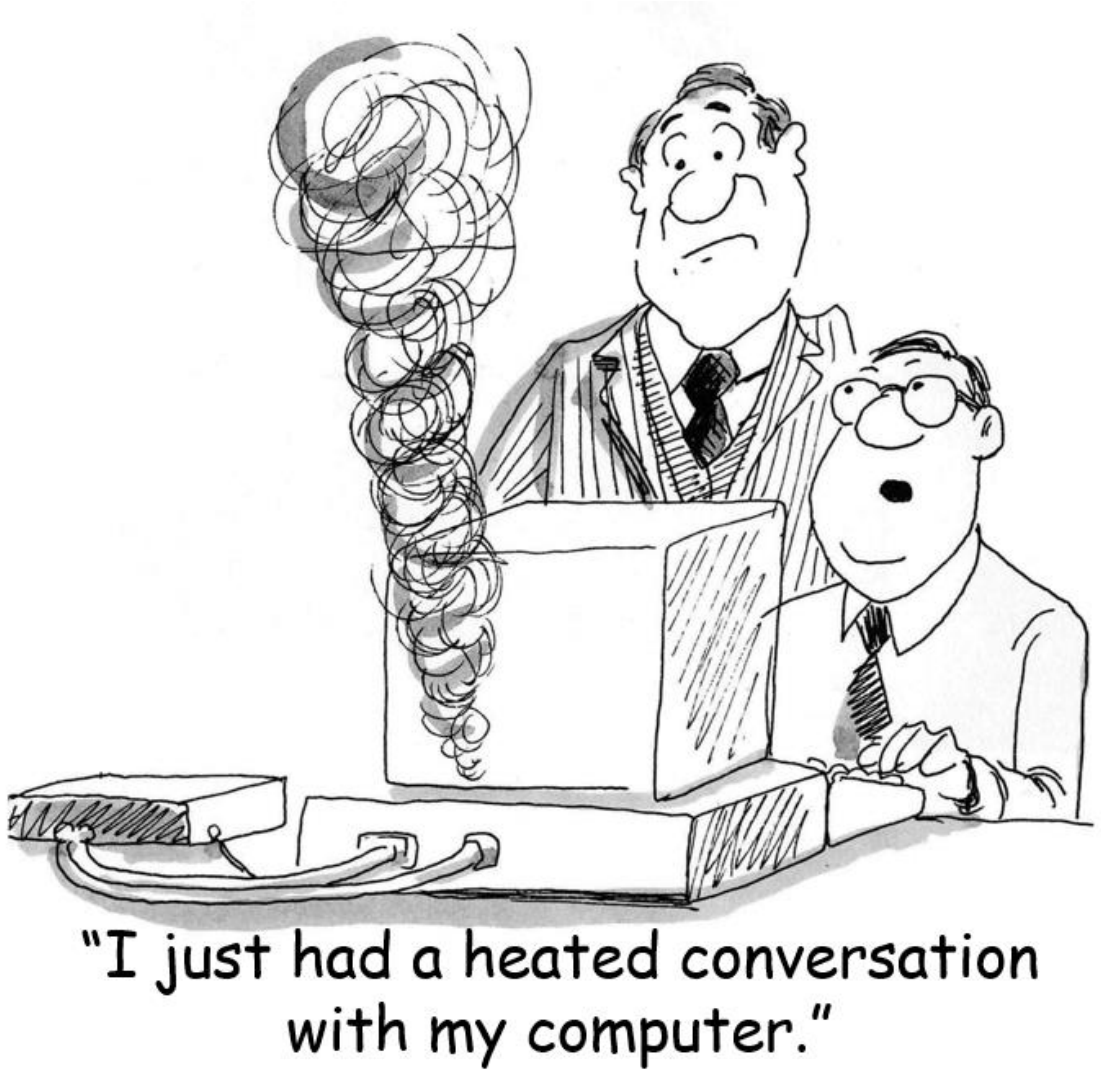
