Recent efforts towards Machine Reading Comprehension: A Literature Review

María Fernanda Mora Alba
Computer Science Department
Instituto Tecnológico Autónomo de México
México, Distrito Federal
maria.mora@itam.mx

Related Work

Problem definition

Natural language understanding (NLU) is considered a subfield of natural language processing (NLP) that deals with machine reading comprehension. Since the 60's there have been efforts towards the advancement in this task. Most of these previous systems share similar components: a lexicon of the language and a parser and grammar rules to break sentences into an internal representation. Building a rich lexicon with an adequate ontology like Wordnet required many person-years [Miller et al.1990].

Regardless of these efforts progress in this area has been slow and NLU is considered an AI-hard problem. Hence, recently, machine learning approaches to reading comprehension have been taken. One long-term goal of machine learning research is to produce methods that are applicable to reasoning and natural language, in particular building an intelligent dialogue agent [Weston et al.2015]. In fact, besides the theoretical interest, there is considerable commercial interest in machine reading comprehension because of its application to gathering, mining and analysis of large scale unstructured data such as news and voice. IBM's Watson and Apple's Siri are examples of this trend.

Question-answering is another subfield of NLP and Information Retrieval that is related to machine comprehension because it aims to build systems that automatically answer questions posed by humans in natural language. Currently, Machine reading comprehension lies in the intersection of natural language processing (NLP), machine learning (ML) and question-answering (QA).

Typically, machine reading comprehension can be defined as the ability of a machine system to read and understand natural language documents at a sufficient level where it is capable of answering questions based on the original text. Unlike tasks like dialogue or summarization, this evaluation approach (i.e. based on questions and answers) is easier to grade and thus makes it an appealing research avenue [Weston et al.2015]. Although there is an evaluation criteria, teaching machines to really understand natural language and be able to correctly answer questions remains a puzzling challenge.

The challenge can be divided into two sequential and interrelated sub-problems. First, machine reading comprehension models are hard on its own: it is very difficult to structure and build models flexible enough to learn to exploit document structure. Humans do not understand language in isolation, the *context* in which sentences and words are understood plays an important role in human comprehension, so the question is if machines can exploit this context -and how-to make predictions about natural language [Hill et al.2015]. Second, these models are trained and evaluated using the available data and assessing how good the proposed model answers to questions about a given text. The dilemma is that for this evaluation to be meaningful, adequate training and testing datasets are essential. A clear example of this limitation is that supervised machine learning approaches have largely been absent due to both the lack of large scale training dataset and the difficulty in structuring statistical models flexible enough to learn to exploit document structure [Hermann et al.2015].

Accordingly, high-quality, real and large datasets play an crucial role to make progress on machine comprehension. The complication is that already existent datasets suffer from shortcomings such as being too small but realistic or large but semi-synthetic, thus not realistic.

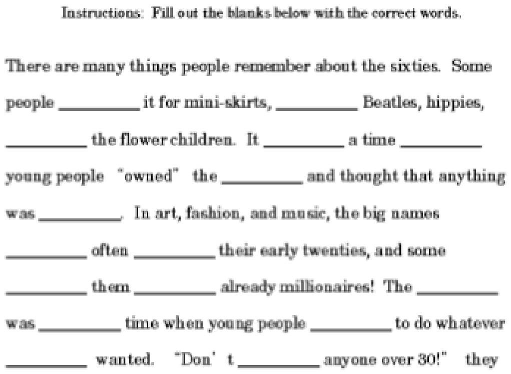
In response to this, recently there has been two major efforts towards the advance of machine reading comprehension: creation of **datasets** and development of **models**.

Datasets

We can classify the reading comprehension datasets by how they generate the questions: cloze style or human annotators.

Cloze style.- While obtaining supervised natural language reading comprehension data has proved difficult, some researchers have explored generating synthetic narratives and queries [Hermann et al.2015]. Such approaches allow the generation of almost unlimited amounts of supervised data and enable researchers to isolate the performance of their algorithms on individual simulated phenomena.

The most representative example of this trend are the cloze style datasets. These datasets are created by removing certain words from chunks of texts. Figure 1 shows an illustration of a text in cloze style.



