Textual resource acquisition A key requirement for high-performing question-answering (QA) systems is access to high-quality reference corpora from which answers to questions can be hypothesized and evaluated. However, the topic of source acquisition and engineering has received very little attention so far. This is because most existing systems were developed under organized evaluation efforts that included reference corpora as part of the task specification. The task of answering Jeopardy!i questions, on the other hand, does not come with such a well-circumscribed set of relevant resources. Therefore, it became procedures to acquire high-quality resources that can effectively support a high-performing QA system. To this end, we developed three procedures, i.e., source acquisition, source transformation, and source expansion. Source acquisition is an iterative development process of acquiring new collections to cover salient topics deemed to be gaps in existing resources based on principled error analysis. Source transformation refers to the process in which information is extracted from existing sources, either as a whole or in part, and is represented in a form that the system can most easily use. Finally, source expansion attempts to increase the coverage in the content of each known topic by adding new information as well as lexical and syntactic variations of existing information extracted from external large collections. In this paper, we discuss the methodology transformation, and expansion of textual resources. We demonstrate the effectiveness of each technique through its impact on candidate recall and on end-to-end QA performance. A key element in high-performing question-answering (QA) systems is access to quality textual resources from which answers to questions can be hypothesized and evaluated. However, the topic of source acquisition and engineering has not received much attention because organized evaluation efforts, such as the Text REtrieval Conference (TREC) [1], Cross Language Evaluation Forum (CLEF) [2], and NTCIR (NII Test Collection for IR Systems Project) [3], have traditionally included a reference corpus as part of the task specification. Although some researchers have used sources in addition to the reference corpus, for the most part, they have focused on investigating how Web data can be used to improve QA performance. Although the breadth of knowledge on the Web and its

redundancy would have presents two problems. First, hardware needs for searching a Web-sized corpus would be very substantial in order for Watson to meet the speed requirement as a competitive Jeopardy!** player. Second, previous experiments have shown that QA performance may not improve or may even degrade if sources are indiscriminately added [4]. Therefore, we adopted as part of the Watson effort the development of a set of well-defined procedures to acquire a high-quality and reasonably sized textual resource that can effectively support a high-performing QA system. We developed three procedures to obtain high-quality textual resources, i.e., source acquisition, source transformation, and source expansion. When developing a new QA system or adapting an existing system to a new domain, relevant sources need to be identified to cover the scope of the new task. We refer to this process as source royalty free without further permission by computer-based and other information-service systems. Permission to republish any other portion of this paper must be obtained from the Editor. J. CHU-CARROLL ET AL. 4 : 1 acquisition, which is an iterative development process of acquiring new collections of documents to cover salient topics deemed to be gaps in existing resources. The acquired sources are examined with respect to characteristics of system components, as well as to the nature of the questions and the answers in the new domain to ensure that they are represented in the most effective manner. Some acquired sources go through a process that we call source transformation, in which information is extracted from the sources, either as a whole or in part, and is represented in a form that the system can most easily use. Finally, whereas source acquisition helps ensure that the system has coverage in salient topics of the domain, source expansion attempts to increase the coverage of each known topic by adding new information, as well as lexical and syntactic variations of existing information. We believe that the methodology that we developed for source acquisition, transformation, and expansion is crucial for providing Watson with the necessary resources to achieve high QA performance. Source acquisition and engineering are relevant for systems and domains beyond Watson and Jeopardy!. Although there are preexisting textual resources for many application domains, their coverage may not

be well aligned with the task. In these cases, source acquisition and expansion are effective techniques for increasing source coverage. Furthermore, regardless of the application, the idea of source transformation is a useful technique to improve the utilization of the available content. Although some of the algorithms that we present in this paper may be tailored to the specific algorithms and data sources that Watson employs, we believe that the overarching ideas and processes that we developed on source acquisition and engineering are general and adaptable to other applications and domains. In this paper, we focus our discussion of source acquisition, transformation, and expansion on unstructured textual resources because the acquisition and the transformation of structured and semistructured resources are described elsewhere [5–7]. Note that although both Watson's sources and the components that use them were constantly changing through our development process, in this paper, we take a retrospective view on the evolution of Watson's sources. To facilitate the assessment of the impact of different source processing algorithms on QA performance, we used the same Watson run time components for all experiments. The components have the same configuration in the experiments except for the sources used for primary search (in which relevant content is retrieved from Watson's sources to identify candidate answers) [9] and supporting passage retrieval (where Watson's sources are used to find additional supporting evidence for each candidate answer) [10]. The results of these experiments demonstrate the effectiveness of each source processing technique on candidate recall and on end-to-end QA The purpose of source acquisition is to identify existing collections with good coverage for the task at hand, i.e., in the case of Watson, for answering Jeopardy! questions. Our initial effort included a domain analysis based on a randomly selected set of Jeopardy! questions to determine the scope and breadth of the problem space. This effort led to the selection of an initial corpus and the establishment of a promising baseline. We subsequently engaged in an iterative error analysis and source acquisition process to fill in gaps in Watson's information sources. The increase in source coverage directly translated into improved overall QA Domain analysis for initial corpus acquisition A cursory examination of Jeopardy! questions reveals that

