Passage Retrieval in a Question Answering System

Juri Pyykkö, Rebecka Weegar, Pierre Nugues

Lund University, Department of Computer Science
S-221 00 Lund, Sweden
juri.pyy@gmail.com, rebecka.weegar@gmail.com, pierre.nugues@cs.lth.se

Abstract

In this paper, we describe a passage retrieval component for a questioning answering system and we evaluate its performance on Swedish documents. We used a corpus of questions and answers transcribed from the Swedish board game *Kvitt eller dubbelt* and, as source for the passages, we used the articles of the Swedish version of Wikipedia. We show that Wikipedia is a suitable knowledge source to answer the game questions. For answers consisting of one word, we could extract passages of text that contained the right answer for 91% of the questions and when these answers corresponded to an entity, we found and ranked correct answers for up to 75% of the questions.

1. Introduction

IBM Watson (Ferrucci, 2012) set a milestone in the field of question answering, winning over all its human contestants. However, IBM Watson is dedicated to English, making the replication of such a system in another language a challenge. This paper investigates if the techniques IBM Watson used could be adapted to Swedish. As question and answer data set, IBM Watson used the *Jeopardy!* quiz show and as knowledge source, the English version of Wikipedia, *inter alia*. In this experiment, we used Swedish equivalents and we measured the performance of passage retrieval in a baseline question-answering system.

2. A Dataset of Questions and Answers

We transcribed a data set of questions and answers from the *Kvitt eller Dubbelt – Tiotusenkronorsfrågan* board game (Thorsvad and Thorsvad, 2005). This game consists of 385 cards divided into seven categories. Each card has a unique number to identify it; a category that specifies the general theme of the question; a difficulty that can either be '–' for normal or '*' for easy; and a card title. A card contains six questions, their answers, and possibly a complement or clarification to the answer. Questions also have value credits that we ignored in our experiments. Table 1 shows an example of question. In addition to the original card content, we annotated the answers with a type derived from Li and Roth (2002); the most notable type being *entity* to designate things.

Property	Value
Card #	164
Category	Djur och natur
Difficulty	–
Card title	I terrängen
Question	Är en myr ett fuktigt eller torrt område i naturen?
Value	5000
Answer	Det är ett fuktigt område
Clarification	(våtmark)
Answer type	location

Table 1: Question example.

3. Passage Retrieval

Most question answering systems feature a text retrieval step that searches passages relevant to the question and orders them by similarity. As base of documents, we used a data dump of the Swedish Wikipedia (Wikimedia Foundation, 2014), where we removed the markup code (Attardi and Fuschetto, 2013). We segmented the articles into paragraphs and we indexed them using Lucene (Apache Software Foundation, 2014).

Given a question, we carried out the passage retrieval, where the passages correspond here to Wikipedia paragraphs, using Lucene's text retrieval functionality. Lucene retrieves and ranks the passages using a combination of a Boolean model and the BM25 vector space model (Zaragoza et al., 2004). The maximum number of matching paragraphs is an adjustable parameter of the search. We used Lucene's language-dependent analyzers to stem words and remove stop words during the indexing and search steps. We applied the Swedish version of these tools to index and search the passages.

4. Results and Evaluation

We first assessed the completeness of Wikipedia relative to the *Kvitt eller dubbelt* questions and the Lucene ranking function to retrieve answer candidates. We used a subset of 1,374 questions together with the card title that we submitted as queries. We considered that an answer was present in a passage if we could find the exact or lowercased string of the answer in the text. Figure 1 shows the results we obtained relative to the number of paragraphs the indexer returns. The figures range from 14% of the answers when setting the cutoff to one paragraph to 74% when keeping the 300 paragraphs most similar to the question. We could not significantly improve this figure with more paragraphs.

Some answers in *Kvitt eller dubbelt* consist of sentences or lists of alternative answers that cannot be found by our naïve matching method. To quantify them, we restricted the data set to answers consisting of only one word. On average, we could retrieve 10-15% more answers (Figure 1).

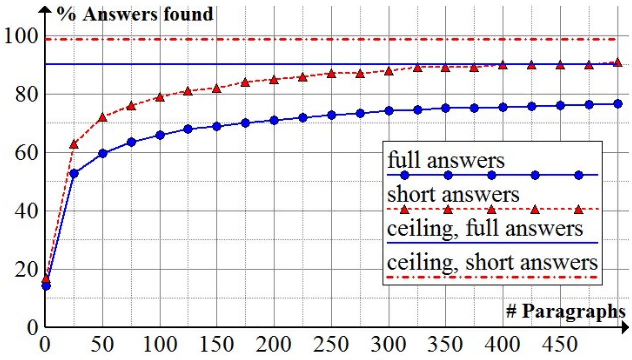

Figure 1: Ceiling of answers present in Wikipedia.