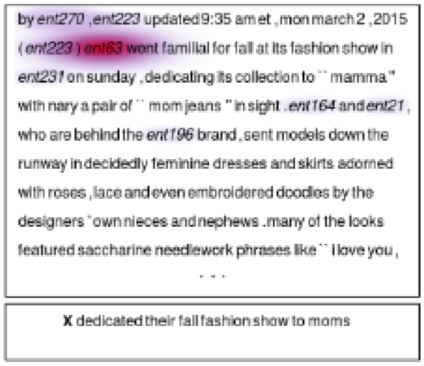
Instructions: Fill out the blanks below with the correct words.

There are many things people remember about the sixties. Some
people ________ it for mini-skirts, ________ Beatles, hippies,
________ the flower children. It ________ a time ________
young people "owned" the ________ and thought that anything
was ________. In art, fashion, and music, the big names
________ often ________ their early twenties, and some
________ them ________ already millionaires! The ________
was ________ time when young people ________ to do whatever
________ wanted. "Don' t ________ anyone over 30!" they

Figure 1: Example of a text in cloze style from [Hadley and Naaykens1999].

The reading comprehension ability is assessed by how well the model is able to replace the missing words. The advantage of these types of datasets is that they can be retrieved or generated automatically, thus they can run very large. Undoubtedly, they have accelerated the research of machine comprehension [Cui et al.2016] . For example, [Hermann et al.2015] created a supervised dataset of this type collecting roughly one million short summaries of the news articles from CNN and Daily Mail: c is a context document, q is a query relating to that document, and a the answer to that query. The summary and paraphrase sentences, with their associated documents, can be readily converted to context–query–answer triples using simple entity detection and anonymisation algorithm [Hermann et al.2015]. The entities were anonymised so that models cannot apply knowledge that is not related to the content of the article (e.g. pointing to an entity as a candidate answer simply because it appears very frequently in the corpus) and therefore favoring text understanding.

The following Figure 2 shows an example of how the cloze-style questions look like: authors constructed the corpus of document-query-answer triples by replacing one entity at a time with a placeholder.



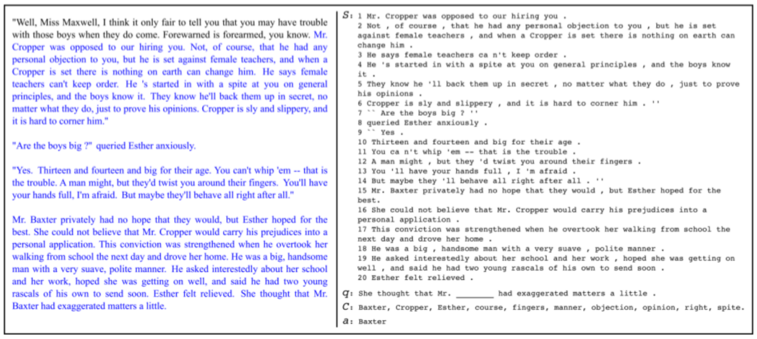
by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015
(*ent223*) *ent63* went familial for fall at its fashion show in
ent231 on sunday , dedicating its collection to `` mamma "
with nary a pair of `` mom jeans " in sight . *ent164* and *ent21* ,
who are behind the *ent196* brand , sent models down the
runway in decidedly feminine dresses and skirts adorned
with roses , lace and even embroidered doodles by the
designers ' own nieces and nephews . many of the looks
featured saccharine needlework phrases like `` i love you ,
. . .

X dedicated their fall fashion show to moms

Figure 2: Example of the cloze-style questions in [Hermann et al.2015]

Another example is the Children's Book Test dataset by [Hill et al.2015] which was generated automatically from books that are freely available in the Project Gutenberg[1]. According to the authors, using children's books guarantees a clear narrative structure, which can make the role of context more preeminent. The machine comprehension test is performed using 20 sentences from a children's story (the context S) that are used to predict the missing word (the answer a) in the 21th sentence (the question q) from a pool of 10 candidate answers (the candidates C).

The following Figure 3 shows an example question: a cloze style question q (right) created from a book passage S (left, in blue), the candidate answers C are both entities and common nouns, and the answer a, which is Baxter.



