Using the IBM Watson Cognitive System in Educational Contexts

Ilianna Kollia
Big Data and Business Analytics Center of Competence
IBM
Athens, Greece
ikollia@gr.ibm.com

Georgios Siolas
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
gsiolas@islab.ntua.gr

***Abstract*— In the current paper we describe how Watson Experience Manager (WEM), an industrial Question Answering (QA) tool developed by IBM, has been used in an educational context at the National Technical University of Athens (NTUA). During the postgraduate course on Data Science, three student teams experimented with WEM's QA capabilities on three different topics, namely, Nutrition, Autism and New York sightseeing. We present WEM, the workflow followed by the teams together with qualitative and quantitative experimental evaluations.**

Keywords*— *Watson Experience Manager, Deep Question Answering, Data Science, Business Software in Higher Education.

I. Introduction

Watson Experience Manager is part of IBM's Watson ecosystem. Watson is a natural language question answering system that became famous after beating two expert human players at the quiz show "Jeopardy" in 2011. Watson essentially determines its answers and associated confidence scores based on the knowledge it has acquired. Watson solutions provide a combination of the following key characteristics:

- Understand questions through natural language. Watson provides natural language processing to help understand the complexities of human communication as a means of user interaction and to extract knowledge from sources of data
- Generate and evaluate hypotheses, answers and supporting evidence. Watson provides hypothesis generation and evaluation by applying advanced analytics to weigh and evaluate a panel of responses based on only relevant evidence.

While WEM was designed mainly as a business solution enabling corporate websites to automate customer support, by answering common questions about products, services, procedures etc., the underlying machine learning technologies, like natural language processing and reasoning, make it particularly suitable as an experimental testbed in an advanced Data Science course. IBM Greece and NTUA worked together to create a new postgraduate course, providing the necessary theoretical background on the aforementioned fields and complementing it with hands on experimentation on the WEM platform.

II. The IBM Watson Cognitive System

A. IBM Watson and Deep QA

IBM Watson is a Question Answering (QA) computing system, built by IBM to apply advanced natural language

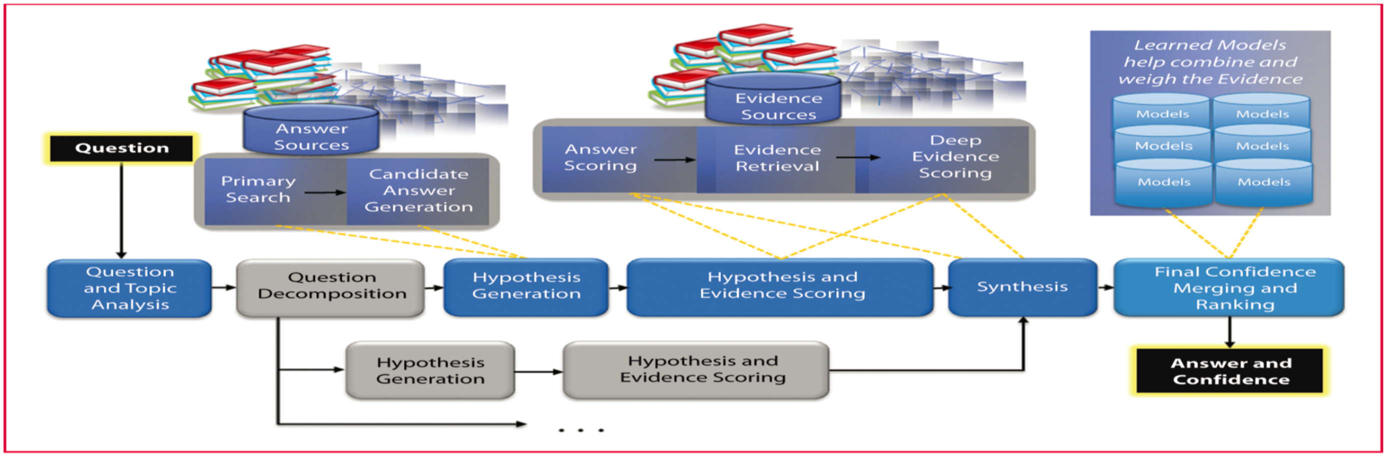


Fig. 1. Deep QA Architecture

processing, information retrieval, knowledge representation, automated reasoning, and machine learning technologies to the field of open domain question answering [4].

