Chapter 1

Literature Survey

Question answering has transformed very much and came a long way in the last five decades. But the past decade or so has shown tremendous interest and growth in this area. Researchers have come up with various ways and have given some good models for this problem.This chapter explores the diverse umbrella of techniques harnessed for question answering systems, auxiliary techniques for IE. It also explores the various datasets and paradigms in cause-effect detection or extraction.

1.1 Major paradigms of Question answering

Now we will see two major paradigms for answering factoid questions.

1.1.1 IR-Based Factoid Question Answering

The IR-based QA relies on the large quantity of the textual data present on the web. When a question is given, the first task is to find the relevant documents or passages. Then systems like feature-based or neural are used to read these documents and find an answer which are spans from the text.

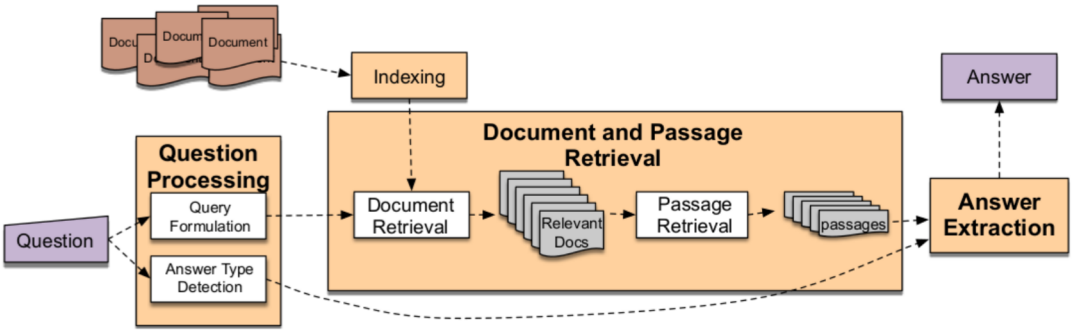

Figure 1.1: IR-Based Factoid QA

1.1.2 Knowledge Based Question Answering

In the second paradigm,knowledge-based question answering, a semantic representation of the query is built by the system. These meaning representations are then used to query databases of facts. The main idea is to answer a natural question by mapping the question to a query over a structured database.

Consider an RDF triple like the following

```
subject predicate   object
Ajay    birth-year  2003
```

This triple can be used to answer text questions like 'When was Ajay born?' or 'Who was born in 2003?'.

Question	Logical form
When was Ada Lovelace born?	`birth-year (Ada Lovelace, ?x)`
What states border Texas?	λ x.state(x) $\wedge$ borders(x,texas)
What is the largest state	argmax(λx.state(x), λx.size(x))
How many people survived the sinking of the Titanic	`(count (!fb:event.disaster.survivors fb:en.sinking_of_the_titanic))`

Figure 1.2: Logical forms for QA

1.1.3 Hybrid System(IBM's Watson)

IBM watson uses a hybrid system in which both textual datasets and structured knowledge bases are used to answer questions. One such example is the DeepQA

system. Here the main idea used by DeepA system is that it tries to find many possible answers from the knowledge bases and text data and then gives a score to each candidate answer based on knowledge sources like geospatial databases, taxonomical classification, or other textual sources.