"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

```
S: 1 Mr. Cropper was opposed to our hiring you .
   2 Not , of course , that he had any personal objection to you , but he is set
   against female teachers , and when a Cropper is set there is nothing on earth can
   change him .
   3 He says female teachers ca n't keep order .
   4 He 's started in with a spite at you on general principles , and the boys know
   it .
   5 They know he 'll back them up in secret , no matter what they do , just to prove
   his opinions .
   6 Cropper is sly and slippery , and it is hard to corner him . ''
   7 `` Are the boys big ? ''
   8 queried Esther anxiously .
   9 `` Yes .
   10 Thirteen and fourteen and big for their age .
   11 You ca n't whip 'em -- that is the trouble .
   12 A man might , but they 'd twist you around their fingers .
   13 You 'll have your hands full , I 'm afraid .
   14 But maybe they 'll behave all right after all . ''
   15 Mr. Baxter privately had no hope that they would , but Esther hoped for the
   best.
   16 She could not believe that Mr. Cropper would carry his prejudices into a
   personal application .
   17 This conviction was strengthened when he overtook her walking from school the
   next day and drove her home .
   18 He was a big , handsome man with a very suave , polite manner .
   19 He asked interestedly about her school and her work , hoped she was getting on
   well , and said he had two young rascals of his own to send soon .
   20 Esther felt relieved .
q: She thought that Mr. ________ had exaggerated matters a little .
C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.
a: Baxter
```

Figure 3: Example of the cloze-style questions in [Hill et al.2015]

A difference between both datasets is that [Hermann et al.2015] focuses more on paraphrasing parts of a text, rather than making inferences and predictions from contexts as in the Children's Book Test dataset by [Hill et al.2015].

It is worthwhile to point out that cloze style datasets does not have 'real' or *factoid questions*. Factoid question answering is the most widely studied task in question answering [Wang2006]. These questions ask to provide concise facts, for example: Where is the Louvre located?, In what year did American civil war take place?, What metal has the highest melting point?, When were William Shakespeare 's twins born?, Which president was unmarried?. According to [Wang2006] the recent research trend is shifting toward more complex types of questions[2]:

- Definitional questions:

  Entity definition questions: "what is epilepsy?", "what are coral reefs?", "how do you measure earthquakes?".

  Biographical questions: "who was Galileo?", "who is Hilary Clinton?".
- List questions: "list five Communist countries", "list the female characters of "The Iliad and the Odyssey"".
- Scenario-based question answering: given a short description of a scenario, the objective is to answer ques-

[1] https://www.gutenberg.org/

[2] The example questions where taken from the pool of questions generated by [Li and Roth2002].

tions about relations between entities mentioned in the scenario.

- Why-type: ask for an explanation or reason, for example, “why is the sun yellow?”, “why is a ladybug helpful?”, “Why does the moon turn orange?”.

The importance of *factoid questions* is that there exist clearly defined and relatively uncontroversial evaluation standards for answering them: usually only one or at most a few correct answers to a given question, and the answer in most cases is a single word token or a short noun phrase, while for other types of questions the evaluation is somewhat controversial [Wang2006].

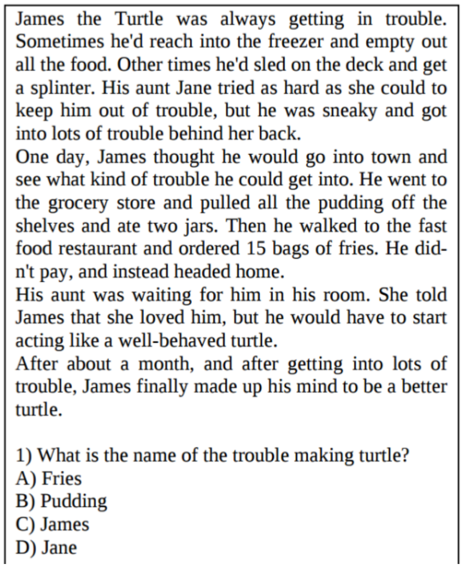
Additionally, some researchers suggested that these datasets require less high-level inference than expected. Also, the path in training and testing phase is nearly the same, making it easier for the machine to learn these patterns instead of reasoning about the meaning of the text [Cui et al.2016]. This is why this datasets are also called *synthetic* or *semi-synthetic*.

[Cui et al.2016] published two similar datasets in Chinese: the People Daily dataset and the Children’s Fairy Tale dataset. The approach is similar to [Hermann et al.2015] and [Hill et al.2015] but there is a human evaluated dataset for testing purpose. So it will be harder for the machine to answer these questions than the automatically generated ones, because the human evaluated dataset is further processed and curated and may not be accordance with the pattern of automatic questions [Cui et al.2016]. This observation leads us to the next type of datasets: *human annotated*.

