A multi-constraint structured hinge loss for named-entity recognition

Hanieh Poostchi
Nine Entertainment Co.
Willoughby NSW 2068, Australia
HPoostchi@nine.com.au

Massimo Piccardi
FEIT, University of Technology Sydney
Broadway NSW 2007, Australia
Massimo.Piccardi@uts.edu.au

Abstract

The negative log-likelihood or cross entropy is the usual training objective of NLP models owing to its versatility and empirical performance. However, training objectives which directly target the performance measure used to evaluate the task have the potential to lead to higher empirical accuracy. For this reason, in this short paper we propose using a multi-constraint structured hinge loss as the training objective of a contemporary named-entity recognition (NER) model. Experimental results over the challenging OntoNotes 5.0 dataset have shown that the proposed objective has been able to achieve an improvement of 0.62 CoNLL score points at a complete parity of testing set-up.

1 Introduction

All NLP models utilise a loss function as minimisation objective for model training. Choosing the most appropriate loss function for a particular task can play an important role in the performance of the trained models at test time and in-field. However, almost invariably the utilised loss function is the negative log-likelihood (NLL), also known as cross entropy. This is due to a number of attractive properties of the NLL such as its smoothness and differentiability in large regions of the parameter space. In addition, training with minimum NLL often leads to models of high empirical accuracy. However, this function is not exempt from shortcomings. To name two, 1) the NLL only rewards the probability of the ground-truth class and does not distinguish between the other classes, and 2) it does not impose explicit margins (or ratios) between the probability assigned to the ground-truth class and those assigned to the other classes. For this reason, other differentiable loss functions are regarded as appealing alternatives or complements to the NLL. Amongst them are the hinge loss (Cortes and Vapnik, 1995) and the REINFORCE loss (Williams, 1992; Ranzato et al., 2016) which both attempt to directly optimise the performance measure used to evaluate the model's accuracy (e.g., the Hamming loss, the CoNLL score, the BLEU score etc). Both these losses can be used for the usual classification at token level or for the joint classification of all the tokens in a sentence (i.e., structured prediction) (Tsochantaridis et al., 2005). Given that targeting the evaluation loss during training may lead to improved performance at test time, in this short paper we explore the use of a structured hinge loss for named-entity recognition (NER). Our main contribution is the introduction of additional constraints between specific labelings aimed at increasing the accuracy of the learned model. Experimental results over a challenging NER dataset (OntoNotes 5.0, which is still far from accuracy saturation) show that the proposed approach has been able to achieve higher accuracy than both the NLL and a conventional structured hinge loss.

2 Related Work

In this section we briefly review the main literature on NER architectures and on structural loss functions. For a broader review of deep learning for NER, the reader can refer to (Li et al., 2018).

A current and well-known approach for NER combines a bidirectional LSTM with a CRF output layer to benefit from both their properties in sequential tagging (Huang et al., 2015). In this approach, the LSTM is used first to process each sentence token-by-token and produce an intermediate representation. Then, the CRF uses the intermediate representation as input to provide the joint prediction of all the labels. Lample et al. (2016) have extended this model with a second, auxiliary LSTM encoding each token character-by-

character to also capture the regularities at character level. More recently, Peters et al. (2017; 2018) have proposed tagLM and ELMo to take advantage of the contextualised embeddings provided by pre-trained neural language models. Several other variants have been proposed since, including the Flair embedddings of Akbik et al. (2018) which currently hold the state-of-the-art accuracy over OntoNotes 5.0 (NB: besides a system hat uses gazetteers as extra resources to increase accuracy (Liu et al., 2019)). However, a bidirectional LSTM-CRF with ELMo embeddings can still be regarded as a very strong baseline for NER and, for this reason, it is used in the rest of this paper.

For what concerns alternative training objectives to the negative log-likelihood, Zhang at al. (2016) have proposed training an LSTM additioned with a linear output layer by using an SVM objective. In their model, the parameters of both the LSTM and the output layer have been learned jointly using a combination of sequence-level and frame-level regularised hinge losses. Similarly, Shi et al. (2016) have proposed adding a structural SVM output layer (Tsochantaridis et al., 2005) to an RNN to improve its discriminative capability. In 2012, Gimpel and Smith (2012) have proposed a structured ramp loss that leverages various styles of margins between predicted labelings. Recently, Edunov et al. (2018) have carried out an extensive review of structured loss functions, including hinge losses, cost-weighted likelihoods and reinforcement learning objectives. In a 2015 computer vision paper, Zhang and Piccardi (2015) have proposed adding extra constraints to a structured hinge loss to increase its accuracy in a task of activity segmentation in video. Inspired by that approach, in this paper we explore its application to NER, proposing three original combinations of dedicated constraints and margins.

