MT

History and Rule-based systems

Miles Osborne

School of Informatics
University of Edinburgh
miles@inf.ed.ac.uk

January 20, 2012

1 History: MT from 1949 to 1990

2 Rule-based MT

Weaver Memo

> *It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?*

Warren Weaver, 1949

Weaver Memo

The Weaver Memo introduced key ideas:

- Translation can be automatic:
  - If we can break codes, then we can apply similar methods to translation.
- There is a notion of mapping from one language to another language.
  - This mapping can be well defined.

It took 50 years for this view to become widespread

1954 – 1966

Subsequent MT grew out of the Cold War:

- Early systems were concerned with Russian-English.
- Computers were very slow, expensive and primitive.
- Great expectations of making rapid progress.
  - Developments in Computational Linguistics (Chomsky) suggested we can write rules to describe translation

1954 – 1966

Bar-Hillel (MT researcher, early 1960s):

- Fully Automatic High Quality Machine Translation (FAHQMT) is impossible.
  - Argument that translation implies *understanding natural language*.
  - Since we do not know how to understand language, we cannot translate it.
- (More recent view is that we do not always need FAHQMT)

1954 – 1966

ALPAC* Report (1964):

- Evaluated progress in computational linguistics and MT.
- Very skeptical of claims made for MT.
- Humans could do it better and cheaper.

Resulted in (US) MT funding drying-up
*Automatic Language Processing Advisory Committee

1970s – 1990

MT research continued outside of the US:

- Driven by commercial (non-military) needs.
- Desire for cost-effective translation, possibly semi-automated.
- Cheap PCs increased demand.
- *Rule-based* systems emerged:
  - Systran (still going now)
  - Japanese systems.
  - EUROTRA

A Simple Rule-based System

Overview:

source sentence $\overset{\text{transform}}{\rightarrow}$ representation $\overset{\text{transform}}{\rightarrow}$ target sentence

- There may be multiple transformations
- The intermediate representation is sometimes called an *interlingua*
- This is called *transfer-based* MT

Most rule-based systems take this approach (but are vastly more complex)

A Simple Rule-based System

Suppose we are translating from German to English:

Drehen Sie den Knopf eine Position zurück
Turn you the button one position back (gloss)
Turn the button back one position

A Simple Rule-based System

Minimally, we need:

- A dictionary mapping each German word to English word(s).
- Rules to represent the German sentence structure.
- Rules to represent the English sentence structure.
- Rules to relate these two structures together
- A method for selecting one alternative translation over another one.

A Simple Rule-based System

First lookup simple part-of-speech information for each word:

- *knopf* $\rightarrow$ noun
- *ein* $\rightarrow$ det

- Use a dictionary.
- POS information can disambiguate
- Rules can be used to extend coverage.

A Simple Rule-based System

Parse the source sentence:

- (NP *den Knopf*) is the object of *zurückdrehen*
- Parsing allows us to relate source words together.
- Parsing can be partial –we may only need to focus upon parts of the sentence

A Simple Rule-based System

Translate German words into English:

- *knopf* (category=noun) → *button* (category=noun)
- *ein* (category=determiner) → *a* (category=determiner)

Again, we can use a dictionary

A Simple Rule-based System

For the verb *zurückdrehen* we need deal with syntax:

- *If there is an imperative verb X followed by the NP* sie, *the translation is the translation X*
- *Turn back the button a postion*
- **Turn back you the button one position*

These rules are manually created
We may also reorder the sentence using more rules

A Simple Rule-based System

Finally we may map dictionary entries into inflected forms:

- Singular to plural
- Correct form of verbs (etc)

Comments

This discussion is sketchy, but:

- A lexical step translates words in isolation
- Words are related to each other using syntax
- Translation involves writing rules.
  - Building rule-based systems requires skilled labour
- Hard to deal with ambiguity

Comments

Rule-based systems are no longer fashionable:

- Probabilistic approaches are now the dominant paradigm.
- Rules are useful for highly regular translation:
  - Numbers (*one* $\rightarrow \ldots$, *3,14* $\rightarrow$ 3.14)
  - Dates (21/10/08 $\rightarrow$ 10/21/08)
  - Unknown words?

Summary

- MT has been ongoing for half a century.
- Early optimism (bubble) burst.
- Subsequent developments more sober (new bubble?).
- Rule-based systems were the dominant paradigm until the 1990s.