# What or Who is Multilingual Watson?

**Keith Cortis**
IBM Ireland
keithcor@ie.ibm.com

**Urvesh Bhowan**
IBM Ireland
URVESHBH@ie.ibm.com

**Ronan Mac an tSaoir**
IBM Ireland
ronan.mcateer@ie.ibm.com

**D.J. McCloskey**
IBM Ireland
dj_mccloskey@ie.ibm.com

**Mikhail Sogrin**
IBM Ireland
SOGRIMIK@ie.ibm.com

**Ross Cadogan**
IBM Ireland
ROSSCADO@ie.ibm.com

**Abstract**

IBM Watson is an intelligent open-domain question answering system capable of answering questions posed in natural language. However, the system originally developed to compete against human players on Jeopardy! is heavily reliant on English language components, such as the English Slot Grammar parser, which impacts multilingual extensibility and scalability. This paper presents a working prototype for a multilingual Watson, introducing the major challenges encountered and their proposed solutions.

## 1 Introduction

IBM Watson is an intelligent open-domain question answering (QA) system capable of answering questions posed in natural language (Ferrucci, 2012). The open-domain QA problem is one of the most challenging in computer science and artificial intelligence, as it touches on aspects of information retrieval, Natural Language Processing (NLP), knowledge representation, machine learning and reasoning. Since Jeopardy! in 2011 (Ferrucci et al., 2010), Watson has been applied successfully to other domains, such as healthcare, finance and customer engagement. However, like Jeopardy!, these application areas deal exclusively with English data. The system is therefore heavily reliant on English NLP components, in particular, the rule-based English Slot Grammar (ESG) parser used throughout. While the ESG parser performs well and has limited multilingual capabilities, the required grammar or slot rules are language-specific, impacting scalability and deployment speed. This paper presents some of the major challenges and proposed solutions to extend Watson to support any natural languag. We introduce a robust cross-lingual method for identifying crucial characteristics in a question (such as the Lexical Answer Type), which shows similar expressiveness to the hand-crafted English-based implementation. We outline a system for detecting the same named entities across text in multiple languages, and our current effort to train a multilingual Watson system using Wikipedia data for demonstration at Coling 2014. The demo setup and a brief discussion of results is also presented.

## 2 Overview of Watson Architecture

This section presents an overview of the DeepQA architecture for multilingual Watson. In **Question Analysis** a detailed syntactic and semantic analysis is performed on the input question. Unstructured text is converted to structured information for use in later components. This process uses a series of NLP technologies, such as a statistically trained natural language parser, and components for named entity recognition, anaphora resolution and relation extraction. The Lexical Answer Type and Focus are important examples, described later. The **Hypothesis Generation** phase produces all possible candidate answers for a given question. Watson searches its corpora for relevant content. Example sources include unstructured knowledge, such as Wikipedia and structured resources including DBpedia and PRISMATIC. Potential answers to the question are generated from the retrieved content as unscored hypotheses. In **Supporting Evidence Retrieval**, the system gathers additional supporting evidence for each hypothesis by searching for occurrences of the candidate answer in the context of analysed question data.



*Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 95–99, Dublin, Ireland, August 23-29 2014.

**Hypothesis and Evidence Scoring** uses many scoring algorithms to determine the relevance of retrieved candidate answers. Each scorer, whether context dependent or independent, produces a measure of how well the evidence supports a candidate answer for a question. The **Final Merging and Ranking** phase merges equivalent candidate answers and uses a machine learning model to rank the final merged set of answers accordingly.

## 3 NLP and Parsing in Multilingual Watson

Syntactic parsing plays an important role throughout the major stages in Watson architecture, from question analysis, to primary search and answer scoring. We aim to investigate the impact of deep syntactic parsing compared to shallow methods for named entities, temporal and geographic references, etc. It is thought that accurate determination of the roles these aspects play requires a deeper analysis of sentence structure. For example, in the original system, the rules which detect the question focus and LAT are heavily dependent on the deep syntactical parse. Our experiments have show comparable results in this task with shallow methods. However, in corpus ingestion aspects such as building syntactic frames in order to learn axioms, such as "is _a(liquid, fluid)" and vice-versa, the subject-verb-object directionality of a statement is paramount.

Watson uses the rule-based ESG parser (McCord et al., 2012), which performs exceptionally well in terms of parse quality and throughput, and defines our parsing benchmark. The XSG formalism underpinning ESG supports new languages through the generation of language-specific grammar or slot rules. This activity requires significant effort from highly skilled linguists for each new language. As the system further evolves for new domains and use-cases, such rules must be revised and extended, resulting in a long term requirement for this specialised skillset. To address this skill requirement and enable a more scalable approach, a move towards statistical parsing methods is being investigated. We identified the following attributes of our ideal parsing technology. It should be multilingually capable, fast (compared to XSG, currently 2 orders of magnitude faster than current statistical parsers); highly accurate (comparable with XSG); easily extensible to new languages with low effort (relative to XSG); easy (and fast) to train on a new language or domain; memory efficient; robust to noisy/ungrammatical input; support a rich set of annotation features (ESG has 70+); facilitate overriding of biases in training data.

