Looping Pipelining approach to Knowledge Base Population in Optimized Implementation

Fun Chan, Ka Kan Lo, Wai Yeung Chung

Machine Reading Co.

This document provides a description of our first attempt of the system deployed to run the TAC2017 cold start track. The system is composed of various components of syntactic, semantic and knowledge base modules inter-linking to each other for the output. Some of the output is pipelined to the input of the next module, some are routed to provide additional information for the running of the particular module. The system is composed of the following modules in execution.

Syntax module: The syntax module is composed of an unsupervised syntax parser. The parser is run on the documents of the English section of the LDC corpus – to create the statistics of linking of words and phrases, and to generate parse tree for every sentences in every document. Two fast and efficient parsers built are deployed to generate parse trees to be used in the next phase.

Semantic module: The semantic modules, built on the input of syntax module, creates the pseudo logical expression of the sentences for inferences on the inter-relationships of potential entities (organization, locations …). The pseudo logical expression is used to build inference rules for various relations required to detect in the evaluation.

Semantic role labeling module: The semantic role labeling module, built on the input from parse tree and semantic module – generalized the various semantic role of predicates encountered in the corpus and helps to establish the linking of relation between entities.

Coreference module: Coreference module – links the entities and the pre-defined set of pronoun in English languages together, using both pre-defined sets of heuristics and the clues learnt from the underlying parse tree and semantic logical expression.

Regular expression module: The regular expression module detects number, URL links, and other commonly encountered multi word expression to enrich the set of entities to be used in linking.

Gazetteer module: The gazetteer module, which is built on the running of the above algorithms on other corpus before the evaluation, provides information on the set of real world entities (countries, cities, geopolitical organization and alike) and a form structured knowledge base.

Relation Extraction and tagging module: The relation extraction and tagging module extracts the targeted relation and tags the entities, using a model built from text in web corpus.

Implementation: The system is run on Linux with data structures and algorithms aggressively optimized in term of execution speed and memory footprint.