NLP Project: Paperwork

**TECHNICAL BACKGROUND**

**Natural Language Processing: The Basics**

Natural Language Processing (NLP) is a field of study which focuses on the interactions involving human languages and computers. Specifically, it aims to provide a way for computers to analyze and comprehend human languages in an intelligent and useful way. Concepts in computer science, artificial intelligence and computational linguistics are what comprises most of NLP. It differs from common word processor operations in the sense that NLP analyzes the word for its meaning (and not only for its structure). However, Natural Language Processing remains limited in terms of completely bridging the gap between interactions among people and computers for the computer still required several processes (such as text pre-processing techniques) in order for it to comprehend the message presented (a statement expressed in a human language). The term “natural language” refers to a stricter subset of the said human languages - NLP does not allow anomalies which can often occur in human languages.

NLP covers a wide array of topics including speech recognition, context reference issues, discourse planning and generation, and syntactic and semantic analysis and processing. It provides assistance in the fulfillment of several other processes such as text mining, machine translation, and automated question-answering by executing tasks like automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, topic segmentation, topic extraction, Part-Of-Speech (POS) tagging, stemming and lemmatization, etc. The term “processing” is not restricted to “interpretation” only. It pertains to a combination of both “interpretation” and “generation”, which is why systems possessing the said capability (NLP) are not limited to getting a grasp of the natural languages but are also given the ability to produce their own. A system’s skill that could count as an example of a natural language processing capability would be developing a decent conversation in pure human language. Additionally, computer systems that can convert human languages to computer languages and vice-versa are already existing. Translation programs were also made possible by NLP. Likewise, grammar and spelling checkers are also programmed following the NLP algorithm. Lastly, a computer that can read human languages (in publications such as books) is also a product of natural language processing.

**The Corpus: What it is and its Significance to NLP**

A corpus is a large body of structured texts, taken from newspaper articles, books, and even from webpages in the World Wide Web, that may either be written or spoken in one (monolingual corpus) or multiple (multilingual corpus) human languages. Nowadays, text corpora are often stored and processed electronically (using databases and the like). A corpus can either by *open* or *closed*. An open corpus does not contain all the data that may be related to a particular field. On the other hand, a closed corpus claims to have all, or nearly all, of the data regarding the said field (in other words, further data can still be added in order to gain a better understanding of a currently improving field).

Various types of corpora are often used by grammarians and lexicographers in order for them to better understand the descriptions presented by a language, and researchers (geared towards linguistics) to use them as training sets for Natural Language Processing related tasks (or for further developing current NLP tools).