Watson uses IBM's DeepQA software which is illustrated in Fig. 1 and the Unstructured Information Management Architecture (UIMA) framework [1]. UIMA provides a platform for integrating diverse collections of text, speech, and image analytics independently of algorithmic approaches, programming languages, or underlying domain model.

The DeepQA architecture [5] defines various stages of analysis in a processing pipeline and every stage admits multiple implementations that can produce alternative outcomes. At each stage, alternatives are independently pursued as part of a massively parallel computation. DeepQA never assumes that any component perfectly understands the question. On the contrary, many candidate answers are proposed by searching different resources, based on different interpretations of the question. A commitment to an answer is deferred the more evidence is gathered and analyzed for each answer and each alternative path through the system. DeepQA applies many algorithms that analyze evidence along different dimensions, such as: type classification, time, geography, popularity, passage support, source reliability, and semantic relatedness.

This produces hundreds of features or scores, each indicating the degree to which evidence supports an answer according to one of these dimensions. All the features for a candidate answer must be combined into a single score representing the probability of the answer being correct. DeepQA trains statistical machine learning algorithms on prior sets of questions and answers to learn how to optimally weight each of the hundreds of features relative to one another. These weights are used at runtime to balance all of the features when combining the final scores for candidate answers to new questions. The final result of the process is a ranked list of candidate answers, each with a final confidence score representing the likelihood of the answer being correct based on the analysis of all its supporting evidence. If the top answer's confidence is above a threshold, Watson will return the answer, otherwise it will not.

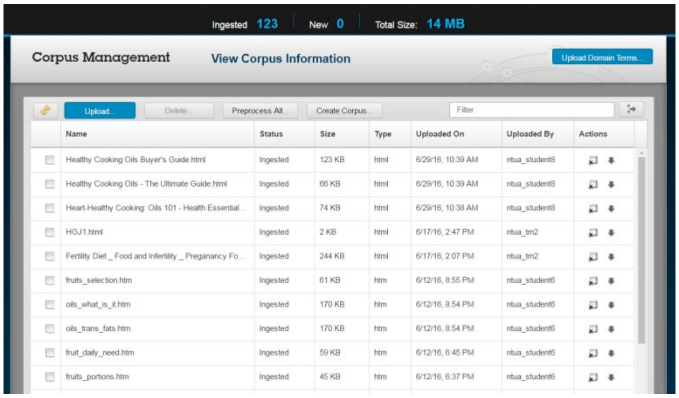
To be more specific, Watson goes through the following process, as illustrated in Fig. 1:
1. **Question decomposition**: When the question is presented, Watson parses it in order to extract its major features.
2. **Hypothesis Generation**: Watson searches the corpus (which consists of unstructured knowledge) for passages that might contain a valuable response.
3. **Hypothesis and evidence scoring**: Watson compares the language of the question and the language of all potential responses with specific reasoning algorithms. Each one of these algorithms executes a different comparison (e. g: search for the matching of terms and synonyms, examine the temporal and spatial features) and then produces one or more scores that indicate the response's degree of relevance (inference) to the question.
4. **Synthesis**: Each resulting score is weighted against a statistical model that captures the algorithm's performance at the establishment of inference between two similar passages for that domain, during Watson's training period. This statistical model is then used to summarize a level of confidence as Watson's metric of evidence that the candidate answer is inferred by the question.
5. **Final Confidence Merging and Ranking**: The process (steps 1-4) is repeated for each of the candidate answers until Watson finds responses that are stronger candidates than the others and finally returns a ranked list of them followed by their confidence score.

B. *Application Development in Watson*

As stated in the introduction, in the context of the course, the solutions were developed using Watson Experience Manager, a web based tool where users can upload and test their data[2].

To develop an application a workflow of three main phases is required to be performed in WEM: Corpus Management/ Ingestion, Training and Testing.

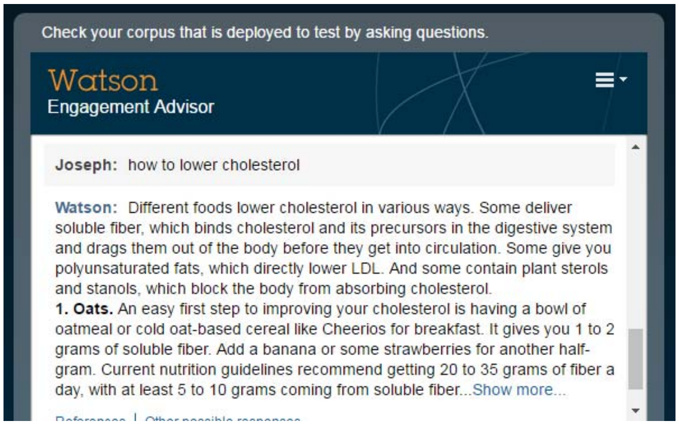
Ingestion refers to the breakdown of an unstructured corpus of documents. These can have various supported formats, such as HTML pages, pdf or word documents, XML documents and others. Once uploaded to the WEM platform, they are analyzed on their grammar and context, and broken down in segments. Fig. 2 illustrates the WEM's UI for document ingestion.



