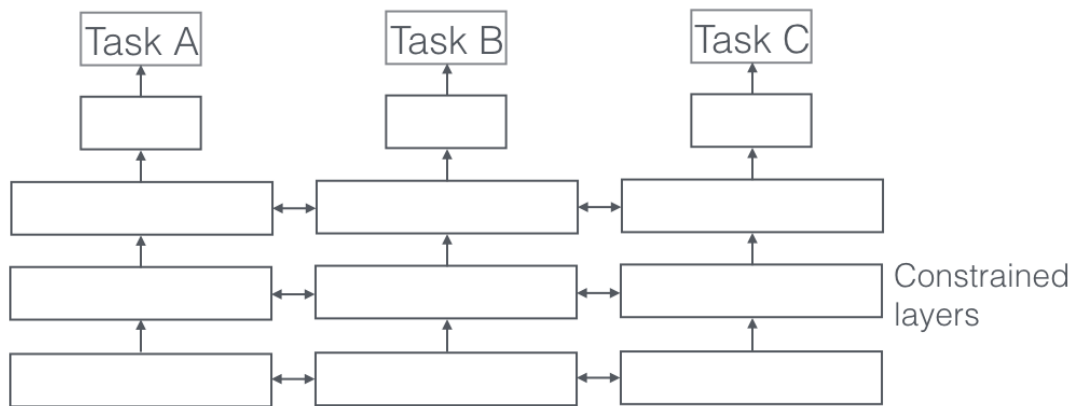


24 Sep 2017



## Multi-Task Learning Objectives for Natural Language Processing 🐦

In a [previous blog post](#), I discussed how multi-task learning (MTL) can be used to improve the performance of a model by leveraging a related task. Multi-task learning consists of two main components: a) The architecture used for learning and b) the auxiliary task(s) that are trained jointly. Both facets still have a lot of room for improvement. In addition, multi-task learning has the potential to be a key technique on the path to more robust models that learn from limited data: Training a model to acquire proficiency in performing a wide range of NLP tasks would allow us to induce representations, which should be useful for transferring knowledge to many other tasks, as outlined in [this blog post](#).

On the way to this goal, we first need to learn more about the relationships between our tasks, what we can learn from each, and how to combine them most effectively. Most of the existing theory in MTL has focused on homogeneous tasks, i.e. tasks that are variations of the same classification or regression problem, such as classifying individual MNIST digits. These guarantees, however, do not hold for the heterogeneous tasks to which MTL is most often applied in Natural Language Processing (NLP) and Computer Vision.

There have been some recent studies looking into when multi-task learning between different NLP tasks works but we still do not understand very well which tasks are useful. To this end, as inspiration, I will give an overview in the following of different approaches for multi-task learning for NLP. I will focus on the second component of multi-task learning; instead of discussing *how* a model is trained, as most architectures only differ in which layers they share, I will concentrate on the auxiliary tasks and objectives that are used for learning.

This post has two main parts: In the first part, I will talk about artificial tasks that can be used as auxiliary objectives for MTL. In the second part, I will focus on common NLP tasks and discuss which other NLP tasks have benefited them.

## Artificial auxiliary objectives

Multi-task learning is all about coming up with ways to add a suitable bias to your model. Incorporating artificial auxiliary tasks that cleverly complement your target task is arguably one of the most ingenious and fun ways to do MTL. It is a feature-engineering of sorts: instead of engineering the features, you are engineering the auxiliary task you optimize. Similarly to feature engineering, domain expertise is therefore required as we will see in the following:

**Language modelling** Language modelling has been shown to be beneficial for many NLP tasks and can be incorporated in various ways. Word embeddings pre-trained by word2vec have been shown to be beneficial -- as is known, word2vec approximates the language modelling objective; language models have been used to pre-train MT and sequence-to-sequence models [3]; contextual language model embeddings have also been found useful for many tasks [4]. In this context, we can also treat language modelling as an auxiliary task that is learned together with the main task. Rei (2017) [2] shows that this improves performance on several sequence labelling tasks.

