Putting a Value on Comparable Data

Kevin Knight
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA, 90292 USA
knight@isi.edu

Invited Talk

Abstract

Machine translation began in 1947 with an influential memo by Warren Weaver. In that memo, Weaver noted that human code-breakers could transform ciphers into natural language (e.g., into Turkish)

- without access to parallel ciphertext/plaintext data, and
- without knowing the plaintext language's syntax and semantics.

Simple word- and letter-statistics seemed to be enough for the task. Weaver then predicted that such statistical methods could also solve a tougher problem, namely language translation.

This raises the question: can sufficient translation knowledge be derived from comparable (non-parallel) data?

In this talk, I will discuss initial work in treating foreign language as a code for English, where we assume the code to involve both word substitutions and word transpositions. In doing so, I will quantitatively estimate the value of non-parallel data, versus parallel data, in terms of end-to-end accuracy of trained translation systems. Because we still know very little about solving word-based codes, I will also describe successful techniques and lessons from the realm of letter-based ciphers, where the non-parallel resources are (1) enciphered text, and (2) unrelated plaintext. As an example, I will describe how we decoded the *Copiale* cipher with limited "computer-like" knowledge of the plaintext language.

The talk will wrap up with challenges in exploiting comparable data at all levels: letters, words, phrases, syntax, and semantics.

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, page 1,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.