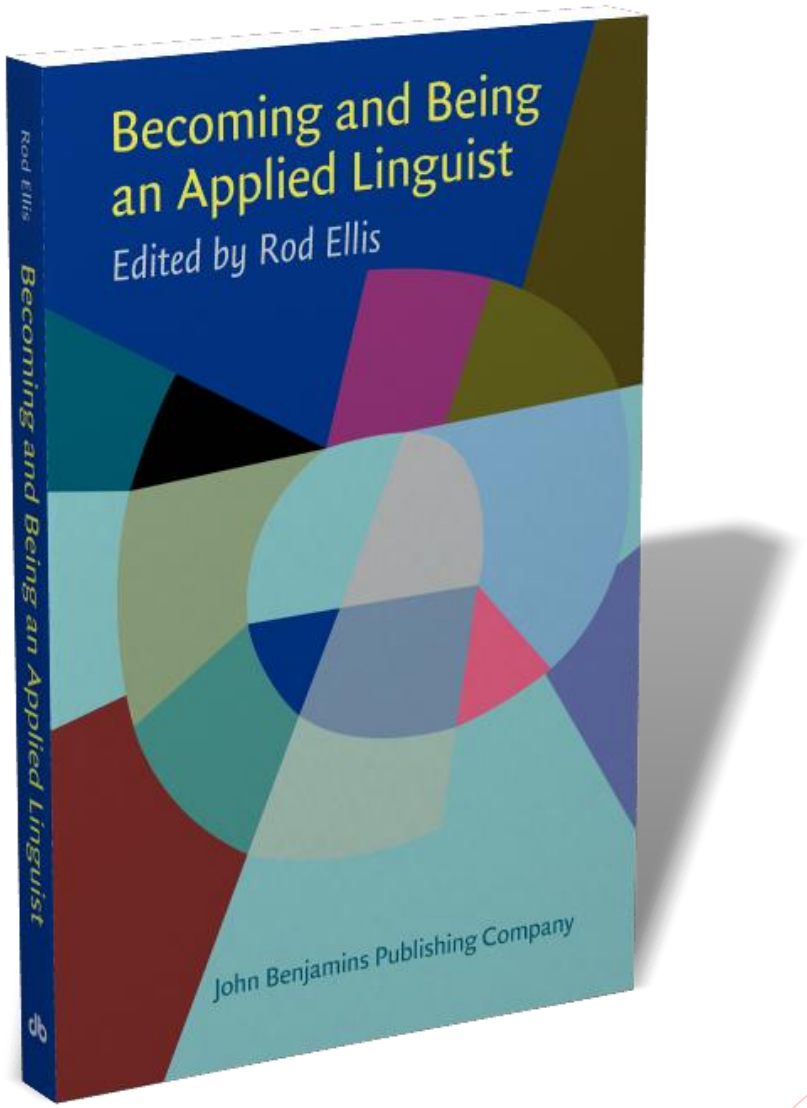
Products of Computational Linguistics