**Conditioning the initial state** The initial state of a recurrent neural network is typically initialized to a 0 vector. According to a [lecture by Hinton in 2013](#), it is beneficial to learn the initial state just like any other sets of weights. While a learned state will be more helpful than a 0 vector it will be independent of the sequence and thus unable to adapt. Weng et al. (2017) [1] propose to add a suitable bias to the initial encoder and decoder states for NMT by training it to predict the words in the sentence. In this sense, this objective can essentially be seen as a *language modelling objective for the initial state* and might thus be helpful for other tasks. Similarly, we can think of other task-specific biases that could be encoded in the initial state to aid learning: A sentiment model might benefit from knowing about the general audience response to a movie or whether a user is more likely to be sarcastic while a parser might be able to leverage prior knowledge of the domain's tree depth or complexity.

**Adversarial loss** An auxiliary adversarial loss was first found to be useful for domain adaptation [5, 6], where it is used to learn domain-invariant representations by rendering the model unable to distinguish between different domains. This is typically done by adding a gradient reversal layer that reverses the sign of the gradient during back-propagation, which in turn leads to a maximization rather than a minimization of the adversarial loss. It is not to be confused with adversarial examples [7], which significantly increase the model's loss typically via small perturbations to its input; adversarial training [8], which trains a model to correctly classify such examples; or Generative Adversarial Networks, which are trained to generate some output

representation. An adversarial loss can be added to many tasks in order to learn task-independent representations [9]. It can also be used to ignore certain features of the input that have been found to be detrimental to generalization, such as data-specific properties that are unlikely to generalize. Finally, an adversarial auxiliary task might also help to combat bias and ensure more privacy by encouraging the model to learn representations, which do not contain information that would allow the reconstruction of sensitive user attributes.

**Predicting data statistics** An auxiliary loss can also be to predict certain underlying statistics of the training data. In contrast to the adversarial loss, which tries to make the model oblivious to certain features, this auxiliary task explicitly encourages the model to predict certain data statistics. Plank et al. (2016) [10] predict the log frequency of a word as an auxiliary task for language modelling. Intuitively, this makes the representation predictive of frequency, which encourages the model to not share representations between common and rare words, which benefits the handling of rare tokens. Another facet of this auxiliary task is to predict attributes of the user, such as their gender, which has been shown to be beneficial for predicting mental health conditions [49] or other demographic information [51]. We can think of other statistics that might be beneficial for a model to encode, such as the frequency of POS tags, parsing structures, or entities, the preferences of users, a sentence's coverage for summarization, or even a user's website usage patterns.

**Learning the inverse** Another auxiliary task that might be useful in many circumstances is to learn the inverse of the task together with the main task. A popular example of this framework is CycleGAN [43], which can generate photos from paintings. An inverse auxiliary loss, however, is applicable to many other tasks: MT might be the most intuitive, as every translation direction such as English->French directly provides data for the inverse direction, as Xia et al. (2016) [44] demonstrate. Xia et al. (2017) [45] show that this has applications not only to MT, but also to image classification (with image generation as its inverse) and sentiment classification (paired with sentence generation). For multimodal translation, Elliott and Kádár (2017) [19] jointly learn an inverse task by predicting image representations. It is not difficult to think of inverse complements for many other tasks: Entailment has hypothesis generation; video captioning has video generation; speech recognition has speech synthesis, etc.

**Predicting what should be there** For many tasks, where a model has to pick up on certain features of the training data, we can focus the model's attention on these characteristics by encouraging it explicitly to predict them. For sentiment analysis, for instance, Yu and Jiang (2016) [20] predict whether the sentence contains a positive or negative domain-independent sentiment word, which sensitizes the model towards the sentiment of the words in the sentence. For name error detection, Cheng et al. (2015) [50] predict if a sentence contains a name. We can envision similar auxiliary tasks that might be useful for other tasks: Predicting whether certain entities occur in a sentence might be useful for relation extraction; predicting whether a headline contains certain lurid terms might help for clickbait detection, while predicting whether an emotion word

occurs in the sentence might benefit emotion detection. In summary, this auxiliary task should be helpful whenever a task includes certain highly predictive terms or features.