	Name	Status	Size	Type	Uploaded On	Uploaded By	Actions
□	Healthy Cooking Oils Buyer's Guide.html	Ingested	123 KB	html	6/29/16, 10:39 AM	ntua_student8	
□	Healthy Cooking Oils - The Ultimate Guide.html	Ingested	66 KB	html	6/29/16, 10:39 AM	ntua_student8	
□	Heart-Healthy Cooking: Oils 101 - Health Essential...	Ingested	74 KB	html	6/29/16, 10:38 AM	ntua_student8	
□	HGJ1.html	Ingested	2 KB	html	6/17/16, 2:47 PM	ntua_tm2	
□	Fertility Diet _ Food and Infertility _ Preganancy Fo...	Ingested	244 KB	html	6/17/16, 2:07 PM	ntua_tm2	
□	fruits_selection.htm	Ingested	61 KB	htm	6/12/16, 8:55 PM	ntua_student6	
□	oils_what_is_it.htm	Ingested	170 KB	htm	6/12/16, 8:54 PM	ntua_student6	
□	oils_trans_fats.htm	Ingested	170 KB	htm	6/12/16, 8:54 PM	ntua_student6	
□	fruit_daily_need.htm	Ingested	59 KB	htm	6/12/16, 6:45 PM	ntua_student6	
□	fruits_portions.htm	Ingested	45 KB	htm	6/12/16, 6:37 PM	ntua_student6	

Fig. 2. Document Ingestion

After ingestion a training phase follows, where a number of questions are matched with potential answers that Watson has extracted from the corpus. It is therefore important that the corpus has been ingested properly and for this some preprocessing of documents may be needed. We now describe this phase in more detail. After a question is provided it must either be matched with an answer or be clustered with other questions that share an answer. In either case the answer is picked by the Subject Matter Expert (SME), with the help of some suggestions that Watson has created from the ingested corpus. Watson gives a partially filled bar underneath each answer that shows the system's confidence that this answer matches the question. Watson can learn similar questions by creating clusters of differently phrased variations of the same question, which have the same answer. We say that in this phase SMEs build the ground truth in the form of question-answer pairs. It is suggested [3] that training questions should either be descriptive, yes/no or procedural. Descriptive questions are questions whose answers are definitions, explanations, or descriptions, Yes/no questions are questions whose answers are either true or false. Finally, procedural questions are questions whose answers are a series of steps to accomplish a specific task.


Check your corpus that is deployed to test by asking questions.
Watson
Engagement Advisor

Joseph: how to lower cholesterol

Watson: Different foods lower cholesterol in various ways. Some deliver soluble fiber, which binds cholesterol and its precursors in the digestive system and drags them out of the body before they get into circulation. Some give you polyunsaturated fats, which directly lower LDL. And some contain plant sterols and stanols, which block the body from absorbing cholesterol.
1. Oats. An easy first step to improving your cholesterol is having a bowl of oatmeal or cold oat-based cereal like Cheerios for breakfast. It gives you 1 to 2 grams of soluble fiber. Add a banana or some strawberries for another half-gram. Current nutrition guidelines recommend getting 20 to 35 grams of fiber a day, with at least 5 to 10 grams coming from soluble fiber...Show more...

Fig. 3. Question Answering Interface

The testing phase starts after a training set of sufficient size has been provided to Watson. During this phase SMEs ask Watson questions in a prompt, similar to the ones that appear to the application's users in the final version and the experts check the quality of answers returned by Watson (Fig. 3). During this phase two kinds of questions are asked; variation and blind questions. Variation questions are rephrases of questions used in training phase and are used to assess how well trained the system is to what is considered as ground truth. Blind questions are not similar to any of the questions used in training phase (i.e., their answer has not been specified during training) but the answers to these questions exist inside the ingested corpus. These are used to assess the system's performance to questions that are not part of its ground truth.