3 Methodology

In this section, we first briefly review the structured hinge loss (3.1) and the utilised scoring function (3.2), and then introduce the proposed approach (3.3).

3.1 Structured hinge loss

Given a token sequence, $x = \{x_1 \dots x_t \dots x_T\}$, we note with $y = \{y_1 \dots y_t \dots y_T\}$ a labeling, i.e. a sequence of corresponding labels, one per token. We also assume to have a scoring function, $F(x, y; w)$ or $F(x, y)$ for brevity, which is able to assign a compatibility score to any such (x, y) pair. This function is completely defined by its set of parameters, w, and it is a structured predictor if the score of a labeling is computed jointly rather than independently for each label. Given these assumptions, the goal of a *structured hinge loss* is simply to ensure that the ground-truth labeling, y^g, for a given x is assigned a score larger than that of any other labeling, $y \neq y^g$, by a chosen margin, K:

$$F(x, y^g) - F(x, y) \geq K \quad \forall y \neq y^g \tag{1}$$

It is often useful to impose a margin that is the larger the more the labeling differs from the ground truth, and this can be achieved by setting the margin to be the evaluation loss ("margin rescaling" (Tsochantaridis et al., 2005)):

$$F(x, y^g) - F(x, y) \geq \Delta(y^g, y) \quad \forall y \neq y^g \tag{2}$$

However, the number of distinct labelings is exponential in the length of the sequence and it may not be possible to find a set of parameters which is able to satisfy all the constraints. In that case, the constraints are relaxed by introducing a non-negative term, $\xi \geq 0$, in Eq. 2 to minimally satisfy all the constraints:

$$F(x, y^g) - F(x, y) \geq \Delta(y^g, y) - \xi \ \forall y \neq y^g$$

It is easy to see that the value of ξ is set by the most violated of the constraints, with y^* its corresponding labeling:

$$\xi = \max_y [-F(x, y^g) + F(x, y) + \Delta(y^g, y)] \tag{3}$$

$$y^* = \operatorname*{argmax}_y [F(x, y) + \Delta(y^g, y)] \tag{4}$$

where we have omitted the first term in Eq. 4 since it does not depend on y. Note that since the search domain includes y^g, and $\Delta(y^g, y^g) = 0$, the above guarantees that $\xi \geq 0$. Eq. 3 is known as the *structured* hinge loss because of the interdependencies between the individual labels inside the scoring function and, possibly, the evaluation loss. In turn, the solution of Eq. 4 is known as the "loss-augmented" inference and is the crux of structured hinge loss minimisation. Given a training set, $\{x^i, y^i\}, i = 1 \dots N$, the training objective is therefore:

$$w^* = \operatorname*{argmin}_w \sum_{i=1}^{N} \xi^i(w) \tag{5}$$

While the minimisation in Eq. 5 can be easily entertained by automated differentiation, the inference of the most-violating labelings must be performed externally with a dedicated algorithm.

3.2 Scoring function

The scoring function, $F(x, y; w)$, has been implemented as a BiLSTM-CRF (Lample et al., 2016), a popular NER model using a bidirectional LSTM as its feature layer and a CRF as its output layer. Its scoring function can be expressed as:

$$F(x, y; w) = \sum_{t=2}^{T} w_{y_{t-1}, y_t} + \sum_{t=1}^{T} f(y_t; w) \quad (6)$$

where $w_{i,j}$ are the transition weights for transitioning from label $y_{t-1} = i$ to label $y_t = j$, and $f(y_t; w)$ denotes the score assigned to label y_t by the BiLSTM layer. At its turn, the BiLSTM layer is organised as a bidirectional LSTM with trainable word and character embeddings as its inputs. At initialisation, the word embeddings can be assigned with either random or pre-trained values. At inference time, $\text{argmax}_y F(x, y; w)$ is provided by the Viterbi algorithm. For further details, please refer to (Lample et al., 2016).