Human annotated.- These datasets are created totally or partially by humans with the purpose of creating real questions. The goal is to build technology that actually understands stories and paragraphs on a meaningful level, as opposed to using information retrieval methods and the redundancy of the web or knowledge repositories (such as Wikipedia) to find the answers. The questions of these datasets are mainly *factoid questions*.

For example, [Richardson, Burges, and Renshaw2013] constructed through crowdsourcing the MCTest dataset consisting of short fictional stories, together with four associated factoid questions and four candidate answers. The fictional approach imply that the answer can be found only in the story itself. This feature allows to focus on the high-level goal of open-domain machine comprehension, instead of the previous work that focused on limited-domain datasets, or on solving a more restricted task (e.g., open-domain relation extraction). The dataset is open-domain, yet restricted to concepts and words that a 7 year old is expected to understand. The ability to perform causal or counterfactual reasoning, inference among relations, or just basic understanding of the world in which the passage is tested through multiple-choice answers, thus providing an objective metric to evaluate the future models.

The following Figure 4 shows an randomly generated sample of a story written by a worker of this experiment.



James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

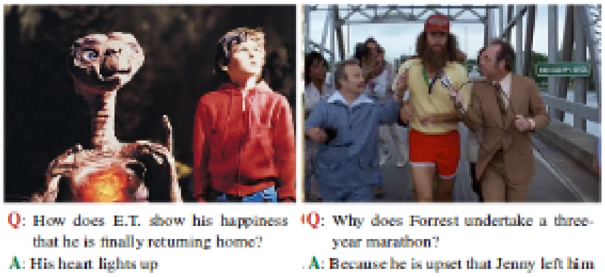
1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

Figure 4: Example of the fictional stories and questions in [Richardson, Burges, and Renshaw2013]

Although the authors claim that its approach is scalable, the dataset is totally generated by humans, thus making real scalability prohibitive. In fact the authors only generated 500 stories and 2000 questions. Another weakness of this dataset is that the stories were written to be understandable by a child in grade school, potentially preventing a model from really performing natural language comprehension.

Another human annoted dataset is the MovieQA created by [Tapaswi et al.2015] and is novel because it uses multiple sources of information such as video clips, plots, subtitles, scripts and DVS of movies for question answering. According to the authors, the dataset consists of 14,944 questions about 408 movies with high semantic diversity. The questions range from simpler *Who did What to Whom*, to *Why and How certain events occurred*. The questions have a set of five possible answers, a correct one and four deceiving answers provided by human annotators. There are also candidate answers provided.

Figure 5 shows two single frame examples of the MovieQA dataset.



Q: How does E.T. show his happiness that he is finally returning home?
A: His heart lights up

Q: Why does Forrest undertake a three-year marathon?
A: Because he is upset that Jenny left him

Figure 5: Two single frame examples from the MovieQA dataset of [Tapaswi et al.2015]

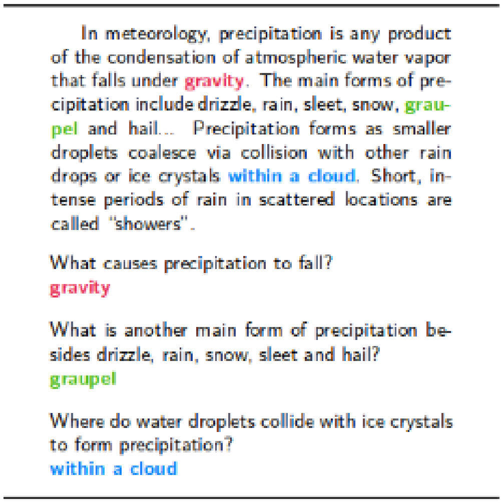
While some questions can be answered using vision or dialogs alone, most require both: vision can be used to locate the scene set by the question and semantics extracted from dialogs can be used to answer it. The goal of this dataset is semantic understanding over long temporal data [Tapaswi et al.2015]. So MovieQA poses a more difficult problem than machine reading comprehension alone. This dataset has the same scalability disadvantages than the previous one.

As we have discussed, the **cloze style** approach is scalable but synthetic, and the **human annotation** approach is more realistic but not scalable.