III. Applications

In this section we describe the applications that were developed as part of the graduate course at NTUA by three separate teams. Each team consisted of 4-6 members and has followed its own approach in order to develop an application on a chosen topic according to WEM's guidelines [3]. For each team we first describe the approach followed and then we provide the achieved experimental results.

A. Nutrition

The first application focused on nutrition. The definition of nutrition is given below:

Definition
We define nutrition (or nourishment) as the science that interprets the interaction of nutrients and other substances in food in relation to maintenance, growth, reproduction, health and disease of an organism. The field of nutrition has gained much attention recently due to critical debates and controversial theories on healthy eating. Lack of nutrition knowledge and food misinformation are both issues to be addressed.

1) Approach

All team members were involved in the WEM platform during each phase of development, therefore the task assignment was topic-oriented. Since the topic of nutrition is quite complex the team was guided by its definition (given above) to narrow it and divide it into subtopics to minimize topic overlapping between team members. The division resulted in four wide subtopics that are briefly described below:
1. **Meat, Grains, Dairy**: Provides information (daily portion, nutrients, health benefits and risks) about animal source food (meat, dairy) and grain food products (refined and whole).
2. **Fruits, Vegetables, Oils**: Refers to the nutrients of vegetables, fruits and oils, their variety of colors and types, as well as their recommended intake.
3. **Drinks**: Includes healthy beverage options (analyzing their ingredients and benefits), a classification of drinks to consume in moderation (soft drinks, artificially sweetened drinks, alcohol) and their association with chronic conditions.
4. **Disease Prevention**: There are a number of nutrition-related chronic diseases (obesity, diabetes, cardiovascular diseases, cancer, osteoporosis, dental disease, anemia). Dietary recommendations are made for their prevention and treatment.

After completing the subject's division, team members organized the project in two phases:
Phase 1: This phase was completed in 4 stages:
1. Collection of Representative Questions: Team members collected questions that represented the types of questions that users who are interested in a healthy diet might ask. Specific types of questions (descriptive, yes/no and procedural) were used that Watson can answer.
2. Selection of Documents: Team members selected documents whose content provided answers to the representative questions. It was ensured that documents were comprehensible by Watson.
3. Creation of Corpus: Corpus is the term that describes the data that is available to Watson. When we create the corpus, the content of the documents is processed so that Watson can retrieve sections of the content quickly to answer questions.
4. Ground Truth Development/Training: It is the process of grouping similar questions and answering questions by matching answers that explicitly link expected input (question) to specific output (answer).

Each team member separately followed the workflow suggested in Phase 1, focusing on his/her assigned subtopic.

Phase 2: This phase started by the time all team members have completed their duties as defined in Phase 1. In this phase, all members worked collaboratively, following the steps:

1. Questions approval: Team members reviewed the question-answer pairs in order to validate that all pairs are matched correctly (no overlaps/conflicts).

2. Testing: Team members ran a few testing cycles in which they asked Watson to answer a number of questions which did not belong to the set of questions used in ground truth development in order to view the system's performance.

3. Analyze results: Each testing cycle returned a number of results which were analyzed. Some of them helped team members identify changes needed to be performed on ground truth and/or corpus. These changes were evaluated by new testing cycles.

2) Experimental Results

The first step for the creation of a question answering system is gathering the corpus. There are a lot of good sources for nutrition on the web. The main sites the team members chose to use were Wikipedia, the World Health Organization, the American Diabetes Association, Nutrition Score, the Livestrong Foundation and Cleveland Clinic. The documents that were ingested into Watson were all in html format.

After collecting the documents a pre-processing step took place in which some "cleaning" was performed on some of the documents. When downloading an html web page there is a lot of "garbage" along with the useful information. For WEM the text and the html header tags (used for document segmentation) are mainly useful. The rest (contents, comments, ads, hyperlinks, scrips) are less useful and they can make the training procedure more difficult as they add noise to the final model. Because of this the team selected to pre-process a small part of the corpus creating well-structured html documents that answer questions clearly and precisely.

The next step was to upload the gathered documents to Watson. The final uploaded corpus consists of 123 html documents, having a total size of 14MB.

After adding documents and ingesting them into Watson, the team trained the system creating around 400 question-answers pairs. After the training procedure was complete the platform was deployed and tested.

During the testing process 101 questions were introduced. These questions were split into two categories. The first category, consisting of 81 questions, included variations of the questions that were part of the training group. The second category consisted of 20 blind questions. In order to demonstrate the results each time Watson answered one of the questions, the expert applied a value to that answer. This value demonstrated the correctness of the answer according to the expert's opinion. The answer was considered correct, if Watson produced the document segment that was intended to be linked to this particular question, partially correct if the general topic regarding the question was correct and wrong if the answer was entirely out of topic. The results are demonstrated in the pie chart in Fig. 4.