Lecture 4

Why do we need Computational Linguistics?



What practical results does Applied Linguistics provide for society?



word processing; text processing; text generation;

natural dialogue creation; language understanding etc.

Computational Linguistics

Applied Linguistics Systems

Text preparation

Information retrieval

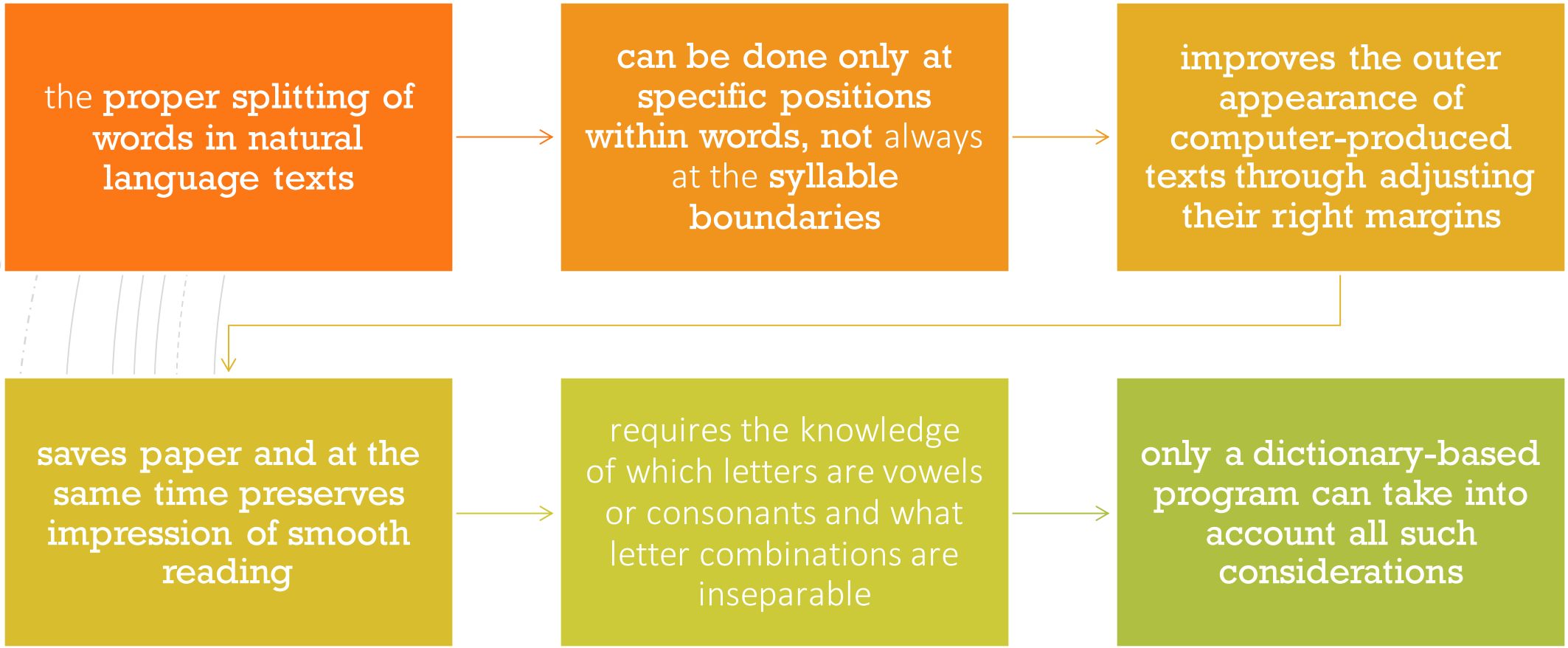
Extraction of factual data

- Automatic translation
- Natural language interfaces
- Text generation
- Natural language understanding
- Optical character recognition, speech recognition, etc.

Text Preparation (text editing)

- Automatic hyphenation
- Spell checking
- Grammar checking
- Style checking
- Referencing

Automatic Hyphenation



Spell Checking

the detection and correction of typographic and orthographic errors in the text at the level of word occurrence considered out of its context

deals with **typos, or typographic errors** as well as **spelling** errors

the user can correct this string in any preferable way— manually or with the help of the program

proposes a set of existing words, which are similar enough (in some sense) to the given corrupted word, and the user can then choose one of them as the correct version of the word, without re-typing it in the line

Grammar Checkers

detection and correction of grammatical errors by taking into account adjacent words in the sentence

grammar errors are those violating, for example, the syntactic laws or the laws related to the structure of a sentence

only a complete syntactic analysis (parsing) of a text could provide an acceptable solution for grammar checkers

a program based on a simplistic approach would toomay frequently give false alarms where there is no error in fact

since the author of the text is the only person that definitely knows what he or she meant to write, the final decision must always be left up to the user, whether to make a correction suggested by the grammar checker or to leave the text as it was

Style Checkers

the stylistic errors are those violating the laws of use of correct words and word combinations in language, in general or in a given literary genre

play a didactic and prescriptive role for authors of texts

should use a dictionary of words supplied with their usage marks, synonyms, information on proper use of prepositions, compatibility with other words, etc

should also use automatic parsing, which can detect improper syntactic construction

the larger the average length of a word, sentence or paragraph, the more difficult the text is to read, according to those simplest stylistic assessments

can only tell the user that the text is too complicated (awkward) for the chosen genre, but usually cannot give any specific suggestions as to how to improve the text

