# Learning Neural Sequence-to-Sequence Models from Weak Feedback with Bipolar Ramp Loss

**Laura Jehl*** **Carolin Lawrence***
Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{jehl, lawrence}@cl.uni-heidelberg.de

**Stefan Riezler**
Computational Linguistics & IWR
Heidelberg University
69120 Heidelberg, Germany
riezler@cl.uni-heidelberg.de

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

## Abstract

In many machine learning scenarios, supervision by gold labels is not available and consequently neural models cannot be trained directly by maximum likelihood estimation. In a weak supervision scenario, metric-augmented objectives can be employed to assign feedback to model outputs, which can be used to extract a supervision signal for training. We present several objectives for two separate weakly supervised tasks, machine translation and semantic parsing. We show that objectives should actively discourage negative outputs in addition to promoting a surrogate gold structure. This notion of bipolarity is naturally present in ramp loss objectives, which we adapt to neural models. We show that bipolar ramp loss objectives outperform other non-bipolar ramp loss objectives and minimum risk training on both weakly supervised tasks, as well as on a supervised machine translation task. Additionally, we introduce a novel token-level ramp loss objective, which is able to outperform even the best sequence-level ramp loss on both weakly supervised tasks.

## 1 Introduction

Sequence-to-sequence neural models are standardly trained using a maximum likelihood estimation (MLE) objective. However, MLE training requires full supervision by gold target structures, which in many scenarios are too difficult or expensive to obtain. For example, in semantic parsing for question-answering it is often easier to collect gold answers rather than gold parses (Clarke et al., 2010; Berant et al., 2013; Pasupat and Liang, 2015; Rajpurkar et al., 2016, inter alia). In machine translation, there are many domains for which no gold references exist, although cross-lingual document-level links are present for many multilingual data collections.

In this paper we investigate methods where a supervision signal for output structures can be extracted from weak feedback. In the following, we use *learning from weak feedback*, or *weakly supervised learning*, to refer to a scenario where output structures generated by the model are judged according to an external metric, and this feedback is used to extract a supervision signal that guides the learning process. Metric-augmented sequence-level objectives from reinforcement learning (Williams, 1992; Ranzato et al., 2016), minimum risk training (MRT) (Smith and Eisner, 2006; Shen et al., 2016) or margin-based structured prediction objectives (Taskar et al., 2005; Edunov et al., 2018) can be seen as instances of such algorithms.

In natural language processing applications, such algorithms have mostly been used in combination with *full supervision tasks*, allowing to compute a feedback score from metrics such as BLEU or F-score that measure the similarity of output structures against gold structures. Our main interest is in *weak supervision tasks* where the calculation of a feedback score cannot fall back onto gold structures. For example, matching proposed answers to a gold answer can guide a semantic parser towards correct parses, and matching proposed translations against linked documents can guide learning in machine translation.

In such scenarios the judgments by the external metric may be unreliable and thus unable to select a good update direction. It is our intuition that

*Both authors contributed equally to this publication.



Transactions of the Association for Computational Linguistics, vol. 7, pp. 233–248, 2019. Action Editor: Phil Blunsom.
Submission batch: 11/2018; Revision batch: 1/2019; Published 5/2019.
© 2019 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

a more reliable signal can be produced by not just encouraging outputs that are good according to weak positive feedback, but also by actively discouraging bad structures. In this way, a system can more effectively learn what distinguishes good outputs from bad ones. We call an objective that incorporates this idea a *bipolar* objective. The bipolar idea is naturally captured by the structured ramp loss objective (Chapelle et al., 2009), especially in the formulation by Gimpel and Smith (2012) and Chiang (2012), who use ramp loss to separate a *hope* from a *fear* output in a linear structured prediction model. We employ several ramp loss objectives for two weak supervision tasks, and adapt them to neural models.

First, we turn to the task of semantic parsing in a setup where only question-answer pairs, but no gold semantic parses, are given. We assume a baseline system has been trained using a small supervised data set of question-parse pairs under the MLE objective. The goal is to improve this system by leveraging a larger data set of question-answer pairs. During learning, the semantic parser suggests parses for which corresponding answers are retrieved. These answers are then compared to the gold answer and the resulting weak supervision signal guides the semantic parser towards finding correct parses. We can show that a bipolar ramp loss objective can improve upon the baseline by over 12 percentage points in F1 score.

Second, we use ramp losses on a machine translation task where only weak supervision in the form of cross-lingual document-level links is available. We assume a translation system has been trained using MLE on out-of-domain data. We then investigate whether document-level links can be used as a weak supervision signal to adapt the translation system to the target domain. We formulate ramp loss objectives that incorporate bipolar supervision from relevant and irrelevant documents. We also present a metric that allows us to include bipolar supervision in an MRT objective. Experiments show that bipolar supervision is crucial for obtaining gains over the baseline. Even with this very weak supervision, we are able to achieve an improvement of over 0.4% BLEU over the baseline using a bipolar ramp loss.

Finally, we turn to a fully supervised machine translation task. In supervised learning, MLE training in a fully supervised scenario has also been associated with two issues. First, it can cause *exposure bias* (Ranzato et al., 2016) because during training the model receives its context from the gold structures of the training data, but at test time the context is drawn from the model distribution instead. Second, the MLE objective is agnostic to the final evaluation metric, causing a *loss-evaluation mismatch* (Wiseman and Rush, 2016). Our experiments use a similar setup as Edunov et al. (2018), who apply structured prediction losses to two fully supervised sequence-to-sequence tasks, but do not consider structured ramp loss objectives. Like our predecessors, we want to understand whether training a pre-trained machine translation model further with a metric-informed sequence-level objective will improve translation performance by alleviating the above-mentioned issues. By gauging the potential of applying bipolar ramp loss in a full supervision scenario, we achieve best results for a bipolar ramp loss, improving the baseline by over 0.4% BLEU.

In sum, we show that bipolar ramp loss is superior to other sequence-level objectives for all investigated tasks, supporting our intuition that a bipolar approach is crucial where strong positive supervision is not available. In addition to adapting the ramp loss objective to weak supervision, our ramp loss objective can also be adapted to operate at the token level, which makes it particularly suitable for neural models as they produce their outputs token by token. A token-level objective also better emulates the behavior of the ramp loss for linear models, which only update the weights of features that differ between hope and fear. Finally, the token-level objective allows us to capture token-level errors in a setup where MLE training is not available. Using this objective, we obtain additional gains on top of the sequence-level ramp loss for weakly supervised tasks.

## 2 Related Work

Training neural models with metric-augmented objectives has been explored for various NLP tasks in supervised and weakly supervised scenarios. MRT for neural models has previously been used for machine translation (Shen et al., 2016) and semantic parsing (Liang et al., 2017; Guu et al., 2017).[1] Other objectives based on classical

[1]Note that Liang et al. (2017) refer to their objective as an instantiation of REINFORCE, however they build an average over several outputs for one input and thus the objective more accurately falls under the heading of MRT.