SQuAD.-

The Stanford Question and Answering Dataset (SQuAD)[3] [Rajpurkar et al.2016] was built in mind to overcome these deficiencies. SQuAD is formed by 100,000+ question-answer pairs based on 500+ Wikipedia articles (specifically 107,785 and 536). The questions and answers were annotated through a mechanical turk. The questions are designed to bring answers which can be defined as a *span*, or segment of the corresponding passage or context.

Figure 6 shows an example posed by [Rajpurkar et al.2016]. In order to answer the question "what causes precipitation to fall?" one might first locate the relevant part of the passage "precipitation ... falls under gravity", then reason that "under" refers to a cause and not to a location, and thus determine the correct answer: "gravity". This exemplifies how machines require both understanding of natural language and knowledge about the world to achieve reading comprehension.



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 6: Example of passage, question and answer pairs in SQuAD

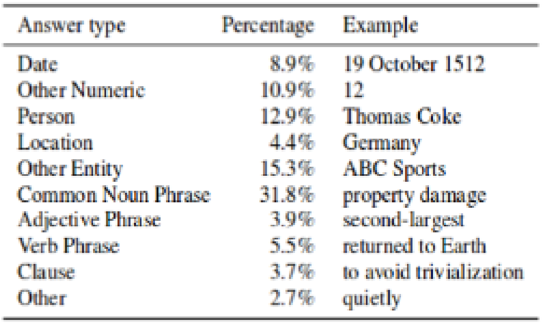
The following text shows a randomly sampled passage, question and answer from SQuAD extracted by the author of this Literature Review. It is interesting to note that the question is generated using very different syntactic symbols. For example, the words "example" and "strongly" doesn't appear in the passage. So the machine must infer that "strongly" is a synonym of "powerful" and that "churches and cathedrals" are examples of the Gothic style.

Passage: It is in the great churches and cathedrals and in a number of civic buildings that the Gothic style was expressed most powerfully, its characteristics lending themselves to appeals to the emotions, whether springing from faith or from civic pride.
Question: What is an example of where the Gothic style is expressed most strongly?
Answer: churches and cathedrals

SQuAD provides a challenging dataset for building, testing and evaluating machine comprehension models and systems for three main reasons:

1. *No candidate answers are provided:* instead of a predefined list of answer choices such as [Richardson, Burges, and Renshaw2013] and [Tapaswi et al.2015], in SQuAD all the possible spans in the passages are candidate answers thus needing to cope with a fairly large number of candidates (as many as the number of words in the given paragraph), presumably making the task harder. Regardless that [Hermann et al.2015] can be thought as also lacking question candidates, this is not the case as the answers in this dataset are enumerated entities (see Figure 2).

   [Rajpurkar et al.2016] argues that while questions with span-based answers are more constrained than the more interpretative questions found in more advanced standardized tests, they found a rich diversity of questions and answer types in SQuAD. This richness can be appreciated in Figure 7 that shows that SQuAD consists of a large number of answers beyond proper noun entities: 19.8 % is numeric, 32.6 % are proper nouns, 31.8 % are common noun phrases and 15.8% are made up of adjective phrases, verb phrases, clauses and other types. This is contrary to [Hermann et al.2015], where all the answers are named entities.



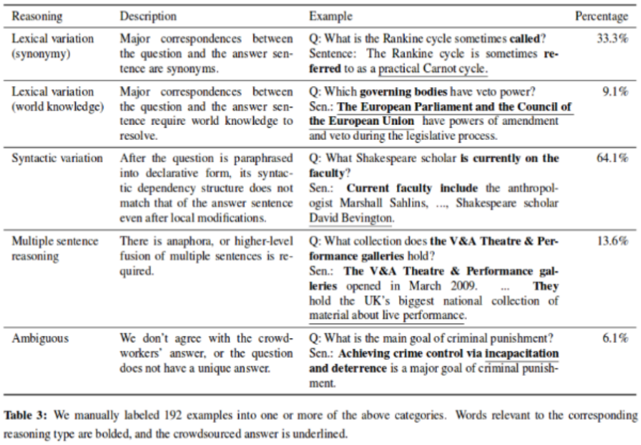
Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Figure 7: Answer types found in SQuAD

To quantify the richness -and difficulty-, the authors performed two tasks: 1. stratified the questions by difficulty

[3] https://stanford-qa.com/

and 2. developed techniques based on distances in dependency trees. To perform the first, they sampled 192 examples, and then manually labeled the examples with the categories shown in Figure 8. The results show that all examples have some sort of lexical or syntactic divergence between the question and the answer in the passage.



Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Table 3: We manually labeled 192 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are bolded, and the crowdsourced answer is underlined.

Figure 8: Assessment of question difficulty in SQuAD

To perform the second, they developed an automatic method based on distances in dependency trees to quantify the *similarity* between question and the sentences containing the answer by measuring syntactic divergence between them. The histogram of this metric can be seen in Figure 9. Intuitively, the higher the syntactic divergence between question and answer, the more difficult the question.

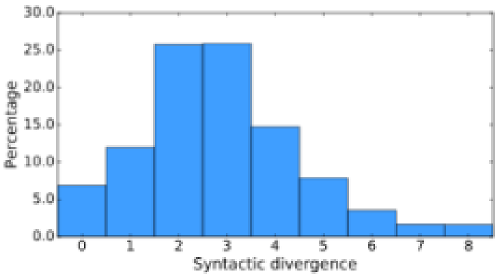


Figure 9: Histogram of syntactic divergence

Both tasks also allow to stratify the dataset, a key component in the modeling phase. The span constraint also comes with the additional benefit that span-based answers are easier to evaluate than free-form answers. [Richardson, Burges, and Renshaw2013] argued that having objective metrics for the evaluation phase is a crucial step towards machine reading comprehension.

2. *A correct answer to a question can be any sequence of tokens from the given text:* instead of having a single token as an answer such as in close style datasets or in [Richardson, Burges, and Renshaw2013] human annotated dataset, in SQuAD the answers can be composed of sequences of tokens. For example, in Figure 6 above, the last question has a multiple token answer (i.e. "within a cloud"). Questions whose answers span multiple tokens are more challenging than those with single-token answers [Wang and Jiang2016]. These sequences of tokens can be quite similar, thus making more difficult the recognition of the correct answer. The evaluation of the models is performed with this criteria, so it is more difficult to achieve a good performance. It is worthwhile to mention that another difference between SQuAD and the cloze style datasets is that in SQUaD the answers are *entailed* by the passage, while in the cloze style datasets the answers are merely *suggested*.

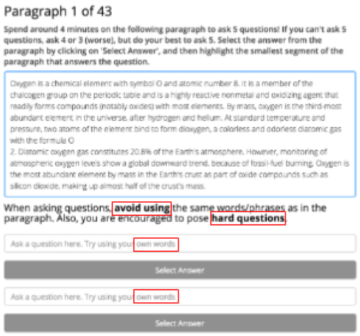
3. *Questions and answers in SQuAD were created by humans, hence they are more realistic*: unlike other datasets such as [Hermann et al.2015] and [Hill et al.2015], whose questions and answers were created automatically and synthetically, SQuAD's were created by humans through crowdsourcing.

Moreover, ¿99% of the questions are *factoid questions*, thus allowing relatively uncontroversial evaluation standards for answering them.

In order to obtain a high-quality dataset, the collection process was divided in three stages:

- Curating passages: to retrieve high-quality articles, [Rajpurkar et al.2016] used Project Nayuki's Wikipedia's internal PageRanks to obtain the top 10000 articles of English Wikipedia to we sample at random 536 articles uniformly. From each of these articles, they extracted individual paragraphs larger than 500 characters obtaining 23,215 paragraphs covering a wide range of topics, from musical celebrities to abstract concepts.

- Crowdsourcing question-answers on those passages: in this phase they employed crowdworkers to create questions through a Daemo platform with Amazon Mechanical Turk as its backend. Candidates were required to have a 97% HIT acceptance rate, a minimum of 1000 HITs, and lived in the United States or Canada. They were asked to ask and answer in 4 minutes up to 5 questions on the content of that paragraph and paid 9 USD per hour. The questions had to be entered in a text field, and the answers had to be highlighted in the paragraph as shown in Figure 10.



Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O
2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words
Select Answer
Ask a question here. Try using your own words
Select Answer

Figure 10: SQuAD's crowd-facing web interface

Is it worthwhile noting that the interface encourages the worker to use his own words to ask the questions, to avoid using the same words as in the paragraph (also disabling copy-paste functionality on the paragraph) and to ask hard questions. This improves SQuAD's quality and therefore its ability to meaningfully evaluate machine reading comprehension. The task was reviewed favorably by crowdworkers, receiving positive comments on Turkopticon.