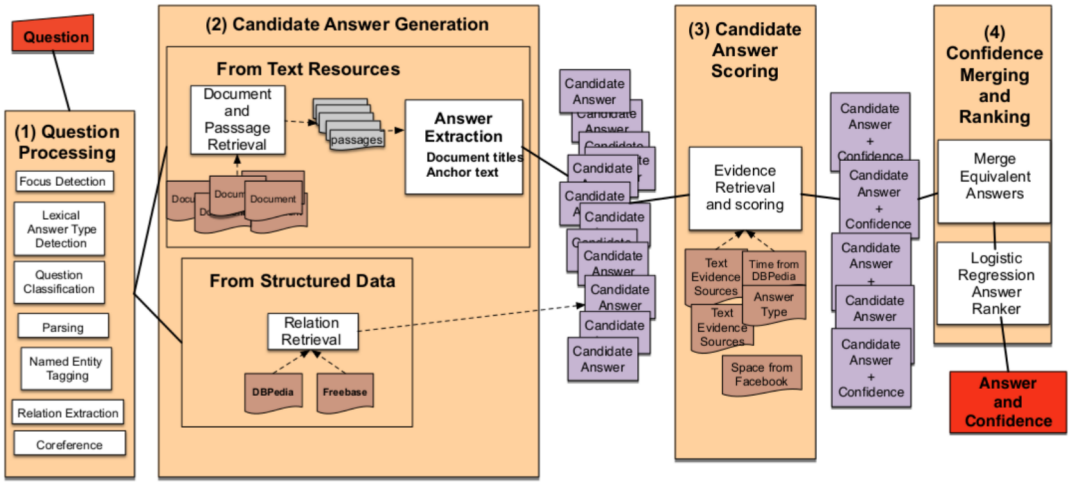

Figure 1.3: IBM Watson QA Model

1.2 Auxiliary Techniques for IE

1.2.1 Embedding: Language token representations

This section presents various helper techniques and tools used widely in conjunction with IE models to enhance the overall performance of the task under consideration.

Word2vec

In two consecutive papers published namely,[Mikolov et al., 2013a] and [Mikolov et al., 2013b], a capability of embeddings to represent words of a language as continuous vectors of much lesser dimension than that of tex- tual corpus (usually in millions) have been showcased in word analogy tasks. Motivation behind Word2Vec is to capture both syntactic and semantic rela- tionships among words through a representation that can encode words with similar meanings

closely.

Glove

[Pennington et al., 2014] proposed GloVe (Global Vectors) as an approach to represent words in a language in continuous vector space that overcame inherent problem of Skip-gram model which could not take global context into consideration. The aim of this model is to learn word representations which capture both local as well as global characteristics of the corpus.

ELMO

[Peters et al., 2018] proposes a solution in the form of deep contextualized embeddings that learn both syntactic and semantic associations along with an ability to handle underlying polysemy.

BERT

BERT [Devlin et al., 2018] combines bidirectional dynamic context from ELMo with attention-based long range context capturing capability of trans- formers [Vaswani et al., 2017].

1.2.2 Dependency Parsing

The syntactic structure of a sentence in a natural language is defined in terms of its constituent words. Dependencies among these words are represented using labeled directed edges. These labels are derived from a fixed set of grammatical relation types and hence are known as typed dependency struc- tures. Head of an entire sentence is marked with a root label. The root label often approximates the semantic association between its arguments.

Example ; James ate some cheese whilst thinking about the play.

Figure 1.4: AllenNLP Dependency Parse Tree

Demo AllenNLP[1]

