Writing Systems, Transliteration and Decipherment

Kevin Knight (USC/ISI)
Richard Sproat (CSLU/OHSU)

Description

Nearly all of the core data that computational linguists deal with is in the form of text, which is to say that it consists of language data written (usually) in the standard writing system for the language in question. Yet surprisingly little is generally understood about how writing systems work. This tutorial will be divided into three parts. In the first part we discuss the history of writing and introduce a wide variety of writing systems, explaining their structure and how they encode language. We end this section with a brief review of how some of the properties of writing systems are handled in modern encoding systems, such as Unicode, and some of the continued pitfalls that can occur despite the best intentions of standardization. The second section of the tutorial will focus on the problem of transcription between scripts (often termed "transliteration"), and how this problem—which is important both for machine translation and named entity recognition—has been addressed. The third section is more theoretical and, at the same time we hope, more fun. We will discuss the problem of decipherment and how computational methods might be brought to bear on the problem of unlocking the mysteries of as yet undeciphered ancient scripts. We start with a brief review of three famous cases of decipherment. We then discuss how techniques that have been used in speech recognition and machine translation might be applied to the problem of decipherment. We end with a survey of the as-yet undeciphered ancient scripts and give some sense of the prospects of deciphering them given currently available data.

Outline

First hour:

- History of writing
- Survey of writing systems and how they work
- Modern encodings

Second hour:

- Problems of transcription (transliteration)
- Generative models of transcription

Break

Proceedings of NAACL HLT 2009: Tutorials, pages 15–16,
Boulder, Colorado, June 2009.

- More on generative models of transcription
- Discriminative models

Third Hour

- Famous cases of decipherment
- Prospects for "autodecipherment"
- What's left to decipher?

Target Audience

This tutorial will be of interest to anyone who wishes to have a better understanding of how writing (the form of language that most computational linguists deal with) works, and how such problems as transcription (transliteration) and decipherment are approached computationally.

Bios

Kevin Knight is a Research Associate Professor in Computer Science at the University of Southern California, a Senior Research Scientist and Fellow at the USC/Information Sciences Institute, and Chief Scientist at Language Weaver. Dr. Knight received a Ph.D. from Carnegie Mellon University in 1992, and a bachelor's degree from Harvard University. His current interests include better statistical machine translation through linguistics, and he is also working on exploiting cryptographic techniques to solve hard translation problems.

Richard Sproat received his Ph.D. in Linguistics from the Massachusetts Institute of Technology in 1985. Since then he has worked at AT&T Bell Labs, at Lucent's Bell Labs and at AT&T Labs – Research, before joining the faculty of the University of Illinois, and subsequently the Oregon Health & Science University. Sproat has worked in numerous areas relating to language and computational linguistics, including syntax, morphology, computational morphology, articulatory and acoustic phonetics, text processing, text-to-speech synthesis, writing systems, and text-to-scene conversion.