ChiLab4It System in the QA4FAQ Competition

Arianna Pipitone, Giuseppe Tirone, Roberto Pirrone
DIID - Dipartimento dell'Innovazione Industriale e Digitale -
Ingegneria Chimica, Gestionale, Informatica, Meccanica
Università degli Studi di Palermo
{arianna.pipitone, giuseppe.tirone, roberto.pirrone}@unipa.it

Abstract

English. ChiLab4It is the Question Answering system (QA) for Frequently Asked Questions (FAQ) developed by the Computer-Human Interaction Laboratory (ChiLab) at the University of Palermo for participating to the QA4FAQ task at EVALITA 2016 competition. The system is the versioning of the QuASIt framework developed by the same authors, which has been customized to address the particular task. This technical report describes the strategies that have been imported from QuASIt for implementing ChiLab4It, the actual system implementation, and the comparative evaluations with the results of the other participant tools, as provided by the organizers of the task. ChiLab4It was the only system whose score resulted to be above the experimental baseline fixed for the task. A discussion about future extensions of the system is also provided.

Italiano. *ChiLab4It è il sistema di Question Answering (QA) usato per rispondere alle Frequently Asked Questions (FAQs), sviluppato dal Laboratorio di Interazione Uomo-Macchina (Chilab) dell'Università degli Studi di Palermo allo scopo di partecipare al task QA4FAQ nella competizione EVALITA 2016. Il sistema è una versione del framework QuASIt, sviluppato dagli stessi autori e che è stato ridefinito per il task in questione. Il report descrive le strategie che hanno consentito di realizzare ChiLab4It a partire da QuASIt, l'effettiva implementazione del sistema e le valutazioni comparative con gli altri team che hanno partecipato al task, così come sono state rese note dagli organizzatori. ChiLab4It è stato l'unico sistema a superare la baseline sperimentale fissata per il task. Nella parte conclusiva del report, verranno altresì discussi i possibili sviluppi futuri del sistema.*

1 Introduction

This technical report presents ChiLab4It (Pipitone et al., 2016a), the QA system for FAQ developed by the ChiLab at the University of Palermo to attend the QA4FAQ task (Caputo et al., 2016) in the EVALITA 2016 competition (Basile et al., 2016). The main objective of such a task is answering to a natural language question posed by the user by retrieving the more relevant FAQs, among those in the set provided by the Acquedotto Pugliese society (AQP) which developed a semantic retrieval engine for FAQs, called *AQP Risponde*[1]. Such an engine is based on a QA system; it opens new challenges about both the Italian language usage and the variability of language expressions by users.
The background strategy of the proposed tool is based on the cognitive model described in (Pipitone et al., 2016b); in this work the authors present QuASIt, an open-domain QA system for the Italian language, that can be used for both multiple choice and essay questions. When a support text is provided for finding the correct answer (as in the case of text comprehension), QuASIt is able to use this text and find the required information.
ChiLab4It is the customized version of QuASIt to the FAQ domain; such a customization was the result of some restrictions applied on the whole

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

functionalities of QuASIt. The intuition was to consider the FAQ as *support text*; the more relevant FAQ will be the one whose text will best fit the user's question, according to a set of matching strategies that keep into account some linguistic properties, such as typology and syntactic correspondences. The good performances obtained in the evaluations demonstrate the high quality of the idea, although the current linguistic resources for the Italian are not exhaustive. This report is organized as follow: in section 2 the QuASIt system is presented, and in section 3 the ChiLab4It system is described as a restriction of QuASIt. In section 4 the results of ChiLab4It are shown according to the evaluation test bed provided by the competition organizers. Finally, future works are discussed in section 5.

