MANY improvements for WMT’11

Loïc Barrault
LIUM, University of Le Mans
Le Mans, France.
FirstName.LastName@lium.univ-lemans.fr

Abstract

This paper describes the development operated into MANY for the 2011 WMT system combination evaluation campaign. Hypotheses from French/English and English/French MT systems were combined with a new version of MANY, an open source system combination software based on confusion networks decoding currently developed at LIUM. MANY has been updated in order to optimize decoder parameters with MERT, which proves to find better weights. The system combination yielded significant improvements in BLEU score when applied on system combination data from two languages.

1 Introduction

This year, the LIUM computer science laboratory participated in the French-English system combination task at WMT’11 evaluation campaign. The system used for this task is MANY[1] (Barrault, 2010), an open source system combination software based on Confusion Networks (CN).

For this year evaluation, rather more technical than scientific improvements have been added to MANY. The tuning process has been improved by using MERT (Och, 2003) as a replacement of the numerical optimizer Condor (Berghen and Bersini, 2005). The impact of such change is detailed in section 3.

After the evaluation period, some experiments have been performed on the English-French system combination task. The results are presented in the section 5. Before that, a quick description of MANY, including recent developments, can be found in section 2.

2 System description

MANY is a system combination software (Barrault, 2010) based on the decoding of a lattice made of several Confusion Networks (CN). This is a widespread approach in MT system combination (Rosti et al., 2007; Shen et al., 2008; Karakos et al., 2008; Rosti et al., 2009). MANY can be decomposed in two main modules. The first one is the alignment module which actually is a modified version of TERp (Snover et al., 2009). Its role is to incrementally align the hypotheses against a backbone in order to create a confusion network. Those confusion networks are then connected together to create a lattice. This module uses different costs (which corresponds to a match, an insertion, a deletion, a substitution, a shift, a synonym and a stem) to compute the best alignment and incrementally build a confusion network. In the case of confusion network, the match (substitution, synonyms, and stems) costs are considered when the word in the hypothesis matches (is a substitution, a synonyms or a stems of) at least one word of the considered confusion sets in the CN.

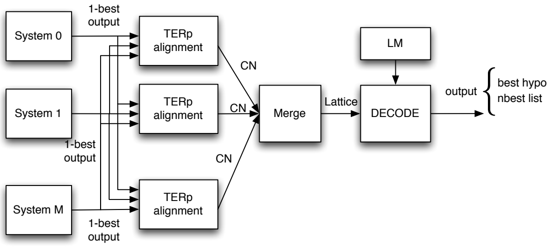

Figure 1: System combination based on confusion network decoding.

The second module is the decoder. This decoder is based on the token pass algorithm and it accepts as input the lattice previously created. The probabilities computed in the decoder can be expressed as follow :

[1] MANY is available at the following address http://www-lium.univ-lemans.fr/~barrault/MANY

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 135–139,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

$$log(P_W) \quad = \quad \sum_i \alpha_i \, log\Big(h_i(t)\Big) \qquad (1)$$

where t is the hypothesis, the α_i are the weights of the feature functions h_i. The following features are considered for decoding:

- The language model probability: the probability given by a 4-gram language model.
- The word penalty: penalty depending on the size (in words) of the hypothesis.
- The null-arc penalty: penalty depending on the number of null-arcs crossed in the lattice to obtain the hypothesis.
- System weights: each word receive a weight corresponding to the sum of the weights of all systems which proposed it.

3 Tuning

As mentioned before, MANY is made of two main modules: the alignment module based on a modified version of TERp and the decoder. Considering a maximum of 24 systems for this year evaluation, 33 parameters in total have to be optimized. By default, TERp costs are set to 0.0 for match and 1.0 for everything else. These costs are not correct, since a shift in that case will hardly be possible. TERp costs are tuned with Condor (a numerical optimizer based on Powell's algorithm, (Berghen and Bersini, 2005)). Decoder feature functions weights are optimized with MERT (Och, 2003). The 300-best list created at each MERT iteration is appended to the n-best lists created at previous iterations. This proves to be a more reliable tuning as shown in the following experiments.