- Obtaining additional answers: the complete dataset was splitted into train, development and test subsets. To get a proxy of human performance on SQuAD and to make the evaluation phase more robust, [Rajpurkar et al.2016] collected at least 2 additional answers for each question in the development and test sets. In the secondary answer generation task, each worker was shown only the questions along with the paragraphs of an article, and asked to select the shortest span in the paragraph that answered the question. If a question was not answerable by a span in the paragraph, workers were asked to submit the question without marking an answer. The recommended speed was 5 questions for 2 minutes, and paid at the same rate. Over the development and test sets, 2.6% of questions were marked unanswerable by at least one of the additional workers.

As have been shown, SQuAD promises to be a scalable, realistic and high-quality dataset in which models can be trained and evaluated on its machine reading comprehension capabilities. Hence, in the following section we will discuss the second recent major effort towards the advance of machine reading comprehension: the **development of models**. We will orientate the section towards building a model for the SQuAD dataset, as it represents the most recent advance in machine reading comprehension.

Models

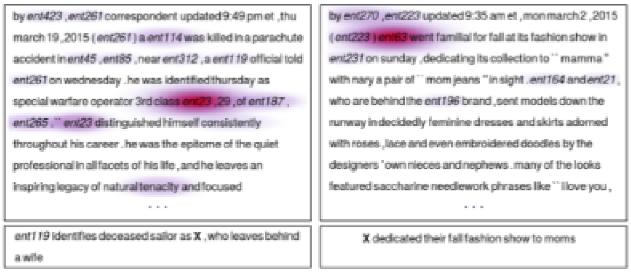
We can classify the machine reading comprehension models into **traditional approaches** and **novel approaches**. Once these approaches have been discussed, the current SQuAD **baseline** and **state-of-the-art** will be presented.

Traditional approaches.- These models are based on either hand engineered grammars, or information extraction methods of detecting predicate argument triples that can later be queried as a relational database [Hermann et al.2015]. Usually a pipeline of NLP models is built; they make heavy use of linguistic annotation, structured world knowledge and semantic parsing.

As of the data they used to be trained, approaches in Computational Linguistics have failed to manage the transition from synthetic data to real environments, as such closed worlds inevitably fail to capture the complexity, richness, and noise of natural language [Hermann et al.2015]. In accordance with these findings, in this document we focus only on the *novel approaches*.

Novel approaches.- Work on scalable data such as *cloze style* data together with advances of applying end-to-end neural network models in NLP, has shown that neural network based models hold promise for modelling reading comprehension.

[Hermann et al.2015] used *recurrent neural networks together with attention based mechanisms* to estimate the probability that a word type a from document d answers query q : $p(a|d,q)$. Vector embeddings of a document and query pair (d,q) are needed. They came up with three different Long Short Term Memory deep learning models (LSTM)[4]: the Deep LSTM Reader, the Attentive Reader and The Impatient Reader. Figure 11 shows the attention heat maps from the Attentive Reader for two correctly answered queries.



by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused
. . .

ent119 identifies deceased sailor as **X** , who leaves behind a wife

by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,
. . .

X dedicated their fall fashion show to moms

Figure 11: Attention Reader heatmaps

The models differ in the direction of how they propagate the dependencies over long distances. The Attentive and Impatient readers to propagate and integrate semantic information over long distances. They learned to read real documents and answer complex questions with minimal prior knowledge of language structure. The major shortcoming is that they predict a single token. But the answers in SQuAD contain multiple tokens, so, as it is, this approach is unfeasible. However, the end-to-end training is worthwhile and the attention mechanism used is just one instantiation of a very general idea which can be further exploited in SQuAD. [Chen, Bolton, and Manning2016] develop a similar model with some modifications than enhance its performance.

[4] Long Short Term Memory network (LSTM) is a type of recurrent network that can keep long-range dependencies, unlike recurrent networks that suffer from exploding or vanishing gradients with such dependencies

Another end-to-end training approach is the *sequence-to-sequence neural models*, which have been successfully applied to many NLP tasks [Yang, Salakhutdinov, and Cohen2016]. These models are very flexible, it is possible to generate single or multiple-token answers, thus suitable for SQuAD. Although SQuAD is larger than most currently available reading comprehension datasets, sequence-to-sequence models usually are built on datasets with a greater scale than the one provided by SQuAD. However this is still an approach that can be explored.