Referencing

the references from any specific word give access to the set of words semantically related to the former, or to words, which can form combinations with the former in a text

nowadays it is performed with linguistic tools of two different kinds: autonomous on-line dictionaries and built-in dictionaries of synonyms

helps the author of a text to create more correct, flexible, and idiomatic texts

only an insignificant part of all thinkable word combinations are really permitted in a language, so that the knowledge of the permitted and common combinations is a very important part of linguistic competence of any author

various complex operations are needed, such as automated reduction of the entered words to their dictionary forms, search of relevant words in the corresponding linguistic database, and displaying all of them in a form convenient to a non-linguistic user

English nouns, verbs, adjectives, and adverbs were divided into synonymy groups, or synsets. Several semantic relations were established between synsets: antonymy (reference to the "opposite" meaning), hyponymy (references to the subclasses), hyperonymy (reference to the superclass), meronymy (references to the parts), holonymy (reference to the whole), etc. Semantic links were established also between synsets of different parts of speech.

Information Retrieval (IRS)

- Information retrieval systems (IRS) are designed to search for relevant information in large documentary databases.

Information Retrieval Systems (IRS)

- "… the query is still a set of words; the system first tries to find the documents containing all of these words, then all but one, etc., and finally those containing only one of the words. Thus, the set of keywords is considered in a step-by-step transition from conjunction to disjunction of their occurrences. The results are ordered by degree of relevance, which can be measured by the number of relevant keywords found in the document. The documents containing more keywords are presented to the user first …"

IRS Characteristics

recall

- the ratio of the number of relevant documents found divided by the total number of relevant documents in the database

precision

- the ratio of the number of relevant documents divided by the total number of documents found

Automatic Abstracting

- The border between *auxiliary* and meaningful words cannot be strictly defined. Moreover, there exist many term-forming words like *system*, *device*, etc., which can seldom be used for information retrieval because their meaning is too general.

Extracting the Factual Data (Topical Summarization)

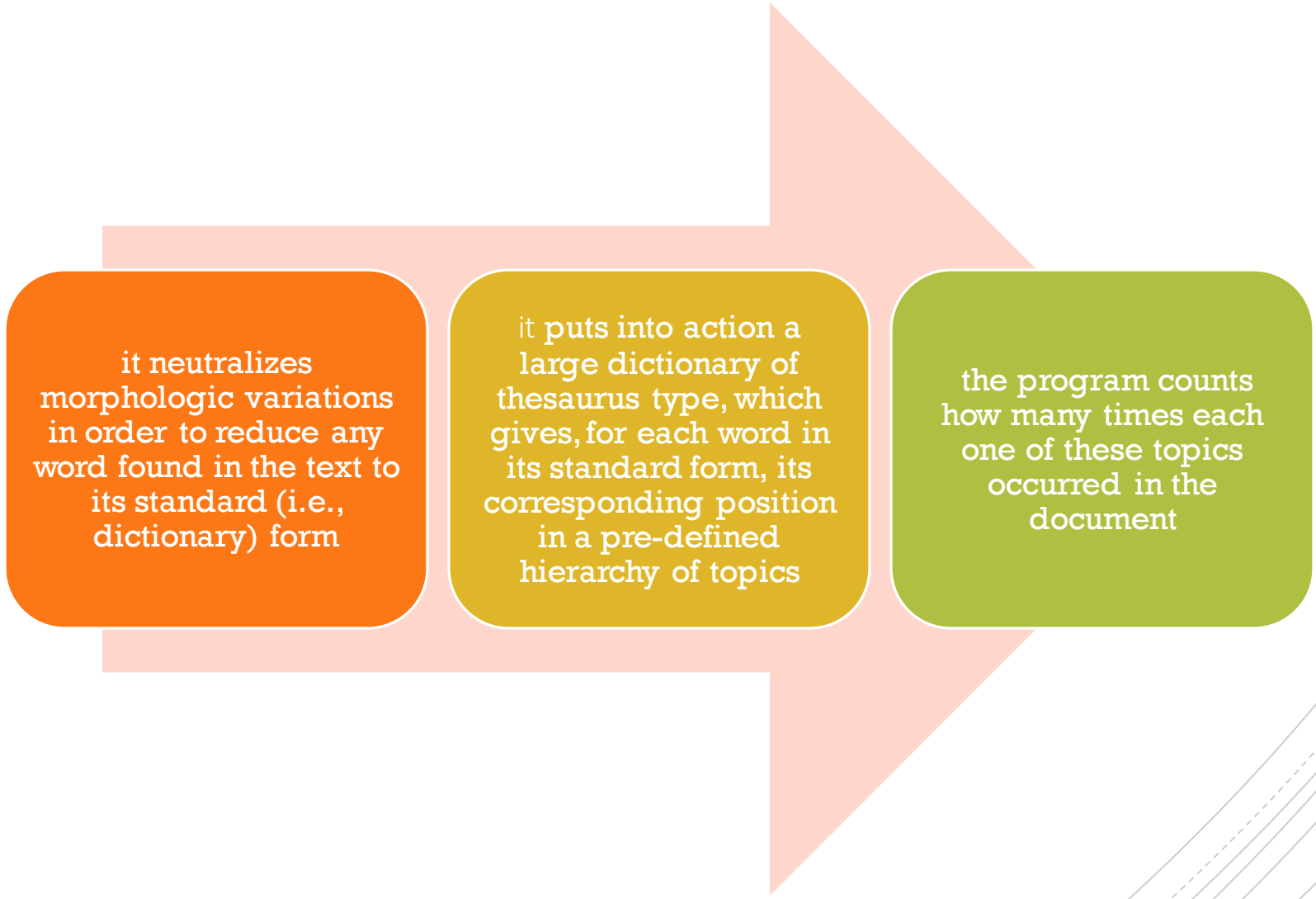
to classify the documents by their main topics

