal size for both cases. The authors showed that the impact of gradually introducing a new system for combination becomes lower as the number of systems increases. Anyway the best result is obtained when all the systems are combined.

Approach based on N-best re-ranking is presented in (Li, Duan, Zhang, Li, & Zhou, 2009) using different way. The authors propose a method allowing to improve machine translation accuracy by leveraging translation consensus between MT systems: the hypothesis are re-ranked using augmented log-linear models with translation consensus based features.

The next Section presents the Driven Decoding concept were only the 1-bests provided by auxiliary systems are used in order to improve a primary system. An important issue to highlight is that, in the DD concept, auxiliary transcripts may be provided by black-box systems.

3 Overview of the Driven Decoding Concept

3.1 Driven Decoding

As said in the introduction part, our DDA implementation consists in adding several feature functions to the log-linear model before N-Best list re-ordering. Practically, after N-Best lists are generated by a primary system, additional scores are added to each line of the N-Best list file. Theses additional scores correspond to the distance between the current hypothesis decoded (called H) and the auxiliary translation available (T) : d(T,H). Let's say that for the LIA primary system, 2 auxiliary translations are available (from LIG and GOOGLE); in that case, 2 distance scores are added. The distance metric used in our experiments is described in the next Section and then N-Best reordering and combination processes are detailed.

3.2 Distance Metric used

We propose to use the BLEU as distance between systems. The BLEU score is the geometric mean of n-gram precision. Higher BLEU score suggest better translation quality. BLEU is used as evaluation metric and allows one to provide informations on the similarity between multiple systems. For DDA at sentence level we use a smoothed BLEU as presented in (Lin & Och, 2004) in order to get meaningful BLEU scores. For future experiments, additional scores related to other metrics could be used.

3.3 N-Best Reordering and Combination

The system combination is based on the 500-best outputs generated by the LIA primary system. Each N-best list is associated with a set of 14 scores: 1 LM score, 5 translation model scores, 1 distance-based reordering score, 6 lexicalized reordering scores and the word penalty. In addition we introduce distance metric scores for each sentence.

The score combination weights are optimized in order to maximize the BLEU score at the sentence level by using Margin Infused Relaxed Algorithm (MIRA) (Hasler, Haddow, & Koehn, 2011). We used the tool kbmira provided in the Moses package. The choice of MIRA against MERT (Minimum Error Rate Training) is motivated by a greater stability during our preliminary experiments. The MIRA parameters used are 100 iterations with a 0.001 C-parameter. For decoding, a global score is computed for each sentence (i.e. the log-linear score combination) and sentences are reordered according to the final combined score.

4 Baseline Systems

4.1 Used data

Both LIG and LIA systems were built using all the French and English data supplied for the WMT 2011 workshop, apart from the Gigaword monolingual corpora released by the LDC. Table 1 sums up the used data and introduces designations that we follow in the remainder of this paper to refer to corpora. Four corpora were used to build

CORPORA	DESIGNATION	SIZE (SENTENCES)
English-French Bilingual training		
News Commentary v6	*news-c*	116 k
Europarl v6	*euro*	1.8 M
UN corpus	*UN*	12 M
10^9 corpus	*giga*	23 M
English Monolingual training		
News Commentary v6	*mono-news-c*	181 k
Shuffled News Crawl corpus (from 2007 to 2011)	*news-s*	25 M
Europarl v6	*mono-euro*	1.8 M
Development		
newstest2008 + newssyscomb2009	*dev*	2,553
newstest2009	*tuning-mt-LIG-LIA*	2,525
Test		
newstest2010	*test10*	2,489
newstest2011	*test11*	3,005

Table 1. Used corpora to design LIA and LIG systems (from WMT 2011 evaluation campaign)

translation models: *news-c*, *euro*, *UN* and *giga*, while three were employed to train language models (LMs). Two bilingual corpora were devoted to model tuning: *tuning-mt-LIG-LIA* was used for the development of our two systems (Minimum Error Rate Training process (Och, 2003) for LIG and LIA systems), whereas *dev* was used to tune the weights for driven decoding. *test10* and *test11* were finally put aside to evaluate the driven decoding method.

