Task on computational linguistic model and data analysis

This is a data processing task for computational linguistic model. To build a computational linguistic model, it requires to give quantitative research on word frequency analysis. Each group will be given an article, and the students need to complete the following three steps:

Online version of the article: <https://www.cs.helsinki.fi/u/jilu/dataset/ComputationalLinguistics.txt>

1. Find the top-3 frequent 1 word and 2 word phrases in the article. You cannot ignore the stop word. In general this task does not require the computer programming skill to finish it. Each student in a group can process each page independently and then combine together: the task could be done as efficiently as possible. Of course, it would be faster to program or use online tool to count.
2. Use some online tool like TagCrowd to visualize the article. (e.g. https://tagcrowd.com/)
3. Use some online text analysis tool to analyze the document and make interesting finding, such as the average sentence length (words) or readability of the article. (e.g. <http://textalyser.net/> or https://voyant-tools.org/ )

**Computational linguistics**

Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.

Traditionally, computational linguistics was performed by computer scientists who had specialized in the application of computers to the processing of a natural language. Today, computational linguists often work as members of interdisciplinary teams, which can include regular linguists, experts in the target language, and computer scientists. In general, computational linguistics draws upon the involvement of linguists, computer scientists, experts in artificial intelligence, mathematicians, logicians, philosophers, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists, among others.

Computational linguistics has theoretical and applied components. Theoretical computational linguistics focuses on issues in theoretical linguistics and cognitive science, and applied computational linguistics focuses on the practical outcome of modeling human language use.

Computational linguistics is often grouped within the field of artificial intelligence, but actually was present before the development of artificial intelligence. Computational linguistics originated with efforts in the United States in the 1950s to use computers to automatically translate texts from foreign languages, particularly Russian scientific journals, into English. Since computers can make arithmetic calculations much faster and more accurately than humans, it was thought to be only a short matter of time before they could also begin to process language. Computational and quantitative methods are also used historically in attempted reconstruction of earlier forms of modern languages and subgrouping modern languages into language families. Earlier methods such as lexicostatistics and glottochronology have been proven to be premature and inaccurate. However, recent interdisciplinary studies which borrow concepts from biological studies, especially gene mapping, have proved to produce more sophisticated analytical tools and more trustful results.

When machine translation (also known as mechanical translation) failed to yield accurate translations right away, automated processing of human languages was recognized as far more complex as had originally been assumed. Computational linguistics was born as the name of the new field of study devoted to developing algorithms and software for intelligently processing language data. The term "computational linguistics" itself was first coined by David Hays, founding member of both the Association for Computational Linguistics and the International Committee on Computational Linguistics. When artificial intelligence came into existence in the 1960s, the field of computational linguistics became that sub-division of artificial intelligence dealing with human-level comprehension and production of natural languages.

In order to translate one language into another, it was observed that one had to understand the grammar of both languages, including both morphology (the grammar of word forms) and syntax (the grammar of sentence structure). In order to understand syntax, one had to also understand the semantics and the lexicon (or 'vocabulary'), and even something of the pragmatics of language use. Thus, what started as an effort to translate between languages evolved into an entire discipline devoted to understanding how to represent and process natural languages using computers.

Nowadays research within the scope of computational linguistics is done at computational linguistics departments, computational linguistics laboratories, computer science departments, and linguistics departments. Some research in the field of computational linguistics aims to create working speech or text processing systems while others aim to create a system allowing human-machine interaction. Programs meant for human-machine communication are called conversational agents.

Modern computational linguistics is often a combination of studies in computer science and programming, math, particularly statistics, language structures, and natural language processing. Combined, these fields most often lead to the development of systems that can recognize speech and perform some task based on that speech. Examples include speech recognition software, such as Apple's Siri feature, spellcheck tools, speech synthesis programs, which are often used to demonstrate pronunciation or help the disabled, and machine translation programs and websites, such as Google Translate and Word Reference.

Computational linguistics can be especially helpful in situations involving social media and the Internet. For example, filters in chatrooms or on website searches require computational linguistics. Chat operators often use filters to identify certain words or phrases and deem them inappropriate so that users cannot submit them. Another example of using filters is on websites. Schools use filters so that websites with certain keywords are blocked from children to view. There are also many programs in which parents use Parental controls to put content filters in place. Computational linguists can also develop programs that group and organize content through Social media mining. An example of this is Twitter, in which programs can group tweets by subject or keywords. Computational linguistics is also used for document retrieval and clustering. When you do an online search, documents and websites are retrieved based on the frequency of unique labels related to what you typed into a search engine. For instance, if you search "red, large, four-wheeled vehicle," with the intention of finding pictures of a red truck, the search engine will still find the information desired by matching words such as "four-wheeled" with "car".

Computational linguistics can be divided into major areas depending upon the medium of the language being processed, whether spoken or textual; and upon the task being performed, whether analyzing language (recognition) or synthesizing language (generation).

Speech recognition and speech synthesis deal with how spoken language can be understood or created using computers. Parsing and generation are sub-divisions of computational linguistics dealing respectively with taking language apart and putting it together. Machine translation remains the sub-division of computational linguistics dealing with having computers translate between languages. The possibility of automatic language translation, however, has yet to be realized and remains a notoriously hard branch of computational linguistics.

Some of the areas of research that are studied by computational linguistics include:

Computational complexity of natural language, largely modeled on automata theory, with the application of context-sensitive grammar and linearly bounded Turing machines.

Computational semantics comprises defining suitable logics for linguistic meaning representation, automatically constructing them and reasoning with them.

Computer-aided corpus linguistics, which has been used since the 1970s as a way to make detailed advances in the field of discourse analysis.

Design of parsers or chunkers for natural languages

Design of taggers like POS-taggers

Machine translation as one of the earliest and most difficult applications of computational linguistics draws on many subfields.

Simulation and study of language evolution in historical linguistics, glottochronology.