January 2, 2015

1 Further Changes

1.1 Ignoring modifiers preceded by *to*

The earlier version ignores extractions if the path has any modifier like increase, decrease, rose etc.. The idea was to ignore sentences that only communicate change in quantities. For example, *The internet user percent has increased by 3.4 over the last year due to telecom reforms.*

However, vanilla implementation of this idea leads to false negatives. Consider: *The population of India* ***increased to*** *1.3 billion.*

Thus, presence of *to diffuses* the modifiers in current version.

1.2 Keywords modifying words on the shortest path

There were some cases in which the keywords were not on the shortest path, but instead modified words that were present on the shortest path, we consider such extractions.

1.3 Handling Redundant Extractions and Spurious pairs

For a given augmented entity phrase E[1], and a given augmented relation phrase R, Let

- N_1, N_2, ..., N_i be the numbers such that modified shortest path hypothesis leads to extractions $R_C(E, R, N_1)$, $R_C(E, R, N_2)$, ... , $R_C(E, R, N_n)$ where R_C is the relation code.
- Let $|E, N_i|$ be the shortest distance from the central entity of the entity phrase to the *ith* number.
- Let $|E, N_k| \leq |E, N_i|$ $\forall i \in [1, n]$.

Then, we ignore all the extractions except $R_C(E, R, N_k)$.

An example is the sentence *India's population sits at 1.3 billion, whereas Pakistan's population has been hovering around 1.1 billion.* Which now leads to a correct extraction:

- POP(pakistan, population, 1.1 billion)
- POP(india, population, 1.3 billion)

[1] the augmented phrases are as defined in the earlier writeup