Overview of the EVALITA 2016 Question Answering for Frequently Asked Questions (QA4FAQ) Task

Annalina Caputo[1] and **Marco de Gemmis**[2,5] and **Pasquale Lops**[2]

Francesco Lovecchio[3] and **Vito Manzari**[4]

[1] ADAPT Centre, Dublin
[2] Department of Computer Science, University of Bari Aldo Moro
[3] Acquedotto Pugliese (AQP) S.p.a. [4] Sud Sistemi S.r.l. [5] QuestionCube S.r.l.
[1]annalina.caputo@adaptcentre.ie
[2]{marco.degemmis,pasquale.lops}@uniba.it
[3]f.lovecchio@aqp.it [4]manzariv@sudsistemi.it
[5]marco.degemmis@questioncube.com

Abstract

English. This paper describes the first edition of the Question Answering for Frequently Asked Questions (QA4FAQ) task at the EVALITA 2016 campaign. The task concerns the retrieval of relevant frequently asked questions, given a user query. The main objective of the task is the evaluation of both question answering and information retrieval systems in this particular setting in which the document collection is composed of FAQs. The data used for the task are collected in a real scenario by AQP Risponde, a semantic retrieval engine used by Acquedotto Pugliese (AQP, the Organization for the management of the public water in the South of Italy) for supporting their customer care. The system is developed by QuestionCube, an Italian startup company which designs Question Answering tools.

Italiano. *Questo lavoro descrive la prima edizione del Question Answering for Frequently Asked Questions (QA4FAQ) task proposto durante la campagna di valutazione EVALITA 2016. Il task consiste nel recuperare le domande più frequenti rilevanti rispetto ad una domanda posta dall'utente. L'obiettivo principale del task è la valutazione di sistemi di question answering e di recupero dell'informazione in un contesto applicativo reale, utilizzando i dati provenienti da AQP Risponde, un motore di ricerca semantico usato da Acquedotto Pugliese (AQP, l'ente per la gestione dell'acqua pubblica nel Sud Italia). Il sistema è sviluppato da QuestionCube, una startup italiana che progetta soluzioni di Question Answering.*

1 Motivation

Searching within the Frequently Asked Questions (FAQ) page of a web site is a critical task: customers might feel overloaded by many irrelevant questions and become frustrated due to the difficulty in finding the FAQ suitable for their problems. Perhaps they are right there, but just worded in a different way than they know.

The proposed task consists in retrieving a list of relevant FAQs and corresponding answers related to the query issued by the user.

Acquedotto Pugliese (AQP) developed a semantic retrieval engine for FAQs, called AQP Risponde[1], based on Question Answering (QA) techniques. The system allows customers to ask their own questions, and retrieves a list of relevant FAQs and corresponding answers. Furthermore, customers can select one FAQ among those retrieved by the system and can provide their feedback about the perceived accuracy of the answer.

AQP Risponde poses relevant research challenges concerning both the usage of the Italian language in a deep QA architecture, and the variety of language expressions adopted by customers to formulate the same information need.

The proposed task is strongly related to the one recently organized at Semeval 2015 and 2016 about Answer Selection in Community Question Answering (Nakov et al., 2015). This task helps to automate the process of finding good answers to new questions in a community-created discussion forum (e.g., by retrieving similar questions in

[1] http://aqprisponde.aqp.it/ask.php

the forum and by identifying the posts in the answer threads of similar questions that answer the original one as well). Moreover, the QA-FAQ has some common points with the Textual Similarity task (Agirre et al., 2015) that received an increasing amount of attention in recent years.

The paper is organized as follows: Section 2 describes the task, while Section 3 provides details about competing systems. Results of the task are discussed in Section 4.