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

structured prediction losses have been used for both machine translation and summarization (Edunov et al., 2018), as well as semantic parsing (Iyyer et al., 2017; Misra et al., 2018). Objectives inspired by REINFORCE have, for example, been applied to machine translation (Ranzato et al., 2016; Norouzi et al., 2016), semantic parsing (Liang et al., 2017; Mou et al., 2017; Guu et al., 2017), and reading comprehension (Choi et al., 2017; Yang et al., 2017).[2]

Misra et al. (2018) are the first to compare several objectives for neural semantic parsing. For semantic parsing, they find that objectives employing structured prediction losses perform best. Edunov et al. (2018) compare different classical structured prediction objectives including MRT on a fully supervised machine translation task. They find MRT to perform best. However, they only obtain larger gains by interpolating MRT with the MLE loss. Neither Misra et al. (2018) nor Edunov et al. (2018) investigate objectives that correspond to the bipolar ramp loss that is central in our work.

The ramp loss objective (Chapelle et al., 2009) has been applied to supervised phrase-based machine translation (Gimpel and Smith, 2012; Chiang, 2012). We adapt these objectives to neural models and adapt them to incorporate bipolar weak supervision, while also introducing a novel token-level ramp loss objective.

## 3 Neural Sequence-to-Sequence Learning

Our neural sequence-to-sequence models utilize an encoder-decoder setup (Cho et al., 2014; Sutskever et al., 2014) with an attention mechanism (Bahdanau et al., 2015). Specifically, we employ the framework NEMATUS (Sennrich et al., 2017). Given an input sequence $x = x_1, x_2, \ldots x_{|x|}$, the probability that a model assigns for an output sequence $y = y_1, y_2, \ldots y_{|y|}$ is given by $\pi_w(y|x) = \prod_{j=1}^{|y|} \pi_w(y_j|y_{<j}, x)$. Using beam search, we can obtain a sorted $k$-best list $\mathcal{K}(x)$ of most likely to least likely outputs and we define the most likely output as $\hat{y} = \arg\max_{y \in \mathcal{K}(x)} \pi_w(y|x)$.

[2]We do not use REINFORCE because its updates are based on only one sampled model output, which can lead to high variance. Because it is possible for us to obtain feedback for more than one model output, we employ the more robust MRT that calculates an average over several outputs.

**Maximum Likelihood Estimation (MLE).** Prior to employing metric-augmented objectives, we assume that a model has been pre-trained with a maximum likelihood estimation (MLE) objective. Given inputs $x$ and gold structures $\bar{y}$, the parameters of the neural network are updated using Stochastic Gradient Descent (SGD) with mini-batches of size $M$, leading to the following objective:

$$\mathcal{L}_{MLE} = -\frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{|\bar{y}|} \log \pi_w(\bar{y}_{m,j}|\bar{y}_{m,<j}, x_m). \tag{1}$$

**Minimum Risk Training (MRT).** We compare our ramp loss objectives to MRT (Shen et al., 2016), which uses an external metric to assign rewards to model outputs. Given an input $x$, $S$ outputs are sampled from the model distribution and updates are performed based on the following MRT objective:

$$\mathcal{L}_{\text{MRT}} = -\frac{1}{M} \sum_{m=1}^{M} \frac{1}{S} \sum_{s=1}^{S} \pi_w(y_{m,s}|x_m)\delta(y_{m,s}), \tag{2}$$

where $\delta(y_{m,s})$ is the reward returned for $y_{m,s}$ by the external metric, and $\pi_w(y_{m,s}|x_m)$ is a distribution over outputs that is normalized over $S$ samples and can be controlled for sharpness by a temperature parameter.[3] Following Shen et al. (2016), we use a baseline term $b(x_m)$ that acts as a control variate for variance reduction of the stochastic gradient (Williams, 1992; Greensmith et al., 2004) and allows negative updates for rewards smaller than the baseline. We compute this term by sampling $S'$ outputs from the model distribution such that $b(x) = -\frac{1}{S'} \sum_{s'=1}^{S'} \delta(y_{s'})$.

**Ramp Loss Objectives.** Our ramp loss objectives can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{RAMP}} = & \frac{1}{M} \sum_{m=1}^{M} \pi_w(y_m^-|x_m) \\ & - \frac{1}{M} \sum_{m=1}^{M} \pi_w(y_m^+|x_m), \end{aligned} \tag{3}$$

where $y^-$ is a *fear* output that is to be discouraged and $y^+$ is a *hope* output that is to be encouraged.

[3]We follow the implementation of MRT in NEMATUS with its default settings, including de-duplication of samples and setting the temperature parameter to $\alpha = 0.005$. In case of fully supervised MT where the question arises whether to include the reference in the sample, we choose not to include it in order to be comparable with Edunov et al. (2018) who also do not include it.



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

| Name | $y^+$ | $y^-$ |
|---|---|---|
| RAMP | $\arg\max_{y \in \mathcal{P}(x)} \pi_w(y\|x)$ | $\arg\max_{y \in \mathcal{N}(x)} \pi_w(y\|x)$ |
| RAMP1 | $\hat{y}$ | $\arg\max_{y \in \mathcal{N}(x)} \pi_w(y\|x)$ |
| RAMP2 | $\arg\max_{y \in \mathcal{P}(x)} \pi_w(y\|x)$ | $\hat{y}$ |

Table 1: Configurations for $y^+$ and $y^-$ for semantic parsing. We abbreviate $\mathcal{P}(x) = \mathcal{K}(x) : \delta(y) = 1$, which is the most likely output in the $k$-best list $\mathcal{K}(x)$ that leads to the correct answer, and $\mathcal{N}(x) = \mathcal{K}(x) : \delta(y) = 0$, which is the most likely output in the $k$-best list $\mathcal{K}(x)$ that leads to the wrong answer.

Intuitively, $y^-$ should be an output which has high probability, but receives a bad reward from the external metric. Analogously, $y^+$ should be an output which has high probability and receives a high reward from the external metric. The concrete instantiations of $y^-$ and $y^+$ depend on the underlying task and are thus deferred to the respective sections below (see Tables 1, 4, and 7). The RAMP loss defined in equation (3) has been introduced as equation (8) in Gimpel and Smith (2012). This loss naturally incorporates a bipolarity principle by including both *hope* and *fear* into one objective. An alternative formulation of ramp loss can be given by favoring the current model prediction, that is, setting $y^+ = \hat{y}$, and searching for a *fear* output. This has been called "cost-augmented decoding'' and been formalized in equation (6) in Gimpel and Smith (2012). This loss dates back to the ''margin-rescaled hinge loss" of Taskar et al. (2004) and will be called RAMP1 in the following. The converse approach has been called "cost-diminished decoding" and been formalized in equation (7) in Gimpel and Smith (2012). Here the model prediction is penalized by setting $y^- = \hat{y}$ and searching for a *hope* output. This objective has been called "direct loss" in Hazan et al. (2010), and will be called RAMP2 in the following.