3.3 The proposed multi-constraint structured hinge loss

Rather than constraining the optimisation problem with an exponential number of constraints, the structured hinge loss minimisation only considers the constraint setting the value of the loss:

$$\xi = [-F(x, y^g) + F(x, y^*) + \Delta(y^g, y^*)] \quad (7)$$

While such an approach makes the constrained minimisation feasible, we speculate that the addition of other constraints – either between the ground-truth labeling and other labelings, or between the other labelings themselves – may eventuate in a more performing model. To this aim, we have created a new set of labelings by arbitrarily introducing false positives in the ground-truth labelings of the training set. As false positives, we have decided to change the "Outside" label immediately preceding the first ground-truth entity of the training sentences into a "B-ORG" label. This is an altogether arbitrary change that creates mildly incorrect labelings: as such, we expect the scoring function to assign them scores lower than the corresponding ground truths, yet higher than more incorrect labelings. We note these new labelings as $u_i, i = 1 \dots N$, reserving the subscript position for the sample index henceforth. Given these extra labelings, we propose three versions of a multi-constraint training loss:

- **Hinge-yu**: in this loss, we impose an extra constraint between the altered ground-truth labeling, u_i, and the remaining labelings. However, function $\Delta(u_i, y)$ cannot be used as margin since it expects to have a true labeling as its first argument. Therefore, following (Zhang and Piccardi, 2015) we set the margin to be $(\Delta(y_i, y) - \Delta(y_i, u_i))$. The labeling returned by the loss-augmented inference with this margin is the same as in the standard case (y_i^*) since the the second term in the margin ($\Delta(y_i, u_i)$) does not depend on y.

- **Double Hinge**: in this loss, we instead impose an extra constraint between the ground-truth labeling, y_i, and the altered ground truth, u_i. As margin, we can naturally use $\Delta(y_i, u_i)$.

- **Discounted Margin**: in this loss, we again impose an extra constraint between the altered ground-truth labeling, u_i, and the remaining labelings. As margin, we use the regular loss function, $\Delta(u_i, y)$, but "discounted" by a small discount factor since u_i is not an actual ground truth.

As evaluation loss for the margin, we have simply used the Hamming loss, since it naturally decomposes over the individual tokens of its arguments and it allows us to easily touch up the standard Viterbi algorithm to provide the required loss-augmented inference. Extending the loss-augmented inference to other, more specialised evaluation measures such as the CoNLL and MUC scores (Nadeau and Sekine, 2007) could be the scope of future work.

4 Experiments and results

4.1 Experimental set-up

We have carried out experiments over a challenging NER dataset, OntoNotes v5.0, which was first introduced in CoNLL 2012 as a shared task (Pradhan et al., 2012, 2013). This English dataset contains multi-token entities from 18 different categories, including amongst others, person, facility,

Table 1: The compared training objectives.

NLL	$l_{NLL} = -\sum_{i=1}^{N} \log p(y_i \mid x_i)$
Hinge Loss	$l_{Hinge} = \sum_{i=1}^{N}[-F(x_i, y_i) + F(x_i, y_i^*) + \Delta(y_i, y_i^*)]_+$ $y_i^* = \text{argmax}_y F(x_i, y) + \Delta(y_i, y)$
Hinge-yu	$l_{Hinge-yu} = \sum_{i=1}^{N}[-F(x_i, y_i) + F(x_i, y_i^*) + \Delta(y_i, y_i^*)]_+$ $+ \sum_{i=1}^{N}[-F(x_i, u_i) + F(x_i, y_i^*) + \Delta(y_i, y_i^*) - \Delta(y_i, u_i)]_+$ $y_i^* = \text{argmax}_y F(x_i, y) + \Delta(y_i, y)$
Double Hinge	$l_{DoubleHinge} = \sum_{i=1}^{N}[-F(x_i, y_i) + F(x_i, y_i^*) + \Delta(y_i, y_i^*)]_+$ $+ \sum_{i=1}^{N}[-F(x_i, y_i) + F(x_i, u_i) + \Delta(y_i, u_i)]_+$ $y_i^* = \text{argmax}_y F(x_i, y) + \Delta(y_i, y)$
Discounted Margin	$l_{DiscountedMargin} = \sum_{i=1}^{N}[-F(x_i, y_i) + F(x_i, y_i^*) + \Delta(y_i, y_i^*)]_+$ $+ \sum_{i=1}^{N}[-F(x_i, u_i) + F(x_i, y_i^*) + df \times \Delta(u_i, y_i^*)]_+$ $y_i^* = \text{argmax}_y F(x_i, y) + \Delta(y_i, y)$

Table 2: Comparison of the CoNLL scores for the OntoNotes 5.0 dataset with the different training objectives.