During experiments, data from WMT'09 evaluation campaign are used for testing the tuning approach. *news-dev2009a* is used as development set, and *news-dev2009b* as internal test, these corpora are described in Table 1.

NAME	#sent.	#words	#tok
news-dev2009a	1025	21583	24595
news-dev2009b	1026	21837	24940

Table 1: WMT'09 corpora : number of sentences, words and tokens calculated on the reference.

For the sake of simplicity, the five best systems (ranking given by score on dev) are considered only. Baseline systems performances on dev and test are presented in Table 2.

Corpus	Sys0	Sys1	Sys2	Sys3	Sys4
Dev	18.20	17.83	20.14	21.06	17.72
Test	18.53	18.33	20.43	21.35	18.15

Table 2: Baseline systems performance on WMT'09 data (%BLEU).

The 2-step tuning protocol applied on *news-dev2009a*, when using MERT to optimize decoder feature functions weights provides the set of parameters presented in Table 3.

Costs:	Del	Stem	Syn	Ins	Sub	Shift
	0.87	0.91	0.94	0.90	0.98	1.21
Dec.:	LM weight		Word pen.		Null pen.	
	0.056		0.146		0.042	
Wghts.:	Sys0	Sys1	Sys2	Sys3	Sys4	
	-0.03	-0.21	-0.23	-0.28	-0.02	

Table 3: Parameters obtained with tuning decoder parameters with MERT.

Results on development corpus of WMT'09 (used as test set) are presented in Table 4. We can observe that 2-step tuning provides almost +0.9 BLEU point improvement on development corpus which is well reflected on test set with a gain of more than 0.8 BLEU. By using MERT, this improvement is increased to reach almost +2 BLEU point on dev corpus and +1.7 BLEU on test.

System	Dev	Test
Best single	21.06	21.35
MANY (2010)	**22.08**	**22.28**
MANY-2steps (2010)	**21.94**	**22.09**
MANY-2steps/MERT (2011)	**23.05**	**23.07**

Table 4: System Combination results on WMT'09 data (%BLEU-cased).

There are two main reasons for this improvement. The first one is the use of MERT which make use of specific heuristics to better optimize toward BLEU score. The second one is the fully log-linear interpolation of features functions scores operated into the decoder (previously, the word and null penalties were applied linearly).

4 2011 evaluation campaign

A development corpus, *newssyscombtune2011*, and a test set, *newssyscombtest2011*, described in Table 5, were provided to participants.

NAME	#sent.	#words	#tok
newssyscombtune2011	1003	23108	26248
newssyscombtest2011	2000	42719	48502

Table 5: Description of WMT'11 corpora.

Language model: The English target language models has been trained on all monolingual data provided for the translation tasks. In addition, LDC's Gigaword collection was used for both languages. Data corresponding to the development and test periods were removed from the Gigaword collections.

Sys. #	BLEU	TER	Sys. #	BLEU	TER
Sys0	29.86	52.46	Sys11	27.23	53.48
Sys1	29.74	51.74	Sys12*	26.82	54.23
Sys2	29.73	52.90	Sys13	26.25	55.60
Sys3	29.58	52.73	Sys14*	26.13	55.65
Sys4*	29.39	52.91	Sys15	25.90	55.69
Sys5	28.89	53.74	Sys16	25.45	56.92
Sys6	28.53	53.27	Sys17	25.23	56.09
Sys7*	28.31	54.22	Sys18	23.63	60.25
Sys8*	28.08	54.47	Sys19	21.90	63.65
Sys9*	27.98	53.92	Sys20	21.77	60.78
Sys10	27.46	54.60	Sys21	20.97	64.00
			Sys22	16.63	65.83
MANY-5sys				**31.83**	**51.27**
MANY-10sys				**31.75**	51.91
MANY-allsys				**30.75**	54.33

