

Introduction to Computational Linguistics

Frank Richter

fr@sfs.uni-tuebingen.de.

**Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen
Germany**

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

A bit of Philosophy of Science

- **Theory:**

A set of statements that determine the format and semantics of descriptions of phenomena in the purview of the theory

- **Methodology:**

An effective theory comes with an explicit methodology for acquiring these descriptions

- **Application:**

A theory associated with a methodology can be applied to tasks for which the methodology is appropriate.

Scientific Strategies

- **Method Oriented Approach:**

devise or import a tool, a procedure or a formalism, apply it to a task and develop it further. Then (optionally) see whether it works for additional tasks

- **Task oriented Approach:**

select a task; devise or import a method or several methods for its solution; integrate the methods as required to improve performance.

Machine Translation

What makes Machine Translation an important application area to study:

- historically first application area, and for at least a decade the only application area, of computational linguistics

Machine Translation

What makes Machine Translation an important application area to study:

- historically first application area, and for at least a decade the only application area, of computational linguistics
- requires all steps relevant to linguistic analysis of input sentences and linguistic generation of output sentences

Machine Translation

What makes Machine Translation an important application area to study:

- historically first application area, and for at least a decade the only application area, of computational linguistics
- requires all steps relevant to linguistic analysis of input sentences and linguistic generation of output sentences
- hence, machine translation is scientifically one of the most challenging and most comprehensive tasks in computational linguistics

The Purposes of Translation

- **Information Acquisition:**
 - e.g. Gather information on scientific articles or newspapers written in a foreign language.

The Purposes of Translation

- **Information Acquisition:**

- e.g. Gather information on scientific articles or newspapers written in a foreign language.

- **Information Dissemination:**

- e.g. Translation of technical manuals, legal texts, weather reports, etc.

The Purposes of Translation

- **Information Acquisition:**

- e.g. Gather information on scientific articles or newspapers written in a foreign language.

- **Information Dissemination:**

- e.g. Translation of technical manuals, legal texts, weather reports, etc.

- **Literary Translation:**

- e.g. Translation of novels, poems, etc.

Relating Translation Purposes to MT

- **Information Acquisition:**
 - involves translation from a foreign to a native language

Relating Translation Purposes to MT

● **Information Acquisition:**

- involves translation from a foreign to a native language
- typically used by non-linguists with little or no linguistic competence in the source language

Relating Translation Purposes to MT

● Information Acquisition:

- involves translation from a foreign to a native language
- typically used by non-linguists with little or no linguistic competence in the source language
- pre-processing of the input not feasible due to lack of linguistic competence by the user in the source language

Relating Translation Purposes to MT

● Information Acquisition:

- involves translation from a foreign to a native language
- typically used by non-linguists with little or no linguistic competence in the source language
- pre-processing of the input not feasible due to lack of linguistic competence by the user in the source language
- may require special-purpose lexica

Relating Translation Purposes to MT

● Information Acquisition:

- involves translation from a foreign to a native language
- typically used by non-linguists with little or no linguistic competence in the source language
- pre-processing of the input not feasible due to lack of linguistic competence by the user in the source language
- may require special-purpose lexica
- low-quality translation is tolerable

Relating Translation Purposes to MT(2)

- **Information Dissemination:**
 - involves translation from a native to a foreign language

Relating Translation Purposes to MT(2)

● **Information Dissemination:**

- involves translation from a native to a foreign language
- pre- and post-processing of the input feasible due to linguistic competence by the translator in the source language

Relating Translation Purposes to MT(2)

● **Information Dissemination:**

- involves translation from a native to a foreign language
- pre- and post-processing of the input feasible due to linguistic competence by the translator in the source language
- may involve sublanguage with restricted vocabulary; e.g. translation of weather reports

Relating Translation Purposes to MT(2)

● **Information Dissemination:**

- involves translation from a native to a foreign language
- pre- and post-processing of the input feasible due to linguistic competence by the translator in the source language
- may involve sublanguage with restricted vocabulary; e.g. translation of weather reports
- often involves special terminologies stored in a terminology database; e.g. for translation of technical manuals

Relating Translation Purposes to MT(2)

● Information Dissemination:

- involves translation from a native to a foreign language
- pre- and post-processing of the input feasible due to linguistic competence by the translator in the source language
- may involve sublanguage with restricted vocabulary; e.g. translation of weather reports
- often involves special terminologies stored in a terminology database; e.g. for translation of technical manuals
- purely human translation for such tasks can be time-consuming, inconsistent, or tedious.

Relating Translation Purposes to MT(3)

- **Literary Translation**

- requires stylistic elegance, often involves metaphorical and metonymic language

Relating Translation Purposes to MT(3)

● **Literary Translation**

- requires stylistic elegance, often involves metaphorical and metonymic language
- abundance of highly-trained human translators

Relating Translation Purposes to MT(3)

● **Literary Translation**

- requires stylistic elegance, often involves metaphorical and metonymic language
- abundance of highly-trained human translators
- task rarely performed by machine translation