ADAPT.DCU at TREC LiveQA: A Sentence Retrieval based Approach to Live Question Answering

Dasha Bogdanova, Debasis Ganguly, Jennifer Foster, Ali Hosseinzadeh Vahid

ADAPT centre, School of Computing, Dublin City University
Dublin, Ireland

{dbogdanova,dganguly,jfoster,avahid}@computing.dcu.ie

Abstract

This paper describes the work done by ADAPT centre at Dublin City University towards automatically answering questions for the TREC LiveQA track. The system is based on a sentence retrieval approach. In particular, we first use the title of a new question as a query so as to retrieve a ranked list of conceptually similar questions from an index of previously asked on "Yahoo! Answers". We then extract the best matching sentences from the answers of the retrieved questions. In order to construct the final answer, we combine these sentences with the best answer of the top ranked (most similar to the query) question. When no pre-existing questions with sufficient similarity with the new one can be retrieved from the index, we output an answer from a candidate set of pre-generated answers based on the domain of the question.

1 Introduction

The task of automated Question Answering (QA) has been frequently addressed previously, including the TREC competitions of 1999-2004. However, most existing work focused only on factoid questions, that usually require a named or a numerical entity as an answer. The research on answering non-factoid questions, such as manner or reason questions (e.g. a factoid question *Who is the prime minister of Ireland?* versus a non-factoid *How is the prime minister of Ireland elected?*), is rather piecemeal.

Several attempts towards non-factoid question answering were made. For example, Higashinaka and Isozaki (2008) present a learning-to-rank approach to answer Japanese *why* questions. The work in Surdeanu et al. (2011) address the problem of ranking answers to non-factoid *how* questions from Yahoo! Answers. The authors use a wide variety of features including translational, similarity and web correlation features. Several other studies focus on the task of answer reranking for non-factoid *how* and *why* questions, including (Jansen et al., 2014; Sharp et al., 2015; Fried et al., 2015). However in neither of the mentioned studies the questions were coming live from real users.

The TREC 2015 LiveQA track, unlike previous QA tracks, involves answering real questions from Yahoo! Answers in real time. each participant needed to submit a web service application that on receiving a question responds with an answer of no more than 1000 characters. The answer had to be provided within 60 seconds. The questions, being sampled from a stream of real Yahoo! Answers questions, were much more diverse than in past QA tracks. In fact, the questions included not only factoid but also manner, opinion, advice and many other types of questions. All questions submitted to the systems had a title, a body (if any), and a user-reported category from the following list: *Arts & Humanities*, *Beauty & Style*, *Computers & Internet*, *Health*, *Home & Garden*, *Pets*, *Sports* and *Travel*.

This paper describes our participation in the TREC 2015 LiveQA track. We undertake a sentence retrieval approach over an indexed collection of 4.48M existing Yahoo! Answers questions[1] crawled in 2007. In particular, we use the title of a new question as the query to retrieve a ranked list of *conceptually similar* questions previously asked on the forum and then extract the best matching sentences from the answers. When no pre-existing questions with a sufficient degree of overlap with the new one can be retrieved from the index, we output an answer from a candidate set of pre-generated answers based on the domain of the question.

2 Approach

In this section, we describe our approach to live question answering in detail. We start by describing the data that is used to construct the archived index of previously asked questions on Yahoo! Answers and then follow it up with a description of how the index is used to retrieve similar questions and extract answer snippets from them.

2.1 Index Construction

To build our index, we use the L6 dataset[2] of Yahoo! Answers. This data set contains about 4,48M questions along with their answers. We use Lucene,[3] an open-source information retrieval system implemented in Java, to build up the index. We represent each document as a set of individual fields, each field comprising a set of terms. The content of each individual field is extracted from the respective XML tags of each document. The field-based indexing ensures that the contributions from the similarities of each field can be combined to constitute the overall similarity value between a new question and the existing ones. Document collection statistics are shown in Table 1.

Field Name	Description	Vocab Size
MainCategory	Top level category name of the question	179
SubCategory	Sub category name	1546
Category	Category description	2919
Title	Title of a question	945,708
Body	Body of a question	601,862
BestAnswer	The text of the best answer for a question	2,039,651
AllAnswers	Concatenated text for all (but the best) answers for a question	5,123,702

Table 1: Summary of the individual fields of the indexed documents comprising the Yahoo! Answers collection. The total number of documents is $4,483,032$.