2 Task Description: Dataset, Evaluation Protocol and Measures

The task concerns the retrieval of relevant frequently asked questions, given a user query. For defining an evaluation protocol, we need a set of FAQs, a set of user questions and a set of relevance judgments for each question. In order to collect these data, we exploit an application called AQP Risponde, developed by QuestionCube for the Acquedotto Pugliese. AQP Risponde provides a back-end that allows to analyze both the query log and the customers' feedback to discover, for instance, new emerging problems that need to be encoded as FAQ. AQP Risponde is provided as web and mobile application for Android[2] and iOS[3] and is currently running in the Acquedotto Pugliese customer care. AQP received about 25,000 questions and collected about 2,500 user feedback. We rely on these data to build the dataset for the task. In particular, we provide:

- a knowledge base of 406 FAQs. Each FAQ is composed of a question, an answer, and a set of tags;
- a set of 1,132 user queries. The queries are collected by analyzing the AQP Risponde system log. From the initial set of queries, we removed queries that contains personal data;
- a set of 1,406 pairs $< query, relevantfaq >$ that are exploited to evaluate the contestants. We build these pairs by analyzing the user feedback provided by real users of AQP Risponde. We manually check the user feedback in order to remove noisy or false feedback. The check was performed by two experts of the AQP customer support.

We provided a little sample set for the system development and a test set for the evaluation. We did not provide a set of training data: AQP is interested in the development of unsupervised systems because AQP Risponde must be able to achieve good performance without any user feedback. Following, an example of FAQ is reported:

Question "Come posso telefonare al numero verde da un cellulare?" *How can I call the toll-free number by a mobile phone?*

Answer "È possibile chiamare il Contact Center AQP per segnalare un guasto o per un pronto intervento telefonando gratuitamente anche da cellulare al numero verde 800.735.735. Mentre per chiamare il Contact Center AQP per servizi commerciali 800.085.853 da un cellulare e dall'estero è necessario comporre il numero +39.080.5723498 (il costo della chiamata è secondo il piano tariffario del chiamante)." *You can call the AQP Contact Center to report a fault or an emergency call without charge by the phone toll-free number 800 735 735...*

Tags *canali, numero verde, cellulare*

For example, the previous FAQ is relevant for the query: "Si può telefonare da cellulare al numero verde?" *Is it possible to call the toll-free number by a mobile phone?*

Moreover, we provided a simple baseline based on a classical information retrieval model.

2.1 Data Format

FAQs are provided in both XML and CSV format using ";" as separator. The file is encoded in UTF-8 format. Each FAQ is described by the following fields:

id a number that uniquely identifies the FAQ

question the question text of the current FAQ

answer the answer text of the current FAQ

tag a set of tags separated by ","

Test data are provided as a text file composed by two strings separated by the *TAB* character. The first string is the user *query id*, while the second string is the text of the user query. For example: "1 Come posso telefonare al numero verde da un cellulare?" and "2 Come si effettua l'autolettura del contatore?".

[2] https://play.google.com/store/apps/details?id=com.questioncube.aqprisponde&hl=it

[3] https://itunes.apple.com/it/app/aqp-risponde/id1006106860

2.2 Baseline

The baseline is built by using Apache Lucene (ver. 4.10.4)[4]. During the indexing for each FAQ, a document with four fields (*id, question, answer, tag*) is created. For searching, a query for each question is built taking into account all the question terms. Each field is boosted according to the following score *question=4*, *answer=2* and *tag=1*. For both indexing and search the *ItalianAnalyzer* is adopted. The top 25 documents for each query are provided as result set. The baseline is freely available on GitHub[5] and it was released to participants after the evaluation period.

2.3 Evaluation

The participants must provide results in a text file. For each query in the test data, the participants can provide 25 answers at the most, ranked according by their systems. Each line in the file must contain three values separated by the TAB character: $< queryid >< faqid >< score >$.

Systems are ranked according to the accuracy@1 (c@1). We compute the precision of the system by taking into account only the first correct answer. This metric is used for the final ranking of systems. In particular, we take into account also the number of unanswered questions, following the guidelines of the CLEF ResPubliQA Task (Peñas et al., 2009). The formulation of c@1 is:

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n}) \tag{1}$$

where n_R is the number of questions correctly answered, n_U is the number of questions unanswered, and n is the total number of questions.