Our investigations indicate that meeting all of these requirements will be a challenge in any single parsing formalism. Statistical dependency parsing allowing for non-projective trees appears to be a good fit for the variation in language structure that we will need to support (McDonald et al., 2013). Experiments with MSTParser and the Eisner algorithm in Italian have shown promise from a quality point of view. We have also identified a language-independent formal representation of a parse. McDonald et al. (2013) present a harmonized set of dependency labels for multilingual parsing which has been adapted for other typologically diverse languages, such as Chinese and Finnish and encourages convergence for reuse. Our initial investigations using this set for English, Spanish, French, Brazilian Portugese and German text, suggest minimal modification is required to adapt for use in question answering. Modification of existing pipeline components to a more streamlined dependency structure than the XSG formalism, will form part of ongoing research.

Treebanks of parse data with part-of-speech and dependency labels will be required in order to train the chosen parser. There are several examples of existing Treebanks for the chosen languages, such as the IULA Spanish LSP Treebank[1]. However, the context of the data included is very rarely representative of question-answering scenarios. For example, in the IULA corpus, there are less than 10 questions. We will therefore be supplementing this training data with our own hand-annotated corpora, in order to increase the validity of the trained parser for use in question answering. As mentioned previously, we will also be adapting these resources to use the UniPos and UniDep part-of-speech and dependency label sets. In parallel to the

[1] http://www.iula.upf.edu/recurs01_tbk_uk.htm

investigations for multilingual parsing, we have also considered the requirements of a parser-independent system which can leverage shallow parse data in order to perform reasonably well at the same task. The multilingual LAT detection component which builds on part-of-speech data to identify noun phrases and associated head words and modifiers, may be additionally used to generate a simple linked parse structure without dependency labels. As the number of supported languages in the system grows, it is important to have a meaningful baseline upon which to build, even while a parser or appropriate training data for the new language is still being prepared. The parser-independent capability will be particularly useful in this context.

## 4 Other Challenges Faced in Multilingual Watson

### 4.1 Multilingual Lexical Answer Type (LAT)

For the Jeopardy! challenge, one of the most critical elements in the system design was the recognition of what is termed the LAT (Ferrucci et al., 2010). This is typically a noun or noun-phrase in the question which identifies the type of answer required, without any attempt to evaluate its semantics. Similarly influential, the Focus is the part of the sentence that, if replaced with the answer, makes the question a standalone statement. For example, in the question "What countries share a border with China?", the LAT is "countries", and the Focus is "What countries". Replacing this piece of text with the answer becomes a valid standalone statement, such as "Russia, Mongolia, India, ..., share a border with China". Ferrucci et al. (2010) found that identifying a candidate answer as an instance of a LAT is an important part of the answer scoring mechanism, and a common source of critical errors.

In the original system, a Prolog component was used to match specific English language patterns for various purposes including LAT detection. In a multilingual context, we require a more robust method that retains the same potential expressiveness and performance of a good Prolog implementation, while facilitating cross-lingual pattern recognition. Our multilingual prototype uses lexical features, such as part-of-speech and lemma. Pattern matching rules over these features were developed using the IBM LanguageWare[2] rules engine which provides a comparable level of expression with standard Prolog implementations. While maintaining pipeline accuracy for English, this prototype was also 4 times faster than the original Prolog modules. A statistical method that was originally used in the Jeopardy! pipeline is also being adapted for use in a multilingual context. This approach will reduce the dependency on hard-coded language-specific parsing rules. The use of harmonized Stanford dependencies (McDonald et al., 2013) will further enhance these efforts.

### 4.2 Detecting Concepts across Multiple Languages

One of the most challenging aspects of multilingual NLP is the recognition of identical concepts across text in multiple languages. Wikipedia and Wiktionary provide translations of words from one language to another, however they do not establish language-independent identifiers for concepts. Open Multilingual Wordnet (OMW) project[3] links WordNet style structured resources, in up to 150 languages, to the Princeton Wordnet of English[4]. While Princeton Wordnet is made specifically for English, its numeric ID system is in fact a set of language-independent identifiers, and may be used to relate concepts and words. The OMW project provides links to the same IDs from words in other languages. The Extended version of the OMW dataset additionally links Wiktionary data with these WordNet structured resources, thus greatly improving coverage of vocabulary.

In the multilingual Watson system, we can perform semantic analysis using domain knowledge irrespective of the language of the question, or the domain. This is enabled by a process of concept identification that maps instances of concepts in natural language text to a set of

[2]http://www-01.ibm.com/software/globalization/topics/languageware/
[3]http://compling.hss.ntu.edu.sg/omw/
[4]http://wordnet.princeton.edu/

language-independent identifiers. Our implementation takes inspiration from efforts, such as the OMW and the Unstructured Medical Language System[5] in the biomedical domain, which use alphanumeric labels to identify individual semantic concepts, irrespective of the forms these concepts take in any language. In addition to these unique identifiers, our design incorporates fully qualified URI namespaces for these instances, as proposed by the W3C Semantic Web Standard, in order to distinguish between instances of a concept in various contexts.

