QA4FAQ - Question Answering for Frequently Asked Questions

Task Guideline

Task Description

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 task that we propose is strongly related to the recently organized task at Semeval 2015 and 2016 about Answer Selection in Community Question Answering [1]. 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 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 [2] that received an increasing amount of attention in recent years.

Data Description

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.

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

AQP Risponde is provided as web and mobile application for Android[2] and iOS[3] and is currently running in Acquedotto Pugliese customer care.

AQP received about 25,000 questions and collected about 2,500 user feedbacks. We rely on these data for building the dataset for the task. In particular, we provide:

- a knowledge base of about 470 FAQs. Each FAQ is composed of a question, an answer, and a set of tags;
- a set of user queries. The queries are collected by analyzing the AQP Risponde system log;
- a set of pairs <query, relevant faq> that are exploited to evaluate the contestants. We build these pairs by analyzing the user feedbacks provided by the real users of AQP Risponde. We manually check the user feedbacks in order to remove noisy or false feedbacks.

We will provide a little sample set for the system development and a test set for the evaluation. We will not provide a set of training data: AQP is interested in the development of unsupervised systems since AQP Risponde should be able to provide good performance without any user feedback.

Following, an example of FAQ is reported:

Question: *Come posso telefonare al numero verde da un cellulare?*
Answer: *"E' 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 e necessario comporre il numero +39.080.5723498 (il costo della chiamata e secondo il piano tariffario del chiamante)."*
Tags: *canali, numero verde, cellulare*

The previous FAQ is relevant for the query: *"Si può telefonare da cellulare al numero verde?"*

We will provide a simple baseline based on a classical Information Retrieval model.

Data Format

FAQs will be provided in CSV format using ‘;’ as separator. The file will be 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 ‘,’

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

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

Test data will be 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?*
2 *Come si effettua l'autolettura del contatore?*

The participants must provide results in a text file. For each query in the test data the participants can provide max 25 answers ranked according to their system. Each line in the file must contain three values separated by the TAB character:

<query id> <faq id> <score>

Systems will be ranked according to the accuracy@1 (*c@1*). We will compute the precision of the system taking into account only the first answer. This metric will be used for the final task ranking. In particular, we will take into account also the number of unanswered questions following the guidelines of the CLEF ResPubliQA Task. The formulation of c@1 is:

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

where:

- n_R: number of questions correctly answered;
- n_U: number of questions unanswered;
- n: 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.

Moreover, we will provide other measures, such as: MAP, GMAP and MRR, for further analysis.

If you have any questions, please contact us: qa4faq@gmail.com.

References

[1] Preslav Nakov, Lluís Márquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. SemEval-2015 Task 3: Answer Selection in Community Question Answering. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), June 2015, Association for Computational Linguistics, 269-281, http://www.aclweb.org/anthology/S15-2047.

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria and Janyce Wiebe. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), June 2015, Association for Computational Linguistics, 252-263, http://www.aclweb.org/anthology/S15-2045.