1.2.3 Attention Mechanism

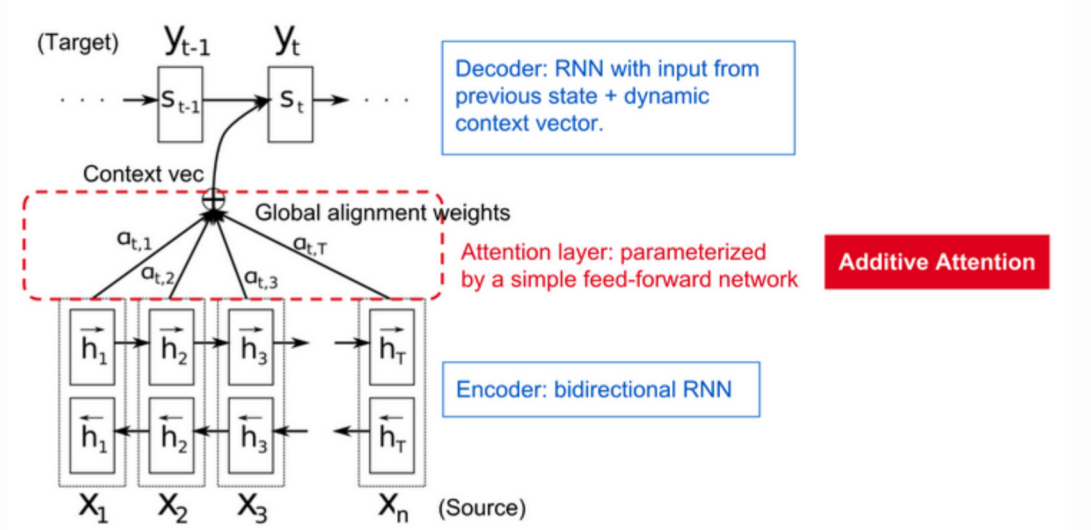

Figure 1.5: Attention Mechanism[Bahdanau et al., 2014]

With a great success achieved by [Bahdanau et al., 2014] in the domain of machine translation, attention mechanisms came to popularity. Attention helps one

[1] `https://demo.allennlp.org/dependency-parsing`

network focus on different parts of given input with different extents Olah and Carter (2016). Self-attention is one of the prime components of deep learning based relationship extraction architectures [Lin et al., 2017].

1.3 Cause-effect

Cause-effect relations exist often in text as a fundamental component of human cognition, and curating cause-effect links from literature aids in the construction of causal networks for prediction tasks. Knowledge-based, statistical machine learning (ML)-based, and deep learning-based approaches are among the existing causality extraction methodologies. Each approach has its own set of benefits and drawbacks. Expert knowledge techniques, for example, are simple to comprehend yet need significant manual subject knowledge and have limited cross-domain applicability. Natural language processing (NLP) toolkits have made statistical machine learning approaches more automated. Feature engineering, on the other hand, is time-consuming, and tool kits can lead to mistake propagation. Deep learning approaches have gotten a lot of interest from NLP academics in recent years because to their tremendous representation learning capacity and quick rise in computer resources.

1.3.1 Datasets

Data, as we all know, is the foundation of every experiment. There are several datasets that have been used in the past to evaluate CE models. We discuss four datasets from the general domain and two datasets from the biomedical area in this part, and characterise them in terms of causality sizes, sources, and accessible conditions.

- **SemEval-2007 Task 4**: SemEval (Semantic Evaluation), the 4th edition of the semantic evaluation event, is a component of it. This job gives you a dataset to use to classify semantic relationships between two nominals. The organisers divided the Cause-Effect examples into 140 training with

52 percent positive data and 80 test with 51 percent positive data within the set of seven relations.

- **SemEval-2010 Task 8**: Unlike its predecessor, SemEval-2007 Task 4, which had a separate binary-labeled dataset for each type of connection, this is a multi-classification task in which each sample's relation label is one of nine types of relations. There are 1,003 training examples with 13 percent positive data and 328 test examples with 12 percent positive data among the 10,717 annotated examples. The dataset's primary drawbacks are the limited sample size and unbalanced condition.

- **PDTB 2.0**: The penndiscourse treebank (PDTB) dataset's second release is the world's biggest annotated corpus of dis-course relations. It contains 72,135 non-causal and 9,190 causal examples drawn from 2,312 stories in the Wall Street Journal (WSJ). In particular, the dataset contains a form of implicit connective known as AltLex (Alternative lexicalization), which is an open class of markers with an unlimited potential. However, the authors save PDTB in a convoluted manner, requiring researchers to utilise tools to transform it into more user-friendly formats. Meanwhile, due of the unmarked entities, this corpus is inaccessible to methods that need entity information.

- **TACRED**: The Text Analysis Conference (TAC), like SemEval, is a series of NLP research assessment seminars. The TAC Relation Extraction Dataset (TACRED) comprises 106,264 news wire and onlinetext items gathered from the TACKBP challenge1 between 2009 and 2014. Person- and organization-oriented related types are indicated in the sentences. The primary drawback of TACRED is the minimal amount of examples supplied for the CE job, with just 269cause of death instances.

