Addendum to *Structured Ramp Loss Minimization for Machine Translation*

Kevin Gimpel and **Noah A. Smith**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{kgimpel,nasmith}@cs.cmu.edu

This note is a brief addendum to Gimpel and Smith (2012). The experiments reported therein using RAMPION did not accumulate the k-best lists across iterations, but rather only used the latest set of k-best lists for each iteration of optimization. They also only used a single ramp loss variant. This works well enough for phrase-based models with small feature sets. However, in subsequent experiments using RAMPION for training hierarchical phrase-based models, syntax-based models, and phrase-based models with large feature sets, we found that it generally works better to accumulate k-best lists across iterations and to use a different ramp loss variant. The provided code for use with Moses accumulates k-best lists by default, as does the implementation in cdec (Dyer et al., 2010). My dissertation (Gimpel, 2012) includes many experimental results comparing ramp loss functions across language pairs, test sets, and systems; we find that two ramp losses—ramp loss 3 and soft ramp loss 2—consistently work best. Optimization algorithms are given in the thesis and are implemented in the RAMPION software release.

Since the dissertation experiments, we have also found that when using a single reference translation, it helps to use the modified sentence-level BLEU approximations from Nakov et al. (2012). When using 4 references, however, this doesn't seem to affect the results very much and standard BLEU+1 works well. Some results are shown below for the single-reference translation tasks from my dissertation experiments. The modified BLEU+1 cost functions do smoothing for all values of n rather than only for $n > 1$, smooth the brevity penalty, and ground the precision term (Nakov et al., 2012). First is English-to-Malagasy phrase-based translation:

loss	cost	tune (LR)	test (LR)
ramp 3	BLEU+1	17.56 (0.94)	15.05 (0.91)
	BLEU+1, smooth all n	17.54 (0.93)	15.10 (0.91)
	BLEU+1, smooth all n, smooth BP	17.93 (0.98)	15.51 (0.95)
	BLEU+1, smooth all n, smooth BP, ground prec.	18.03 (0.99)	15.62 (0.97)
soft ramp 2	BLEU+1	17.30 (0.90)	14.72 (0.89)
	BLEU+1, smooth all n, smooth BP	17.61 (0.94)	15.05 (0.92)
	BLEU+1, smooth all n, smooth BP, ground prec.	17.82 (0.95)	15.25 (0.94)
MERT, run 1	BLEU	18.10 (1.00)	15.52 (0.98)
MERT, run 2	BLEU	18.01 (1.01)	15.39 (0.98)
MERT, run 3	BLEU	18.05 (1.00)	15.34 (0.97)

The table shows the highest BLEU scores reached during tuning and the test BLEU scores obtained using those parameters. In parentheses the ratio of the length of the hypothesis document to the reference translation is shown ("length ratio"; LR). Smoothing the brevity penalty and grounding the precision term lead to

length ratios closer to 1 and higher BLEU scores. The tune and test BLEU scores approach or exceed those obtained by MERT.

Next is German-to-English phrase-based translation:

loss	cost	tune (LR)	test 1 (LR)	test 2 (LR)	test 3 (LR)
ramp 3	BLEU+1	16.21 (0.92)	18.95 (0.91)	20.66 (0.94)	19.16 (0.93)
	BLEU+1, smooth all n, smooth BP	16.57 (0.94)	19.42 (0.93)	21.21 (0.96)	19.60 (0.95)
	BLEU+1, smooth all n, smooth BP, ground p.	16.94 (1.00)	19.76 (0.99)	21.13 (1.02)	19.95 (1.00)
soft ramp 2	BLEU+1	16.53 (0.97)	19.87 (0.95)	21.41 (0.99)	19.80 (0.97)
	BLEU+1, smooth all n, smooth BP	16.86 (1.01)	19.93 (0.99)	21.04 (1.03)	19.91 (1.01)
	BLEU+1, smooth all n, smooth BP, ground p.	16.83 (1.02)	19.93 (1.00)	20.78 (1.03)	19.75 (1.02)
MERT, run 1	BLEU	17.10 (1.00)	19.94 (0.99)	21.04 (1.02)	19.93 (1.01)
MERT, run 2	BLEU	17.17 (1.00)	19.95 (0.98)	21.38 (1.02)	20.03 (1.00)
MERT, run 3	BLEU	16.87 (1.00)	19.81 (0.99)	20.93 (1.02)	19.71 (1.01)

We again see that the modified cost functions lead to a length ratio nearer to 1, although occasionally the translations become too long, harming BLEU scores slightly.

Nonetheless, we have shown for two single-reference phrase-based translation tasks that the modified BLEU+1 cost functions can improve the length ratios and BLEU scores with RAMPION when using ramp loss 3 and soft ramp loss 2. These results complement those using PRO from Nakov et al. (2012). These and other BLEU+1 variants are implemented in the RAMPION v0.2 software release, which is available at `www.ark.cs.cmu.edu/MT`.

References

C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. of ACL*.

K. Gimpel and N. A. Smith. 2012. Structured ramp loss minimization for machine translation. In *Proc. of NAACL*.

K. Gimpel. 2012. *Discriminative Feature-Rich Modeling for Syntax-Based Machine Translation*. Ph.D. thesis, Carnegie Mellon University.

P. Nakov, F. Guzmán, and S. Vogel. 2012. Optimizing for sentence-level BLEU+1 yields short translations. In *Proc. of COLING*.