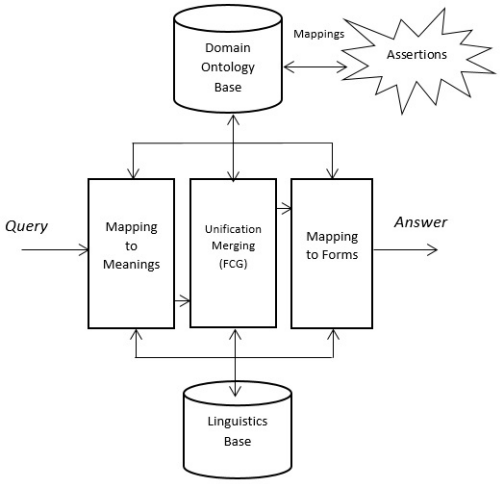

Figure 1: The QuASIt Cognitive Architecture

2 The QuASIt System

The main characteristic of QuASIt is the underlying *cognitive* architecture, according to which the interpretation and/or production of a natural language sentence requires the execution of some cognitive processes over both a perceptually grounded model of the world (that is an ontology), and a previously acquired linguistic knowledge.
In particular, two kinds of processes have been devised, that are the *conceptualization of meaning* and the *the conceptualization of form*.
The conceptualization of meaning allows to associate a sense to perceived forms, that are the words of the user query. A sense is the set of concepts of the ontology that explains the form; such a process is implemented considering the ontology nodes whose labels match best the forms from a syntactic point of view. The set of such nodes is the candidate sub-ontology to contain the answer to produce. The syntactic match is based on a syntactic measure.
The second process associates a syntactic expression to a meaning; it implements the strategies for producing the correct form of an answer, once it has been inferred. The form depends on the way QuASIt can be used, that is in both multiple choice and essay questions. In the case of multiple choice questions, the form must be one of the proposed answers. The system infers the correct answer among the proposed ones using the values of the properties' ranges in the sub-ontology; the answer that better syntactically match such ranges is considered the correct answer. If no answer can be inferred in this way, a support text can be used if available. The support text can be either derived automatically by the system, using the plain text associated to the nodes of the sub-ontology (such as an abstract node in the DBPedia ontology[2]) or provided directly to the questions as in the case of a text comprehension task. In figure 1 the architecture of QuASIt is shown. The ontology and the linguistic knowledge are located respectively in the *Domain Ontology Base* and the *Linguistic Base*. The *Mapping to Meanings* (MtM) and the *Mapping to Forms* (MtF) modules are the components that model the cognitive processes related to the conceptualization of meaning and form respectively. The *Unification Merging* module is essentially the FCG engine (Steels, 2011) that is used to perform query parsing.
The strategy we implemented in ChiLab4It system is based on the QuASIt function that selects the correct answer to multiple choice questions using support text; the intuition was that *a FAQ can be considered a support text* that can be used for retrieving the more relevant FAQ to a user's query. For this reason, in the next subsection, we describe this strategy in detail, and next we show how it was applied in the proposed tool.

2.1 Searching in the support text

Searching in a support text is a possible strategy to deal with unstructured information when an artificial agent is trying to answer a particular question. In this case the agent learns a possible answer by comprehending the text dealing with the question

[2]http://it.dbpedia.org/

topic. Such a process is implemented in QuASIt by the MtF module.

Formally, let $Q = \{q_1, q_2 ... q_n\}$ be the query of the user, and $P = \{p_1, p_2, ... p_m\}$ a sentence in the support text; each element in these sets is a token. P will be considered as much similar as Q when maximizing the following similarity measure m:

$$m = |\Im| - (\alpha l + \beta u)$$

where $\Im = \{p_j | \exists q_i \in Q,\ J(p_j, q_i) > \tau\}$, and $J(p_j, q_i)$ is the Jaro-Winkler distance between a couple of tokens (Winkler, 1990). As a consequence, $\Im \supset Q \cap P$, and $|\Im|$ is the number of matching tokens both in Q and P.

$l = 1 - \frac{|\Im|}{|P|}$ is the number of "lacking tokens" that are tokens belonging to Q that do not match in P, while $u = 1 - \frac{o(Q,\Im)}{|\Im|}$ is the number of "unordered tokens" that is the number of tokens in Q that do not have the same order in $\Im$; here $o(a, b)$ is the function returning maximum number of ordered tokens in a with respect to b.

Both l and u are normalized in the range $[0 \dots 1]$; they are penalty values representing syntactical differences among the sentences. The higher u and l are, the lower is the sentences similarity.

The α and β parameters weight the penalty, and they have been evaluated empirically through experimentation along with τ.

We re-used such strategy in ChiLab4It using different values for α and β parameters depending on which kind of support text we consider during the search, as next explained.

3 ChiLab4It

The basic idea of the proposed tool was to consider a FAQ as a support text. According to the provided dataset, a FAQ is composed by three textual fields: the *question text*, the *answer text* and the *tag set*. For each of these fields we applied the search strategy defined above; in particular we set different α and β parameters for each field in the m measure, depending on linguistics considerations. For this reason, we defined three different parameterized m measures named m_1, m_2 and m_3. Moreover, further improvements were achieved by searching for the synonyms of the words of the query in the answer text. These synonyms were not considered in the QuASIt implementation.

