The Language Resource Spectrum: A Perspective from Google

Ryan McDonald
Google
London, U.K.
ryanmcd@google.com

Abstract

Google extracts a vast amount of knowledge from a wide variety of data for a diverse set of language technologies. This includes everything from morphosyntactic lexical information to the facts that populate the Knowledge Graph. At the core of these systems are language resources, ranging from high quality ones created by trained experts to noisy ones automatically extracted from large data. While early success using machine learning in language technologies relied on the former, recent advances have held the promise that the latter is sufficient to build state-of-the-art systems, potentially negating the need for human intensive resource creation and even linguistic insight. In this talk I will argue – based on empirical evidence from end-user tasks – that resources across this spectrum have an important role to play. In many cases, from machine translation to knowledge extraction, replacing small high quality resources leads to drops in quality that cannot be overcome with cheap noisy data, no matter how abundant. Yet, leveraging massive automatically constructed resources almost always adds an additional layer of signals that improve the quality of language technologies.