there is no well-defined set of Jeopardy! domains. Jeopardy! covers a very broad range of topics, ranging from classical literature, science, history, geography, and politics to arts and entertainment; hence, it is difficult to find a set of domain-specific sources that can provide sufficient coverage for Jeopardy!. Fortunately, although the questions cover many domains, for the most part, they cover popular topics within those domains and represent information of interest to the general audience. This observation compelled us to explore the match between Jeopardy! questions and Wikipedia**, because it is a crowd-sourced large collection of documents on topics of interest to the authors. As a first step, we measured the percentage of questions whose answers are Wikipedia titles. This is a rough estimate of the percentage of Jeopardy! answers that are covered as topics in Wikipedia, assuming that the word sense of the Jeopardy! answer is the same as the Wikipedia title. If we further assume that the information given in the question is present in the Wikipedia document about the answer, this can also give us an approximate upper bound of the coverage of Wikipedia for answering Jeopardy! questions.1 Our investigation shows that on a randomly selected set of 3,500 questions, all but 4.53% of the answers were Wikipedia titles. Of those answers that are not Wikipedia titles, some are multiple answer questions (e.g., BIndiana, Wisconsin, and Ohio[and Bheat and electricity[), some are synthesized answers to puzzle questions (e.g., BTGIF Murray Abrahams[and Blevel devil[), and a small number are verb phrases (e.g., Bget your dog to heel[). On the basis of the results of this analysis, we adopted an August 2010 snapshot of Wikipedia (3.5 million articles and 13 GB of textual content) as our baseline corpus. An experiment using Watson's most comprehensive search and 1We are well aware that this is a strong assumption, which is what led to the source expansion work described later in this paper. However, the assumption is true often enough for us to assess the potential effectiveness of Wikipedia as a corpus for answering Jeopardy! questions. J. CHU-CARROLL ET AL. candidate generation strategies, described in [9], on Wikipedia alone revealed that we achieve a candidate binary recall of 77.1%, with an average of 178 candidates per question. Candidate binary recall is defined as the percentage of questions for which at least one correct

answer is present in the candidate answer list. When ranked by Watson's full suite of answer scorers, these candidates yield an end-to-end accuracy of 59.0%. The candidate binary recall of Wikipedia alone is not sufficient for the QA performance that we were targeting as a competitive Jeopardy! player. A binary recall of 77.1% means that roughly one quarter of all questions failed after candidate generation, before the bulk of the deeper semantic analysis in the QA process takes place. In the next section, we describe the iterative error analysis and the source acquisition process that we performed during Watson's development phase to increase the coverage of our Error analysis for additional source acquisition Our error analysis process for identifying gaps in Watson's sources changed over time as the source coverage improved. Initially, when the candidate recall was low, we focused on identifying systematic failures of certain question classes that corresponded to the lack of sources in particular domains. In the initial stages of Watson's development, its search and candidate generation strategies were less mature; as a result, many of the error cases were due to failures in the search or candidate generation process, i.e., the process of formulating effective search queries to retrieve relevant texts from Watson's sources and of identifying potential candidate answers from those texts. An analysis in early 2008 indicated that candidate recall failures at that time were fairly evenly divided between source and search/candidate generation causes. In this paper, we focus on the steps that we took to address source failures, i.e., instances where the correct answer is not in the sources or is not justified by the sources. For improvements in search and candidate generation strategies, see [8]. In the initial rounds of analyses, we identified several classes of questions for which the system failed, including inverse definition questions (identify the term given one or more of its definitions), quotation questions (either complete a quotation or identify the source of the quote), Bible trivia, book and movie plots, etc. On the basis of these results, we acquired collections that cover these domains, which included, among other sources, Wiktionary**, Wikiquote, multiple editions of the Bible, and popular books from Project Gutenberg**. Once the systematic failures were addressed, we had sources that covered most popular topics in our target domain. However, the candidate