Table 6: Systems performance on *newssyscombtune2011* development data (%BLEU-cased). (* indicate a contrastive run)

Choosing the right number of systems to combine: Table 6 shows the performance of the input systems (ordered by BLEU score computed on *newssyscombtune2011*) and the result of 3 system combination setups. The difference in these setups only reside on the number of inputs to use for combination (5, 10 and all system outputs). Notice that the contrastive runs have not been used when combining 5 and 10 systems. The motivation for this is to benefit from the multi-site systems development which more likely provide varied outputs (*i.e.* different ngrams and word choice). The results show that combining 5 systems is slightly better than 10, but give more than 1 BLEU point improvement compared to combining all systems. Still, the combination always provide an improvement, which was not the case in last year evaluation.

The results obtained by combining 5 and 10 systems are presented in Table 7.

Sys. #	BLEU	TER	Sys. #	BLEU	TER
Sys0	29.43	52.01	Sys6	28.08	53.19
Sys1	29.15	51.30	Sys11	27.24	53.74
Sys2	28.87	52.82	Sys13	26.74	52.92
Sys3	28.82	52.57	Sys15	26.31	54.61
Sys5	28.08	53.19	Sys16	25.23	55.38
MANY (5sys)				**30.74**	**51.17**
MANY (10sys)				30.60	51.39

Table 7: Baseline systems performance on WMT'11 syscomb test data (%BLEU-cased).

Optimizing MANY on *newssyscombtune2011* corpus produced the parameter set presented in Table 8. We can see that the weights of all system are not proportional to the BLEU score obtained on the development corpus. This suggest that a better system selection could be found. This is even more probable since the weight of system Sys2 is positive (which imply a negative impact on each word proposed by this system), which means that when an hypothesis contains a word coming from this system, then its score is decreased.

Costs:	Del	Stem	Syn	Ins	Sub	Shift
	0.90	0.88	0.96	0.97	1.01	1.19
Dec.:	LM weight		Null pen.		Len pen.	
	0.0204		0.26		0.005	
Wghts.:	Sys0	Sys1	Sys2	Sys3	Sys5	
	-0.16	-0.30	0.008	-0.16	-0.09	

Table 8: Parameters obtained after tuning the system parameter using 5 hypotheses.

Table 9 contains the BLEU scores computed between the outputs of the five systems used during combination. An interesting observation is that the system which receive the bigger weight is the one which "distance"[2] against all other system outputs

[2] This "distance" is expressed in terms of ngrams agreement

	Sys0	Sys1	Sys2	Sys3	Sys5	mean
Sys0	-	53.59	62.67	64.60	62.50	60.84
Sys1	53.51	-	54.19	52.42	51.69	**52.95**
Sys2	62.72	54.28	-	65.49	63.09	***61.40***
Sys3	64.63	52.51	65.47	-	61.35	60.99
Sys5	62.55	51.78	63.10	61.37	-	59.70
mean	60.85	**53.04**	***61.36***	60.97	59.66	

Table 9: Cross-system BLEU scores computed on WMT'11 French-English test corpus outputs (%BLEU-cased).

Corpus	syscombtune2011		syscombtest2011	
	BLEU	TER	BLEU	TER
Sys0	35.99	**49.16**	34.36	**49.78**
Sys1	32.99	51.90	30.73	52.52
Sys2	32.41	52.77	29.85	53.61
Sys3	32.40	51.26	30.48	52.20
Sys4	32.30	52.21	31.02	52.49
MANY	**36.81**	49.74	**34.51**	50.54

Table 11: Systems and combination performance on WMT'11 french data (%BLEU-cased).

is the highest, whereas the "closest" system get the smallest weight. This suggests that systems closer to other systems tends to be less useful for system combination. This is an interesting behaviour which has to be explored deeper and validated on other tasks and corpora.