Finally, we introduce a ramp loss objective that can operate on the token level. To be able to adjust individual tokens, we move to log probabilities, so that the sequence decomposes as a sum over individual tokens and it is possible to ignore tokens while encouraging or discouraging others. This leads to the RAMP-T objective:

$$\mathcal{L}_{\text{RAMP-T}} = \tag{4}$$
$$\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{|y_m^-|}\tau_{m,j}^- \log \pi_w(y_{m,j}^-|y_{m,<j}, x_m)$$
$$-\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{|y_m^+|}\tau_{m,j}^+ \log \pi_w(y_{m,j}^+|y_{m,<j}, x_m),$$

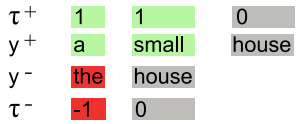


Figure 1: Settings for token-level rewards $\tau+$ and $\tau^-$ for hope output $y^+$ = "*a small house*" and fear output $y^-$ = "*the house*".

where $\tau_{m,j}^+$ and $\tau_{m,j}^-$ are set to 0, 1 or $-1$ depending on the decision whether the corresponding token $y_{m,j}^+/y_{m,j}^-$ should be left untouched, encouraged or discouraged. Concretely, we define:

$$\tau_{m,j}^+ = \begin{cases} 0 & \text{if } y_{m,j}^+ \in y^- \\ 1 & \text{else} \end{cases} \tag{5}$$

and

$$\tau_{m,j}^- = \begin{cases} 0 & \text{if } y_{m,j}^- \in y^+ \\ -1 & \text{else.} \end{cases} \tag{6}$$

With this definition, tokens that appear in both $y^+$ and $y^-$ are left untouched, whereas tokens that appear only in the hope output are encouraged, and tokens that appear only in the fear output are discouraged (see Figure 1 for an example). This more fine-grained contrast allows the model to learn what distinguishes a good output from a bad one more effectively.[4]

## 4 Semantic Parsing

**Ramp Loss Objectives.** In semantic parsing for question answering, natural language questions are mapped to machine readable parses. Such a parse, $y$, can be executed against a database that returns an answer $a$. This answer $a$ can be compared to the available gold answer $\bar{a}$ and the following metric can be defined:

$$\delta(y) = \begin{cases} 1 & \text{if } a = \bar{a} \\ 0 & \text{else.} \end{cases} \tag{7}$$

[4] An implementation of the RAMP objectives can be found at `https://github.com/carhaas/nematus`.



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

For RAMP, $y^+$ is defined as the most probable output in the $k$-best list $\mathcal{K}(x)$ that leads to the correct answer, that is, where $\delta(y) = 1$. In contrast, $y^-$ is defined as the most probable output in $\mathcal{K}(x)$ that does not lead to the correct answer, namely, where $\delta(y) = 0$. The definitions of $y^+$ and $y^-$ for this objective and the related ramp loss objectives can be found in Table 1. If $y^+$ or $y^-$ are found, the parse is cached as a hope or fear output, respectively, for the corresponding input $x$. If at a later point $y^+$ or $y^-$ cannot be found in the current $k$-best list, then previously cached outputs are accessed instead. Should no cached output exist, the corresponding sample is skipped.

**Experimental Setup.** Our experiments are conducted on the NLMAPS v2 corpus (Lawrence and Riezler, 2018), which is a publicly available corpus[5] for geographical questions that can be answered with the OPENSTREETMAP database.[6] The corpus is a recent extension of its predecessor (Haas and Riezler, 2016), which has been used in Kočiský et al. (2016) or Duong et al. (2018).

For each question, the corpus provides both gold parses and gold answers that can be obtained by executing the parses against the database. We take a random subset of 2,000 question-parse pairs to train an initial model $\pi_w$ with the MLE objective. Following Lawrence and Riezler (2018), we take a pre-order traversal of the tree-structured parses to obtain individual tokens. A further 1,843 and 2,000 instances of the corpus are retained for development and test set, respectively. For the remaining 22,766 questions, we assume that no gold parses exist and only gold answers are available. With the gold answers as a guide, the initial model $\pi_w$ is further improved using the metric-augmented objectives of Section 3 and the metric defined in equation (7).

The model has 1,024 hidden units (GRUs) and word embeddings of size 1,000. The optimal learning rate was chosen in preliminary experiments on the development set and is set to 0.1. Gradients are clipped to 1.0 if they exceed a value of 1.0 and the sentence length is capped at 200. In the case of the MRT objectives, we set $S = S' = 10$. For the RAMP objectives the size of the $k$-best list $\mathcal{K}$ is 10. For objectives with minibatches, the size of a minibatch is $M = 80$ and validation on the development set is performed after every 100 updates. For objectives where updates are performed after each seen input, the validation is run after every 8,000 updates, leading to the same number of seen inputs compared to the objectives with minibatches.

For validation and at test time, the most likely parse is obtained after a beam search with a beam of size 12. The obtained parse is executed against the database to retrieve its corresponding answer, which is compared to the available gold answer. We define recall as the percentage of correct answers in the entire set and precision as the percentage of correct answers in the set of non-empty answers. The harmonic mean of recall and precision constitutes the F1 score. The stopping point is determined by the highest F1 score on the development set after 30 validations or 30 days of run time[7] and corresponding results are reported on the test set. To measure statistical significance between models we use an approximate randomization test (Noreen, 1989).

**Experimental Results.** Results using the various ramp loss objectives as well as MRT are shown in Table 2. MRT outperforms the MLE baseline by about 6 percentage points in F1 score. RAMP1 performs worse than MRT, but can still significantly outperform the baseline by 3.05 points in F1 score. RAMP2 performs better than RAMP1, but outperforms MRT only nominally.

In contrast to this, by carefully selecting both a hope and fear parse, RAMP achieves a significant further 5.43 points in F1 score over MRT. By incorporating token-level feedback, our novel objective RAMP-T outperforms all other models significantly and beats the baseline by over 12 points in F1 score. Compared with RAMP, RAMP-T can take advantage of the token-level feedback that allows a model to determine which tokens in the hope output are instrumental to obtain a positive reward but are missing in the fear output. Analogously, it is possible to identify which tokens in the fear output lead to an incorrect parse, rather than also punishing the tokens in the fear output which are actually correct.

MRT is not naturally a bipolar objective. It can only discourage wrong parses if the baseline is larger than 0. Investigating the value of the baseline for 10,000 instances shows that in 37%

[5] https://www.cl.uni-heidelberg.de/statnlpgroup/nlmaps/.

[6] https://www.openstreetmap.org.

[7] The 30-day mark was only hit by RAMP2.