The system should not provide result for a particular question when it is not confident about the correctness of its answer. The goal is to reduce the amount of incorrect responses, keeping the number of correct ones, by leaving some questions unanswered. Systems should ensure that only the portion of wrong answers is reduced, maintaining as high as possible the number of correct answers. Otherwise, the reduction in the number of correct answers is punished by the evaluation measure for both the answered and unanswered questions.

3 Systems

Thirteen teams registered in the task, but only three of them actually submitted the results for the evaluation. A short description of each system follows:

chiLab4It - The system described in (Pipitone et al., 2016a) is based on the cognitive model proposed in (Pipitone et al., 2016b). When a support text is provided for finding the correct answer, QuASIt is able to use this text to find the required information. ChiLab4It is an adaptation of this model to the context of FAQs, in this case the FAQ is exploited as support text: the most relevant FAQ will be the one whose text will best fit the user's question. The authors define three similarity measures for each field of the FAQ: question, answer and tags. Moreover, an expansion step by exploiting synonyms is applied to the query. The expansion module is based on Wiktionary.

fbk4faq - In (Fonseca et al., 2016), the authors proposed a system based on vector representations for each query, question and answer. Query and answer are ranked according to the cosine distance to the query. Vectors are built by exploring the word embeddings generated by (Dinu et al., 2014), and combined in a way to give more weight to more relevant words.

NLP-NITMZ the system proposed by (Bhardwaj et al., 2016) is based on a classical VSM model implemented in Apache Nutch[6]. Moreover, the authors add a combinatorial searching technique that produces a set of queries by several combinations of all the keywords occurring in the user query. A custom stop word list was developed for the task, which is freely available[7].

It is important to underline that all the systems adopt different strategies, while only one system (*chiLab4It*) is based on a typical question answer module. We provide a more detailed analysis about this aspect in Section 4.

[4] `http://lucene.apache.org/`

[5] `https://github.com/swapUniba/qa4faq`

Table 1: System results.

System	c@1
qa4faq16.chilab4it.01	0.4439
baseline	*0.4076*
qa4fac16.fbk4faq.2	0.3746
qa4fac16.fbk4faq.1	0.3587
qa4fac16.NLP-NITMZ.1	0.2125
qa4fac16.NLP-NITMZ.2	0.0168

4 Results

Results of the evaluation in terms of c@1 are reported in Table 1. The best performance is obtained by the *chilab4it* team, that is the only one able to outperform the baseline. Moreover, the *chilab4it* team is the only one that exploits question answering techniques: the good performance obtained by this team proves the effectiveness of question answering in the FAQ domain. All the other participants had results under the baseline. Another interesting outcome is that the baseline exploiting a simple VSM model achieved remarkable results.

A deep analysis of results is reported in (Fonseca et al., 2016), where the authors have built a custom development set by paraphrasing original questions or generating a new question (based on original FAQ answer), without considering the original FAQ question. The interesting result is that their system outperformed the baseline on the development set. The authors underline that the development set is completely different from the test set which contains sometime short queries and more realistic user's requests. This is an interesting point of view since one of the main challenge of our task concerns the variety of language expressions adopted by customers to formulate the information need. Moreover, in their report the authors provide some examples in which the FAQ reported in the gold standard is less relevant than the FAQ reported by their system, or in some cases the system returns a correct answer that is not annotated in the gold standard. Regarding the first point, we want to point out that our relevance judgments are computed according to the users' feedback and reflect their concept of relevance[8]. We tried to mitigate issues related to relevance judgments by manually checking users' feedback. However, this manual annotation process might have introduced some noise, which is common to all participants.

Regarding missing correct answers in the gold standard: this is a typical issue in the retrieval evaluation, since it is impossible to assess all the FAQ for each test query. Generally, this issue can be solved by creating a pool of results for each query. Such pool is built by exploiting the output of several systems. In this first edition of the task, we cannot rely on previous evaluations on the same set of data, therefore we chose to exploit users' feedback. In the next editions of the task, we can rely on previous results of participants to build that pool of results.