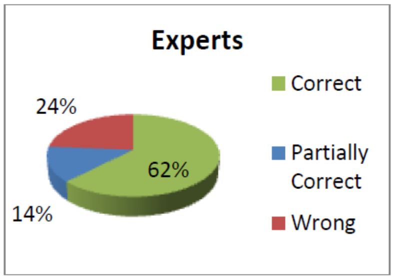


Fig. 4. Correctness of answers returned by Watson

According to the expert's opinion 62% of the answers were correct and 38% of the questions were answered either wrongly or partially correctly.

Regarding the difference in performance when Watson encountered blind questions or variations of the ones used in training the results are depicted in Fig. 5.

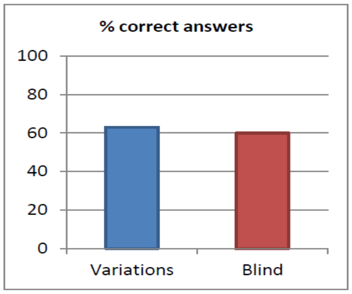


Fig. 5. Percentage of correct answers for variation and blind questions

There is little difference between the two cases and that is mainly because of the way the variation questions were generated. In specific, when creating variations of a question, there is no certain way to judge how difficult the resulting question will be. Instead of generating questions with small errors or the change of a single word, most questions were rephrased entirely and completely new words were used. Thus the results between the blind and variant questions are similar.

B. New York Visit

The main theme of the second application chosen is "New York". The designed application is a general tourist application for first time New York visitors. Thus the domains supported are the most popular touristic sites and information on transportation and culture.

1) Approach

A problem in creating the corpus that the team faced was to define the domains the application would be covering and to distribute the work among team members. Team members had to coordinate well in order to choose the correct documents that would be in line with the application definition. Furthermore, the team paid attention to the fact that the corpus should not contain documents with contradicting information. In order to avoid these problems, the team divided the application in domains. Each domain was then assigned to a team member. Each team member was responsible for gathering the relative texts and producing training and testing questions for this specific domain. Some of the domains extracted were the following:

- Monuments (Statue of Liberty, Central Park, Empire State Building, Golden State Bridge, Natural History Museum)
- Transportation
- Weather
- General

In more detail, the following steps were performed by the team.

Step 1: Decide on the right corpus

In order to better understand what types of questions a tourist would be likely to ask when visiting New York, team members decided to study FAQ (Frequently Asked Questions) sections of various online sources that provide information about New York City. These pages were included as-is in the corpus and also served as seeds for the rest of the data collection process, since they helped the team identify the main topics that are of interest and also understand what questions should be created.

Step 2: Upload corpus

The WEM corpus management tool provides a very straightforward and intuitive way to upload documents and even provides the ability to specify the type of the document (e.g. FAQ, Wikipedia article etc.) if desired. The first documents that the team uploaded to Watson were the ones corresponding to the FAQ sections which were either explicitly defined as such (e.g. https://www.newyorkpass.com/En/faq/) or just organized in a similar way (e.g. http://www.nycgo.com/plan-your-trip/basic-information). The team then proceeded to add various other sources to the corpus, mainly from Wikipedia articles for various attractions, official web pages for New York (NY) monuments and transportation-related information. After uploading a document, the system processes (ingests) it and divides it to multiple segments. In the case of html files, the html tags are used internally to ensure that the segmentation corresponds to different conceptual parts of the page. Corpus creation was the first task performed, but was later enriched with new sources in cases where gaps were identified during Q&A pair creation. In the end the team a total of 53 documents (10MB) were loaded and ingested by Watson, mainly html, although team members experimented to a smaller extent with .doc and .txt files as well.

Step 3: Create questions

Having already studied, collected and uploaded content from various online sources, each member of the team created a number of questions that were either directly found online (corresponding to questions from the FAQ web pages) or were created by team members based on the content that had been added to Watson. More specifically, each team member created an initial set of 20-25 questions.