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

| | | $M$ | % F1 | Δ |
|---|---|---|---|---|
| 1 | MLE | | 57.45 | |
| 2 | MRT | 1 | $63.60 \pm 0.02$ | $+ 6.15$ |
| 3 | RAMP1 | 80 | $60.50 \pm 0.01$ | $+ 3.05$ |
| 4 | RAMP2 | 80 | $64.22 \pm 0.00$ | $+ 6.77$ |
| 5 | RAMP | 80 | $69.03 \pm 0.04$ | $+ 11.58$ |
| 6 | RAMP-T | 80 | $69.87 \pm 0.02$ | $+ 12.42$ |

Table 2: Answer F1 scores on the NLMAPS v2 test set for various objectives, averaged over two independent runs. $M$ is the minibatch size. All models are statistically significant from each other at $p < 0.01$, except the pair (2, 4).

| | | $M$ | % F1 | Δ |
|---|---|---|---|---|
| 1 | MLE | | 57.45 | |
| 2 | MRT | 1 | $63.60 \pm 0.02$ | $+ 6.15$ |
| 3 | MRT NEG | 1 | $65.93 \pm 0.16$ | $+ 8.48$ |
| 4 | RAMP M=1 | 1 | $66.78 \pm 0.21$ | $+ 9.33$ |
| 5 | RAMP | 80 | $69.03 \pm 0.04$ | $+ 11.58$ |

Table 3: Answer F1 scores on the NLMAPS v2 test set for RAMP and the MRT objective as well as two further objectives, which help crystallize the difference between the two former objectives, averaged over two independent runs. $M$ is the minibatch size. All models are statistically significant from each other at $p < 0.01$, except the pair (3, 4).

of the cases the baseline is 0 (i.e., none of the sampled parses leads to the correct answer). As a result, 37% of the time, wrong parses are ignored rather than discouraged. To explore the importance of always discouraging wrong parses, we introduce the objective MRT NEG: it modifies the feedback for parses with a wrong answer to be $-1$ rather than 0, which resembles the fear output that is discouraged in the RAMP objective. With this change, the MRT objective always behaves in a bipolar manner, irrespective of the baseline's value. As a consequence, MRT NEG can significantly outperform MRT by 2.33 points in F1 score (see Table 3). This showcases the importance of utilizing bipolar supervision and it constitutes an important finding compared to previous approaches (Liang et al., 2017; Misra et al., 2018), where the feedback is defined to lie in the range of $[0, 1]$.

However, MRT NEG still falls short of RAMP by 3.1 points in F1 score. This could be because of the different batch sizes, as MRT uses a batch size of 1, whereas RAMP employs a batch size of 80. To ensure that the difference between the objectives does not stem from this difference, we run an experiment with RAMP where the batch size is also set to 1 (i.e., RAMP M=1). Crucially, it still significantly outperforms MRT. At the same time, it does, however, have a lower F1 score than RAMP (see Table 3). This showcases the importance of using a larger minibatch size, so that an average over several inputs is computed before updating. In fact, its F1 score is on par with the MRT NEG objective, which uses the same minibatch size and incorporates bipolar supervision just as RAMP does. However, RAMP M=1 should still be preferred because the RAMP objectives are more efficient than MRT objectives. In the case of MRT, for every training instance $S + S' = 20$ queries need to be executed against the database to obtain an answer and corresponding reward. On the other hand, RAMP has to execute *at most* the 10 queries of the $k$-best list $\mathcal{K}$, but often less if both a correct and an incorrect query are found earlier.

To summarize, RAMP can attribute its success to two factors: First, it discourages parses that receive a wrong answer rather than ignoring them as MRT often does. Second, a larger minibatch size leads to improvements because updates are based on an average over several inputs. Further performance gains can be obtained by using the token-level objective RAMP-T. Finally, RAMP objectives are more efficient because fewer outputs have to be judged.

## 5 Weakly Supervised Machine Translation

**Ramp Loss Objectives.** We consider machine translation (MT) in a weakly supervised domain adaptation setting, where in-domain references are unavailable. In this setting, we obtain weak feedback by matching translation model outputs against cross-lingually linked documents. For each input sentence $x$, we can obtain a set of *relevant* documents $D^+(x) \in D$ where $D$ is a collection of target language documents. Cross-lingual link structures can be found in many multilingual document collections, such as cross-lingual citations in patent documents or product categories in e-commerce data. Our example is links between Wikipedia documents. Instead of a



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

reference translation, we use a relevant document $d^+$ sampled from $D^+(x)$ to guide our search for $y^+$ and $y^-$. As a relevant document provides much weaker supervision than a reference translation, we construct a more informative supervision signal by integrating negative supervision from an irrelevant document $d^-$ sampled from a collection of irrelevant contrast documents. For each input $x$, the bipolar supervision signal then consists of a pair of sampled documents $(d^+, d^-)$.

Unlike semantic parsing for question answering, our task uses a continuous reward $\delta(y) \in [0, 1]$. In fully supervised MT a sentence-level approximation of the BLEU score can serve as the reward. But computing the BLEU score between a translation and a document does not make sense. We therefore propose two different alternative metrics. The first, $\delta_1(y, d)$, computes how well a translation matches a relevant document. The second, $\delta_2(y, d^+, d^-)$ computes how well a translation differentiates between a relevant and an irrelevant document. $\delta_1(y, d)$ is defined as the average $n$-gram precision between a hypothesis and a document, multiplied by a brevity penalty. As we do not have a reference length, we include a brevity penalty term that compares the output length to the input length. This ratio can be modified by a factor $r$ that represents the average length difference between source and target language and which can be computed over the training data:

$$\delta_1(y, d) = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{u_n} c(u_n, y) \cdot \mathbb{1}_{u_n \in d}}{\sum_{u_n} c(u_n, y)} \cdot BP\,, \tag{8}$$

where $u_n$ are the $n$-grams present in $y$, $c()$ counts the occurrences of an $n$-gram in $y$, and $N$ is the maximum order of $n$-grams used. The brevity penalty term is

$$BP = \min\left(1, \frac{r \cdot |y|}{|x|}\right).$$

$\delta_2(y, d^+, d^-)$ is defined as the difference between $\delta_1(y, d^+)$ and $\delta_1(y, d^-)$, subject to a linear transformation to allow values to lie between 0 and 1:

$$\begin{aligned} &\delta_2(y, d^+, d^-) = \\ &0.5 \cdot (\delta_1(y, d^+) - \delta_1(y, d^-) + 1)\,. \end{aligned} \tag{9}$$

Our intuition behind this metric is that it should measure how well a translation differentiates between the relevant and irrelevant document, leading to domain-specific translations being weighted higher than domain-agnostic ones.

Table 4 shows our loss functions for the weakly supervised case. RAMP and RAMP2 define $y^+$ and $y^-$ in the same way as is done in the semantic parsing task, except that the metric $\delta_1(y, d^+)$ is used to match outputs against documents. Like Gimpel and Smith (2012), we include a scaling factor $\alpha$ to trade off the importance of the reward against the model score in determining $y^+$ and $y^-$. Note that these objectives do not include negative supervision from $d^-$. Using the metrics defined above, we formulate two objectives that include $d^-$: RAMP$^-$ defines $y^+$ in the same way as RAMP, but uses a different definition of $y^-$: Instead of using a *fear* output with respect to $d^+$ (i.e., a translation with high probability and low reward $\delta_1(y, d^+)$), we use a *hope* output with respect to $d^-$ (i.e., a translation with high probability and high reward $\delta_1(y, d^-)$). As this translation matches an irrelevant document well, it can be used as a negative output. The same definition of $y^-$ is also used in RAMP1$^-$. Note that this objective does not include positive supervision from $d^+$. Finally, RAMP$_{\delta_2}$ incorporates $d^+$ and $d^-$ in a different way. This objective defines $y^+$ as a hope and $y^-$ as a fear, but uses the joined metric $\delta_2(y, d^+, d^-)$ with respect to the document pair $(d^+, d^-)$.