4.2 LIA system characteristics

LIA system uses phrase-based translation models. All the data were provided by WMT 2011 campaign and data was tokenized with the tokenizer provided. Kneser-Ney discounted LMs were built from monolingual corpora using the SRILM toolkit (Stolcke, 2002), while bilingual corpora were aligned at the word-level using Giza++ (Och & Ney, 2003) or its multi-threaded version MGiza++ (Gao & Vogel, 2008) for the large corpora UN and giga. Phrase Table and lexicalized reordering models were built with Moses (Koehn et al., 2007). Finally, 14 features were used in the phrase-based models: 5 translation model scores, 1 distance-based reordering score, 6 lexicalized reordering score, 1 LM score and 1 word penalty score. Score weights were optimized on the WMT newstest2009 corpus (2525 sentences) according to BLEU thanks to the MERT method. More details on LIA system can be found in (Potet et al., 2011).

4.3 Baseline Performances

Table 2 summarizes the baseline results obtained by LIA system using case-insensitive BLEU for scoring (used everywhere in this paper). The evaluation performance is given on 3 corpora : dev refers to WMT newstest2008 + newssyscomb2009 (2553 sentences in total) ; tst10 refers to WMT newstest2010 (2489 sent.) and tst11 refers to newstest2011 (3005 sent.). For comparison purpose, we also scored the hypotheses given by LIG system and GOOGLE Translate (online system available in February 2012) on the same data sets. Each 1-best hypothesis of these systems will be used as auxiliary translations in the DD experiments. We are aware that there is a risk that the Google system, used in 2012, may actually contain WMT 2011 data in its MT models. This is, however, something we cannot control (but the performance obtained by Google Translate makes us think that this is not the case) ; this is why we use another auxiliary system (LIG) which we control totally. We also propose a combination baseline based on the MANY (Barrault, 2010) combination tool. MANY use a LM in order to decode a CN based on the input hypothesis. We present results based on target LM used in the LIA MT system. The MIRA algorithm is used to optimize the new set of log-linear weights that includes the driven decoding (DD) features (tuning is done on our *dev* set).

5 Experiments and Results

The DD approach described before was used on LIA primary system from which N-Best lists of size 500 (using the -uniq *Moses* option) were generated. As auxiliary translations, we used the 1-best hypotheses from the 2 systems (LIG and GOOGLE) whose performance are given in Table 2 . This Table shows the results obtained for DD of LIA system using LIG and GOOGLE 1-best hypotheses.

system	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 LIG GOOGLE LIA	26.3	30.46	30.6
DDA GOOGLE	26.37	30.16	30.52
DDA LIG	25.71	29.57	29.51
DDA LIG GOOGLE (4)	**26.41**	**30.44**	**30.91**
ORACLE between 1,2,3	*29.5*	*34.0*	*34.63*
ORACLE between 1,2,3,4	*30.0*	*34.7*	*35.2*

Table 2. Baseline performances measured for LIA, LIG and GOOGLE systems, as well as baseline system combination using MANY (open-source sys. combination toolkit) and performances for DDA of LIA System Using LIG and/or GOOGLE 1-best hypotheses

We observe that while the single LIA system is significantly better than the single LIG system, it can be improved using the 1-best from a weaker one (LIG). We

observe a cumulative improvement when two systems are used in the DDA process. DDA applied to LIA system improved approximately BLEU by 1 point compared to the best individual system. When DDA 1-best is added into the oracle, the score is significantly better which shows that DDA brings MT output variety. The results obtained by the DD approach are slightly better to the state of the art MANY combination while being fully different in concept. Moreover, the method presented in this paper is less expensive in terms of computations that the MANY system (which uses a LM while our method does not).

