Introduction to Computational Linguistics

Frank Richter

fr@sfs.uni-tuebingen.de.

Seminar für Sprachwissenschaft

Eberhard Karls Universität Tübingen

Germany

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 suffix dictionary.

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

The Turing Test for Machine Intelligence

From: Turing, A.M. (1950). Computing machinery and intelligence. Mind, 59, 433–460.

> I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." ...
>
> Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The Imitation Game (1)

The new form of the problem can be described in terms of a game which we call the *imitation game*. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either “X is A and Y is B” or “X is B and Y is A.”

The Imitation Game (2)

The interrogator is allowed to put questions to A and B thus:

C: “Will X please tell me the length of his or her hair?”

Now suppose X is actually A, then A must answer. It is A’s object in the game to try and cause C to make the wrong identification. His answer might therefore be

“My hair is shingled, and the longest strands are about nine inches long.”

The Imitation Game (3)

... The ideal arrangement is to have a teleprinter communicating between the two rooms.

... The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as “I am the woman, don’t listen to him!” to her answers, but it will avail nothing as the man can make similar remarks.

The Imitation Game (4)

We now ask the question, “What will happen when a machine takes the part of A in this game?”

Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, “Can machines think?”

Searle's Argument Against Strong AI (1)

According to strong AI, the computer is not merely a tool in the study of the mind; rather the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. (Searle 1980, p. 417).

In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs themselves are the explanations. (Searle 1980, p. 417)

Searle's Argument Against Strong AI (2)

> According to Strong AI, instantiating a formal program with the right input and output is a sufficient condition of, indeed is constitutive of, intentionality. As Newell (1979) puts it, the essence of the mental is the operation of a physical symbol system. (Searle 1980, p. 421)

Reference: Searle, J. R. (1980), 'Minds, Brains, and Programs', Behavioral and Brain Sciences 3, pp. 417-424

Searle's Chinese Room Argument (1)

Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles.

Searle's Chinese Room Argument (2)

Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes.

Searle's Chinese Room Argument (3)

Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch.

Searle’s Chinese Room Argument (4)

Unknown to me, the people who are giving me all of these symbols call the first batch “a script," they call the second batch a “story” and they call the third batch “questions." Furthermore, they call the symbols I give them back in response to the third batch “answers to the questions.” and the set of rules in English that they gave me, they call “the program.”

Searle's Chinese Room Argument (5)

[...] it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. [...]

Is Artificial Intelligence (AI) Possible?

- Can machines think?
- Can machines understand?
- Can machines learn?
- Can machines have emotions?

Definition of CL (1a)

Computational linguistics is the scientific study of language from a computational perspective.

Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "knowledge-based" ("hand-crafted") or "data-driven" ("statistical" or "empirical").

Definition of CL (1b)

Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system.

(taken from the former ACL web pages, `www.aclweb.org`)

Definition of CL (2)

Computational linguistics is the application of linguistic theories and computational techniques to problems of natural language processing.

(from the former website of a British university)

Definition of CL (3)

Computational linguistics is the science of language with particular attention given to the processing complexity constraints dictated by the human cognitive architecture. Like most sciences, computational linguistics also has engineering applications.

(former web page at Trinity College, Dublin;
nowadays, it looks like this:
`https://www.cs.tcd.ie/courses/csll/CSLLcourse/computational_linguistics.php`)

Definition of CL (4)

Computational linguistics is the study of computer systems for understanding and generating natural language.

Ralph Grishman, Computational Linguistics: An Introduction, Cambridge University Press 1986.

Two Approaches in CL

- Rule-Based Systems
  - Explicit encoding of linguistic knowledge
  - Usually consisting of a set of hand-crafted, grammatical rules
  - Easy to test and debug
  - Require considerable human effort
  - Often based on limited inspection of the data with an emphasis on prototypical examples
  - Often fail to reach sufficient domain coverage
  - Often lack sufficient robustness when input data are noisy

Two Approaches in CL

- Data-Driven Systems
  - Implicit encoding of linguistic knowledge
  - Often using statistical methods or machine learning methods
  - Require less human effort
  - Are data-driven and require large-scale data sources
  - Achieve coverage directly proportional to the richness of the data source
  - Are more adaptive to noisy data

Central Goal of the Field

- build psychologically adequate models of human language processing capabilities on the basis of knowledge about the way in which humans acquire, store, and process language.

- build functionally correct models of human language processing capabilities on the basis of knowledge about the world and about language elicited from people and stored in the system.

Application Areas

- machine translation
- speech recognition
- speech synthesis
- man-machine interfaces

Application Areas

- intelligent word processing: spelling correction, grammar correction
- document management
  - find relevant documents in collections
  - establish authorship of documents
  - catch plagiarism
  - extract information from documents
  - classify documents
  - summarize documents
  - summarize document collections