NLP Project: Paperwork

**TECHNICAL BACKGROUND**

**An Introduction to Text Mining**

Also known as *text analytics* or *text data mining*, text mining, as defined by Oxford, involves the practice of taking large collections of written documents into consideration in order to obtain the needed information. Its main objective is to discern relevant data in text and transform it (text) into forms that can prove useful in any data analytics process. This can be accomplished through a variety of methodologies which includes Natural Language Processing (NLP).

Text mining (assisted by NLP) can go further from what a typical search tool can do (search tools usually return results by only matching specific words included in the query); It gives profound information about the text itself (e.g. meanings, subtle messages, etc.), and reveals trends or patterns in the millions of data stored within a particular corpus (dataset) to the user.

**Natural Language Processing: The Basics**

A component of text mining, Natural Language Processing (NLP), in its simplest sense, refers to the utilization of a computer’s capability to process sentences stated in natural human languages (such as English, Japanese, Korean, etc.) instead of specialized computing languages (Java, C++, C#, etc.). However, NLP still holds significant differences between natural human languages. The term “natural language” actually refers to a stricter subset of the said human languages. They do not function the same way as how people carry-on with their everyday conversations. In fact, a writer once said that “human languages allow anomalies that natural languages cannot allow” (Steedman, p. 229). The phrase “natural language processing” can cover a wide array of topics involving speech recognition, context reference issues, discourse planning and generation, and syntactic and semantic analysis and processing. The term “processing” in the similar phrase 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. They provide assistance to people with disabilities such as blindness in terms of using personal computers. Translation programs were also made possible by NLP. There are applications in the market having the capability to translate one human language to another. 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. With respect to the interpretation side, NLP involves the following subtopics: signal processing, syntactic analysis, semantic analysis, and pragmatics. Signal processing is basically acquiring spoken words and turning them into text. Syntactic analysis, on the other hand, focuses on the grammar composition of the sentences. Semantic analysis deals with how the words and sentences refer to their corresponding worldly elements, or rather their own meaning in a particular spoken human language. Finally, the term pragmatics concerns how a specific word functions in an individual’s day-to-day conversations with one another. (Introduction to Natural Language Processing, 2006).

Public information dissemination across the Internet triggered the emergence of research areas under the field of NLP such as Information Extraction, Information Retrieval, and Question-Answering. (Mendes & Antunes).

**Software Applications**

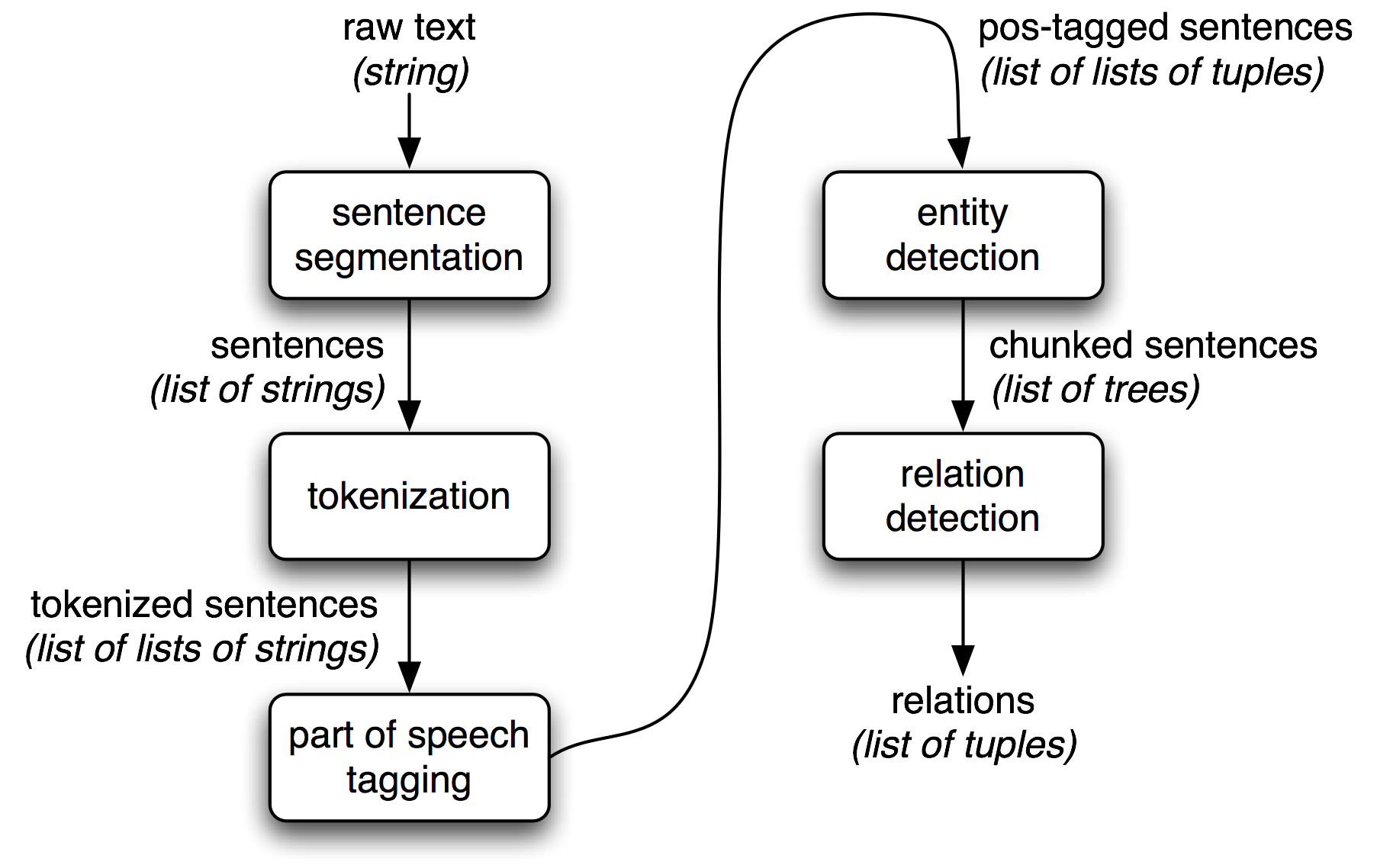
The following applications were developed based on NLP: Thinknowlogy – a grammar-based software, making use of the Natural Laws of Intelligence in grammar to be able to develop a certain intelligence through natural language in software, Bitextor – an application bearing the objective of generating translation memories with the help of multilingual Web sites as a source, TagAligner – a software used to generate translation memories from two XHTML tagged files, RelEx - is an English-language semantic dependency relationship extractor that can identify subject, object, indirect object, etc. between a group of words in a sentence, and Wintermute – a framework of applications using neural networking to learn about its host. (13 Projects Tagged "Natural Language Processing", 2015).

**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 System: How does it work?**

Sources:

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