6 Deeper analysis of the DDA behavior

The Table 4 and Figure 1, show the distances between (DDA LIG+GOOGLE), (DDA LIG), (DDA GOOGLE) and all baseline systems. The similarities have been computed on test11 set. However they are very stable between dev, tst10 and tst11. We observe both for LIG and GOOGLE systems that DD increases the BLEU similarity. When DD uses only the GOOGLE system the result deviates from the LIG system. But when DD uses only the LIG system the deviation against the GOOGLE system is negligeable. This behavior can be explained by the fact that LIG and LIA systems are trained on similar data: the hypotheses are similar while those of GOOGLE differ more. The Figure 1 explicits the behavior introduced by the DDA: DD allows to move the space of the hypotheses relatively to the auxiliary systems. It is interesting to note that even the LIG system (worst system a priori) is close to the DDA hypothesis. The use of a BLEU similarity between systems allows one to find a consensus between hypothesis.

In the case of a full DDA (LIG + GOOGLE) we observe that the DD result is close to both LIG and GOOGLE.

By comparing the Oracle performance between the 3 single systems (ORACLE 1-best) with the performance obtained when combining the 3 systems + DDA (ORACLE 1-bests+DDA), we observe that DDA finds new hypotheses , originally not included in the 3 original systems.

system	LIA	DDA full	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 full	77.2	100	79.68	90.96

Table 3. Similarities (using BLEU metric) intersystems.

7 Further validation on a speech translation task

In order to validate the DD approach, we present another set of experiments on a speech translation task (translation of TED talks). In this case, 4 individual speech recogni-

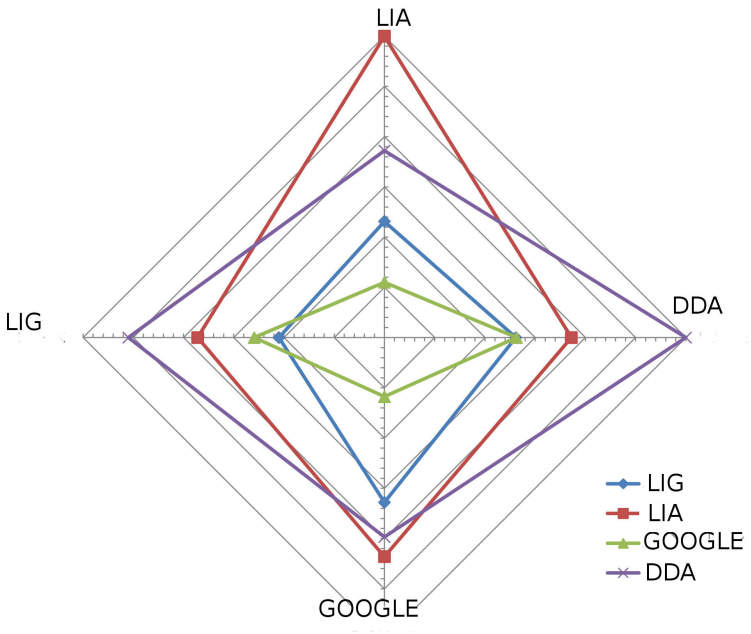

Fig. 1. Similarity between all the systems. The used metric is BLEU: for each vertex the associated system is used as reference compared to the others systems.

tion systems were available to provide their transcription (named systems 0, 1, 2 and 3 further). We propose to apply the DD method to a multiple source translation task (the multiple sources being the multiple ASR outputs) in order to improve final translation. Experiments are carried out on IWSLT 2011 data for the English-to-French speech translation task. We used four ASR and a ROVER between these four ASR. More details on the experimental setup (MT training data, MT systems) are provided in (Lecouteux, Besacier, & Blanchon, 2011). For the IWSLT 2011 evaluation (Lecouteux et al., 2011), we experimented two types of SLT combination: source combination (combine ASR outputs before translation) and target combination (combining MT outputs after translation). We propose here a third method using DD concept presented in this paper. The new system combination is based on the 500-best translated outputs generated from each ASR source system (the -distinct option of Moses is used to ensure that the hypotheses produced for a given sentence are all different inside a N-best list). All N-best lists are merged in a single one and for each hypothesis line, a BLEU distance score is computed against 1-best translation outputs of systems 0,1,2,3 and ROVER. So, in this case, driven decoding (DD) consists in introducing 5 additional scores before N-best re-ranking. The score combination weights are then optimized as presented in Section 3.3.

