

Introduction

Web search engines are useful tools for locating related information from keywords or questions. However, finding answers from complex query such as “*What are the relationship between garlic and bacteria?*” seems out of their abilities. While search engines based on relations can easily return the answer of that question such as:

Garlic attacks (3) bacteria and viruses
garlic can kill (2) many types of bacteria
Raw garlic kills (2) many kinds of fungus and bacteria
garlic can kill (2) certain strains of bacteria
garlic inhibited (2) the growth of bacteria

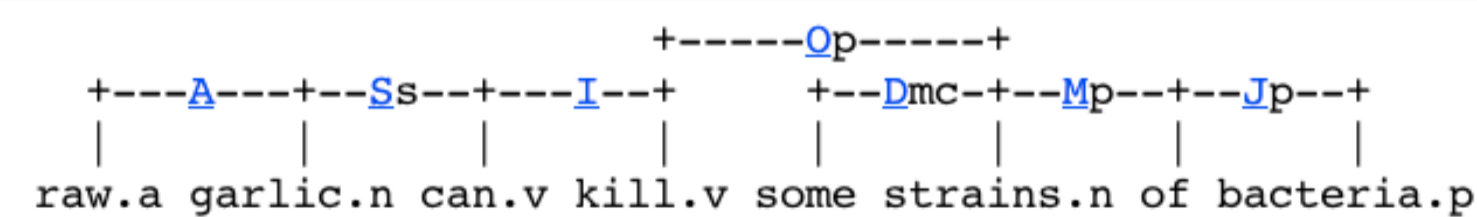
Definition of relation

A relation is considered as a tuple $t=(e_i, r_{ij}, e_j)$, where e_i and e_j are strings denoted as entities, and r_{ij} is a string denoted as the relationship between them.

As the example above, there are four tuples representing the relationship between Garlic and Bacteria:

- t1 = (Garlic, attacks, bacteria)
- t2 = (Garlic, kills, bacteria)
- t3 = (Garlic, can kill, bacteria)
- t4 = (Garlic, inhibited, bacteria)

Each component of a tuple has a certain relationship with others and link-grammar model can be used to describe that relation in each sentence. Based on this model, content of tuples will be extracted such as a tuple $t = (\text{garlic, kill, bacteria})$ from the following sentence.



RELATION EXTRACTION

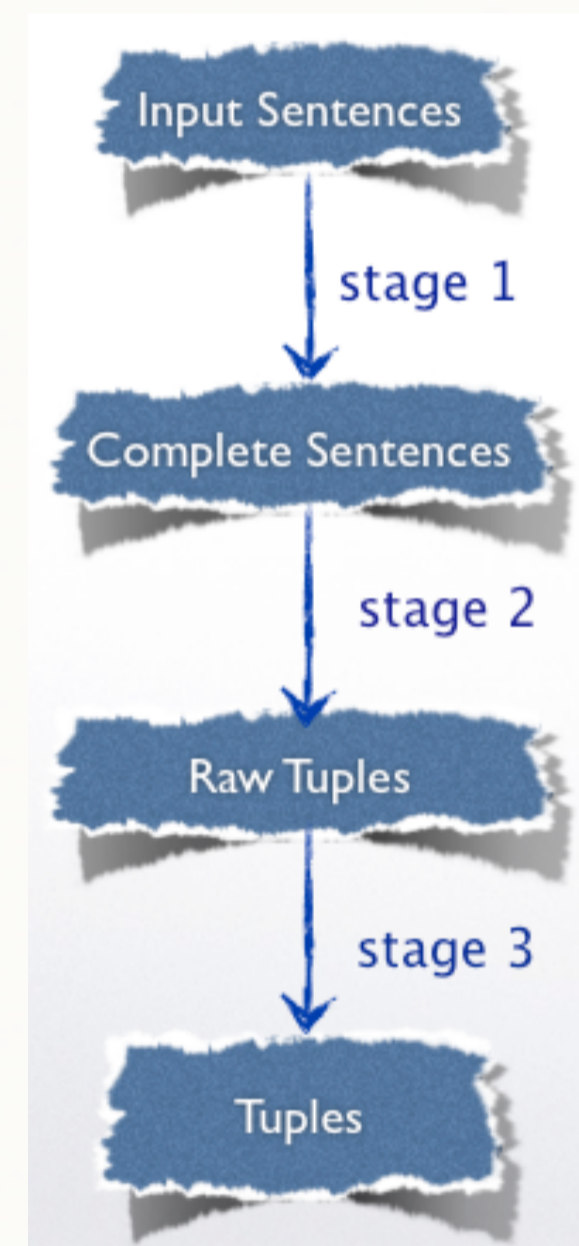
FROM WEB-SCALE UNSTRUCTURED TEXT

Dat Huynh, Master by Research, Faculty of Information Sciences and Engineering

Supervisors: Dat Tran, Wanli Ma

The purpose of this research is to find the method for retrieving a large number of relations among objects from the internet with high accuracy. Those relations can be used as input data for building semantic applications or ontology structures on the web.

Methodology



Bill Gate was born in october 28, 1995 in Seattle, Washington. He is an American business magnate, chairman of Microsoft. He is ranked consistently one of the world's wealthiest people and the wealthiest overall as of 2009.

Bill Gate was born in october 28, 1995.
 Bill Gate was born in Seattle, Washington.
 Bill Gate is an American business magnate, chairman of Microsoft.
 Bill Gate is ranked consistently one of the world's wealthiest people and the wealthiest overall as of

([Bill Gate], [was born], [in october 28, 1995]).
 ([Bill Gate], [was born], [Seattle, Washington]).
 ([Bill Gate], [is], [an American business magnate,])
 ([Bill Gate], [is], [chairman of Microsoft]).
 ([Bill Gate], [is ranked], [one of the world's wealthiest people])

([Bill, Gate], [was, born], [october 28, 1995]).
 ([Bill, Gate], [was, born], [in, Seattle, Washington]).
 ([Bill Gate], [is-a], [American, business, magnate,])
 ([Bill Gate], [is-a], [Microsoft, chairman]).
 ([Bill Gate], [is, ranked], [world, wealthiest, people])
 ([Bill Gate], [is, ranked], [2009, wealthiest, overall])

Stage 2:

EXTRACTING

The extraction procedure mainly uses deep linguistic method to build a parser for manipulating each sentence. It results main components of tuples.

For examples, with link grammar parser, we can analyze the sentence “*Bill Gates was born in october 28, 1995 in Seattle, Washington.*” and result three raw tuples: ([Bill Gates], [was born], [in october 28, 1995]), ([Bill Gates], [was born], [in Seattle, Washington]).

Stage 3:

OPTIMIZING

This stage will analyze initial tuples to restructure tuples. The new structure of tuple not only offers an easy way of manipulating data but also still keeps meaning of the original data. It is suitable to optimize tuples by identifying the identifiers focusing on the same entities and take advantages redundancy-based methods to reduce the unnecessary tuples. The output of stage 3 is passed through indexing system to support for querying tasks.

Stage 1:

NORMALIZING

The unstructured input text is divided into sentences. Some non-sense sentences will be eliminated by heuristic algorithm. The others containing referring works such as *he, she, it, they, him, her,...* will be substituted by the referred words to make clear their meanings. The output sentences will be considered as potential sentences.

Conclusion

The contribution of this research is to propose a method for extracting relations between objects on the web. The output of this research is to build a corpus of relations, providing input data for semantic applications such as:

- question-answering applications
- relationship query applications
- factoid query applications
- unnamed-item query applications.