Step 4: Provide answers to questions

Question creation does not need to be immediately followed by the provision of an answer. Questions can be created and stored, waiting to be answered by the same or other team members when the appropriate document becomes available in the corpus or when a member that knows the right answer chooses to provide it. For the purposes of the course, questions were created after updating the corpus, so each member already knew what the answer should be and where it could be found. Essentially, the corpus was the driving force for the team for the Q&A creation. After submitting a question, the system opens a 3-step wizard through which the user can provide an answer. The first step lets the user group the newly created question with a previous one. The initial questions created by each member were designed to be unique and independent, so the person creating them proceeded to the second step, i.e. to “Match an Answer”. In this step, Watson performs a lexical matching in order to retrieve the documents that are more probable to contain the answer to the expressed question. The documents are ranked and presented to the user, each of them accompanied with a confidence score, represented as a colour bar, expressing Watson certainty regarding the relevance of the document to the question. In cases when the correct document is not retrieved, there is an extra search field where the user can provide more terms in order to refine the search results. Optionally, the user can then continue to the 3rd step, i.e. the “Specify an Answer” option. In this last step the user can highlight a certain part of the selected document in order to provide a more targeted answer to the expressed question. For all the initially created questions, each team member provided both the document, as well as the exact passage to be used as an answer.

Step 5: Create question clusters

As discussed earlier, when creating a new question, Watson first tries to identify other similar questions and gives the user the ability to link the new question to an existing one, thus creating question clusters. These clusters are meant to handle the fact that in everyday life people may ask the same question in terms of meaning and expected answer, but with different ways in terms of phrasing and word selection. It becomes evident that creating question clusters is of paramount importance for the application’s efficiency and should not be considered an optional step. Therefore, for each of the initial unique questions, the team created multiple variations which were grouped with the questions to which they corresponded. For most questions 3-4 variations were created by team members, depending on the original question and the ability to create meaningful paraphrases. It should be noted that in these variations team members tried to teach Watson some synonyms. Some of the question variations were kept for the system testing phase, which will be described later. Overall, the team created a total of 396 questions (unique and variations) for training.

Step 6: Review and approve Q&A pairs

The final step in system's training is to review and approve the created questions. By default, when first created, each question has a status of "Needs Review" in order to ensure that the training data is actually correct and suitable for the application's purpose. In a real life scenario, the domain experts can use the review functionalities to actually express concern for the validity or completeness of certain answers, or to reject some of the Q&A pairs. Since no domain expertise could be claimed from team members, they decided not to evaluate the correctness of the provided answers themselves; instead they decided to rely on the usage of popular online sources about New York. What they actually evaluated in this step was the phrasing of the questions and the accuracy of the provided answers. More specifically, each team member randomly reviewed approximately 70 Q&A pairs (not created by him/her) and evaluated how well the question was expressed and whether the provided answer was appropriate (in terms of accuracy and completeness, not whether the statement was true). After correcting some minor issues that were identified, all 396 were approved and made available for Watson's training.

For the testing phase the team adopted the following workflow, as depicted in discrete steps:

Step 1: Choose the number (percentage) of questions to be used for testing

The team decided to consider 75% of the total amount of questions prepared for the "New York" application as training questions and leave the rest 25% questions for testing. According to literature, this is considered to be a satisfactory training-testing ratio.

Step 2: Assure that the questions to be used for testing are randomly selected

In order to actually test the trained system and assess the quality of the answers provided, team members had to make sure that they were not biased in the selection of the testing questions. Each member of the team randomly chose a number of cluster questions from the initial set of gathered questions that were excluded from training, coming up with around 120 testing questions that were rephrases of some initial training questions.

Step 3: Consider both variation and blind questions

The testing questions mentioned above were variations of questions used in training. The results of these testing questions can help understand how to modify ground truth or the corpus to improve the system's performance. Also a smaller number of blind questions (around 30) were prepared in order to assess the system's response to other questions that are not part of its ground truth. The team members ensured that the answers of such blind questions were included in the corpus but the system was not trained to answer them or variations of them.

Step 4: Test these questions and note "Confidence Level" and appropriateness of the answer

As a last step of the testing workflow and after making sure that the "New York" project was deployed to test, team members had to go to the test tool of WEM and "ask" the chosen testing questions. For each one of these questions, they noted the "Confidence Level" the system returned for the provided answer as well as the appropriateness/usefulness as evaluated by team members. These elements were afterwards used to obtain a holistic overview of the results acquired and measure their accuracy.

2) Experimental Results

In this section we present the experimental results that the team evaluates the tool they created. The resulting answers to the testing questions (both variations and blind questions) are classified based on their usefulness as "Useful" or "Not Useful". As "Useful" the team considers the responses of text that include the answer while as "Not Useful" the team considers the responses of text that do not include the answer.