Table 7 shows the results for speech translation (for 3 different test sets of the IWSLT

evaluation campaign for speech translation - see more in (Lecouteux et al., 2011)) of:

- Single ASR system outputs (ASR-Sys0, ASR-Sys1, ASR-Sys2, ASR-Sys3)
- ROVER between single systems (ROVER 0123)
- Target combination, using CN decoding method explained in (Lecouteux et al., 2011)
- Driven Decoding as explained above (DD Sys0123R)

We observe a slight but significant improvement provided by DD, while this method only rerank N-best lists (the target combination method explained in (Lecouteux et al., 2011) is more complex and is based on the building of a confusion network from translation N-best lists). Another benefit of DD is the simplicity to perfuse auxiliary informations into the N-bests which are further rescored according to all the systems.

system	WER (test 2010) 1664 sent	BLEU (dev) 934 sent	BLEU (tst2010) 1664 sent	BLEU (tst2011) 818 sent
ASR-Sys0	17.1	18.94	21.95	26.41
ASR-Sys1	18.2	18.2	21.17	25.51
ASR-Sys2	17.4	19.36	22.08	25.83
ASR-Sys3	15.3	19.86	22.67	26.48
ROVER 0123	14.1	20.24	23.27	26.89
Target Combination	-	20.55	23.49	27.27
DDA Sys0123R	-	20.7	23.73	27.4

Table 4. WER and BLEU for each ASR systems and DD system, in the framework of IWSLT 2011 evaluation (Speech Translation of TED Talks)

Conclusion and perspectives

We have proposed a preliminary adaptation of driven decoding (DD) for machine translation. This method allows an efficient combination of machine translation systems, by rescoring the log-linear model at the N-best list level according to auxiliary systems: the basis technique is essentially guiding the search using previous system outputs. Different configurations were evaluated on the WMT 2011 evaluation corpus. The results show that the approach allows a significant improvement in BLEU score. Moreover the proposed approach yields similar (and even slightly better) results than state of the art combination methods. A validation of DD for a multisource (speech) translation task was also presented at the end of this paper. Also, the DD approach was recently validated ”live” for two more evaluation campaigns (results not reported here):

- An Arabic-to-French SMT evaluation campaign (TRAD campaign) where the use of a Google auxiliary translation significantly improved the performance of the primary LIG system.

– The IWSLT 2012 SMT evaluation campaign (english-to-french task) where, again, the use of an auxiliary translation significantly improved the performance of the primary system. More experimental details can be found in (Besacier, Lecouteux, Azouzi, & Luong Ngoc, 2012).

Our future work will focus on a DD integrated into the machine translation decoder: the interest is to perfuse all information sources into the primary search algorithm and to enable the evaluation of competing hypotheses while taking into account all the constraints and knowledge available.

Finally, another track will be the use of confidence measures generated for auxiliary systems to weight their contribution to the driven decoding.

References

Bangalore, S. (2001). Computing consensus translation from multiple machine translation systems. In *In proceedings of ieee automatic speech recognition and understanding workshop (asru-2001* (pp. 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., & Luong Ngoc, Q. (2012, dec). 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., & 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* (pp. 49–57). Columbus, OH, USA.

Hasler, E., Haddow, B., & Koehn, P. (2011). Margin infused relaxed algorithm for moses. In *The prague bulletin of mathematical linguistics* (p. 96:69-78).

Hildebrand, A. S., & 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*. New York: Cambridge University Press. Retrieved from `get-book.cfm?BookID=49784`

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Herbst, E. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics (acl), companion volume* (pp. 177–180). Prague, Czech Republic.

Lecouteux, B., Besacier, L., & Blanchon, H. (2011). Lig english-french spoken language translation system for iwslt 2011. In *Iwslt 2011.*

Lecouteux, B., Linares, G., Estève, Y., & 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., & 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., & 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., & 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., & 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., & Lefevre, F. (2011, jul). 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., & Dorr, B. (2007). Combining outputs from multiple machine translation systems. In *In proceedings of the north american chapter of the association for computational linguistics human language technologies* (pp. 228–235).

Rosti, A.-v., Matsoukas, S., & Schwartz, R. (2007). 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.