recall failures suggested that there were many cases where the specific facts in the question were not in Watson's sources, although the topics were present. To increase coverage, we acquired additional existing general-purpose collections such as other encyclopedias and newswire articles, as well as specific corpora such as additional dictionaries and In order to evaluate the impact of the new sources that we acquired, we performed candidate recall regression tests to determine whether the addition of these sources led to better coverage of Jeopardy! questions/answers. This evaluation was done for two purposes. First, we wanted to ensure that the new sources made additional contributions beyond the existing sources. Second, we needed to verify that the new corpora did not introduce too much noise and cause previously retrieved relevant passages to drop out of the search hit list. More specifically, the regression test was performed by comparing all the candidates produced for a set of questions with two versions of Watson's sources, i.e., with and without the new sources. In the analysis of the regression test results, we focused on those questions where a correct answer was either gained or lost in the candidate list because of the new content. A correct answer was gained because of new content if a correct answer previously not in the candidate answer list was produced from text in the new sources. A correct answer was lost because of new content if a previously found correct answer was no longer produced, and the search hit list was populated by texts from the new sources; this suggested that the new sources may contain too much noise to be effective. If the new sources produced more gains than losses, they were adopted as part of Watson's sources. Because we want to achieve high QA performance using a high-quality and reasonably sized corpus, we need to best use the sources that we have acquired. We observed that not all sources are created equal. Some of our corpora contain title-oriented documents whose titles are concepts or entities, and the document texts provide information about those concepts or entities. For example, encyclopedia articles are, by nature, title-oriented and so are documents consisting of dictionary entries. On the other hand, newswire article titles typically express a current event or an opinion, such as BMandela visits Philippines in first Asian tour[and BDVDs are season's

bargains[. This distinction turns out to be significant in the effectiveness of the search strategies that we developed. The observed effectiveness of title-oriented sources not only gives us a preference in source acquisition but also motivates us to reengineer the sources that we Utilizing title-oriented sources There are several interesting properties about title-oriented sources. First, if a question gives a number of properties about the answer and a title-oriented document matches the question text well, then the answer is likely to be the J. CHU-CARROLL ET AL. 4 : 3 title of that document. For example, consider the following TENNIS: The first U.S. men's national singles championship, played right here in 1881, evolved into this New York City tournament. The question contains multiple facts about the correct answer, including the following: 1) it is the first U.S. men's singles championship; 2) it was first played in 1881; and 3) it is now a tournament played in New York City. These facts, although disjoint, are highly likely to be mentioned in a title-oriented document about the correct answer BU.S. Open[. Second, if salient entities in the question have title-oriented documents about them, then there is a good chance that the answer to the question is in one of those documents. For instance, consider the question BAleksander Kwasniewski became the president of this country in 1995[. The answer to the question is in the first sentence of the Wikipedia article on Aleksander Kwasniewski: BAleksander Kwasniewski is a Polish socialist politician who served as the President of Poland from 1995 to 2005[. We made use of the characteristics of title-oriented sources by adopting a three-pronged search strategy: document search, title-in-clue (TIC) passage search, and passage search. Passage search, which extracts relevant short passages from a text corpus, is the traditional search strategy adopted in most existing QA systems. Watson additionally adopts the document search and TIC passage search strategies, which are motivated by the title-oriented nature of some of Watson's corpora. Document search is effective because of the first interesting property previously discussed, whereas TIC passage search is intended to leverage the second property. Further details on the motivation and the implementation of these strategies can be found in [8]. To demonstrate the impact of the two search strategies based on title-oriented documents, we

evaluated the separate contributions of these strategies and compared them with the baseline of using passage search alone. Note that all three search strategies go against the same title-oriented corpus (Wikipedia). Our results, presented in Table 1, show that using the traditional passage search approach on this corpus, the candidate binary recall is only 64.7%. Contrast this traditional approach with our three-pronged search strategy that additionally uses the title-oriented nature of the corpus, which achieves an overall binary recall of 77.1%. For more details on the impact of the various search approaches on system performance, see [8]. Generating title-oriented pseudo-documents In order to maximize Watson's use of its sources, we converted some of our non-title-oriented sources, such as Shakespeare's works, the Bible, and song lyrics, into title-oriented sources. These resources were chosen because the existing title-oriented sources did not have sufficient coverage of these topics. For example, although there are Wikipedia articles for some of the well-known songs, none of them has the complete lyrics of each song. The same is true about Shakespearean characters and their lines. Non-title-oriented sources were transformed on the basis of the content of each source and the likely relationship that may be sought between the content and potential answers. For example, we converted the complete works of Shakespeare into several sets of title-oriented documents: an author-centric set where each document is a work by Shakespeare and the title is BWilliam Shakespeare[; a work-title-centric set where each document is the same as the previous set but the title is the name of the work; and a character-centric set where the lines spoken by each Shakespeare play character were extracted to form a document whose title consisted of the name of the character. Similarly, song lyrics were converted into two sets of documents, a song-title-centric collection and an artist-centric Larger sources, such as the Bible and classic books from Project Gutenberg, were processed in a similar fashion but were additionally divided into documents roughly 40 KB to 50 KB in size to improve search speed. To maximize text coherence, in dividing the document text, we preserved paragraph boundaries at all times and attempted to preserve natural chapter boundaries when possible. Although we found the process of generating