In Fig. 6,7,8 we can see the results of testing. With blue the "Useful" responses are indicated while with red the "Not Useful" ones are indicated. In all figures we have three columns, each one for the returned confidence level the Watson system provides when answering a question.

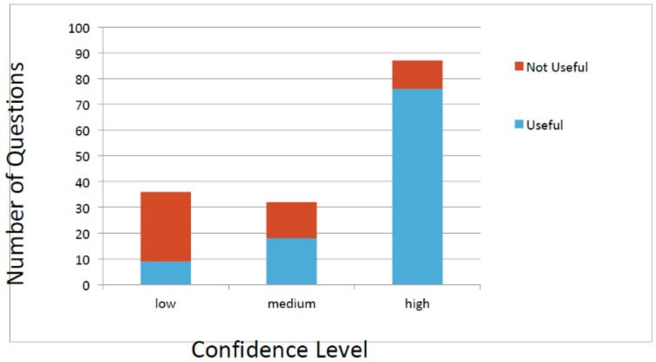


Fig. 6. Overall results of testing

As we can see on the overall results in Fig. 6 responses with high confidence tend to be correct and the lower the confidence, the less possibility the responses have to be "Useful". In general, following the way we defined usefulness, we can see that the tool provides good results with 66% accuracy (103 "Useful" and 52 "Not Useful" answers).

As far as the other two figures are concerned (Fig. 7 and 8), we observe the same behavior on both variation questions and blind questions – the higher the confidence, the more useful the response. However, as we can see blind questions have a higher miss rate than the variation questions with more than half responses marked as “Not Useful”. The difference is to be expected since the system is not trained on those questions.

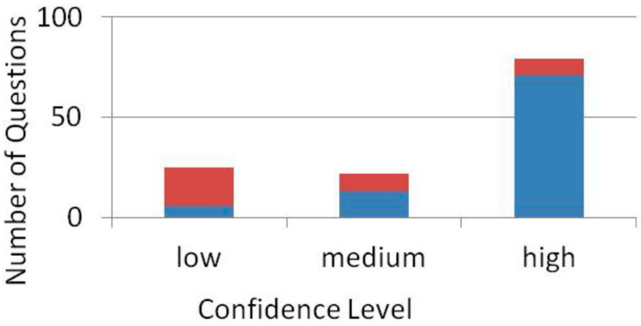


Fig. 7. Results on variation questions

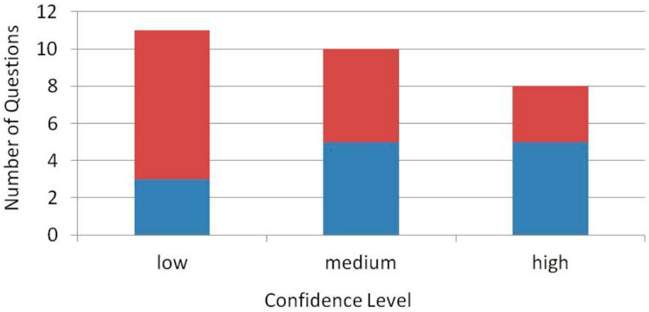


Fig. 8. Results on blind questions

C. Autism

The third developed application focused on autism. We begin with a brief introduction to the topic of autism. While it is certainly an interesting topic in and by itself, our main intent is to examine exactly what kind of information a user would be interested in. The Wikipedia article on autism begins with the following paragraph:

> Autism is a neurodevelopmental disorder characterized by impaired social interaction, verbal and nonverbal communication, and restricted and repetitive behavior. Parents usually notice signs in the first two years of their child's life. These signs often develop gradually, though some children with autism reach their developmental milestones at a normal pace and then regress. The diagnostic criteria require that symptoms become apparent in early childhood, typically before age three. While autism is highly heritable, researchers suspect both environmental and genetic factors as causes. In rare cases, autism is strongly associated with agents that cause birth defects. Controversies surround other proposed environmental causes; for example, the vaccine hypotheses have been disproven.

1) Approach

The team broke down the broad topic of autism in three segments, namely: causes, symptoms and treatment of autism, and work on each was undertaken by a team of two students (six in total). Each student was involved in all procedures and operations. Autism is a fairly easy topic to research, with plenty of online sources and QA forums, which were used for the corpus and from which training questions were also extracted. Some academic documents were also included, mainly to test the system’s capabilities.