**Experimental Setup.** We test our objectives on a weakly supervised English–German Wikipedia translation task first proposed in Jehl and Riezler (2016). In-domain training data are 10,000 English sentences with relevant German documents sampled from the WikiCLIR corpus (Schamoni et al., 2014).[8] The task includes a small in-domain development and test set (dev: 1,712 sentences, test: 1,526 sentences), each consisting of four Wikipedia articles on diverse subjects. Irrelevant documents $d^-$ are sampled from the German side of the News Commentary[9] data set, which contains document boundary information.

Byte-pair encoding (Sennrich et al., 2016) with 30,000 merge operations is applied to all source and target data. Sentences longer than 80 words

[8] WikiCLIR annotates both a stronger *mate* relation when there is a direct cross-lingual link between documents and a weaker *link* relation when a there is a bidirectional link between a German mate document and another German document. The experiments reported here use the *mate* relation.

[9] http://casmacat.eu/corpus/news-commentary.html



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

| Loss | $y^+$ | $y^-$ |
|---|---|---|
| RAMP | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \delta_1(y, d^+))$ | $\arg\max_y \pi_w(y\|x) + \alpha(1 - \delta_1(y, d^+))$ |
| RAMP$^-$ | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \delta_1(y, d^+))$ | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \delta_1(y, d^-))$ |
| RAMP1$^-$ | $\hat{y}$ | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \delta_1(y, d^-))$ |
| RAMP2 | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \delta_1(y, d^+))$ | $\hat{y}$ |
| RAMP$_{\delta_2}$ | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \delta_2(y, d^+, d^-))$ | $\arg\max_y \pi_w(y\|x) + \alpha(1 - \delta_2(y, d^+, d^-))$ |

Table 4: Configurations for $y^+$ and $y^-$ for weakly supervised MT adaptation. $\hat{y}$ is the highest-probability model output. $\pi_w(y|x)$ is the probability of $y$ under the model. The $\arg\max_y$ is taken over the $k$-best list $\mathcal{K}(x)$. $\alpha$ is a scaling factor regulating the influence of the metric compared to the model probability. $\delta_1$ and $\delta_2$ are metrics defined with respect to relevant and irrelevant documents $d^+$ and $d^-$ (see Eq. 8 and 9).

are removed from the training set. Our neural MT model uses 500-dimensional word embeddings and hidden layer dimension of 1,024. Encoder and decoder use GRU units. An out-of-domain model is trained on 2.1 million sentence pairs from Europarl v7 (Koehn, 2005), News Commentary v10, and the MultiUN v1 corpus (Eisele and Chen, 2010). The baseline (MLE) is trained using the MLE objective and ADADELTA (Zeiler, 2012) for 20 epochs. We train on batches of 64 and use dropout for regularization, with a dropout rate of 0.2 for embedding and hidden layers and 0.1 for source and target layers. Gradients are clipped if their norm exceeds 1.0.

The metric-augmented objectives are trained using SGD. All hyperparameters are chosen on the development set. For the ramp loss objectives, we use a learning rate of 0.005, $\alpha = 10$, and a $k$-best size of 16. We compare ramp loss to MRT using both $\delta_1(y, d^+)$ and $\delta_2(y, d^+, d^-)$ as the external cost function, denoted as MRT$_{\delta_1}$ and MRT$_{\delta_2}$, respectively. MRT is trained using a learning rate of 0.05, $S = 16$, and $S' = 10$. For testing and validation, translations are obtained using beam search with a beam size of 16. Results are validated every 200 updates and training is run for 25 validations. The stopping point is determined by the BLEU score (Papineni et al., 2001) on the development set. We report scores computed with Moses'[10] `multi-bleu.perl` on tokenized, truecased output. Results are averaged over 2 runs.

**Experimental Results.** Results for the different objectives can be found in Table 5. The ramp losses RAMP, RAMP1$^-$, and RAMP2, which do not incorporate bipolar supervision from $d^+$ *and* $d^-$ (lines 2, 3, and 4) actually deteriorate

[10] `https://github.com/moses-smt/mosesdecoder`.

| | | $M$ | % BLEU | $\Delta$ |
|---|---|---|---|---|
| 1 | MLE | 64 | 15.59 | |
| 2 | RAMP | 40 | $15.03 \pm 0.01$ | $-0.56$ |
| 3 | RAMP1$^-$ | 40 | $15.12 \pm 0.02$ | $-0.47$ |
| 4 | RAMP2 | 40 | $15.19 \pm 0.01$ | $-0.40$ |
| 5 | MRT$_{\delta_1}$ | 1 | $15.37 \pm 0.04$ | $-0.22$ |
| 6 | MRT$_{\delta_2}$ | 1 | $15.70 \pm 0.04$ | $+0.11$ |
| 7 | RAMP$^-$ | 40 | **15.85** $\pm 0.02$ | $+0.26$ |
| 8 | RAMP$_{\delta_2}$ | 40 | **15.86** $\pm 0.04$ | $+0.27$ |
| 9 | RAMP$^-$-T | 40 | **16.03**$^*$ $\pm 0.02$ | $+0.44$ |
| 10 | RAMP$_{\delta_2}$-T | 40 | **15.84** $\pm 0.02$ | $+0.25$ |

Table 5: BLEU scores for weakly supervised MT experiments. **Boldfaced** results are significantly better than the baseline at $p < 0.05$ according to `multeval` (Clark et al., 2011). $^*$ marks a significant difference over RAMP$^-$.

in performance. This shows that supervision from only $d^+$ or only $d^-$ is insufficient. The deteriorating effect is strongest for RAMP, which uses $d^+$ to select both $y^+$ and $y^-$. We explain this by the fact that $d^+$ is an imperfect label. Trying to push the model to perfectly reproduce $d^+$ will not lead to a good translation. The same observation holds true for MRT$_{\delta_1}$. This objective only includes the reward $\delta_1(y, d^+)$. Compared with the RAMP objectives, the decrease for MRT$_{\delta_1}$ is smaller.

On the other hand, MRT$_{\delta_2}$, which incorporates bipolar supervision, produces a nominal improvement over the MLE baseline. This objective is outperformed by RAMP$^-$ and RAMP$_{\delta_2}$. Both objectives produce a small, but significant, improvement of 0.3% BLEU over the MLE baseline. This result shows that bipolar supervision is crucial for success in this weak supervision scenario. It also shows that unlike MRT, for the bipolar ramp loss it does not matter whether $\delta_1$ or $\delta_2$ is used, as they both capture the same idea. The superiority of these objectives over MRT shows



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

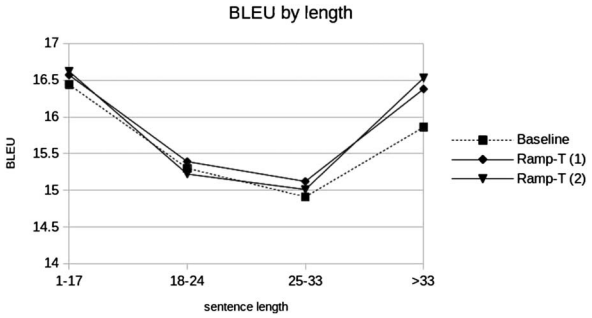


Figure 2: BLEU scores by sentence length for the MLE Baseline and the RAMP$^-$-T runs.