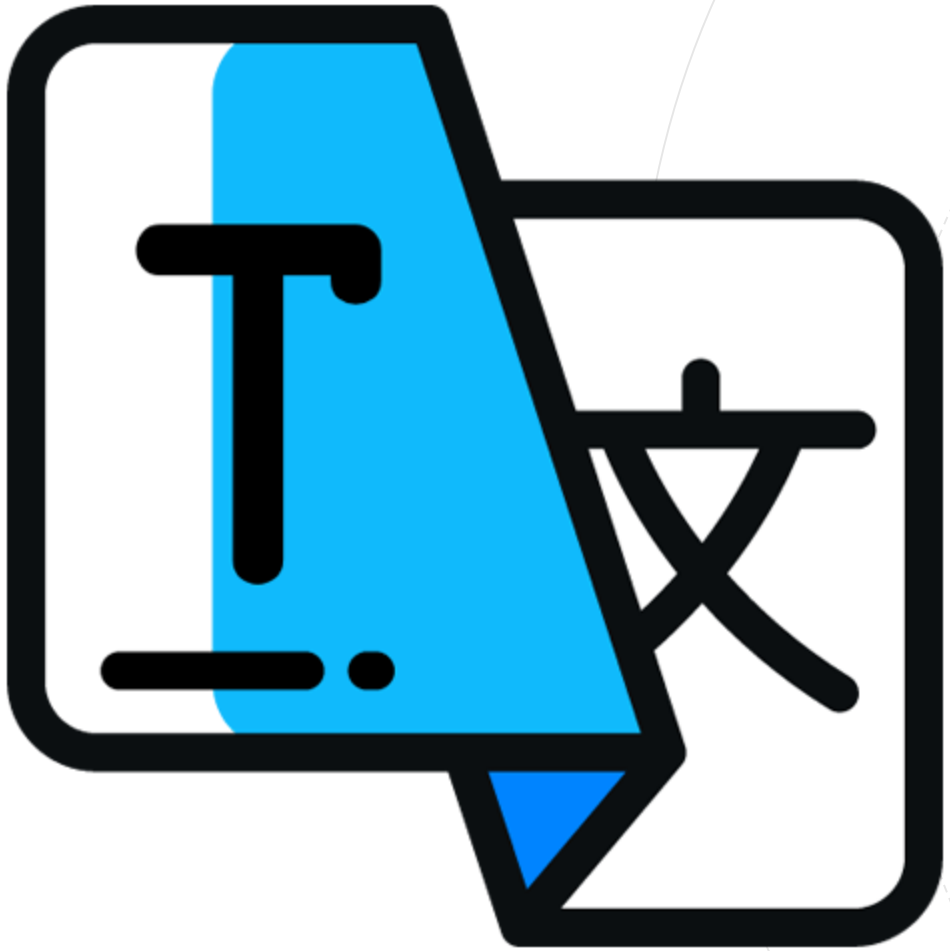
to deliver by Internet the documents on a specific subject to the user