Each team created about 40 questions and gave each one 4 syntactic variants, that is, rephrasing the same question or even changing its scope to slightly broader or narrower, but setting its answer to be the same as the other variants during the training phase. The corpus each team picked varied, while some teams experimented with some pre-processing of the corpus in order to aid Watson ingest it. Moreover, each team provided Watson with a number of variants to questions that it had been trained with, and some blind questions that hadn’t occurred in training.

2) Experimental Results

Information about corpus creation and collection of questions is given below for each category of topics (i.e., causes, symptoms, treatment).

Causes: The corpus to train Watson on questions related to the causes of autism was built mainly by html versions of academic papers published in scientific magazines, pages from Wikipedia, PubMed and Quora. On the other hand, the questions used to train the system were mainly from Quora, other QA sites and a few common sense questions. Each introduced question was clustered with some variations of itself mainly built using synonyms and grammatical variations of the original question. In total 100 questions (30 core questions and 2-5 variations for each core question) were created for training. The system was tested using variations of the questions not introduced in the training stage, yet very similar to the ones used for training. Out of 20 questions, 5 were incorrectly answered with an accuracy of 75%.

Symptoms: Text corpus was obtained by several webpages, including Wikipedia, specialized articles and scientific papers. Text parts were separated from the respective html pages either manually or by using the Mozilla firebug tool. The majority of the corpus documents consisted of well paragraphed texts each of which had a header indicating quite clearly the subject of the respective paragraph. Those headers were used as a base to derive a large number of training questions. A total of 21 documents have been ingested by Watson for this subtopic.

A total set of 160 questions consisting from 40 core ones and 4 variation questions for each core question were used to train Watson. The system was tested using variation questions. Out of 20 questions, 9 were answered correctly and precisely, for 7 questions the answer contained a larger text which included the correct answer, 4 were completely wrong, i.e., an accuracy of 80% (assuming that the 7 questions, for which Watson returned a large text containing the answer, were correctly answered).

Treatment: A total set of 160 questions consisting from 40 core ones and 4 variation questions for each core question were used to train Watson on treatment related corpus. The system was tested using variation questions. Out of 20 questions 13 were correctly answered yielding an accuracy of 65%.

As we can see the results varied from team to team; they are presented consolidated in Table 1. In this table one can additionally see the accuracy when each team posed blind questions to Watson. Note that the total number of documents ingested by Watson was 150 (with total size 20MB) and the number of questions collected by all teams were around 400.

Table 1: Results per category

	Causes	Symptoms	Treatment
Training	30 clusters of mixed (2-5) variants, 100 questions total	40 clusters of 4 variants each, 160 questions total	40 clusters of 4 variants each, 160 questions total
Testing	20 variants of training questions, 6 new unlearned questions	20 variants of training questions, 6 new unlearned questions	20 variants of training questions, 6 new unlearned questions
Results on learned questions	15 correct answers, 5 wrong	9 correct and precise, 7 returned a larger text that included the correct answer, 4 wrong	13 correct answers, 7 wrong
Results on unknown questions	2 correct answers, 4 wrong	2 correct, 1 correct text lacking information, 3 wrong	4 correct answers, 3 wrong
Accuracy % (trained, unknown)	75%, 33%	80% (or 45%), 50%	65%, 57%

The symptoms team has provided two scores for accuracy in the case of variation questions, because, for some questions, while the answer was contained in the text response of Watson, the text was 3 pages long.

IV. Conclusions

In the current work we have presented a successful usage of the IBM Watson system in educational contexts. Three applications have been developed by university teams, aggregating content and using deep learning and question answering methodologies on the Watson system in the domains of health care and tourism. In our forthcoming work we will extend the developed systems targeting their usage in real life environments.

Acknowledgment

The presented applications have been developed in the framework of the postgraduate course “Advanced Topics in Data Science” organized by the National Technical University of Athens with the support of IBM.

References

[1] D. A. Ferrucci, “Introduction to ‘This is Watson’”, IBM Journal of Research and Development, vol. 56, no. ¾, Paper 1, pp. 1:1-1:15, May/Jul. 2012.

[2] IBM Watson Ecosystem Getting Started Guide, Version 1.1, IBM, July 2014.

[3] Understanding an IBM Watson Engagement Advisor Solution, Version 1.0, IBM Corporation, 2013.

[4] Strzalkowski, T., and Harabagiu, S., Advances in Open-Domain Question-Answering, Berlin: Springer, 2006.

[5] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, Building Watson: An Overview of the DeepQA Project, AI Mag., vol. 31, no. 3, pp. 59–79, Fall 2010.