Given the previously defined variables $\Im$, l and u, the α and β parameters were set according to the following considerations:

- *question text*; the α and β parameters are the same of QuASIt, that is $\alpha = 0.1$ and $\beta = 0.2$. This choice is based solely on linguistic motivations; in fact, considering that the support text is a question such as the user query, both sentences to be matched will have interrogative form. As a consequence, both l and u influence the final match. The final measure is:

$$m_1 = |\Im| - (0.1 * l + 0.2 * u)$$

- *answer text*; the search is iterated for each sentence in the text. In this case, the α and β parameters are zero ($\alpha = 0$ and $\beta = 0$). This is because the answer text has a direct form, so the order of tokens must not be considered; moreover, a sentence in the answer text owns more tokens than the query, so this information is not discriminative for the final match.

  In this case, *the search is extended to the synonyms of the words in the query* except to the synonyms of the stop-words; this extension has improved significantly the performances of the system. Empirical evaluations demonstrated that there were not the same improvements when the synonyms were considered for the other parts of a FAQ (question text and tag set) because in these cases the synonyms increase uselessly the number of irrelevant FAQs retrieved by the system.

  Formally, let Σ be the *σ-expansion* set (Pipitone et al., 2014) that contains both the words and the synonyms of such words in the $Q - S_w$ set, being Q the user query as previously defined and S_w the set of stop-words:

$$\Sigma = \{\sigma_i \mid \sigma_i = synset(q_i) \ \wedge \ q_i \in Q - S_w\}$$

  Let's define $S = \{S_1, S_2, \dots, S_N\}$ the set of sentences in the answer text. We defined the M set that contains the m_{s_i} measures computed with $\alpha = 0$ and $\beta = 0$ in m, for each sentence $S_i \in S$ with the σ-expanded query:

$$M = \{m_{s_i} \mid m_{s_i} = |\Im_i|\}$$

  where

$$\Im_i = \{p_j \in S_i \cap \Sigma \mid \exists q_k \in Q,\ J(p_j, q_k) > \tau\}$$

  The final similarity measure m_2 will be the maximum value in M:

$$m_2 = max\ \{m_{s_i} \mid m_{s_i} = |\Im_i|\}$$

- *tag set*; the α and β parameters are zero ($\alpha = 0$ and $\beta = 0$) also in this case. This is because the tags in the set do not own a particular linguistic typology, so the information related to both the order of tokens and the lacking ones must not to be considered. As already explained, the synonyms are not included in this search. As consequence:

$$m_3 = |\Im|$$

where $\Im$ is the previously defined intersection among the query of the user and the set of tags.

A query will be considered as much similar as a FAQ when maximizing the sum of the measures defined previously, so the final similarity value is:

$$m_{faq} = m_1 + m_2 + m_3$$

These values were ordered, and the first 25 FAQs were outputted for a single query as required by the task.

3.1 The architecture

In figure 2 the architecture of ChiLab4It is shown; the input is the query of the user, while the output is the list of the first 25 relevant FAQs. The sources became the *FAQ base* and the *Wiktionary* source from which the provided FAQ dataset and the synonyms are respectively queried.
The white module of such an architecture is the MtF module as implemented in QuASIt. The dark modules are the integrations that have been applied to the MtF module for customizing it to the FAQ domain; in particular, such integrations regard both the σ-expansion of the query and the setting of the analytic form (including parameters) of the m measure depending on the FAQ field.
The first integration is implemented by the *σ module*, that returns the Σ set for the query of the user retrieving the synset from Wiktionary[3].
Parameters and the measure settings are performed by the *FAQ Ctrl* module which is encapsulated into the main MtF module; it retrieves the FAQ from the *FAQ base* and customizes the m measure according to the analyzed field (m_1 for the question text, m_2 for the answer text, m_3 for the tag set). The MtF module computes such measures referring to the σ-expanded query, and finally the m_{faq} value is computed and memorized by the

[3]https://it.wiktionary.org/

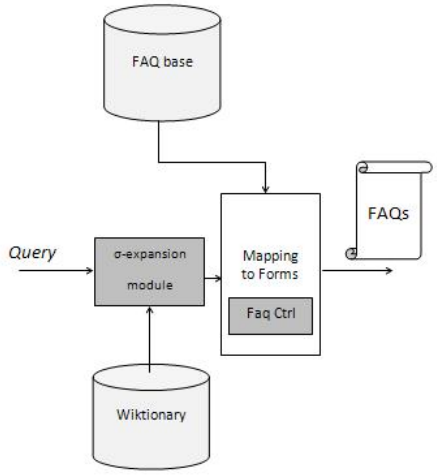

Figure 2: The ChiLab4It Architecture

FAQ Ctrl for tracing the id of the FAQ with the highest value.