## Joint training of existing NLP tasks

In this second section, we will now look at existing NLP tasks, which have been used to improve the performance of a main task. While certain tasks such as chunking and semantic tagging have been found to be useful for many tasks [60], the choice whether to use a particular auxiliary task largely depends on characteristics of the main task. In the following, I will thus highlight different strategies and rationals that were used to select auxiliary tasks for many common tasks in NLP:

**Speech recognition** Recent multi-task learning approaches for automatic speech recognition (ASR) typically use additional supervision signals that are available in the speech recognition pipeline as auxiliary tasks to train an ASR model end-to-end. Phonetic recognition and frame-level state classification can be used as auxiliary tasks to induce helpful intermediate representations. Toshniwal et al. (2017) [11] find that positioning the auxiliary loss at an intermediate layer improves performance. Similarly, Arik et al. (2017) [12] predict the phoneme duration and frequency profile as auxiliary tasks for speech synthesis.

**Machine translation** The main benefit MTL has brought to machine translation (MT) is by jointly training translation models from and to different languages: Dong et al. (2015) [13] jointly train the decoders; Zoph and Knight (2016) [14] jointly train the encoders, while Johnson et al. (2016) [15] jointly train both encoders and decoders; Malaviya et al. (2017) [16] train one model to translate from 1017 languages into English.

Other tasks have also shown to be useful for MT: Luong et al. (2015) [17] show gains using parsing and image captioning as auxiliary tasks; Niehues and Cho (2017) [18] combine NMT with POS tagging and NER; Wu et al. (2017) [55] jointly model the target word sequence and its dependency tree structure.

**Multilingual tasks** Similarly to MT, it can often be beneficial to jointly train models for different languages: Gains have been shown for dependency parsing [22, 28], named entity recognition [23], part-of-speech tagging [24], document classification [25], discourse segmentation [26], and sequence tagging [27].

**Language grounding** For grounding language in images or videos, it is often useful to enable the model to learn causal relationships in the data. For video captioning, Pasunuru and Bansal (2017) [30] jointly learn to predict the next frame in the video and to predict entailment, while Hermann et al. (2017) [46] also predict the next frame in a video and the words that represent the visual state for language learning in a simulated environment.

**Semantic parsing** For a task where multiple label sets or formalisms are available such as for semantic parsing, an interesting MTL strategy is to learn these formalisms together: To this end, Guo et al. (2016) [31] jointly train on multi-typed treebanks; Peng et al. (2017) [32] learn three semantic dependency graph formalisms simultaneously; Fan et al. (2017) [33] jointly learn different Alexa-based semantic parsing formalisms; and Zhao and Huang (2017) [57] jointly train a syntactic and a discourse parser.

**Representation learning** For learning general-purpose representations, the challenge often is in defining the objective. Most existing representation learning models have been based on a single loss function, such as predicting the next word [34] or sentence [35] or training on a certain task such as entailment [36] or MT [37]. Rather than learning representations based on a single loss, intuitively, representations should become more general as more tasks are used to learn them. As an example of this strategy, Hashimoto et al. (2017) [59] jointly train a model on multiple NLP tasks, while Jernite et al. (2017) [38] propose several discourse-based artificial auxiliary tasks for sentence representation learning.

**Question answering** For question answering (QA) and reading comprehension, it is beneficial to learn the different parts of a more complex end-to-end model together: Choi et al. (2017) [52] jointly learn a sentence selection and answer generation model, while Wang et al. (2017) [56] jointly train a ranking and reader model for open-domain QA.

**Information retrieval** For relation extraction, information related to different relations or roles can often be shared. To this end, Jiang (2009) [29] jointly learn linear models between different relation types; Yang and Mitchell (2017) [53] jointly predict semantic role labels and relations; Katiyar and Cardie (2017) [58] jointly extract entities and relations; and Liu et al. (2015) [39] jointly train domain classification and web search ranking.

