Introduction to Computational Linguistics

Frank Richter

fr@sfs.uni-tuebingen.de.

Seminar für Sprachwissenschaft

Eberhard Karls Universität Tübingen

Germany

Course Goals

- First (largely non-technical) introduction to the field of computational linguistics and its history

- Survey of natural language processing applications

- In-depth look at machine translation as a means to illustrate the major tasks for natural language processing

- Presentation of tools and resources needed for natural language processing applications

The Name of the Field

- Computational Linguistics

- Natural Language Processing

- Human Language Technology

- Language Engineering

The History of the Field

- Not surprisingly, the history of Computational Linguistics is closely connected to the development of the digital computer.

- The theoretical foundations of Computational Linguistics are in the field of mathematical linguistics and formal language theory and in the field of information theory.

- The first application of Computational Linguistics was in the area of machine translation. Therefore, the first professional organization was called *Association for Machine Translation and Computational Linguistics*.

The History of the Digital Computer

1939   John J. Atanasoff designs a prototype for the
ABC (Atanasoff-Berry Computer) with the help of
graduate student Clifford Berry at Iowa State
College. In 1973 a judge ruled it the
first automatic digital computer.

1941   Colossus computer is designed by Alan M. Turing
and built by Thomas H. Flowers and
Maxwell H.A. Newman.

The History of the Digital Computer (2)

1941   Konrad Zuse builds the Z3 computer in Germany, the first calculating machine with automatic control of its operations.

1946   ENIAC (Electronic Numerical Integrator and Computer), with 18,000 vacuum tubes, is dedicated at the U. of Pennsylvania. It was 8 by 100 feet and weighed 80 tons. It could do 5,000 additions and 360 multiplications per second.

The First NLP Application

Bi-lingual Dictionaries for Word-to-Word Machine Translation

1947 Donald Booth and D.H.V. Britten worked out a detailed code for realizing dictionary translation on a digital computer.

1948 R.H. Richens worked out a stem-affix encoding with a longest-match strategy for stem identification and translation.

Richens' Dictionary Encoding

An Example: Latin *amat* = love (3rd pers. sing.)

Dictionary Encoding:

stem dictionary: a-m
suffix dictionary: -at

Search Strategy:

for a given input (e.g. *amat*) find longest prefix (in this case *am*) that matches the stem dictionary;
match the remainder of the string (in this case *at*) against the stem dictionary.

Note: this is a direct precursor to later finite-state-encodings of computational lexica.