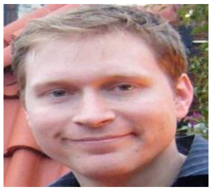
Hal Daumé III, UMD; Marine Carpuat, CNRC; Alex Fraser, University of Stuttgart; Chris Quirk, Microsoft Research

Domain Adaptation in Machine Translation: Findings from the 2012 Johns Hopkins University Summer Workshop

Hal Daumé III is Assistant Professor of Computer Science at the University of Maryland, with a joint appointment in Linguistics. He was previously an assistant professor in the School of Computing at the University of Utah. His primary research interest is in developing new learning algorithms for prototypical problems arising in the context of language processing and artificial intelligence. This includes topics like structured prediction, domain adaptation and unsupervised learning as well as multilingual modeling and affect analysis. He associates himself with conferences like ACL, ICML, NIPS and EMNLP. He earned his PhD at the University of Southern California with a thesis on structured prediction for language, supervised by Daniel Marcu. In 2003, he worked with Eric Brill in machine learning and applied statistics at Microsoft.Research. Prior to that, he studied math—mostly logic—at Carnegie Mellon University.

Marine Carpuat is a researcher at the National Research Council Canada, where she works on natural language processing and statistical machine translation. Marine is particularly interested in designing translation models that capture word meaning in context, and in making translation tools more useful for human translators and users. Before moving to NRC, Marine was a postdoctoral researcher at Columbia University. She received a PhD in Computer Science from the Hong Kong University of Science & Technology in 2008, under the supervision of Dekai Wu. She also earned a MPhil in Electrical Engineering from HKUST, and an engineering degree from the French Grande Ecole Supélec.



Chris Quirk is a Senior Researcher in the Natural Language Processing group at Microsoft Research. His primary focus is statistical syntax-based machine translation and related technologies such as parsing and machine learning. He completed his BS in Computer Science and Mathematics from Carnegie Mellon University in 2000 prior to joining Microsoft. He has acted as an area chair at ACL (2010) and EMNLP (2009, 2012).

ABSTRACT: Statistical Machine Translation (SMT) is now the common, perhaps dominant, paradigm for machine translation. Like most statistical learning approaches, SMT works under the assumption that the distribution of the training data matches the distribution of the test data, so that translation rules learned on parallel corpora are representative of translations needed at test time. However, this assumption rarely holds in practice: parallel corpora in the domain and language pair of interest are often too small to train domain-specific models, while models trained on large amounts of unrelated corpora do not necessarily match new test domains. As a result, SMT systems perform poorly when applied on new domains. Yet many consumers could benefit from improved domain specific translation: this includes enterprise content producers who need products-specific translations, and content consumers with access to ever increasing sources of data. In this talk, we provide an overview of research conducted during the Johns Hopkins University summer workshop to understand how domain differences are manifested in the translation task and how they affect SMT models. We first investigate techniques for quantifying the difference between domains, from the perspective of translation quality (measured by BLEU score), across a variety of domains and data sets. Next, we present several statistical techniques for adapting statistical machine translation systems using feature rich models. Finally, we explore the problem of identifying translations for new terms, and discovering translations for words whose meanings may shift across translations.