2.2 Retrieval

Given a new question, we use the title[4] of the new question as the query to retrieve a ranked list of similar questions from the index. While retrieving from the index, we make use of the Field-based Language Modeling (FLM) with Jelinek Mercer similarity (Zhai and Lafferty, 2001) as shown in Equation 1. In Equation 1, λ_i is the weight assigned to the i^{th} field, $i = 1, \ldots, F$, F being the number of fields, $P(t|C)$ is the maximum likelihood estimate of sampling the term t from the collection, and $P(t|f_i, d)$ is the probability of sampling the term from the field f_i of document d.

[1] L6 dataset suggested as the training set by the task organizers and available on request http://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[2] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[3] http://lucene.apache.org/

[4] In our initial experiments, we also used the body of a question for query formulation but it produced worse results.

$$P(d|q) = \prod_{t \in q}(1 - \sum_{i=1}^{F} \lambda_i) P(t|C) \sum_{i=1}^{F} \lambda_i P(t|f_i, d) \tag{1}$$

The fields that we use in particular for the retrieval are the “MainCategory”, “Title” and the “Body” fields, setting equal weights of 0.2[5] for each. For constructing queries, we use the “Title” field of the new question only, because after some initial experiments we noticed that including terms from the “Body” field of the new question is often prone to introducing query drift.

After obtaining the ranked list of questions from the set of indexed questions, we explore three possible strategies for formulating an answer to the new question by using the information extracted from these similar questions.

1. Using the best answer of the most similar question as the answer to the new question.

2. Extracting sentences with highest similarity with the query, i.e. the title of the new question, and then concatenating them together.

3. A combination of the two approaches, where the final answer contains the first two sentences of the output of the first approach followed by the output of the second approach.[6]

In order to obtain sentences from the answers that are most similar to the query, we build an in-memory index of the sentences extracted from the “BestAnswer” and the “AllAnswers” fields from the top 10 retrieved set of documents. For sentence splitting, we use the Stanford NLP toolkit.[7] The retrieved ranked sentences are then included in the generated answer in decreasing order of their similarity with the query. Too short sentences, i.e. the ones less than 10 characters, are discarded.

2.3 Pre-generated Responses

We observed that many questions were looking for an advice or approval rather than information. For instance, questions such as *Am I pretty?*, *is it okay to wear leggings to work?*, etc. are advice seeking in nature lacking definite and precise answers. Thus, when no pre-existing questions with a sufficient similarity with the new one can be retrieved from the index, we output an answer from a candidate set of pre-generated answers based on the domain of the question. These canned answers are not informative but rather comforting. For example, the pre-generated response for the Beauty & Style category was *Don't worry about this! You are beautiful!* To estimate the similarity between the retrieved question and the query, which we use an approximation for the reliability of the retrieved answer, we calculate the following value:

$$rel(query, top\_q) = max(nsimqq(query, top\_q), cos(query, top\_a)) \tag{2}$$

where top_q and top_a are the top retrieved question and its best answer respectively; cos is the cosine similarity; and $nsimqq$ is the normalized BM25 similarity, i.e.

$$nsimqq(query, top\_q) = \frac{BM25(query, top\_q)}{BM25(query, query)} \tag{3}$$

If the certainty value $rel(query, top\_a)$ (see Equation 2) is lower than a predefined threshold th (which after initial experiments was set to 0.2), we first check if the asked question follows the *yes/no* pattern. Our approach of checking whether a question is objective type is simple and computationally effective. More specifically, we check if a question starts with *do/does/are/am/is* etc. In this case we

[5] This value was tuned based on initial experiments

[6] We decide to extract first two sentences of the best answer, as according to our observations, in most cases they contain most useful information.

[7] `http://nlp.stanford.edu/nlp/`

simply reply *yes* or *no*. If the question does not follow this pattern, we opt for returning a pre-generated response associated with the top-level category of the question. The list of pre-generated responses was prepared manually, for each category there are 1-3 canned responses. In case there are two or more canned responses for the category, the answer is chosen randomly.

3 Experiments

For the purpose of developing our system, we used the dataset of 1000 questions provided by the task organizers as a semi-official development set. This dataset was crawled from Yahoo! Answers in 2013. It contains questions from the predefined eight categories (the number of questions per category varies from 26 *(Home & Garden*) to 296 (*Health*)).