**Chunking** Chunking has been shown to benefit from being jointly trained with low-level tasks such as POS tagging [40, 41, 42].

**Miscellaneous** Besides the tasks mentioned above, various other tasks have been shown to benefit from MTL: Balikas and Moura (2017) [21] jointly train coarse-grained and fine-grained sentiment analysis; Luo et al. (2017) [47] jointly predict charges and extract articles; Augenstein and Søgaard (2017) [48] use several auxiliary tasks for keyphrase boundary detection; and Isonuma et al. (2017) [54] pair sentence extraction with document classification.












## Conclusion

I hope this blog post was able to provide you with some insight with regard to which strategies are employed to select auxiliary tasks and objectives for multi-task learning in NLP. As I mentioned before, multi-task learning can be very broadly defined. I have

tried to provide as broad of an overview as possible but I still likely have omitted many relevant approaches. If you are aware of an approach that provides a valuable perspective that is not represented here, please let me know in the comments below.













## References

1. Weng, R., Huang, S., Zheng, Z., Dai, X., & Chen, J. (2017). Neural Machine Translation with Word Predictions. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. [DOI](#)
2. Rei, M. (2017). Semi-supervised Multitask Learning for Sequence Labeling. In Proceedings of ACL 2017. [DOI](#)
3. Ramachandran, P., Liu, P. J., & Le, Q. V. (2016). Unsupervised Pretraining for Sequence to Sequence Learning. arXiv Preprint arXiv:1611.02683. [DOI](#)
4. Peters, M. E., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 1756–1765). [DOI](#)
5. Ganin, Y., & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. In Proceedings of the 32nd International Conference on Machine Learning. (Vol. 37). [DOI](#)
6. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research, 17, 1–35. <http://www.jmlr.org/papers/volume17/15-239/source/15-239.pdf> [DOI](#)
7. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In ICLR 2014. Retrieved from <http://arxiv.org/abs/1312.6199> [DOI](#)
8. Miyato, T., Dai, A. M., & Goodfellow, I. (2016). Virtual Adversarial Training for Semi-Supervised Text Classification. Retrieved from <http://arxiv.org/abs/1605.07725> [DOI](#)
9. Liu, P., Qiu, X., & Huang, X. (2017). Adversarial Multi-task Learning for Text Classification. In ACL 2017. Retrieved from <http://arxiv.org/abs/1704.05742> [DOI](#)
10. Plank, B., Søgaard, A., & Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. [DOI](#)

11. Toshniwal, S., Tang, H., Lu, L., & Livescu, K. (2017). Multitask Learning with Low-Level Auxiliary Tasks for Encoder-Decoder Based Speech Recognition. Retrieved from <http://arxiv.org/abs/1704.01631> 
12. Phoneme duration and frequency profile 
- Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... Shoeybi, M. (2017). Deep Voice: Real-time Neural Text-to-Speech. In ICML 2017.
13. Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-Task Learning for Multiple Language Translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (pp. 1723–1732). 
14. Zoph, B., & Knight, K. (2016). Multi-Source Neural Translation. NAACL, 30–34. Retrieved from <http://arxiv.org/abs/1601.00710> 
15. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... Dean, J. (2016). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv Preprint arXiv:1611.0455. 
16. Malaviya, C., Neubig, G., & Littell, P. (2017). Learning Language Representations for Typology Prediction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Retrieved from <http://arxiv.org/abs/1707.09569> 
17. Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task Sequence to Sequence Learning. In arXiv preprint arXiv:1511.06114. Retrieved from <http://arxiv.org/abs/1511.06114> 
18. Niehues, J., & Cho, E. (2017). Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning. In WMT 2017. Retrieved from <http://arxiv.org/abs/1708.00993> 
19. Elliott, D., & Kádár, Á. (2017). Imagination improves Multimodal Translation. Retrieved from <http://arxiv.org/abs/1705.04350> 
20. Yu, J., & Jiang, J. (2016). Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP2016), 236–246. Retrieved from <http://www.aclweb.org/anthology/D/D16/D16-1023.pdf> 
21. Balikas, G., & Moura, S. (2017). Multitask Learning for Fine-Grained Twitter Sentiment Analysis. In International ACM SIGIR Conference on Research and Development in Information Retrieval 2017. 

22. Duong, L., Cohn, T., Bird, S., & Cook, P. (2015). Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), 845–850. [↗](#)
23. Gillick, D., Brunk, C., Vinyals, O., & Subramanya, A. (2016). Multilingual Language Processing From Bytes. NAACL, 1296–1306. Retrieved from <http://arxiv.org/abs/1512.00103> [↗](#)
24. Fang, M., & Cohn, T. (2017). Model Transfer for Tagging Low-resource Languages using a Bilingual Dictionary. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). [↗](#)
25. Popescu-belis, A. (2017). Multilingual Hierarchical Attention Networks for Document Classification. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. [↗](#)
26. Braud, C., Lacroix, O., & Søgaard, A. (2017). Cross-lingual and cross-domain discourse segmentation of entire documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. [↗](#)
27. Yang, Z., Salakhutdinov, R., & Cohen, W. (2016). Multi-Task Cross-Lingual Sequence Tagging from Scratch. [↗](#)
28. Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., & Smith, N. A. (2016). One Parser, Many Languages. Transactions of the Association for Computational Linguistics, Vol. 4, Pp. 431–444, 2016, 4, 431–444. Retrieved from <http://arxiv.org/abs/1602.01595> [↗](#)
29. Jiang, J. (2009). Multi-task transfer learning for weakly-supervised relation extraction. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, (August), 1012–1020. <https://doi.org/10.3115/1690219.1690288> [↗](#)
30. Pasunuru, R., & Bansal, M. (2017). Multi-Task Video Captioning with Video and Entailment Generation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). [↗](#)
31. Guo, J., Che, W., Wang, H., & Liu, T. (2016). Exploiting Multi-typed Treebanks for Parsing with Deep Multi-task Learning. Retrieved from <http://arxiv.org/abs/1606.01161> [↗](#)
32. Peng, H., Thomson, S., Smith, N. A., & Allen, P. G. (2017). Deep Multitask Learning for Semantic Dependency Parsing. In ACL 2017. Retrieved from <https://arxiv.org/pdf/1704.06855.pdf> [↗](#)



33. Fan, X., Monti, E., Mathias, L., & Dreyer, M. (2017). Transfer Learning for Neural Semantic Parsing. ACL Repl4NLP 2017. Retrieved from <http://arxiv.org/abs/1706.04326> 
34. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. NIPS. 
35. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-Thought Vectors, (786). Retrieved from <http://arxiv.org/abs/1506.06726> 
36. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 
37. Mccann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in Translation: Contextualized Word Vectors. 
38. Jernite, Y., Bowman, S. R., & Sontag, D. (2017). Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning. Retrieved from <http://arxiv.org/abs/1705.00557> 
39. Liu, X., Gao, J., He, X., Deng, L., Duh, K., & Wang, Y.-Y. (2015). Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval. NAACL-2015, 912–921. 
40. Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing. Proceedings of the 25th International Conference on Machine Learning - ICML '08, 20(1), 160–167. <https://doi.org/10.1145/1390156.1390177> 
41. Søgaard, A., & Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 231–235. 
42. Ruder, S., Bingel, J., Augenstein, I., & Søgaard, A. (2017). Sluice networks: Learning what to share between loosely related tasks. arXiv Preprint arXiv:1705.08142. Retrieved from <http://arxiv.org/abs/1705.08142> 
43. Zhu, J., Park, T., Efros, A. A., Ai, B., & Berkeley, U. C. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. 
44. Xia, Y., He, D., Qin, T., Wang, L., Yu, N., Liu, T.-Y., & Ma, W.-Y. (2016). Dual Learning for Machine Translation. In Advances in Neural Information Processing Systems 29 (NIPS 2016) (pp. 1–9). Retrieved from <http://arxiv.org/abs/1611.00179> 

45. Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., & Liu, T. (2017). Dual Supervised Learning. In ICML. [↗](#)
46. Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., ... Phil Blunsom. (2017). Grounded Language Learning in a Simulated 3D World. Retrieved from <https://arxiv.org/pdf/1706.06551.pdf> [↗](#)
47. Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to Predict Charges for Criminal Cases with Legal Basis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Retrieved from <http://arxiv.org/abs/1707.09168> [↗](#)
48. Augenstein, I., & Søgaard, A. (2017). Multi-Task Learning of Keyphrase Boundary Classification. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Retrieved from <http://arxiv.org/abs/1704.00514> [↗](#)
49. Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-Task Learning for Mental Health using Social Media Text. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Retrieved from <http://m-mitchell.com/publications/multitask-clinical.pdf> [↗](#)
50. Cheng, H., Fang, H., & Ostendorf, M. (2015). Open-Domain Name Error Detection using a Multi-Task RNN. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 737–746). [↗](#)
51. Roy, D. (2017). Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 478–483). [↗](#)
52. Choi, E., Hewlett, D., Uszkoreit, J., Lacoste, A., & Berant, J. (2017). Coarse-to-Fine Question Answering for Long Documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 209–220). [↗](#)
53. Yang, B., & Mitchell, T. (2017). A Joint Sequential and Relational Model for Frame-Semantic Parsing. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. [↗](#)
54. Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., & Sakata, I. (2017). Extractive Summarization Using Multi-Task Learning with Document Classification. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2091–2100). [↗](#)
55. Wu, S., Zhang, D., Yang, N., Li, M., & Zhou, M. (2017). Sequence-to-Dependency Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 698–707). [↗](#)

56. Wang, S., Yu, M., Guo, X., Wang, Z., Klinger, T., & Zhang, W. (2017). R<sup>3</sup>: Reinforced Reader-Ranker for Open-Domain Question Answering. [\[PDF\]](#)
57. Zhao, K., & Huang, L. (2017). Joint Syntacto-Discourse Parsing and the Syntacto-Discourse Treebank. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. [\[PDF\]](#)
58. Katiyar, A., & Cardie, C. (2017). Going out on a limb : Joint Extraction of Entity Mentions and Relations without Dependency Trees. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 917–928). [\[PDF\]](#)
59. Hashimoto, K., Xiong, C., Tsuruoka, Y., & Socher, R. (2017). A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Retrieved from <http://arxiv.org/abs/1611.01587> [\[PDF\]](#)
60. Bingel, J., & Søgaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In EACL. Retrieved from <http://arxiv.org/abs/1702.08303> [\[PDF\]](#)



Welcome to Disqus! Discover more great discussions just like this one. We're a lot more than comments.

Get Started

Dismiss ✕

0 Comments

Blog

 Jun-Feng Tian ▾

 Recommend

 Share

Sort by Best ▾



Start the discussion...

Be the first to comment.

#### ALSO ON BLOG

##### Deep Learning for NLP Best Practices

11 comments • 3 months ago

Avatar **Sebastian Ruder** — Hey Stefan, I'm glad the articles were helpful. Good luck with your thesis! :)

##### An Overview of Multi-Task Learning for Deep Neural Networks

22 comments • 5 months ago

Avatar **Kirtane Chiron** — Thank you for putting this all together. Really excited to see how this direction unfolds and if it can be applied to

##### An overview of gradient descent optimization algorithms

2 comments • 2 years ago

Avatar **Sebastian Ruder** — Hey Liam, Thanks for asking. Sure, feel free to translate the post. It would be great if you could link back to the

##### Highlights of EMNLP 2016: Dialogue, deep learning, and more

2 comments • a year ago

Avatar **Sebastian Ruder** — Hey Domyoung, glad you found it helpful. :)