title-oriented pseudo-documents to improve overall QA performance, this source transformation process was largely the result of manual error analysis to identify source weaknesses. We believe that there is value in developing a procedure to automatically identify salient topics from a collection of non-title-oriented documents and to build title-oriented pseudo-documents on these topics from the non-title-oriented content. For example, some sources contain metadata that encode the relevant topics of each article. These metadata can be used to extract salient sentences about the topic that form Effectiveness of leveraging title-oriented sources. J. CHU-CARROLL ET AL. the basis for constructing pseudo-documents for that topic. We leave this as a possible direction for future work. Reorganizing title-oriented documents In addition to creating title-oriented documents from non-title-oriented collections, we also transformed the representation of some original title-oriented documents to make them more effective for QA. For example, in Wiktionary, a crowd-sourced dictionary resource, alternative spellings of the same term are listed as separate entries and are cross-referenced. We found that the definitions of these cross-referenced entries are often not identical, as in the example below. A set of structured activities. A leaflet listing information about a play, a game, or A performance of a show or other broadcasts on radio . . . A planned sequence of events. A sheet or a booklet that lists a schedule of events. A presentation that is broadcast on radio or television. . . . Although the definitions in one document apply equally well to the title of the cross-referenced document, with Bprogram[and Bprogramme[as two separate entries in Wiktionary, neither document is a great match for the hypothetical Jeopardy! clue BA set of structured activities or a planned sequence of events[. To increase the effectiveness of these title-oriented documents, we merged entries that are cross-referenced as a result of being alternative spellings of one another and randomly selected the shorter variation as the title for the merged document. We observed another potential source improvement that pertains to all dictionary resources. As detailed in [9], Watson employs a secondary search process, called supporting passage retrieval, to find additional evidence to help score each candidate answer. These supporting passages are used

by our context-dependent answer scorers to evaluate the goodness of the match between the passage and the question. The logical-form answer scorer, in particular, attempts to align the predicate-argument structure for the question with that for the passage to determine the likelihood that the candidate answer from that passage is the correct answer. Both the supporting passage retrieval and the logical-form answer-scoring processes work best on complete sentences that include both the candidate answer and the terms in the question. Dictionary entries, on the other hand, are represented as title-oriented documents where the document title is the dictionary term and the document text contains the definitions of the term. Although this representation is very effective for retrieving the relevant documents (dictionary entries) given a definition and therefore generating the correct answer as a candidate, it is not ideal for supporting passage retrieval and context-dependent answer scoring. To remedy this problem, we automatically rewrote each definition to include the dictionary term in a complete sentence, as shown below, where the italicized text is added to each definition. A program is a set of structured activities. A program is a leaflet listing information about a play, a game, or other activities. A program is a performance of a show or other broadcasts on radio or television. . . . The rewrite process takes into consideration the part of speech and plurality of the dictionary entry to ensure that the augmented definitions are grammatical. Although encyclopedia and dictionary sources are a good fit for many Jeopardy! questions and their title-oriented nature can be effectively used for QA, all encyclopedias and dictionaries suffer from incompleteness of data. They may not contain all the facts about a topic, and they usually lack redundancy when describing a well-known fact, i.e., encyclopedia articles and dictionary entries rarely repeat the same information in different ways. Lexical and syntactic variations facilitate the extraction and the validation of answers because if the same information is conveyed in different ways, this increases the chances that the system can successfully match one representation against the question [10]. To increase the coverage and the variation of encyclopedia articles and dictionary entries, we performed source expansion on these documents by selectively acquiring additional knowledge from It is

easy to assume that adding large amounts of Web data to a corpus will guarantee improvements in the end-to-end performance of a QA system. However, previous experiments have shown that QA system accuracy does not monotonically increase with the size of the underlying corpus. For example, Clarke et al. [4] found that large crawls of more than 50 GB were required to outperform the 3-GB reference corpus used in the TREC QA task and that QA system performance degraded if the crawl exceeded approximately 500 GB. We found that by using J. CHU-CARROLL ET AL. 4 : 5 discriminating statistical models, we can reduce the size of the retrieved Web data by two orders of magnitude and filter out noise that may adversely affect QA performance. In this section, we discuss how we select the set of documents to expand from a given collection. We then outline our source expansion algorithm to generate a pseudo-document for each selected document. For further details on the source expansion process, see [11]. Selecting high-relevance seed documents Of our title-oriented sources, Wikipedia alone contains 3.5 million documents. Obviously, some of these documents cover more popular topics than others, and expanding more popular topics such as BBarack Obama[is likely to yield more impact than expanding obscure topics such as BSaskatchewan Highway 627[, because most Jeopardy! answers are popular entities (see below). Therefore, we focus our source expansion process on the most popular topics in our title-oriented sources. We used the hyperlink metadata present in Wikipedia to determine popularity for encyclopedia topics. We consider documents that are more often referenced by other documents in Wikipedia to be more popular than those that have fewer references. The popularity of dictionary entries was estimated on the basis of their frequencies in a large collection of English documents across a variety of topics and genre. We sorted the documents by popularity in descending order and plotted the number of documents versus the number of those documents that are relevant for answering Jeopardy! questions. A document is considered relevant if its title is the answer to one of 32,000 randomly selected Jeopardy! questions outside of our train and test sets. Figure 1 shows the relevance curves for both Wikipedia and Wiktionary documents. It can be seen that the popularity-based relevance