In parallel with this concept ID system, we have developed a lexicon expansion framework that incorporates pluggable transformation modules to generate alternative forms for lexicon entries. This facilitates the increased coverage of semantic concepts in the chosen domain text, which remain linked to their respective namespace-qualified unique identifiers.

### 4.3 Machine Learning Challenges

The original Watson system uses a cascade of multiple trained machine-learning models to decide if a candidate answer is correct. In each cascade, questions are categorised and routed to models trained for different types of English Jeopardy! questions. Training these models requires questions with known answers for the different question types. However, the same Jeopardy! style question characteristics may not apply or be evident in different languages. To address this, we simplified the hand-crafted model routing in the multilingual system to make no *a priori* assumptions about the question type. While this requires good model generalisation over a potentially broad range of questions, this is offset by the smaller but highly-focused feature set used in this multilingual system. Initial features were chosen by ranking scorers whose output showed the highest correlation to the correct class on experiments with English questions.

## 5 Multilingual Watson on Wikipedia-based Questions

### 5.1 Searching over Wikipedia

Ingestion is the process of transforming documents for use by Watson. Raw Wikipedia XML documents[6] are transformed into the TREC standard[7]. These TREC files must conform to the UTF-8 character encoding scheme. The TREC-formatted documents are then transformed into Lucene[8] search indices. During the TREC transformation process, text normalisation and character replacement was being conducted for English text. All corpora text in Unicode was normalised to ASCII, e.g., for the term *Japón*, the accent was stripped from the *ó* character, thus normalising to *Japon*. In addition, characters with particular ISO8859 codes, were replaced, such as $\pi$ with the character *n*. Since the Jeopardy! pipeline handles only ASCII character encoding, this prevents the system from generating correct answers which contain non-ASCII characters, dramatically lowering recall on multilingual questions. These issues are resolved in the multilingual Watson system, which uses Unicode in the Normalisation Form Compatibility Composition.

### 5.2 Wikipedia-based Questions and Answers

We used 3732 English questions with known answers (originally gathered by the Watson team) as our question base (split into 3359 training and 373 test questions). These questions were machine translated to Spanish, French and Brazilian Portugese using IBM's n.Fluent Translation service[9] . The test set was manually reviewed to correct any translation errors, and questions deemed unanswerable (where Watson had no means of retrieving the correct answer from the source Wikipedia corpus) were removed. To assist in identifying which questions are unanswerable, the MediaWiki API[10] (an open web API service providing access to Wikipedia meta-data) was

[5] http://www.nlm.nih.gov/research/umls/
[6] Obtained from: http://dumps.wikimedia.org/
[7] For the Text REtrieval Conference (TREC) standard see: http://trec.nist.gov/
[8] http://lucene.apache.org/
[9] http://www-03.ibm.com/press/us/en/pressrelease/28887.wss
[10] http://www.mediawiki.org/wiki/API

used to filter questions whose answers could not be mapped to an article title in the Wikipedia source corpus.The MediaWiki API also has a cross-language aspect which provides useful redirect information between Wikipedia article titles, which we used to supplement our answers, e.g., "JFK" in English redirects to "John F. Kennedy" in Spanish. The manual curation of the translated English questions for English, Spanish, French and Brazilian Portuguese ensure that the same question set is used across the different languages, and that these common questions are all answerable with respect to their respective Wikipedia corpus.

## 6 Demo Setup and Discussion of Results

The setup of the demo will include a multilingual QA system (for several languages, such as English, Spanish, French, Brazilian Portuguese, etc.). The participants attending the Coling conference will be able to ask the multilingual Watson QA system a question via its web user interface. The system will then attempt to answer the question in real-time, and will return a list of the five top-ranked candidate answers and their confidence scores. The confidence score for each candidate answer represents the likeliness that it is correct, based on the analysis of all supporting evidence gathered by the system. Any supporting evidence can also be examined for a given answer, such as the passage or document hits from the search process.

For disclosure purposes we are unable to provide details on our multilingual experimental results. Therefore, we will briefly discuss our initial baseline results and the improvements made in our current system. Our initial baseline is based on the results using the English test questions, which achieved very high (near perfect) recall and high accuracy rates. In terms of multilingual Watson, our initial recall results were very low compared to English. As a result, recall became the primary focus of our multilingual investigation. Recall was improved by around 6% with the full Unicode text normalisation support and parser-independent changes (discussed in Sections 3 and 5.1 respectively). In addition, the answer curation discussed in Section 5.2 improved recall by 9%. Other language specific improvements and customisation to the primary search components in multilingual Watson, in particular, the Lucene analyser and search query, resulted in a further 29% increase in recall. These combined efforts produced comparable recall rates for Spanish, French and Brazilian Portuguese test questions to the English questions. The next area of our investigation will be focused on accuracy improvements.

## References

David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79.

David A. Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3):235–249.

Michael C. McCord, J. William Murdock, and Branimir Boguraev. 2012. Deep parsing in watson. *IBM Journal of Research and Development*, 56(3):3.

Ryan T. McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B. Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97. The Association for Computer Linguistics.