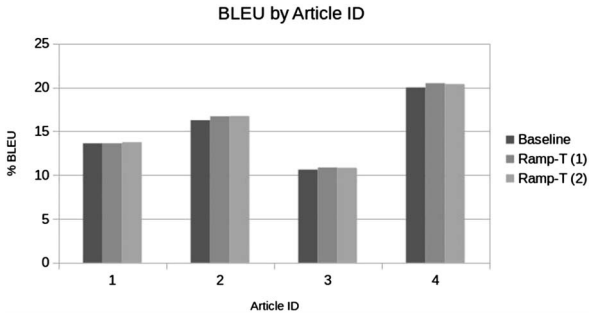


Figure 3: BLEU scores by Wikipedia article for the MLE Baseline and the RAMP$^-$-T runs.

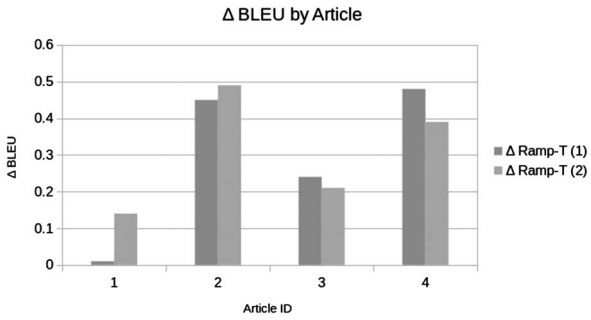


Figure 4: Improvements in BLEU scores by Wikipedia article for the RAMP$^-$-T runs.

again the success of intelligently selecting positive and negative outputs. Another small, but significant, improvement is produced by the token-level variant RAMP$^-$-T, leading to the best overall result.

To summarize, we find that for this task, which uses very weak supervision from document-level links, small improvements can be obtained. To achieve these improvements, it is imperative to use objectives that include bipolar supervision from $d^+$ and $d^-$. This finding holds for both ramp loss and MRT. The best overall result is obtained using ramp loss in the token-level variant.

**Analysis of Translation Results.** As the improvements in the translation experiments are very small, we conduct a small-scale analysis to better determine the nature of the gains. Our analysis is inspired by Bentivogli et al. (2016). We compare the weakly supervised MLE baseline to the best experiment in this setting, which uses the bipolar token-level ramp loss RAMP$^-$-T.

We first analyze the performance by sentence length. We separate the translations into source length brackets and score each bracket separately. The brackets represent quartiles of the source length distribution, ensuring an approximately equal amount of sentences in each bracket. Results are shown in Figure 2. For all systems, we observe a drop in performance up to an input length of 33. Surprisingly, BLEU scores increase again for the top bracket (source length $> 33$). For this bracket, we also see the biggest gap between MLE and RAMP$^-$-T of 0.52 and 0.67% BLEU for the two runs. This increase is mitigated by much weaker increases in the bottom brackets. A possible explanation for the weaker performance of MLE in the top bracket is the observation that hypotheses produced by the MLE system are longer than for RAMP$^-$-T. For the top bracket, hypothesis lengths exceed reference lengths for all systems. However, for MLE this over-generation is more severe at 106% of the reference length, compared to RAMP$^-$-T at 102%, potentially causing a higher loss in precision.

As our test set consists of parallel sentences extracted from four Wikipedia articles, we can examine the performance for each article separately. Figure 3 shows the results. We observe large differences in performance according to article ID. These are probably caused by some articles being more similar to the out-of-domain training data than others. Comparing RAMP$^-$-T and MLE, we see that RAMP$^-$-T outperforms MLE for each article by a small margin. Figure 4 shows the size of the improvements by article. We observe that margins are bigger on articles with better baseline performance. This suggests that there are challenges arising from domain mismatch that are not addressed by our method.

Lastly, we present an examination of example outputs. Table 6 shows an example of a long sentence from Article 2, which describes the German town of Schüttorf. This article is originally in German, meaning that our model is



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

| | |
|---|---|
| Source | Towards the end of the 19th century, a strong textile industry was developing itself in Schüttorf with several large local businesses (Schlikker & Söhne, Gathmann & Gerdemann, G. Schümer & Co. and ten Wolde, later Carl Remy; today's RoFa is not one of the original textile companies, but was founded by H. Lammering and later taken over by Gerhard Schlikker jun., Levert Rost and Wilhelm Edel; |
| MLE | Ende des 19. Jahrhunderts, eine starke Textilindustrie, die sich in *Ettorf* mit mehreren großen lokalen Unternehmen (Schlikker & Söhne, Gathmann & *Geréann*, G. *Schal* & Co. und *zehn Wolde*, später Carl Remy) entwickelt hat; die heutige RoFa ist nicht einer der ursprünglichen Textilunternehmen, sondern wurde von H. Lammering *[gegründet]* und später von Gerhard *Schaloker Junge*, Levert Rost und Wilhelm Edel übernommen. |
| RAMP$^-$-T | Ende des 19. Jahrhunderts **entwickelte sich** *[in Schüttorf]* eine starke Textilindustrie mit mehreren großen lokalen Unternehmen (Schlikker & Söhne, Gathmann & **Gerdemann**, G. *Schal* & Co. und *zehn Wolde*, später Carl Remy; die heutige RoFa ist nicht **eines der ursprünglichen Textilunternehmen**, sondern wurde von H. Lammering *[gegründet]* und später von Gerhard *Schaloker Junge*, Levert Rost und Wilhelm Edel übernommen. |
| Reference | gegen Ende des 19. Jahrhunderts entwickelte sich in Schüttorf eine starke Textilindustrie mit mehreren großen lokalen Unternehmen (Schlikker & Söhne, Gathmann & Gerdemann, G. Schümer & Co. und ten Wolde, später Carl Remy, die heutige RoFa ist keine ursprüngliche Textilfirma, sondern wurde von H. Lammering gegründet und später von Gerhard Schlikker jun., Levert Rost und Wilhelm Edel übernommen.) |

Table 6: MT example from Article 2 in the test set. All translation errors are underlined. Incorrect proper names are also set in *cursive*. Omissions are inserted in brackets and set in cursive *[like this]*. Improvements by RAMP$^-$-T over MLE are marked in **boldface**.

back-translating from English into German. The reference contains some awkward or even ungrammatical phrases such as *"was developing itself"*, a literal translation from German. The example also illustrates that translating Wikipedia involves handling frequent proper names (there are 11 proper names in the example). Both models struggle with translating proper names, but RAMP$^-$-T produces the correct phrase *"Gathmann & Gerdemann"*, while MLE fails to do so. The RAMP$^-$-T translation is also fully grammatical, whereas MLE incorrectly translates the main verb phrase *"was developing itself"* into a relative clause, and contains an agreement error in the translation of the noun phrase *"one of the original textile companies"*. Although making fewer errors in grammar and proper name translation, RAMP$^-$-T contains two deletion errors and MLE only contains one. This could be caused by the active optimization of sentence length in the ramp loss model.