Finally, in Table 2 we report some information retrieval metrics for each system[9]. In particular, we compute Mean Average Precision (MAP), Geometrical-Mean Average Precision (GMAP), Mean Reciprocal Rank (MRR), Recall after five (R@5) and ten (R@10) retrieved documents. Finally we report the success_1 that is equal to c@1, but without taking into account answered queries. We can notice that on retrieval metrics the baseline is the best approach. This was quite expected since an information retrieval model tries to optimize retrieval performance. Conversely, the best approach according to success_1 is the *chilab4it* system based on question answering, since it tries to retrieve a correct answer in the first position. This result suggests that the most suitable strategy in this context is to adopt a question answering model, rather than to adapt an information retrieval approach. Another interesting outcome concerns the system *NLP-NITMZ.1*, which obtains an encouraging success_1, compared to the c@1. This behavior is ascribable to the fact that the system does not adopt a strategy that provides an answer for all queries.

5 Conclusions

For the first time for the Italian language, we propose a question answering task for frequently asked questions. Given a user query, the participants must provide a list of FAQs ranked by relevance according to the user need. The collection

[6] `https://nutch.apache.org`

[7] `https://github.com/SRvSaha/QA4FAQ-EVALITA-16/blob/master/italian_stopwords.txt`

[8] Relevance is subjective.

[9] Metrics are computed by the latest version of the trec_eval tool: `http://trec.nist.gov/trec_eval/`

Table 2: Results computed by using typical information retrieval metrics

System	**MAP**	**GMAP**	**MRR**	**R@5**	**R@10**	**success_1**
chilab4it	0.5149	0.0630	0.5424	0.6485	0.7343	0.4319
baseline	0.5190	0.1905	0.5422	0.6805	0.7898	0.4067
fbk4faq.2	0.4666	0.0964	0.4982	0.5917	0.7244	0.3750
fbk4faq.1	0.4473	0.0755	0.4781	0.5703	0.6994	0.3578
NLP-NITMZ.1	0.3936	0.0288	0.4203	0.5060	0.5879	0.3161
NLP-NITMZ.2	0.0782	0.0202	0.0799	0.0662	0.1224	0.0168

of FAQs was built by exploiting a real application developed by QuestionCube for Acquedotto Pugliese. The relevance judgments for the evaluation are built by taking into account the user feedback.

Results of the evaluation demonstrated that only the system based on question answering techniques is able to outperform the baseline, while all the other participants reported results under the baseline. Some issues pointed out by participants suggest exploring a pool of results for building more accurate judgments. We plan to implement this approach in future editions of the task.

Acknowledgments

This work is supported by the project “Multilingual Entity Liking” funded by the Apulia Region under the program FutureInResearch.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalara, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Divyanshu Bhardwaj, Partha Pakray, Jereemi Bentham, Saurav Saha, and Alexander Gelbukh. 2016. Question Answering System for Frequently Asked Questions. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Erick R. Fonseca, Simone Magnolini, Anna Feltracco, Mohammed R. H. Qwaider, and Bernardo Magnini. 2016. Tweaking Word Embeddings for FAQ Ranking. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Preslav Nakov, Lluıs Marquez, Walid Magdy, Alessandro Moschitti, James Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. *SemEval-2015*, page 269.

Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forăscu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. 2009. Overview of ResPubliQA 2009: question answering evaluation over European legislation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 174–196. Springer.

Arianna Pipitone, Giuseppe Tirone, and Roberto Pirrone. 2016a. ChiLab4It System in the QA4FAQ Competition. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Arianna Pipitone, Giuseppe Tirone, and Roberto Pirrone. 2016b. QuASIt: a Cognitive Inspired Approach to Question Answering System for the Italian Language. In *Proceedings of the 15th International Conference on the Italian Association for Artificial Intelligence 2016*. aAcademia University Press.