Memory networks architecture [Weston, Chopra, and Bordes2014] is an alternative end-to-end approach that focus on the *memorization* process with which recurrent neural networks are known to struggle. According [Weston, Chopra, and Bordes2014] memory networks reason with inference components combined with a long-term memory component; they learn how to use these jointly. The core idea behind these networks is to combine the successful learning strategies developed in the machine learning literature for inference with a memory component that can be read and written to. They used and evaluated these models for question answering (QA) tasks, not reading comprehension tasks, but they showed an interesting reasoning power (e.g. understanding the intention of verbs) that can be further explored for the SQuAD challenge. [Hill et al.2015], [Sukhbaatar et al.2015], [Kumar et al.2015] also used these kind of networks. It is important to point out that currently these models suffer from lack of scalability on large datasets [Wang and Jiang2016]. But we don't know a priori if this approach will work with SQuAD, because usually QA datasets run larger than reading comprehension ones, so the lack of scalability may not apply to SQuAD.

Pointer Networks have been adopted in a few studies in order to copy tokens from the given passage as answers [Kadlec et al.2016], [Trischler et al.2016].

Baseline.- [Rajpurkar et al.2016] proposed the first baseline model over SQuAD but it is below human performance by more than 35 percentual points (F1 score of 51% vs 86.8 %). And 40.4% in exact match vs 82.3% of human performance. The proposed model is a logistic regression built with handcrafted features. [Rajpurkar et al.2016] found that the model performance is very sensitive to the following features:

- *Lexicalized and dependency tree path features:* these are the features that contribute in greater proportion to the performance of the model.
- *Answer types:* the model performs better on answers regarding number and entities, while human performance is more uniform.
- *Syntactic divergence between the question and the sentence containing the answer:* the performance of the model decreases with with increasing divergence while human's performance remains almost constant.

State-of-the-art.- In the past months there have been advances on the SQuAD challenge outperforming the **baseline**.

[Wang and Jiang2016] proposed two new end-to-end neural network models for machine comprehension, which combine match-LSTM and Pointer Networks to handle the special properties of the SQuAD dataset. They used match-LSTM to match a question and a given passage and Pointer Network in a different way than [Kadlec et al.2016] and [Trischler et al.2016] to generate answers that contain multiple tokens from the given passage. They achieved the state-of-the-art performance of an exact match score of 59.5% and an F1 score of 70.3% on the unseen test dataset, hence outperforming [Rajpurkar et al.2016] baseline model.

References

Chen, D.; Bolton, J.; and Manning, C. D. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858.*

Cui, Y.; Liu, T.; Chen, Z.; Wang, S.; and Hu, G. 2016. Consensus attention-based neural networks for chinese reading comprehension. *arXiv preprint arXiv:1607.02250.*

Hadley, G., and Naaykens, J. 1999. Testing the test: Comparing semac and exact word scoring on the selective deletion cloze. *THE EFFECTS OF THREE LEARNING STRATEGIES ON EFL VOCABULARY ACQUISITION...* 1(1):63.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, 1693–1701.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301.*

Kadlec, R.; Schmid, M.; Bajgar, O.; and Kleindienst, J. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547.*

Kumar, A.; Irsoy, O.; Su, J.; Bradbury, J.; English, R.; Pierce, B.; Ondruska, P.; Gulrajani, I.; and Socher, R. 2015. Ask me anything: Dynamic memory networks for natural language processing. *arXiv preprint arXiv:1506.07285.*

Li, X., and Roth, D. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 1–7. Association for Computational Linguistics.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. J. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography* 3(4):235–244.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR* abs/1606.05250.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 3, 4.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, 2440–2448.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. Movieqa: Understanding stories in movies through question-answering. *arXiv preprint arXiv:1512.02902*.

Trischler, A.; Ye, Z.; Yuan, X.; and Suleman, K. 2016. Natural language comprehension with the epireader. *arXiv preprint arXiv:1606.02270*.

Wang, S., and Jiang, J. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Wang, M. 2006. A survey of answer extraction techniques in factoid question answering. *Computational Linguistics* 1(1).

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Yang, Z.; Salakhutdinov, R.; and Cohen, W. 2016. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*.