Summary of *Making Watson Fast*

Qishun Yu

CM 2043

The *Making Watson Fast* article describes the mechanism of the world-famous question answering game machine Watson and the design concept behind the scene. This article was published in 2012, years after Watson won the million prized Jeopardy! game, by E. A. Epstein, M. I. Schor, B. S. Iyer, A. Lally, E. W. Brown, and J. Cwiklik.

The authors first explain the initial Watson application, which typically took 1 to 2 hours to answer a question running with a single processor core. Then, this paper describes how they scaled Watson to use thousands of processor cores, working in parallel, to respond in less than 3 seconds on average. On the other hand, experience built on previous program also contributes significantly on the strategies like Unstructured Information Management Architecture, DeepQA analyzes, CAS, etc.

The principle structure is called CAS (Common Analysis Structure), organized the data into work units called that are passed between annotators, and specifying the flow that each CAS takes through all possible annotators. Questions pass through primary search CAS, after primary analysis, Watson gives a primary search result and candidate answers. Eventually search result, candidate answers, and supporting evidence go through final merging and ranking, the highest scored result is the answer.

With the UIMA structure, questions analysis is conducted by 100 different analytic components. Algorithm development was focused on optimizing accuracy, and experiments required the results from thousands of test questions in order to be statistically significant. Researchers focused entirely on improving accuracy and confidence measures. Question analysis is passed down to scaleout controller thus the result generated can be finally be merged and ranked.

In summary, the success of Deep Blue was the result of both incredible Analysis Structure and Information Management Architecture built by brilliant coders.