Abstract

Human Language Technologies use two main paradigms: rules and statistical models. Rule-based methods are more traditional, statistical tools relying on huge textual and spoken corpora are considered rather modern. Both are, however, sequential methods. The first elements of parallel processing are available today. The (near) future’s language technology should rely on parallel language processing, as human brain works in a parallel way, as well.

Keywords: computational linguistics; brain research; language technologies

1. Human Language Technologies (HLT)

HLT is a branch of information science that deals with written and spoken natural language information. It concerns with gathering, manipulating, storing, retrieving, and classifying written and spoken natural language information. HLT applications support computer users via understanding and producing information in human languages. Natural language consists of speech and written texts; according to this, HLT consists of speech technologies and language technologies. Current fields of HLT – among others – are: information extraction, text mining, machine(-assisted) translation, intelligent search, speech recognition, text-to-speech systems, dialogue systems, etc. The key problem is that synchronous description of a language is not a trivial task: we need a good model for that. A formal model is a frame that helps to describe human languages. Human language descriptions – traditional or contemporary – are always made for another human being: supporting language acquisition of another human being, or providing information on language to another human being. Machines are different: humans and machines cannot wink to each other. The (hidden) task of our age is describing language for the computer – instead of a human.


2. The rise of the statistical paradigm in HLT

A general question arises: can machines find out what the linguist knows? To answer this question, we have to see what do the linguists know? Linguists usually work on indentifying principles encoded in spoken and written utterances on the basis of their own knowledge of languages. Everything they know are described and prepared for other human beings. This information is not enough for a computer, because it does not have the background knowledge of 5-20-50-90 years based on “living as human being”. Findings of the last two decades show that a collection of original spoken and written materials contain almost everything about language that may be needed to understand their content. Texts are available for anyone and even for a computer, without the help of linguists, so the consequence is to omit linguists and provide machines with the original utterances themselves (texts and spoken signals). This approach lead to the
wide usage of statistical approaches: noisy channel model, naive Bayes theorem and other statistical technologies were used first in modeling speech understanding, and then borrowed by machine translation researchers. What is needed here is a huge amount of source (input) and target (output) language “products” made by humans and a tool (called decoder) that is able to calculate output for any input. Today anyone can find open source tools on the web for this purpose, but unfortunately, there is still not enough data in certain areas. Statistical modeling of less spoken languages (with less data) provides low quality. Domain-specific applications also meet the sparse data problem, even for well-elaborated languages. Rule-based applications work well if the topic is narrower: fewer ambiguities arise, but they are language dependent, need direct human interaction, thus, they are relatively slow and expensive. Statistical approaches are language independent, need automatic processing only, thus, they are cheaper. Thus, hybridization is in the air: using beneficial features of both paradigms.

3. Psychological and neurological reality: parallel approaches instead of the sequential ones

According to researchers of psycholinguistics both rule-based and statistical methods are present in the human information processing. Rules describe competence and statistical knowledge describes performance in the Chomskyan sense. One of the possible reasons of high performance in human language processing is that humans use both systems: we use a sort of statistics: for everyday use, and rules when the sparse data problem arises. The rule-based subsystem and the statistical subsystem run parallel, not sequentially. The approaches so far have been fast, but all of them are nothing else but sequential emulations. Today, however, we do not know the right way how to realize parallelism in language technology applications. The processor development trend can easily be seen: there are more and more dual-, tri-, quad-, hexa-, octo-core chips to ones with tens or even hundreds of cores. On the other hand, human brain is difficult to understand, but what we know about it is that it does not work sequentially – in spite of our existing HLT algorithms. Results of brain research and neurology should be combined with the language technology world, that is, HLT should go closer to The Human Brain Project, which is a FET Flagship Pilot Action on understanding the way the human brain works. We have to model the functionality of the brain while “confronted” with natural language, the processes going on in the brain, both functionally and spatially, and finding its quantitative characteristics. It is going to be the future core technology for language understanding.

References