rankings significantly outperform random baselines, which are illustrated by For Watson, we selected the top ranked encyclopedia and dictionary documents as seeds and generated pseudo-documents for each topic on the basis of relevant text nuggets retrieved from the Web. Details of the parameter settings for this process and the resulting impact of expanding these documents on Watson’s end-to-end QA performance are discussed in the section on experimental Source expansion algorithm The input of the source expansion algorithm is a collection of documents in which each document contains information about a distinct topic. We refer to these documents as seed documents or simply seeds, and we refer to the entire collection as the seed corpus. Examples of preexisting seed corpora that are suitable for source expansion are title-oriented collections such as encyclopedias and dictionaries. For each seed, a new pseudo-document is generated, which contains salient information retrieved from large external sources. By expanding the seeds, we gather additional relevant content about their topics, as well as paraphrases of information already covered in the original Seed documents are expanded in a four-stage pipeline, illustrated in Figure 2 using the Wikipedia article about Tourette syndrome as an example. For each seed, the source expansion process retrieves related documents from a large external source, such as the Web (retrieval stage in Figure 2). The retrieved documents are split into paragraph-length text nuggets (extraction stage), and their relevance to the topic of the seed is estimated using a statistical model (scoring stage). Finally, a pseudo-document on that seed topic is compiled from the most relevant text nuggets (merging stage). In the following, we describe each pipeline stage in more detail. 1. RetrievalVFor each seed document, the retrieval component generates a query, performs a Yahoo!** search for related content, and fetches up to 100 Web pages. For encyclopedic seed documents, we use the document titles as queries and retrieve all resulting HTML and text documents. For dictionary seed documents, we add the word Bdefine[to the dictionary term and retrieve only those documents that contain the term in their title or URL to focus search results on definitions of these terms. Relevance of Wikipedia and Wiktionary seeds for the Jeopardy! task. (Used with permission from N. Schlaefer, J. Chu-Carroll, E.

Nyberg, J. Fan, W. Zadrozny, and D. Ferrucci. BStatistical source expansion for question answering,[in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, 2011.) J. CHU-CARROLL ET AL. 2. ExtractionVThe extraction component splits the retrieved Web documents into paragraph-length text nuggets. For HTML documents, structural markup can be used to determine boundaries. Typical text nuggets are HTML paragraphs, list items, or table cells. They range in length from short fragments (e.g., Bborn: 1968[) to narratives of multiple sentences. Since, in the merging stage, individual nuggets are selected and added to the pseudo-document, the text nuggets should ideally be self-contained and be either entirely relevant or irrelevant. Longer text nuggets are more likely to be self-contained but are often only partially relevant. We found that nuggets based on structural markup represent a good tradeoff between the two properties. 3. ScoringVThe core component of our source expansion process is a statistical model that scores the relevance of extracted text nuggets with respect to the topic of the seed document. In order to train the model, we first extracted a set of text nuggets from the top 100 Web pages for each of 15 Wikipedia seed documents. These text nuggets are then manually labeled on the basis of a binary classification of relevant or irrelevant with respect to the original seed document. We identified features of text nuggets that are predictive of their relevance to a seed document. The features can be broadly categorized into three classes: topicality features (features that relate the content of the text nugget to the content of the seed document, e.g., by comparing the word distributions of the nugget and the seed using cosine similarities or language modeling techniques), search features (features based on Web search results, such as Yahoo! search rank), and surface features (lexical features extracted from the text nugget alone, such as nugget length). Among the most predictive features are topicality scores estimated using language models and term-frequency–inverse-document-frequency (tf–idf) term weighting.2 Furthermore, our initial experiments showed that text nuggets are not independent of one another; rather, they are more likely to be relevant if surrounded by other relevant nuggets. To capture this

relationship, we added to the feature set of each text nugget features of directly adjacent nuggets. The annotated data set and features are used to train a logistic regression model that assigns to each new text nugget an estimate of its relevance to a given seed document. 4. MergingVThe merging component ranks all the text nuggets for a given seed document by their relevance scores in descending order. A filter reduces lexical redundancy by removing nuggets whose keywords are entirely subsumed by higher ranking nuggets or the seed. Nuggets are also dropped if their relevance scores are below an absolute threshold or if the total character length of all nuggets exceeds a threshold relative to the length of the seed. The remaining nuggets are compiled into a pseudo-document that forms a separate document from the seed on the same topic. 2The tf–idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection. Source expansion pipeline on the Wikipedia seed for Tourette syndrome. (Used with permission from N. Schlaefer, J. Chu-Carroll, E. Nyberg, J. Fan, W. Zadrozny, and D. Ferrucci. BStatistical source expansion for question answering,[in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, 2011.) J. CHU-CARROLL ET AL. 4 : 7 Source expansion examples To better understand how source expansion improves QA performance, consider the question BThis is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)[.3 The expanded version of the Wikipedia article about Tourette syndrome in Figure 2 contains the following nuggets, which originated from different Web pages and, in conjunction, almost perfectly cover the question terms (in italics below). 1. Rare neurological disease that causes repetitive motor and vocal tics. 2. The first symptoms are usually involuntary movements (tics) of the face, the arms, the limbs, or the trunk. 3. Tourette syndrome is a neurological disorder characterized by repetitive stereotyped involuntary movements and vocalizations called tics. 4. The person afflicted may also swear or shout strange words, grunt, bark, or make other loud sounds. The expanded document is retrieved by Watson, enabling it to correctly answer the