1.3.2 Knowledge-based approaches

Pattern-based methods and rule-based approaches are two types of knowledge-based CE systems. Some pattern-based systems employ pre-defined graphical patterns or phrases to convey language patterns (e.g.,thanks to, because,lead to). Patterns, on the other hand, can be discovered via sentence structure analysis, such as lexico-semantic and syntactic analysis. These structural analyses result in improved performance as well as the ability to derive implicit causality. Some rule-based systems, like pattern-based methods, use a collection of patterns or templates to explicitly detect potential causation. To investigate causality, some rule-based systems apply a collection of processes or heuristic algorithms to the syntactic structure of phrases.

Patterns are used by Garcia et al. [Garcia, 1997] and Khoo et al. [Khoo et al., 2000] to find explicitly expressed causal linkages inside a single sentence. COATIS, a technique developed by Garcia et al., extracts causation from French texts using lexico-syntactic patterns and 23 explicit causal verbs such as pro-voke, disturb, consequence, and lead to. Khoo et al. [Khoo et al., 2000] present a method for determining causation from medical textual datasets. As language hints, they employ medical-specific cause information, such as common causal phrases for depression, schizophrenia, AIDS, and heart disease.

1.3.3 Statistical machine learning-based approaches

Knowledge-based techniques require more manual preset patterns than statistical machine learning-based approaches. They usually use third-party NLP tools (e.g., Stanford CoreNLP, Spacy, Stanza) to generate a set of features for large amounts of labelled data, and then perform classification using machine learning algorithms (e.g., support vector machine (SVM), maximum entropy (ME), naive bayes (NB), and logistic regression (LG)). We'll go through how statistical machine learning techniques are employed in CE systems in the next paragraph.

Girju[Girju and Moldovan, 2002] presents a methodology for detecting causal

links in a QA system a year later. This model focuses on the most common explicit intra-sentential causality patterns, NP1, verb, NPs¿, where the verb is a simple causative, and then uses a decision tree to confirm that those patterns correspond to causation (DT). Blanco et al.[Blanco et al., 2008] only find causality in the form of VerbPhrase, relator, Cause¿, where relator is one of the following: because, since, after, or as.

The technique of [Lin et al., 2009] uses four types of characteristics, including production rules, dependency rules, word pairs, and contextual, with a ME to detect implicit discourse relations in PDTB in order to overcome the problem of a lack of explicit key-words.

1.3.4 Deep learning-based approaches

Deep learning algorithms map words and characteristics into low-dimensional dense vectors, which helps relieve the feature sparsity problem, as opposed to knowledge-based and statistical ML models. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory are the most common deep learning models (LSTM). Furthermore, in deep learning models, using attention mechanisms to selectively concentrate on significant items while ignoring others makes deep learning models more successful. Unsupervised pre-training language models (PTMs), such as BERT, that yield contextualised embeddings for each token, increase performance on many NLP tasks afterwards.

On the SemEval 2010 task 8 corpus, Xu et al. [Xu et al., 2015] used LSTM to learn relation representations along the words on the shortest dependency path (SDP) and achieved an F-score of 84 percent. Wang et al. [Wang et al., 2016] propose a CNN with multi-level attention mechanisms to collect entity- and relation-specific information. To extract relationships in TACRED, Zhang et al.[Zhang et al., 2018] propose a dependency tree-based GCN model.

To extract causality, Ponti and Korhonen [Ponti and Korhonen, 2017] build a feedforward neural network (FNN) model. As an enhanced feature set, they in-

tegrate positional and event-related characteristics with the fundamental lexical feature. Man et al.[Bowman et al., 2015], Guo et al. [Gupta et al., 2018], and Jin et al. [Jin et al., 2020] employ LSTM to capture long-dependency causality, taking use of the fact that it can handle the problem of learning long-range dependencies from a series of words. Jin et al. [Jin et al., 2020] employ CNN to extract key characteristics from input samples, then BiLSTM to extract more contextual semantic information between cause and effect. To extract causality, KPonti and Korhonen [Ponti and Korhonen, 2017] build a feedforward neural network (FNN) model.