Decipherment

Kevin Knight
USC/ISI
4676 Admiralty Way
Marina del Rey CA 90292
knight@isi.edu

Abstract

The first natural language processing systems had a straightforward goal: decipher coded messages sent by the enemy. This tutorial explores connections between early decipherment research and today's NLP work. We cover classic military and diplomatic ciphers, automatic decipherment algorithms, unsolved ciphers, language translation as decipherment, and analyzing ancient writing as decipherment.

1 Tutorial Overview

The first natural language processing systems had a straightforward goal: decipher coded messages sent by the enemy. Sixty years later, we have many more applications, including web search, question answering, summarization, speech recognition, and language translation. This tutorial explores connections between early decipherment research and today's NLP work. We find that many ideas from the earlier era have become core to the field, while others still remain to be picked up and developed.

We first cover classic military and diplomatic cipher types, including complex substitution ciphers implemented in the first electro-mechanical encryption machines. We look at mathematical tools (language recognition, frequency counting, smoothing) developed to decrypt such ciphers on proto-computers. We show algorithms and extensive empirical results for solving different types of ciphers, and we show the role of algorithms in recent decipherments of historical documents.

We then look at how foreign language can be viewed as a code for English, a concept developed by Alan Turing and Warren Weaver. We describe recently published work on building automatic translation systems from non-parallel data. We also demonstrate how some of the same algorithmic tools can be applied to natural language tasks like part-of-speech tagging and word alignment.

Turning back to historical ciphers, we explore a number of unsolved ciphers, giving results of initial computer experiments on several of them. Finally, we look briefly at writing as a way to encipher phoneme sequences, covering ancient scripts and modern applications.

2 Outline

1. Classical military/diplomatic ciphers (15 minutes)
   - 60 cipher types (ACA)
   - Ciphers vs. codes
   - Enigma cipher: the mother of natural language processing
     - computer analysis of text
     - language recognition
     - Good-Turing smoothing
2. Foreign language as a code (10 minutes)
   - Alan Turing's "Thinking Machines"
   - Warren Weaver's Memorandum
3. Automatic decipherment (55 minutes)
   - Cipher type detection
   - Substitution ciphers (simple, homophonic, polyalphabetic, etc)
     - plaintext language recognition
       - how much plaintext knowledge is needed

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 3–4,
Sofia, Bulgaria, August 4-9 2013.

        * index of coincidence, unicity distance, and other measures
      - navigating a difficult search space
        * frequencies of letters and words
        * pattern words and cribs
        * EM, ILP, Bayesian models, sampling
      - recent decipherments
        * Jefferson cipher, Copiale cipher, civil war ciphers, naval Enigma
    - Application to part-of-speech tagging, word alignment
    - Application to machine translation without parallel text
    - Parallel development of cryptography and translation
    - Recently released NSA internal newsletter (1974-1997)

4. *** Break *** (30 minutes)

5. Unsolved ciphers (40 minutes)
    - Zodiac 340 (1969), including computational work
    - Voynich Manuscript (early 1400s), including computational work
    - Beale (1885)
    - Dorabella (1897)
    - Taman Shud (1948)
    - Kryptos (1990), including computational work
    - McCormick (1999)
    - Shoeboxes in attics: DuPonceau journal, Finnerana, SYP, Mopse, diptych

6. Writing as a code (20 minutes)
    - Does writing encode ideas, or does it encode phonemes?
    - Ancient script decipherment
      - Egyptian hieroglyphs
      - Linear B
      - Mayan glyphs
      - Ugaritic, including computational work
      - Chinese Nüshu, including computational work
    - Automatic phonetic decipherment
    - Application to transliteration

7. Undeciphered writing systems (15 minutes)
    - Indus Valley Script (3300BC)
    - Linear A (1900BC)
    - Phaistos disc (1700BC?)
    - Rongorongo (1800s?)

8. Conclusion and further questions (15 minutes)

3 About the Presenter

Kevin Knight is a Senior Research Scientist and Fellow at the Information Sciences Institute of the University of Southern California (USC), and a Research Professor in USC's Computer Science Department. He received a PhD in computer science from Carnegie Mellon University and a bachelor's degree from Harvard University. Professor Knight's research interests include natural language processing, machine translation, automata theory, and decipherment. In 2001, he co-founded Language Weaver, Inc., and in 2011, he served as President of the Association for Computational Linguistics. Dr. Knight has taught computer science courses at USC for more than fifteen years and co-authored the widely adopted textbook *Artificial Intelligence*.