question. Note that this article was expanded, along with many other Wikipedia articles in the source expansion process, without knowing yet which articles will be relevant for answering questions. Although our experimental evaluation below shows that our source expansion process yields significant end-to-end performance improvement, manual examination of the pseudo-documents reveals that they do include noisy text nuggets that may be better removed or associated with a different topic. For example, for the topic Aldous Huxley, the following nugget, which is arguably more relevant to Thomas Henry Huxley, was selected. Thomas Henry Huxley (1825-1895)VEnglish biologist, who wrote on biology as a specialist and as a popularizer. He also published books on education, philosophy, and theology. . . The writer Aldous Huxley (1894-1963) was his brother. We expect that enhancements in the features selected for nugget scoring will help address the issue of nugget topic versus the pseudo-document topic, as in the example above, as well as other issues such as facts versus opinions. We leave these extensions for future work. We conducted experiments to measure the impact of source acquisition, transformation, and expansion. The baseline used for comparison is Watson's end-to-end QA performance using Wikipedia only as a source for search and candidate generation [8], as well as supporting passage retrieval [9]. Our first experiment evaluates the impact of source acquisition and transformation by incorporating all the sources that we acquired based on our iterative error analysis process. Some of these sources were additionally transformed into title-oriented sources and/or reorganized to be best used by Watson. Our second experiment adds to the corpus in the first experiment the results of source expansion. In Watson's full source expansion process, we selected the top 300,000 most popular Wikipedia documents and the 100,000 most popular Wiktionary documents as seeds (indicated by vertical lines in Figure 1). We expanded these Wikipedia and Wiktionary seed documents, as well as a variety of smaller encyclopedias obtained during the source Our test set contains 3,508 previously unseen Jeopardy! questions. These questions are from a randomly selected set of 66 Jeopardy! games, from which we excluded audio/visual questions and selected special questions. The excluded special questions have made-up

phrases as answers, such as BTom Cruise Control[in the BBEFORE AND AFTER[category or Blevel devil[in the BRHYME TIME[category [12]. These answers are not expected to be directly found and/or justified from textual resources and are therefore not of interest in the evaluation of our source coverage. The metrics that we used to compare the different system configurations include the following: 1) candidate binary recall, i.e., the percentage of questions for which the correct answer was produced as a candidate answer; 2) accuracy, i.e., the percentage of questions correctly answered, and 3) Precision@70, i.e., the precision of the system if it answers only the top 70% of questions for which it is most confident in its top answer. Table 2 shows the results of our experiments. The first column in the table shows our baseline performance using the 3.5 million Wikipedia documents for search, candidate generation, and supporting passage retrieval, along with Watson's full suite of answer scorers for ranking these candidates.4 The second column shows the results of our source acquisition and source transformation processes. After source acquisition and transformation, the corpus size is roughly doubled, and the additional content yielded more than 7% improvement in all three evaluation metrics. Note that in the Wikipedia-only baseline, Watson achieved an accuracy of 59.0% with a 77.1% candidate binary recall. This indicates that for questions with a correct answer in the candidate list, Watson is able to rank the correct answer in first place 76.5% of the time. This percentage increases to 78. 7% in our first experiment. This increase indicates that the new sources not only enabled Watson to generate correct answers as candidates for an additional 7.6% of all questions 3Adapted from the TREC-8 test collection (Question 8). 4Most scorers consult external resources other than Wikipedia for scoring candidate answers. We did not remove those external resources from Watson's answer-scoring process. J. CHU-CARROLL ET AL. but also provided better passages to help Watson score the correct answers more accurately. The last column in the table shows the cumulative results of source acquisition, transformation, and expansion. The source expansion process more than doubles the corpus size of all sources combined before expansion. The performance metrics show that source expansion further improves upon the

results of source acquisition and transformation, with a roughly 3% gain in all metrics. As in the previous experiment, the gain in accuracy is greater than that for candidate binary recall, again demonstrating the dual impact of the newly acquired sources. Note that the evaluation results are order dependent because there is an overlap in the new content added by multiple approaches. We ran an additional experiment in which source expansion was applied to Wikipedia without source acquisition and transformation. Those expanded Wikipedia sources achieved a level of performance comparable with the second column in the table. A high coverage and effective reference corpus is a fundamental element of any high-performing retrieval-based QA system. However, the topic of source acquisition and engineering has received very little attention because organized evaluation efforts, such as TREC [1], CLEF [2], and NTCIR [3], have traditionally included a reference corpus as part of the task specification. Some systems [4, 13, 14] have used supplementary sources, such as the Web, in addition to the given corpus, but not all have found additional sources to be helpful for improving QA performance. Clarke et al. [4] evaluated the performance impact of Web crawls on TREC QA data. They found that large crawls of more than 50 GB were required to outperform the 3 GB reference corpus used in TREC and that performance degraded if the crawl exceeded approximately 500 GB. In contrast, we were able to incrementally improve Watson's performance with 12 GB of selected sources compared with a baseline of 13 GB of Wikipedia documents using an iterative analysis and an acquisition process to selectively add resources in targeted domains to our collection. Although encyclopedias have been used as a source for QA in the past [2, 15, 16], to our knowledge, these efforts used Wikipedia documents in a similar way to how systems leveraged newswire corpora. Our approach differs from theirs by recognizing the unique characteristics of title-oriented documents and how they can be used effectively to increase QA performance. This observation also led us to transform non-title-oriented sources to title-oriented sources to further impact our search and candidate generation strategies that target title-oriented sources. Balasubramanian and Cucerzan [17] proposed an algorithm that, similar to our source expansion approach, generates documents

about given topics from Web content. The usefulness of sentences extracted from Web pages is determined with aspect models built from Web search logs. These aspect models consist of terms that frequently co-occur with a given topic or related topics in the query logs. In contrast, our source expansion method leverages the content of existing seed corpora to model topicality. Furthermore, a variety of techniques have been proposed for document expansion by adding related terms extracted from relevant documents [18], by propagating anchor texts associated with hyperlinks to the target documents [19], and by adding terms in the vicinity of citations in scientific publications to the referenced articles [20]. These approaches primarily focused on adding relevant key words and phrases to the original document. Although this enhances search performance, the additional content produced by these approaches cannot have an impact on answer scoring as we have demonstrated our expanded sources did. In this paper, we have described three procedures for acquiring and enhancing Watson's textual resources to support high-performance QA, i.e., source acquisition, transformation, and expansion. Source acquisition involves a domain analysis for identifying one or more initial corpora well suited for the task. Additional sources have been subsequently acquired based on the results of an iterative error analysis process to cover salient topics deemed to be gaps in existing resources. In Watson, Wikipedia has been adopted as the initial corpus. General-purpose corpora, such as additional encyclopedias, dictionaries, and newswire articles, as well as domain-specific corpora, such as Impact of source acquisition and engineering on end-to-end QA performance. J. CHU-CARROLL ET AL. 4 : 9 Shakespeare's works, the Bible, and classic literature, have been subsequently acquired. Source transformation involves analyzing the system's utilization of its available sources and identifying ways to transform existing content so that it can be more easily used by the system. To this end, we extracted information from select non-title-oriented sources and built title-oriented documents on the basis of the likely relationship that may be sought between the extracted content and potential answers. Finally, we have applied our source expansion procedure to key encyclopedias and dictionaries to build pseudo-documents on popular topics

and dictionary terms based on Web content. These pseudo- documents complement the original seed documents in coverage and support them in providing lexical and syntactic variations of existing content. We expect that this set of procedures can be adapted and used when developing new QA systems or when porting an existing system to a We have evaluated the impact of Watson's source acquisition and engineering procedures on end-to-end QA performance. Overall, this process increased the total corpus size from Wikipedia's original 13 to 59 GB. The additional content resulted in an increase in 10.4% in candidate binary recall and 11.4% in end-to-end QA accuracy. These performance results provide evidence that the new content was effective in both providing increased coverage, as evidenced by the increase in candidate recall, and enabling Watson to better utilize the available information, as evidenced by the improvement in Business Machines Corporation in the United States, other countries, Productions, Inc., Wikimedia Foundation, Michael S. Hart, or Yahoo!, Inc., in the United States, other countries, or both. 1. E. Voorhees, "Overview of the TREC 2002 Question Answering Track," in Proc. 11th Text REtrieval Conf., 2002, 2. D. Giampiccolo, P. Forner, A. Pen˜as, C. Ayache, D. Cristea, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, and R. Suteliffe, "Overview of the CLEF 2007 multilingual QA track," in Advances in Multilingual and Multimodal Information Retrieval, C. Peters, V. Jijkoun, T. Mandl, M. Henning, D. W. Oard, P. Anselmo, V. Petras, and D. Santos, Eds. 3. Y. Sasaki, C. Lin, K. Chen, and H. Chen, "Overview of the NTCIR-6 cross-lingual question answering task," in Proc. 6th 4. C. Clarke, G. Cormack, M. Laszlo, T. Lynam, and E. Terra, "The impact of corpus size on question answering performance," in Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. 5. J. Fan, A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci, "Automatic knowledge extraction from documents," 6. J. W. Murdock, A. Kalyanpur, C. Welty, J. Fan, D. A. Ferrucci, D. C. Gondek, L. Zhang, and H. Kanayama, "Typing candidate 7. A. Kalyanpur, B. K. Boguraev, S. Patwardhan, J. W. Murdock, A. Lally, C. Welty, J. M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, Y. Pan, and Z. M. Qiu, 8. J. Chu-Carroll, J. Fan, B. K. Boguraev, D. Carmel, D. Sheinwald, and C. Welty, "Finding needles in the haystack: Search 9. J. W. Murdock, J. Fan, A. Lally, H. Shima, and B. K. Boguraev, 10. S. Dumais, M. Banko, E. Brill, J.