For internally testing our approach, we select 60 questions from this dataset, trying to make sure that the subset includes questions of different types and different levels of detail. We have manually evaluated the three approaches described in Section 2.2, i.e. (1) extracting the best answer of the most similar question; (2) extracting sentences with highest similarity with the query; and (3) a combination of the two approaches. At the moment of submission, the official evaluation guidelines were not completely clear, so we roughly followed the scheme described in Table 2.

The first approach relies purely on the user-provided best answers, which unfortunately are not always reliable, even if the retrieved question is semantically equivalent to the query (for example, the best answer to the question *I want a phrase 'welcome, sit, goodbye' in all the Indian languages such as Gujarati, Bengali, Assamese, Punjabi...*[8] is *good luck*).

The obvious drawback of combining answers from different sources and/or combining the outputs of several systems is the possible lack of coherence. However, our observation was that answers generated in such way were more likely to contain useful information. We have selected for the final submission the combined system with the highest score of 2.95.

4 Results

During the competition, the systems stayed live for 24 hours and received 1340 questions. Later, some questions were removed, and the evaluation was done on a subset of 1087 questions. The runs were evaluated by NIST assessors, each answer was assigned a score from 1 (bad or unreadable) to 4 (perfect). The following evaluation metrics were calculated:

- $avgScore$(0-3): the average score over all questions transferring the scale to $(0-3)$.
- $succ@i+$: the number of questions with score i or above ($i \in 2..4$) divided by the total number of questions.
- $prec@i+$: the number of questions with score i or above ($i \in 2..4$) divided by number of questions answered by the system.

The runs were ranked according to the $avgScore$ value. The evaluation metrics for our system are reported in Table 3. The $avgScore$ of our system is slightly below the average score computed over all runs submitted to the track. Our system was ranked 11 out of 21 participating systems. The system answered all the questions, so the $succ@1+$ has the maximum value. The percentages of fair answers – $succ@2+$ and $prec@2+$ (which are the same in case of our system since it provided an answer to all questions) are higher than average (0.290 versus 0.262), while the percentages of answers with higher scores – $succ@3+$, $succ@4+$ and $prec@3+$, $prec@4+$ – are below average. One possible explanation for that is that many answers provided by our system did not get scores better than *fair* due to the lack of coherence, discussed in Section 3.

[8] https://answers.yahoo.com/question/index?qid=1006010500264

Score	Meaning	Example
5	Perfect answer	*Contact your local or state dental association and and see if there are any dentists who provide free or reduced cost care for low-income, disabled or senior patients. In some areas, you can reach them now by dialing 2-1-1 for “non-emergency information.” Go to a dental school, if there is one near you, for reduced costs. If you are a senior citizen, call your local Area Agency on Aging or Office on Aging.*
4	Contains the answer but also may contain other non-relevant information	*Contact your local or state dental association and and see if there are any dentists who provide free or reduced cost care for low-income, disabled or senior patients. Call your local health department and ask about health centers that provide dental care for free or on a sliding scale. there aint that much to pulling a tooth. I needed 4 teeth pulled!*
3	Relevant and partially useful answer	*Not all states have D.D.S. programs. I think a different system would be a WONDERFUL idea. In some areas, you can reach them now by dialing 2-1-1 for “non-emergency information.*
2	Relevant, but not useful answer	*Please go to the dentist or wait until you can live with the fact that you will need to be numbed to have the teeth pulled.*
1	Completely irrelevant and useless answer	*you nailed it!*

Table 2: Evaluation guidelines we followed during internal evaluation and example answers to "*I need a tooth pulled now but don’t have the money for it?*"

5 Discussion and Future Work

As reported in Section 4, our automatically generated answer consists of a combination of the best answer of the top retrieved question and most similar sentences from the answer of several similar questions. Combining the output of several systems sometimes results in the final answer not being coherent. To overcome this, the strategy of the CLIP team (Bagdouri and Oard, 2015) can be applied: the decision on the system choice (title versus body; only the best answer versus all answers) are made for each question using a supervised classifier. Bagdouri and Oard (2015) obtain the training data for the classifier via crowd-sourcing.