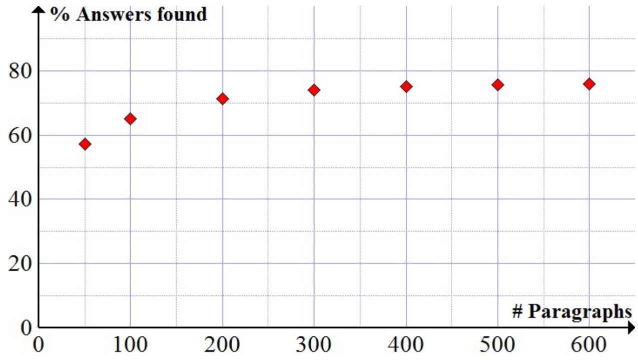

Figure 2: Found answers among tagged candidates.

4.1 Ceiling

We then estimated the percentage of answers to the *Kvitt eller dubbelt* questions that the whole Wikipedia contains. We computed this ceiling using a set of 1,374 questions. As before, we considered a question answerable when we could match its answer in at least one of the returned passages. We used a lowercase text without stemming or lemmatization.

When using all the questions (any type of answer without the clarification field), we could retrieve an answer for 90.1% of the cases. For the 1,013 questions that had one-word answers, the answer was present for all but 13 questions (98.7%). 449 of the questions had answers consisting of one word and were annotated as an entity. For these questions, we could retrieve 98.5% of the answers. The conclusion is that Wikipedia has a decent coverage with a no-answer rate of 10% for questions with any kind of answers; it is of 1.3% for one-word answers (Fig. 1).

4.2 Generating and Ranking Candidates

We evaluated a baseline candidate extraction step to the retrieved passages, where we limited the search to questions where the answer consisted of one word and was classified as an entity. We applied a part-of-speech tagger (Östling, 2013) to the passages and to match the *entity* type, we extracted the nouns from the retrieved passages. The POS tagger lemmatizes the words and we counted the lowercase word lemmas.

Figure 2 shows the percentage of answers found relative to the number of paragraphs retrieved for this setup. When retrieving 300 paragraphs or more, we found about 75% of correct answers. As an example, when setting the paragraph cutoff to 150, the query

> Vad kallas hundens ungar?
> ‘What do you call a baby dog?’

returns 1,783 words of which 384 are distinct nouns. The correct lemmatized answer, *valp*, is present four times and is the 8th most frequent noun in the retrieved text.

Figure 3 shows the cumulative distribution for the ranking of candidate answers with different paragraph cutoffs. Each point in the figure corresponds to the percentage of correct answers having this rank or a better one. With a paragraph cutoff of 50, an answer was found for 57.2% of the questions and the answer was among the top 100 nouns/candidates for 55 percent of the questions. If the search was extended to 400 paragraphs, the respective percentages were 75% and 41%.

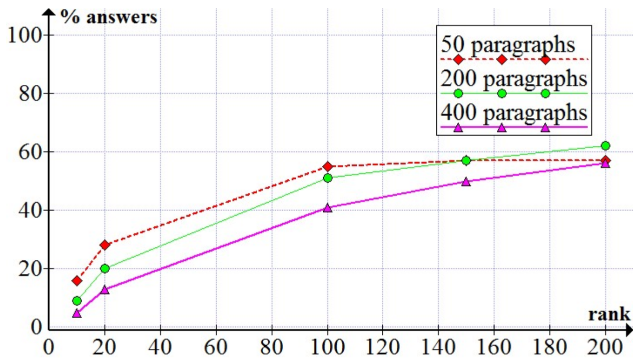

Figure 3: Ranking of answer candidates.

5. Conclusion

We designed an experimental setup to assess the suitability of Wikipedia as underlying knowledge source aimed at answering questions from the *Kvitt eller dubbelt* game. For one-word answers, we managed to extract passages that contained the answer for 91% of the questions and, using baseline information retrieval techniques, we could find and rank the answers for up to 75% of the questions where answers were entities. This hints at the validity of Wikipedia as a reference collection of articles for a question answering system although other collections of documents would have to supplement it for the remaining 25% of unanswered questions.

Acknowledgments

This research was supported by Vetenskapsrådet under grant 621-2010-4800 and the *Det digitaliserade samhället* program.

References

Apache Software Foundation. 2014. Lucene. lucene.apache.org.

Giuseppe Attardi and Antonio Fuschetto. 2013. Wikipedia extractor. Medialab, University of Pisa.

David Angelo Ferrucci. 2012. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1:1 –1:15, May-June.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of COLING '02*, pages 1–7, Taipei.

Robert Östling. 2013. Stagger: an open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

Karin Thorsvad and Hasse Thorsvad. 2005. *Kvitt eller Dubbelt*. SVT Tactic, Stockholm.

Wikimedia Foundation. 2014. Wikipedia. http://dumps.wikimedia.org/.

Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, and Stephen Robertson. 2004. Microsoft Cambridge at TREC-13: Web and HARD tracks. In *Proceedings of TREC-2004*.