## 6 Fully Supervised Machine Translation

Our work focuses on weakly supervised tasks, but we also conduct experiments using a fully supervised MT task. These experiments are motivated on the one hand by adapting the findings of Gimpel and Smith (2012) to the neural MT paradigm, and on the other hand by expanding the work by Edunov et al. (2018) on applying classical structured prediction losses to neural MT.

**Ramp Loss Objectives.** For fully supervised MT we assume access to one or more reference translations $\bar{y}$ for each input $x$. The reward $\text{BLEU}_{+1}(y, \bar{y})$ is a per-sentence approximation of the BLEU score.[11] Table 7 shows the different definitions of $y^+$ and $y^-$, which give rise to different ramp losses. RAMP, RAMP1, and RAMP2 are defined analogously to the other tasks. We again include a hyperparameter $\alpha > 0$ interpolating cost function and model score when searching for $y^+$ and $y^-$. Gimpel and Smith (2012) also include the perceptron loss in their analysis. PERC1 is a re-formulation of the Collins perceptron (Collins, 2002) where the reference is used as $y^+$ and $\hat{y}$ is used as $y^-$. A comparison with PERC1 is not possible for the weakly supervised tasks in the previous sections, as gold structures are not available for these tasks. With neural MT and subword methods we are able to compute this loss for any reference without running into the problem of *reachability* that was faced by phrase-based MT (Liang et al., 2006). However,

[11] We use the BLEU score with add-1 smoothing for $n > 1$, as proposed by Chen and Cherry (2014).



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

| Loss | $y^+$ | $y^-$ |
|---|---|---|
| RAMP | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \text{BLEU}_{+1}(y, \bar{y}))$ | $\arg\max_y \pi_w(y\|x) + \alpha(1 - \text{BLEU}_{+1}(y, \bar{y}))$ |
| RAMP1 | $\hat{y}$ | $\arg\max_y \pi_w(y\|x) + \alpha(1 - \text{BLEU}_{+1}(y, \bar{y}))$ |
| RAMP2 | $\arg\max_y \pi_w(y\|x) - \alpha(1 - \text{BLEU}_{+1}(y, \bar{y}))$ | $\hat{y}$ |
| PERC1 | $\bar{y}$ | $\hat{y}$ |
| PERC2 | $\arg\max_y \text{BLEU}_{+1}(y, \bar{y})$ | $\hat{y}$ |

Table 7: Configurations for $y^+$ and $y^-$ for fully supervised MT. $\hat{y}$ is the highest-probability model output, $\bar{y}$ is a gold standard reference. $\pi_w(y|x)$ is the probability of $y$ according to the model. The $\arg\max_y$ is taken over the $k$-best list $\mathcal{K}(x)$. $\text{BLEU}_{+1}$ is smoothed per-sentence BLEU and $\alpha$ is a scaling factor.

using sequence-level training towards a reference can lead to degenerate solutions where the model gives low probability to all its predictions (Shen et al., 2016). PERC2 addresses this problem by replacing $\bar{y}$ by a surrogate translation that achieves the highest $\text{BLEU}_{+1}$ score in $\mathcal{K}(x)$. This approach is also used by Edunov et al. (2018) for the loss functions which require an oracle. PERC1 corresponds to equation (9), PERC2 to equation (10) of Gimpel and Smith (2012).

**Experimental Setup.** We conduct experiments on the IWSLT 2014 German–English task, which is based on Cettolo et al. (2012) in the same way as Edunov et al. (2018). The training set contains 160K sentence pairs. We set the maximum sentence length to 50 and use BPE with 14,000 merge operations. Edunov et al. (2018) sample 7K sentences from the training set as heldout data. We do the same, but only use one tenth of the data as heldout set to be able to validate often.

Our baseline system (MLE) is a BiLSTM encoder-decoder with attention, which is trained using the MLE objective. Word embedding and hidden layer dimensions are set to 256. We use batches of 64 sentences for baseline training and batches of 40 inputs for training RAMP and PERC variants. MRT makes an update after each input using all sampled outputs and resulting in a batch size of 1. All experiments use dropout for regularization, with dropout probability set to 0.2 for embedding and hidden layers and to 0.1 for source and target layers. During MLE-training, the model is validated every 2500 updates and training is stopped if the MLE loss on the heldout set worsens for 10 consecutive validations.

For metric-augmented training, we use SGD for optimization with learning rates optimized on the development set. Ramp losses and PERC2 use a $k$-best list of size 16. For ramp loss training, we set $\alpha = 10$. RAMP and PERC variants both use a learning rate of 0.001. A new $k$-best list is generated for each input using the current model parameters. We compare ramp loss to MRT as described above. For MRT, we use SGD with a learning rate of 0.01 and set $S = 16$ and $S' = 10$. As Edunov et al. (2018) observe beam search to work better than sampling for MRT, we also run an experiment in this configuration, but find no difference between results. As beam search runs significantly slower, we only report sampling experiments.

The model is validated on the development set after every 200 updates for experiments with batch size 40 and after 8,000 updates for MRT experiments with batch size 1. The stopping point is determined by the BLEU score on the heldout set after 25 validations. As we are training on the same data as the MLE baseline, we also apply dropout during ramp loss training to prevent overfitting. BLEU scores are computed with Moses' `multi-bleu.perl` on tokenized, truecased output. Each experiment is run 3 times and results are averaged over the runs.

**Experimental Results.** As shown in Table 8, all experiments except for PERC1 yield improvements over MLE, confirming that sequence-level losses that update towards the reference can lead to degenerate solutions. For MRT, our findings show similar performance to the initial experiments reported by Edunov et al. (2018), who gain 0.24 BLEU points on the same test set.[12] PERC2 and RAMP2, improve over the

[12]See their Table 2. Using interpolation with the MLE objective, Edunov et al. (2018) achieve +0.7 BLEU points. As we are only interested in the effect of sequence-level objectives, we do not add MLE interpolation. The best model by Edunov et al. (2018) achieved a BLEU score of 32.91%. It is possible that these scores are not directly comparable to ours due to different pre- and post-processing. They also use a multi-layer CNN architecture (Gehring et al., 2017), which has been shown to outperform a simple RNN architecture such as ours.



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

| | | $M$ | % BLEU | $\Delta$ |
|---|---|---|---|---|
| 1 | MLE | 64 | 31.99 | |
| 2 | MRT | 1 | **32.17** $\pm$ 0.02 | + 0.18 |
| 3 | PERC1 | 40 | 31.91 $\pm$ 0.02 | − 0.08 |
| 4 | PERC2 | 40 | **32.22** $\pm$ 0.03 | + 0.23 |
| 5 | RAMP1 | 40 | **32.36**$^{*}$ $\pm$ 0.05 | + 0.37 |
| 6 | RAMP2 | 40 | **32.19** $\pm$ 0.01 | + 0.20 |
| 7 | RAMP | 40 | **32.44**$^{**}$ $\pm$ 0.00 | + 0.45 |
| 8 | RAMP-T | 40 | **32.33**$^{*}$ $\pm$ 0.00 | + 0.34 |

Table 8: BLEU scores for fully supervised MT experiments. **Boldfaced** results are significantly better than MLE at $p < 0.01$ according to `multeval` (Clark et al., 2011). $^{*}$ marks a significant difference to MRT and PERC2, and $^{**}$ marks a difference to RAMP1.