Lin, and A. Ng, “Web question answering: Is more always better?” in Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 11. N. Schlaefer, J. Chu-Carroll, E. Nyberg, J. Fan, W. Zadrozny, and D. Ferrucci, “Statistical source expansion for question answering,” in Proc. 20th ACM Conf. Inf. Knowl. Manage., 12. J. M. Prager, E. W. Brown, and J. Chu-Carroll, “Special questions 13. C. Clarke, G. Cormack, and T. Lynam, “Exploiting redundancy in question answering,” in Proc. 24th Ann. Int. ACM SIGIR 14. B. Katz, J. Lin, D. Loreto, W. Hilderbrandt, M. Bilotti, S. Felshin, A. Fernandes, G. Marton, and F. Mora, “Integrating web-based and corpus-based techniques for question answering,” in Proc. 15. J. Kupiec, “MURAX: A robust linguistic approach for question answering using an online encyclopedia,” in Proc. ACM SIGIR 16. D. Ahn, V. Jijkoun, G. Mishne, K. Muller, M. de Rijke, and S. Schlobach, “Using Wikipedia at the TREC QA track,” in 17. N. Balasubramanian and S. Cucerzan, “Automatic generation of topic pages using query-based aspect models,” in Proc. 18. A. Singhal and F. Pereira, “Document expansion for speech retrieval,” in Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 19. N. Craswell, D. Hawking, and S. Robertson, “Effective site finding using link anchor information,” in Proc. ACM SIGIR Conf. Res. 20. J. O’Connor, “Citing statements: Computer recognition and use Received July 18, 2011; accepted for publication January 12, 2012 J. CHU-CARROLL ET AL. Watson Research Center, Yorktown Heights, NY 10598 USA the Semantic Analysis and Integration Department at the T. J. Watson Research Center. She received the Ph.D. degree in computer science she spent 5 years as a Member of Technical Staff at Lucent Technologies Bell Laboratories. Dr. Chu-Carroll’s research interests are in the area of natural-language processing, more specifically in question-answering and dialogue systems. Dr. Chu-Carroll serves on numerous technical committees, including as program committee co-chair of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 2006 and as general chair of NAACL HLT 2012. Dr. Fan is a Research Staff Member in the Semantic Analysis and Integration Department at the T. J. Watson Research Center, Yorktown University of Texas at Austin in 2006. He is a member of the DeepQA Team that developed the Watson question-answering system, which defeated the

two best human players on the quiz show Jeopardy!. Dr. Fan is author or coauthor of dozens of technical papers on subjects of knowledge representation, reasoning, natural-language processing, and machine learning. He is a member of Association for Computational Linguistics. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (nico@cs.cmu.edu). Mr. Schlaefer is a Ph.D. candidate in the School of Computer Science at Carnegie Mellon University. His research focus is the application of machine learning techniques to natural-language processing tasks. He developed algorithms that enable question-answering (QA) systems to find correct answers even if the original information sources contain little relevant content, as well as a flexible architecture that supports the integration of such algorithms. He is the primary author of OpenEphyra, one of the most widely used open-source QA systems. He also developed a statistical approach for expanding Watson's knowledge sources with related information from the Web, making Fellowships in 2009 and 2010. com). Dr. Zadrozny is a Research Staff Member in the T. J. Watson Research Center. He received an M.S. degree in mathematics from Warsaw University, Warszawa, Poland, in 1976 and a Ph.D. degree in mathematics from the Polish Academy of Science in 1980. He joined processing technologies. In the Watson/DeepQA project, he was responsible for acquiring and preparing textual resources. He also worked on applying this question-answering technology to help-desk support and medical diagnosis. Dr. Zadrozny is author or coauthor of approximately 40 patents and 50 technical papers. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), the Association for Computing Machinery (ACM), and the Association for the Advancement of Artificial Intelligence (AAAI). J. CHU-CARROLL ET AL. 4 : 11