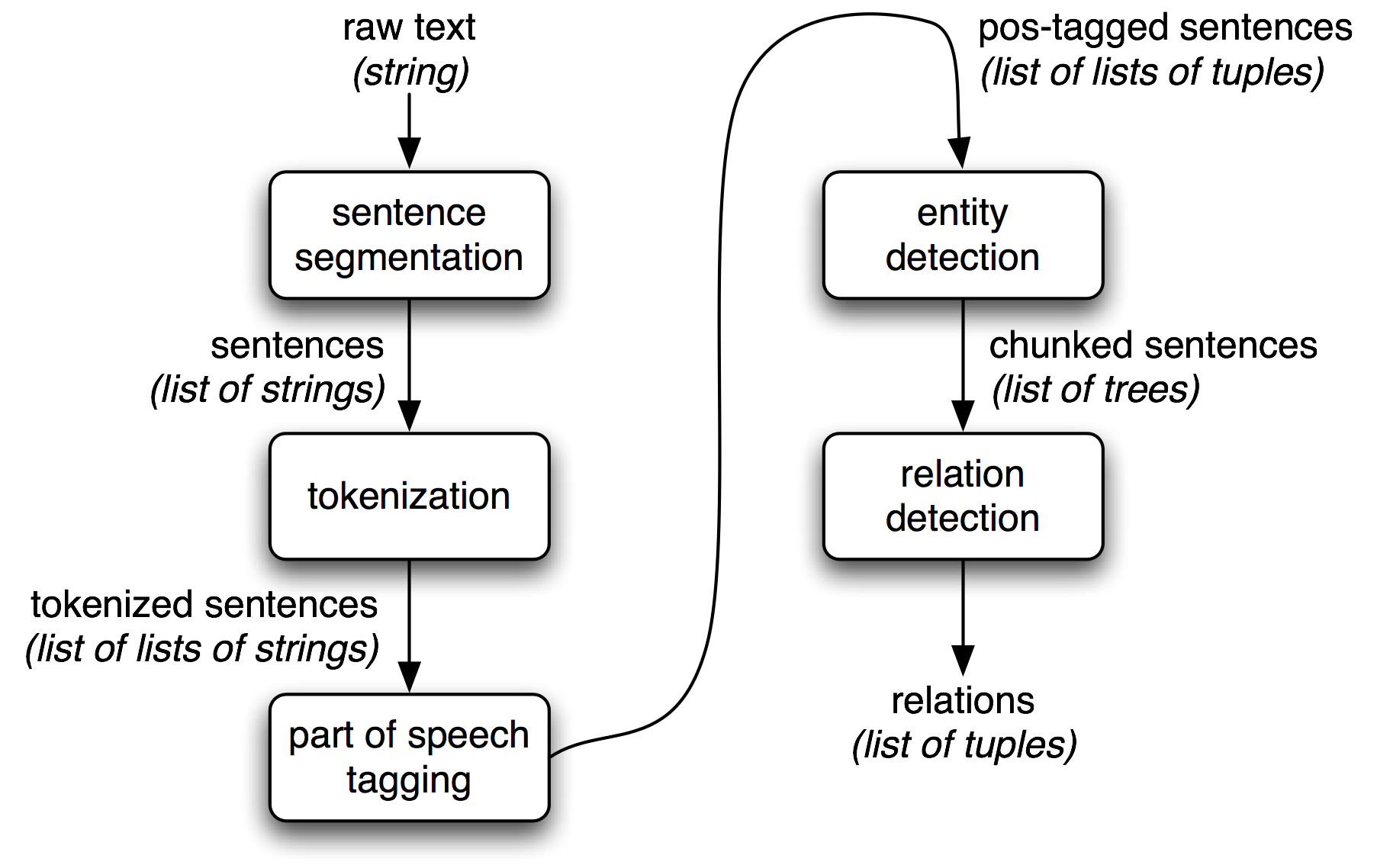
In order to make a corpus beneficial for research, data annotation must be done. Data annotation adds value to the text contained within a particular corpus. It also enables a researcher to extract the annotated data and use it in other applications (such as data that had undergone *POS Tagging* can be used by a lexicographer to find nouns, adverbs, etc. easily in that particular dataset with the help of concordance and use them in dictionaries). Automatic processing and analysis can also be assisted by annotating the dataset.

**Information Extraction, Information Retrieval, Question-Answering: Relationships and Differences**

**Information Extraction**

Information Extraction (IE) is the acquisition of high quality/structured information (where there is a regular and predictable organization of entities and relationships) from unstructured or semi-structured machine-readable documents with the use of the NLP technology. The main goal of IE is to be able to extract chief facts regarding specific events, entities or relationships. Subsequently, the said facts will be automatically inputted into a database. Once stored, this information can then be used to analyze data for trends, give a natural language summary, or simply aid in the process of online accessing (e.g. as basis for search engine-related processes).

**Information Extraction Architecture**



**Information Retrieval**

On the other hand, Information Retrieval (IR) is searching for unstructured documents (typically in the form of text) stored in large repositories (usually accessed via computer systems) which is intended to satisfy an information need. It aims to match queries with documents and measure their relevance in terms of document contents, links between documents and their labels, tags and captions, and clicks and queries. Information Retrieval differs from IE in the sense that the former only involves the extraction of relevant documents while the latter pertains to the acquisition of facts from the documents of interest.

**Question Answering**

Question Answering (QA) is basically answering a question (constructed in a natural language) based on a wide array of question types: fact, list, definition, how, why, hypothetical, semantically constrained, and cross-lingual questions.

**Machine Learning and its Relationship with Natural Language Processing**

Sources:

http://www.expertsystem.com/natural-language-processing-and-text-mining/