Diverse Words, Shared Meanings: Statistical Machine Translation for Paraphrase, Grounding, and Intent

Chris Brockett
Microsoft Research
Redmond, WA
Chris.Brockett@microsoft.com

Abstract

Can two different descriptions refer to the same event or action? Recognising that dissimilar strings are equivalent in meaning for some purpose is something that humans do rather well, but it is a task at which machines often fail. In the Natural Language Processing Group at Microsoft Research, we are attempting to address this challenge at sentence scale by generating semantically equivalent rewrites that can be used in applications ranging from authoring assistance to intent mapping for search or command and control. The Microsoft Translator paraphrase engine, developed in the NLP group, is a large-scale phrasal machine translation system that generates short sentential and phrasal paraphrases in English and has a public API that is available to researchers and developers. I will present the data extraction process, architecture, issues in generating diverse outputs, applications and possible future directions, and discuss the strengths and limitations of the statistical machine translation model as it relates to paraphrasing, how paraphrase is like machine translation, and how it differs in important respects. The statistical machine translation approach also has broad applications in capturing user intent in search, conversational understanding, and the grounding of language in objects and actions, all active areas of investigation in Microsoft Research.

Chris Brockett. 2012. Diverse Words, Shared Meanings: Statistical Machine Translation for Paraphrase, Grounding, and Intent. In *Proceedings of Australasian Language Technology Association Workshop*, pages 3−3.