to automatically index the documents in an IRS

to quickly orient people in a large set of documents, and for other purposes

Topical Summarization





Automatic Translation

Automatic Translation

texts could be translated word by word, so that the only problem would be to create a dictionary of pairs of words: a word in one language and its equivalent in the other

create programs which could understand deeply the meaning of an arbitrary text in the source language, record it in some universal intermediate language, and then reformulate this meaning in the target language with the greatest possible accuracy, so as neither manual preediting of the source text nor manual post-editing of the target text would be necessary

The spirit is willing, but the flesh is weak (Matt. 26:41)

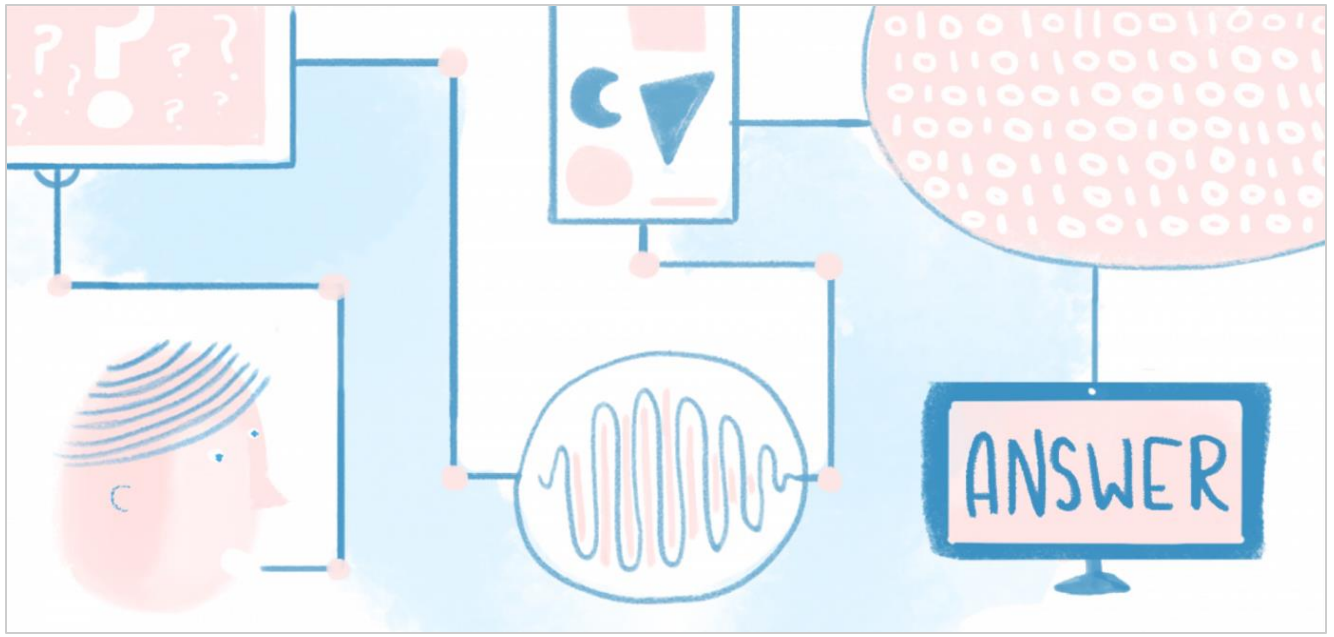
The vodka is strong, but the meat is rotten.

Automatic Translation

Automatic Translation

- "… deep linguistic analysis of the given text is necessary to make the correct choice, on the base on the meaning of the surrounding words, the text, as a whole, and perhaps some extralinguistic information …"

Natural Language Interface



- The task performed by a natural language interface to a database is to understand questions entered by a user in natural language and to provide answers—usually in natural language, but sometimes as a formatted output. Typically, the entered queries, or questions, concern some facts about data contained in a database.