Training objective	CoNLL score
NLL	88.54 ± 0.13
Hinge Loss	88.81 ± 0.06
Hinge-yu	88.88 ± 0.10
Double Hinge	88.85 ± 0.13
Discounted Margin ($df = 0.925$)	$\mathbf{89.16 \pm 0.03}$

organisation, location, product, event, law, date and time, and is split over a training, validations and test sets. For the experiments, we have converted it to the IOB2 tagging scheme.

The experiments have been carried out using the DeLFT[1] implementation of the BiLSTM-CRF. In the experiments, the word embeddings have been initialised with a concatenation of fastText-300d[2] and ELMo-1024d[3] (DeLFT's default). All hyperparameters have also been left to their default values. Each training session has been run until convergence of the evaluation loss over the validation set or a maximum of 20 epochs. In the experiments, we have compared the following training objectives: 1) the NLL/cross entropy; 2) a standard structured hinge loss using the Hamming loss as margin ("Hinge Loss"), with no additional constraints; and 3-5) the three versions of the proposed multi-constraint structured hinge loss presented in Section 3.3. The discount factor, df, for the Discounted Margin approach has been chosen in range $[0.85, 0.95]$ in 0.025 steps over the validation set. All the training objectives are displayed in Table 1.

For evaluation, we have used the CoNLL score, an entity-oriented variant of the F_1 score which is the standard evaluation measure for NER (Nadeau and Sekine, 2007). For every experiment, we have run three independent runs from different random seeds and reported their average score. In addition, for the most noteworthy pairwise comparisons, we have run one-tailed Welch's t-tests to test statistical significance (Hintze, 2019). As shown in (Colas et al., 2019), the Welch's t-test enjoys a good balance between Type I and Type II errors under a variety of assumptions for the underlying score distributions (beyond Gaussian), especially for small sample sizes.

4.2 Results and analysis

Table 2 shows the CoNLL scores achieved by the compared training objectives over the OntoNotes 5.0 test set as average of 3 independent runs. The table shows that even the standard structured hinge loss has achieved a higher score than the NLL ($+0.27$ percentage points). Even if the improvement is mild, the standard deviations over the

[1] https://github.com/kermitt2/delft

[2] https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip

[3] https://s3-us-west-2.amazonaws.com/allennlp/models/elmo/2x4096_512_2048cnn_2xhighway_5.5B/elmo_2x4096_512_2048cnn_2xhighway_5.5B_options.json

three runs are small and the p-value from a one-tailed Welch's t-test is < 0.05, showing that the improvement is statistically significant. In turn, all the versions of the proposed multi-constraint hinge loss have achieved higher scores than both the NLL and the standard structured hinge loss, with the Discounted Margin achieving the highest score. The improvement of the Discounted Margin over the NLL has been $+0.62$ percentage points, with a one-tailed Welch's t-test p-value < 0.01. While this improvement is still somehow limited, we wish to remark that it has leveraged only changes to the loss function in the code, and at a complete parity of model.

5 Conclusion

In this short paper, we have proposed a multi-constraint structured hinge loss to be used as training objective for a named-entity recognition model. The proposed loss enforces additional constraints with respect to the standard structured hinge loss with the aim of improving the test accuracy of the trained model. Experimental results over a challenging NER dataset (OntoNotes 5.0) have showed that the proposed loss has been able to achieve an improvement of 0.62 CoNLL score percentage points over the common negative log-likelihood. In the future, we aim to explore further combinations of constraints and margins, and possibly extend the proposed approach to other tasks.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. 2019. A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms. In *Reproducibility in Machine Learning, ICLR 2019 Workshop*, pages 1–23.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

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 Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 355–364.

Kevin Gimpel and Noah A. Smith. 2012. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 221–231.

Jerry Hintze. 2019. T-test – two-sample. In *NCSS User's Guide II, Chapter 206*, pages 1–18.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris DyerRemi Lebret, and Ronan Collobert. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2018. A survey on deep learning for named entity recognition. *CoRR*, abs/1812.09449.

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 5301–5307.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1756–1765.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning*, Jeju, Korea.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations*, pages 1–16.

Yangyang Shi, Kaisheng Yao, Hu Chen, Dong Yu, Yi-Cheng Pan, and Mei-Yuh Hwang. 2016. Recurrent support vector machines for slot tagging in spoken language understanding. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 393–399.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Guopeng Zhang and Massimo Piccardi. 2015. Structural SVM with partial ranking for activity segmentation and classification. *IEEE Signal Process. Lett.*, 22(12):2344–2348.

Shi-Xiong Zhang, Rui Zhao, Chaojun Liu, Jinyu Li, and Yifan Gong. 2016. Recurrent support vector machines for speech recognition. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5885–5889.