MLE baseline and PERC1, but perform on a par with MRT and each other. Both RAMP and RAMP1 are able to outperform MRT, PERC2, and RAMP2, with the bipolar objective RAMP also outperforming RAMP1 by a narrow margin. The main difference between RAMP and RAMP1, compared to PERC2 and RAMP2, is the fact that the latter objectives use $\hat{y}$ as $y^{-}$, whereas the former use a *fear* translation with high probability and low $\text{BLEU}_{+1}$. We surmise that for this fully supervised task, selecting a $y^{-}$ which has some known negative characteristics is more important for success than finding a good $y^{+}$. RAMP, which fulfills both criteria, still outperforms RAMP2. This result re-confirms the superiority of bipolar objectives compared to non-bipolar ones. Although still improving over MLE, token-level ramp loss RAMP-T is outperformed by RAMP by a small margin. This result suggests that when using a metric-augmented objective on top of an MLE-trained model in a full supervision scenario without domain shift, there is little room for improvement from token-level supervision, while gains can still be obtained from additional sequence-level information captured by the external metric, such as information about the sequence length.

To summarize, our findings on a fully supervised task show the same small margin for improvement as Edunov et al. (2018), without any further tuning of performance (e.g., by interpolation with the MLE objective). Bipolar RAMP is found to outperform the other losses. This observation is also consistent with the results by Gimpel and Smith (2012) for phrase-based MT. We conclude that for fully supervised MT, deliberately selecting a *hope* and *fear* translation is beneficial.

## 7 Conclusion

We presented a study of weakly supervised learning objectives for three neural sequence-to-sequence learning tasks. In our first task of semantic parsing, question-answer pairs provide a weak supervision signal to find parses that execute to the correct answer. We show that ramp loss can outperform MRT if it incorporates bipolar supervision where parses that receive negative feedback are actively discouraged. The best overall objective is constituted by the token-level ramp loss. Next, we turn to weak supervision for machine translation in form of cross-lingual document-level links. We present two ramp loss objectives that combine bipolar weak supervision from a linked document $d^{+}$ and an irrelevant document $d^{-}$. Again, the bipolar ramp loss objectives outperform MRT, and the best overall result is obtained using token-level ramp loss. Finally, to tie our work to previous work on supervised machine translation, we conduct experiments in a fully supervised scenario where gold references are available and a metric-augmented loss is desired to reduce the exposure bias and the loss-evaluation mismatch. Again, the bipolar ramp loss objective performs best, but we find that the overall margin for improvement is small without any additional engineering. We conclude that ramp loss objectives show promise for neural sequence-to-sequence learning, especially when it comes to weakly supervised tasks where the MLE objective cannot be applied. In contrast to ramp losses that either operate only in the undesirable region of the search space (“cost-augmented decoding” as in RAMP1) or only in the desirable region of the search space (“cost-diminished decoding” as in RAMP2), bipolar RAMP operates in both regions of the search space when extracting supervision signals from weak feedback. We showed that MRT can be turned into a bipolar objective by defining a metric that assigns negative values to bad outputs. This improves the performance of MRT objectives. However, the ramp loss objective is still superior as it is easy to implement and efficient to compute. Furthermore, on weakly supervised tasks our novel token-level ramp loss objective RAMP-T can obtain further improvements over its sequence-level counterpart



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

because it can more directly assess which tokens in a sequence are crucial to its success or failure.

## Acknowledgments

The research reported in this paper was supported in part by DFG grants RI-2221/4-1 and RI 2221/2-1. We would like to thank the reviewers for their helpful comments.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*. San Diego, CA.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: A case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, WA.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT$^3$: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy.

Olivier Chapelle, Chuong B. Do, Choon H. Teo, Quoc V. Le, and Alex J. Smola. 2009. Tighter bounds for structured estimation. In *Advances in Neural Information Processing Systems (NIPS)*. Vancouver, Canada.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the 9th Workshop on Statistical Machine Translation*. Baltimore, MD.

David Chiang. 2012. Hope and fear for discriminative training of statistical translation models. *Journal of Machine Learning Research*, 13(1):1159–1187.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar.

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-dine question answering for long documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 2011 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*. Portland, OR.

James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. Driving Semantic Parsing from the World's Response. In *Proceedings of the 14th Conference on Computational Natural Language Learning*. Uppsala, Sweden.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia, PA.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2018. Active learning for deep semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American*



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

*Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*. New Orleans, LA.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*. Valetta, Malta.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Sydney, Australia.

Kevin Gimpel and Noah A. Smith. 2012. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*. Montreal, Canada.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimation in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530.

Kelvin Guu, Panupong Pasupat, Evan Liu, and Percy Liang. 2017. From language to programs: Bridging reinforcement learning and maximum marginal likelihood. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada.

Carolin Haas and Stefan Riezler. 2016. A corpus and semantic parser for multilingual natural language querying of openStreetMap. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*. San Diego, CA.

Tamir Hazan, Joseph Keshet, and David A. McAllester. 2010. Direct loss minimization for structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*. Vancouver, Canada.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada.

Laura Jehl and Stefan Riezler. 2016. Learning to translate from graded and negative relevance information. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit*, volume 5. Phuket, Thailand.

Tomáš Kočiský, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. 2016. Semantic parsing with semi-supervised sequential autoencoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX.

Carolin Lawrence and Stefan Riezler. 2018. Improving a neural semantic parser by counterfactual learning from human bandit feedback. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

Chen Liang, Jonathan Berant, Quoc V. Le, Kenneth D. Forbus, and Ni Lao. 2017. Neural symbolic machines: learning semantic parsers on freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*. Sydney, Australia.

Dipendra Misra, Ming-Wei Chang, Xiaodong He, and Wen-tau Yih. 2018. Policy shaping and generalized update equations for semantic parsing from denotations. In *Proceedings of*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

*the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium.

Lili Mou, Zhengdong Lu, Hang Li, and Zhi Jin. 2017. Coupling distributed and symbolic execution for natural language queries. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Sydney, Australia.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley, New York.

Mohammad Norouzi, Samy Bengio, Zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. 2016. Reward augmented maximum likelihood for neural structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*. Barcelona, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001, BLEU: A method for automatic evaluation of machine translation. Technical Report IBM Research Division Technical Report, RC22176 (W0190-022), Yorktown Heights, NY.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. Beijing, China.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: A toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum Risk Training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. Montreal, Canada.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, Bonn, Germany.

Ben Taskar, Carlos Guestrin, and Daphne Koller. 2004. Max-margin Markov networks. In *Advances in Neural Information Processing Systems (NIPS)*. Vancouver, Canada.



Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 20:229–256.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada.

Matthew D. Zeiler. 2012. ADADELTA: An adaptive learning rate method. *ArXiv e-prints*, cs.LG/1212.5701v1.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00265/1923174/tacl_a_00265.pdf by guest on 07 October 2022