

A Comparison between the IBM Watson DeepQA and Statistical Machine Translation

Nguyen Bach
Language Technologies Institute
Carnegie Mellon University
nbach@cs.cmu.edu

Introduction

The IBM Watson DeepQA system has been recently obtained a lot of public attentions after defeating two “Jeopardy!” champions Ken Jennings and Brad Rutter. Behind the victory is a complicated QA system which has been developed intensively by about 50 researchers in 4 years. Besides QA, IBM is also a strong player in Statistical Machine Translation with many important innovations such as the IBM word alignment models, BLEU score, IBM direct translation models and so on. This essay is trying to understand and recognize the overlapping between the two areas. The technical contents of Watson are based on the Ferrucci et.al. AAAI Fall 2010 paper “*Building Watson: An Overview of the DeepQA project*”. The table below is a mapping between terminologies of the two areas.

IBM Watson DeepQA	Statistical Machine Translation
Content Acquisition	Data Collection
Question Analysis	Source-side Analysis
Hypothesis Generation	Decoding
Soft Filtering	Pruning
Hypothesis and Evidence Scoring	System Combination
Ranking	N-best List Reranking
Confidence Estimation	Confidence Score

Content Acquisition / Data Collection

In Watson, content acquisition is a step to analyze example Jeopardy’s questions and answers, label domain, and collect evidence to support answers. The process can be done manually and automatically.

In SMT, data collection is the step to analyze domain and genre of the translation task, and

collect parallel corpus, for example if the task is to build a SMT system for conversational speech then the collected parallel corpus should be in conversational style. The process is conventionally done manually.

Question/Source Analysis

Question Analysis in Watson attempts to understand what question is asking and performs the initial analyses to determine how the question should be processed by the rest of system. Question Analysis includes shallow parsing, deep parsing, semantic role labeling, co-reference analysis, relation extraction, named entities, and so on.

In SMT, particularly in the syntax-based SMT, source analysis attempts to understand the source sentence by performing syntactic parsing, dependency parsing, named entities, number translation, morphological analysis, tokenization, and so on.

Key distinctions of Watson in term of the way it process input data are Question Classification and Relation Detection.

Watson uses Question Classification to identify what types of question it has been asked, it is a math question, puzzle, factual questions or else. The question type will trigger different models and strategies in the later processing steps. Current state-of-the-art SMT systems generally do not distinguish sentence types and domains. A similar model can be built but to make it works may require additional training data and more time to tune a SMT system toward different types and domains.

Watson uses Relation Detection throughout the question analysis process to understand whether the relations in the question are SVO predicates or semantic relationships between entities. To my knowledge, SMT has not utilized relations in both modeling and decoding.

Hypothesis Generation / Decoding

Hypothesis Generation takes the Question Analysis output and searches on its database to generate potential answer. Hypothesis Generation contains two search steps which are primary and candidate answer generation. Similarly in SMT, decoding is the process to generate and search for the best translation candidate.

The nature of search algorithms in Watson and SMT is probably very different. Watson search task is an information retrieval search task with query formed from question analysis results. Watson is actually using search engines such as Indri and Lucene to obtain evidences. Meanwhile, SMT search algorithms are more felt into classic computer science search problems. Popular SMT search algorithms are the stack-decoding algorithm for phrase-based systems and the CKY bottom-up chart parsing for syntax-based systems.

Soft Filtering / Pruning

It is natural to restrict search space to optimize toward speed and resources. So, both Watson and SMT employ pruning techniques to constraint its own search space.

Watson combines lightweight analysis scores to prune unlikely answer candidates. This pruning strategies may be equivalent to model score pruning approach in SMT which quickly prunes out translation hypothesis using current decoder model scores. SMT probably has more elaborated pruning strategies than Watson, for example cube-pruning or pruning with future cost estimation.

Hypothesis and Evidence Scoring / System Combination

Watson uses more than 50 scoring components to support each answer candidate. Initially I thought 50 scoring components may simply be features in a log-linear model. However, after a carefully reading each scorer seems to be an individual complex scoring system with its own algorithms. So, the Hypothesis and Evidence Scoring component in Watson is very similar to System Combination in SMT in the sense that they both try to combine different systems and come up with only single score after this step.

Hypothesis and Evidence Scoring is probably the most important components in Watson, since it provides a framework to leverage strength and alleviate weakness of different scoring approaches. Similarly, System Combination in SMT is recently becoming a major component in the translation pipeline in which each researcher/group build their own translation system and then later on combine translation output. Both Hypothesis and Evidence Scoring and System Combination have roots from the ensemble learning methods in machine learning. The fundamental idea is the ensemble system yields better result when there is a significant diversity among the models. Therefore, in Watson case with 50 different scorers, i.e math questions scorer, puzzle question scorer, and in SMT case syntax-based systems and phrase-based systems, they all seek to promote diversity among the models they combine.

Ranking and Confidence Estimation / N-best List Reranking and Confidence Score

After all processes, both Watson and SMT need to rank their hypotheses. Ranking and confidence estimation may be in the same component, however, in *Jeopardy!* they are in two different components. The reason is confidence estimation is important to determine the game strategy which basically drives how Watson should play and win the game.

SMT has the n-best list reranking component which plays the same role as ranking component of Watson. Recently, SMT starts developing confidence score models to estimate how reliable

the 1-best translation hypothesis is. The application of confidence score in MT is to provide helpful feedbacks for users, in the way that in which situation users should trust the MT output.

Conclusions

This essay provides a comparison between Watson and SMT in some highlighted components. As an MT researcher, I am fascinated by the way Watson efficiently utilizes semantic information. So far none of the best SMT systems seems to really address the issue of deep understanding source-side sentence before going to translate it. It may be the time to spend efforts on semantic machine translation. Furthermore, preserving semantic relations via translation process is another interesting issue in which SMT may learn from Watson.