The official evaluation assigned the lowest possible value to some of the answers containing arguably relevant information, since the guidelines were looking for usefulness rather than relevance

System	avgscore	succ@1+	succ@2+	succ@3+	succ@4+	prec@2+	prec@3+	prec@4+
ADAPT.DCU	0.444	**1.000**	**0.290**	0.121	0.034	**0.290**	0.121	0.034
All runs	**0.465**	0.925	0.262	**0.146**	**0.060**	0.284	**0.159**	**0.065**

Table 3: Performance of our system in comparison to the average scores computed over all submitted runs.

Title	Body	Our Answer	Score (1-4)
Why does my cat has 2 eyes?	So tell me..	*It is a defect. In show quality cats, it is considered a fault. 2) Cats have TWO eyes, generally, unless they lose one.*	1
Do you like dogs?	-	*DOG – that's GOD spelled backwards say "A dog". There, where a dog looks like you, there is the best place to get a dog.*	1
Do you like Cricket? Your choice of any player?	-	*Brits of course who else? Outside of the Aussies who are by far the best in world.Americans they play baseball, not cricket. Any player!*	1

Table 4: Some examples from final evaluation.

of the answers. Table 4 shows some examples of such questions and answers. One drawback of the official evaluation guidelines is that it did not take into account the fact that for some questions it may be difficult to define what kind of answer could be considered as useful. For example, all the answers to the following question: *Why does my cat have 2 eyes?* received the lowest possible score.

Another drawback of our system is that it relies only on the dataset described in Section 2.1. This dataset was created in 2007, and obviously does not contain answers to questions related to topics ahead of its time, e.g. *Windows 10*. Several other systems also used the *Yahoo! Answers* as the main resource for answer extraction (Bagdouri and Oard, 2015; Nus and Szpekto, 2015). However, these systems made use of larger and more recent datasets, instead of using the L6 one. Using a more recent index will probably increase the performance of the described system.

One of the main advantages of our system is the speed. Even though it was not one of the evaluation metrics, it is worth noting that our system is able to retrieve an answer within 1.546 seconds on average, while the average time for all systems was about 20 seconds. Our system is thus almost 13 times faster than the average response time.

6 Conclusion

We described the work conducted in the ADAPT research centre in DCU for the purpose of participation in the TREC 2015 LiveQA track. In summary, we used a sentence retrieval approach over an indexed collection of previous Yahoo! Answers questions. We used the title of a new question as the query to retrieve a ranked list of similar questions. We then extracted the best matching sentences from all the remaining answers. The final answer was a combination of these sentences with the best answer of the most similar question. When no pre-existing questions with a sufficient similarity with the new one were retrieved, we opted for an answer from a candidate set of pre-generated answers based on the domain of the question.

Our submitted system was ranked 11 (out of 21) with an average score very close to the average score computed over all submitted runs. We believe one of the possible ways to improve the performance of our approach is by including more recent questions into the index (only the L6 dataset prepared in 2007 was used). Another possible improvement is in incorporating a classifier that chooses a retrieval strategy for each incoming question, instead of combining the outputs of several systems that results in incoherent answers.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of ADAPT centre (`www.adaptcentre.ie`) at Dublin City University. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Bagdouri, M. and Oard, D. W. (2015). Clip at trec 2015: Microblog and liveqa. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*.

Fried, D., Jansen, P., Hahn-Powell, G., Surdeanu, M., and Clark, P. (2015). Higher-order lexical semantic models for non-factoid answer reranking. *Transactions of the Association for Computational Linguistics*, 3:197–210.

Higashinaka, R. and Isozaki, H. (2008). Corpus-based question answering for why-questions. In *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pages 418–425.

Jansen, P., Surdeanu, M., and Clark, P. (2014). Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986, Baltimore, Maryland. Association for Computational Linguistics.

Nus, A. and Szpekto, I. (2015). Answering live questions by previously answered questions - yahoo labs at the liveqa track, trec 2015. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*.

Sharp, R., Jansen, P., Surdeanu, M., and Clark, P. (2015). Spinning straw into gold: Using free text to train monolingual alignment models for non-factoid question answering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 231–237, Denver, Colorado. Association for Computational Linguistics.

Surdeanu, M., Ciaramita, M., and Zaragoza, H. (2011). Learning to rank answers to non-factoid questions from web collections. *Comput. Linguist.*, 37(2):351–383.

Zhai, C. and Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM.