David Grangier

Research Scientist / Machine Learning

☎ *(650) 772-8713*
✉ *david@grangier.info*

Machine Learning Research with Practical Impact.

Experience

2018–Present	**Google Brain**, *Mountain View, CA*, Machine Learning Research. Language Modeling, Machine Translation, Speech Separation. Manager: Douglas Eck.
2014–2018	**Facebook AI Research**, *Menlo Park, CA*, Machine Learning Research. Language Modeling, Machine Translation. Manager: Ronan Collobert.
2012–2014	**Microsoft Research**, *Redmond, WA*, Machine Learning Research. Interactive Machine Learning, Machine Teaching, Active Learning. Manager: Patrice Simard.
2008–2011	**NEC Labs**, *Princeton, NJ*, Machine Learning Research. Deep Supervised Learning for Image Segmentation, Learning Text Representation, Learning to Rank. Advisor: Léon Bottou.
2007	**Google Research**, *Mountain View, CA*, Internship in Machine Learning Research. Image Retrieval, Spoken Keyword Spotting. Advisor: Samy Bengio.

Education

2003–2008	**EPFL**, *Lausanne, Switzerland*, PhD in Machine Learning. Machine Learning for Information Retrieval, Learning to Rank. Advisor: Samy Bengio
2002–2003	**Eurecom Institute**, *Sophia Antipolis, France*, Msc in Telecommunications. Specialization in Speech Processing, Graduated with Hitachi Distinction (Best thesis project). Advisor: Christian Wellekens
2000–2003	**ENST Bretagne**, *Brest, France*, Bsc in Telecom Engineering. Graduated with rank 19 of 214.

Notable Projects and Achievements

Present	Natural translation, reliable evaluation for machine translation of high accuracy languages, with Google Translate.
2016 – 2018	Improving neural machine translation at Facebook AI Research, part of the team winning WMT Benchmark for English–German, English–Russian in 2018.
2007 & 2012	Helping deploy my thesis work on image retrieval to improve image search at Google Search (2007, with Samy Bengio) & Microsoft Bing (2012, with Mikhail Parakhin).

Workshops, Tutorials, Courses

2016 **ICML'16 Neural Nets Back to The Future**, Linking pioneer research with current work, co-organizer with John Platt, Leon Bottou and Tomas Mikolov..

2014 **NIPS'14 Workshop on Learning Semantics**, Learning representation & reasoning, co-organizer with Cedric Archambeau, Antoine Bordes, Leon Bottou and Chris Burges..

2011 **Machine Learning for Rankings**, Graduate teaching part of Berkeley/Info 290, UC Berkeley School of Information..

2006 **NIPS'06 Workshop on Learning to Compare Examples**, Metric & distance learning, co-organizer with S. Bengio..

Open Source & Resources

2016–2018 **Fairseq**, Torch library for sequence transduction with Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng and Michael Auli, https://github.com/facebookresearch/fairseq .

2016 **WikiBio**, Biography dataset for structured data to text with Remi Lebret and Michael Auli, https://github.com/rlebret/wikipedia-biography-dataset .

2008 **PAMIR**, C++ image retrieval library with Samy Bengio, https://www.idiap.ch/archive/pamir .

Advised Students

2019–Present **Dan Iter**, Selection of training data for model adaptation.

2018–Present **Daphne Ippolito**, Language modeling for creative story generation.

2020 **Kelly Marchisio**, Supervised vs unsupervised machine translation styles.

2019 **Parker Riley**, Translationese and naturalness in machine translation.

2016–2018 **Angela Fan**, Machine translation and language modeling.

2015 **Rémi Lebret**, Fact-to-text generation.

2015 **Wenlin Chen**, Large-scale language modeling.

2012 **Xiaoxiao Shi**, Gradient boosting consensus.

2011 **Hugo Penedones**, Object classification with 3D pose information at training time.

Languages

English Fluent

French Native

References

Upon request.

Publications

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. Minimum bayes risk decoding with neural metrics of translation quality. *arXiv*, (2111.09388), 2021.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang

Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics (TACL)*, 2021.

Neil Zeghidour, Olivier Teboul, and David Grangier. Dive: End-to-end speech diarization via iterative speaker embedding. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

Lucio Dery, Yann Dauphin, and David Grangier. Auxiliary task update decomposition: The good, the bad and the neutral. In *International Conference on Learning Representation (ICLR)*, 2021.

Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE ACM Transaction on Audio Speech and Language Processing (TASLP)*, 2021.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Aaqib Saeed, David Grangier, Olivier Pietquin, and Neil Zeghidour. Learning from heterogeneous EEG signals with differentiable channel reordering. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

David Grangier and Dan Iter. The trade-offs of domain adaptation for neural language models. *arXiv*, (2109.10274), 2021.

Dan Iter and David Grangier. On the complementarity of data selection and fine tuning for domain adaptation. *arXiv*, (2109.07591), 2021.

Kelly Marchisio, Markus Freitag, and David Grangier. What can unsupervised machine translation contribute to high-resource language pairs? *arXiv*, (2106.15818), 2021.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics (TACL)*, 2020.

Markus Freitag, George Foster, David Grangier, and Colin Cherry. Human-paraphrased references improve neural machine translation. In *Conference on Machine Translation (WMT)*, 2020.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. Translationese as a language in "multilingual" NMT. 2020.

Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch. Towards better storylines with sentence-level language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: long form question answering. In *ACL (1)*, pages 3558–3567. Association for Computational Linguistics, 2019.

Aurko Roy and David Grangier. Unsupervised paraphrasing without translation. In *ACL (1)*, pages 6033–6039. Association for Computational Linguistics, 2019.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pages 7753–7762. Computer Vision Foundation / IEEE, 2019.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT (Demonstrations)*, pages 48–53. Association for Computational Linguistics, 2019.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In *WMT (1)*, pages 53–63. Association for Computational Linguistics, 2019.

Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *BMVC*, abs/1901.07677, 2019.

Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. In *NMT@ACL*, pages 45–54. Association for Computational Linguistics, 2018.

Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *BMVC*, page 299. BMVA Press, 2018.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *EMNLP*, pages 489–500. Association for Computational Linguistics, 2018.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR, 2018.

David Grangier and Michael Auli. Quickedit: Editing text & translations by crossing words out. In *NAACL-HLT*, pages 272–282. Association for Computational Linguistics, 2018.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Classical structured prediction losses for sequence to sequence learning. In *NAACL-HLT*, pages 355–364. Association for Computational Linguistics, 2018.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *WMT*, pages 1–9. Association for Computational Linguistics, 2018.

Jonas Gehring, Michael Auli, David Grangier, and Yann N. Dauphin. A convolutional encoder model for neural machine translation. In *ACL (1)*, pages 123–135. Association for Computational Linguistics, 2017.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 933–941. PMLR, 2017.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR, 2017.

Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. Efficient softmax approximation for gpus. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1302–1310. PMLR, 2017.

Wenlin Chen, David Grangier, and Michael Auli. Strategies for training large vocabulary neural language models. In *ACL (1)*. The Association for Computer Linguistics, 2016.

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *EMNLP*, pages 1203–1213. The Association for Computational Linguistics, 2016.

Yann N. Dauphin and David Grangier. Predicting distributions with linearizing belief networks. In *ICLR*, 2016.

Camille Jandot, Patrice Y. Simard, Max Chickering, David Grangier, and Jina Suh. Interactive semantic featuring for text classification. *arXiv*, (1606.07545), 2016.

Xiaoxiao Shi, Jean-François Paiement, David Grangier, and Philip S. Yu. GBC: gradient boosting consensus model for heterogeneous data. *Statistical Analysis and Data Mining*, 7(3):161–174, 2014.

Patrice Y. Simard, David Maxwell Chickering, Aparna Lakshmiratan, Denis Xavier Charles, Léon Bottou, Carlos Garcia Jurado Suarez, David Grangier, Saleema Amershi, Johan Verwey, and Jina Suh. ICE: enabling non-experts to build models interactively for large-scale lopsided problems. *arXiv*, (1409.4814), 2014.

Xiaoxiao Shi, Jean-François Paiement, David Grangier, and Philip S. Yu. Learning from heterogeneous sources via gradient boosting consensus. In *SDM*, pages 224–235. SIAM / Omnipress, 2012.

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Q. Weinberger. Learning to rank with (a lot of) word features. *Inf. Retr.*, 13(3):291–314, 2010.

Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, pages 163–171. Curran Associates, Inc., 2010.

David Grangier and Iain Melvin. Feature set embedding for incomplete data. In *NIPS*, pages 793–801. Curran Associates, Inc., 2010.

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Corinna Cortes, and Mehryar Mohri. Half transductive ranking. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 49–56. JMLR.org, 2010.

Joseph Keshet, David Grangier, and Samy Bengio. Discriminative keyword spotting. *Speech Communication*, 51(4):317–329, 2009.

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Q. Weinberger. Supervised semantic indexing. In *CIKM*, pages 187–196. ACM, 2009.

Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Corinna Cortes, and Mehryar Mohri. Polynomial semantic indexing. In *NIPS*, pages 64–72. Curran Associates, Inc., 2009.

David Grangier and Samy Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1371–1384, 2008.

David Grangier. *Machine Learning for Information Retrieval*. PhD thesis, Ecole Polytechnique Federale de Lausanne EPFL, 2008.

David Grangier and Samy Bengio. Learning the inter-frame distance for discriminative template-based keyword detection. In *INTERSPEECH*, pages 902–905. ISCA, 2007.

David Grangier, Florent Monay, and Samy Bengio. Learning to retrieve images from text queries with a discriminative model. In *Adaptive Multimedia Retrieval*, volume 4398 of *Lecture Notes in Computer Science*, pages 42–56. Springer, 2006.

David Grangier, Florent Monay, and Samy Bengio. A discriminative approach for the retrieval of images from text queries. In *ECML*, volume 4212 of *Lecture Notes in Computer Science*, pages 162–173. Springer, 2006.

David Grangier and Samy Bengio. A neural network to retrieve images from text queries. In *ICANN (2)*, volume 4132 of *Lecture Notes in Computer Science*, pages 24–34. Springer, 2006.

David Grangier and Samy Bengio. Inferring document similarity from hyperlinks. In *CIKM*, pages 359–360. ACM, 2005.

David Grangier and Alessandro Vinciarelli. Effect of segmentation method on video retrieval performance. In *ICME*, pages 5–8. IEEE Computer Society, 2005.

David Grangier. Information retrieval on noisy text. Master's thesis, Eurecom Institute, France, 2003.