5 MANY for french outputs

After the evaluation period, some experiments have been conducted in order to combine french outputs. The main difference lie in the fact that linguistic resources are not easily or freely available for that kind of language. Therefore, instead of using TERp with *relax*[3] shift constraint, the *strict* constraint was used (shifts occur only when a match is found).

The available data are detailed in the Table 10.

NAME	#sent.	#words	#tok
syscombtune	1003	24659	29171
syscombtest	2000	45372	53970

Table 10: Description of WMT'11 corpora for system combination in french.

The results obtained are presented in Table 11. The BLEU score increase by more than 0.8 point but the TER score decrease by 0.58. The metric targeted during tuning is BLEU, which can explain the improvement in that metric. When dealing with english text, the only case where such behaviour is observed is when combining all systems (see Table 6.

6 MANY technical news

Several improvements have been performed on MANY. The decoder is now based on a fully log-linear model (whereas before, the word and null penalties were applied linearly). Using MERT to tune the decoder parameters is therefore possible and allows to reach bigger improvement compared to using Condor. This is probably due to the fact that MERT uses several heuristics useful for tuning on BLEU score.

In order to facilitate the use of MANY, it has been integrated in the Experiment Management System, EMS - (Koehn, 2010). An experiment can now be setup/modified/re-run easily by modifying a single configuration file. The default behavior of this framework is to perform 3 runs of MERT in parallel (using torque) and take the best optimization run. Apart from avoiding local maximum, the procedure allows to see the variability of the optimization process and report more realistic results (for example, by taking the average).

7 Conclusion and future work

For WMT'11 system combination evaluation campaign, several rather technical improvements have been performed into MANY. By homogenizing the log-linear model used by the decoder and utilizing MERT for tuning, MANY achieves improvements of more than 2 BLEU points on WMT'09 data and about 1.3 BLEU point on *newssyscombtest2011* relatively to the best single system. Moreover, a dry-run operated on french data shows a promising result with an improvement of more than 0.8 BLEU points. This will be further explored in the future.

MANY can benefit from various information. At the moment, the decision taken by the decoder mainly depends on a target language model. This is clearly not enough to achieve greater performances. The next issues which will be addressed within the MANY framework is to estimate good confidence measure to use in place of the systems

[3] Shifts can occur when a match, a stem, a synonym or a paraphrase is found.

priors. These confidences measures have to be related to the system performances, but also to the complementarity of the systems considered.

8 Acknowledgement

This work has been partially funded by the European Union under the EuroMatrix Plus project (http://www.euromatrixplus.net, IST-2007.2.2-FP7-231720)

References

[Barrault, 2010] Barrault, L. (2010). MANY : Open source machine translation system combination. *Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation*, 93:147–155.

[Berghen and Bersini, 2005] Berghen, F. V. and Bersini, H. (2005). CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175.

[Karakos et al., 2008] Karakos, D., Eisner, J., Khudanpur, S., and Dreyer, M. (2008). Machine translation system combination using ITG-based alignments. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*, pages 81–84, Columbus, Ohio, USA.

[Koehn, 2010] Koehn, P. (2010). An experimental management system. *The Prague Bulletin of Mathematical Linguistics*, 94:87–96.

[Och, 2003] Och, F. (2003). Minimum error rate training in statistical machine translation. In *ACL*, Sapporo, Japan.

[Rosti et al., 2007] Rosti, A.-V., Matsoukas, S., and Schwartz, R. (2007). Improved word-level system combination for machine translation. In *Association for Computational Linguistics*, pages 312–319.

[Rosti et al., 2009] Rosti, A.-V., Zhang, B., Matsoukas, S., , and Schwartz, R. (2009). Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task. In *EACL/WMT*, pages 61–65.

[Shen et al., 2008] Shen, W., Delaney, B., Anderson, T., and Slyh, R. (2008). The MIT-LL/AFRL IWSLT-2008 MT System. In *International Workshop on Spoken Language Translation*, Hawaii, U.S.A.

[Snover et al., 2009] Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2009). TER-Plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation Journal*.