3.2 A toy example

In this section we show a toy example with the aim of explaining better the searching process in the support text and how the similarity measure works. Such an example is a real question as retrieved in the data set provided by the organizers.
Let consider the query with $id = 4$, that is: "*a quali orari posso chiamare il numero verde*".
In this case, the Q and the S_w set are:

$$Q = \{A,\ quali,\ orari,\ posso,\ chiamare,\ il,\ numero,\ verde\}$$

and

$$S_w = \{A,\ il\}$$

being "*a*" and "*il*" the stop-words in the question.
The highest measure is computed by ChiLab4It in correspondence to the FAQ with $id = 339$, that is shown in table 1. Considering this FAQ, let compute the three measures for the *question text*, the *answer text* and the *tag set*.
In the first case the support text is the question text of the FAQ, and the P set is:
$P = \{Quali,\ sono,\ gli,\ orari,\ del,\ numero,\ verde\}$ with $|P| = 7$. The m_1 value will be computed considering that the intersection $\Im$ between the question text and the query of the user is:

$$\Im = \{quali,\ orari,\ numero,\ verde\}$$

. The Jaro-Winkler distance is 1 for each word, and $|\Im| = 4$. Also, $l = 1 - \frac{|\Im|}{|P|} = 1 - \frac{4}{7} = 0.428$.

Table 1: The XML description of FAQ 339 as provided in the data set

⟨**faq**⟩

⟨**id**⟩339⟨/**id**⟩

⟨**question**⟩Quali sono gli orari del numero verde?⟨/**question**⟩

⟨**answer**⟩Il servizio del numero verde assistenza clienti AQP 800.085.853 e attivo dal lunedi al venerdi dalle ore 08.30 alle 17.30, il sabato dalle 08.30 alle 13.00; il servizio del numero verde segnalazioni guasto 800.735.735 e attivo 24 ore su 24.⟨/**answer**⟩

⟨**tag**⟩informazioni, orari, numero verde⟨/**tag**⟩

⟨/**faq**⟩

For the calculation of u, we notice that $o(Q, \Im)$ returns 4 because the tokens in Q are all ordered with respect to $\Im$, that means they follow the same sequence in $\Im$. As consequence, $u = 1 - \frac{o(Q,\Im)}{|\Im|} = 1 - \frac{4}{4} = 0$. Substituting all values, m_1 will be:

$$m_1 = |\Im| - (0.1 * l + 0.2 * u) = 3.95$$

In the next step, we consider the answer text; in the FAQ, this text is composed by only one sentence that becomes the new support text P, and the procedure will be applied once. In particular, $S = \{S_1\}$ and $P = S_1 = \{$*Il, servizio, del, numero, verde, assistenza, clienti,...., attivo, 24, ore, su, 24*$\}$ as shown in table 1. In this case, the m_2 measure depends only from the intersection between the σ-expanded query and S_1. In particular, the Σ set is computed unifying the difference set $Q - S_w = \{Quali, orari, posso, chiamare, numero, verde\}$ with the synset from Wiktionary of each such token, so: Σ = {*[[quali], [orari], [posso], [chiamare, soprannominare, chiedere, richiedere], [numero, cifra, contrassegno numerico, matricola, buffone, pagliaccio, elenco, gruppo, serie, classe, gamma, schiera, novero, taglia, misura, attrazione, scenetta, sketch, esibizione, gag, sagoma, macchietta, fascicolo, puntata, dispensa, copia, tagliando, contrassegno, talloncino, titoli, dote, requisito], [verde, pallido, smorto, esangue, acerbo, giovanile, vivace, vigoroso, florido, verdeggiante, lussureggiante, rigoglioso, agricolo, agrario, vegetazione, vigore, rigoglio, freschezza, floridezza, via, avanti, ecologista, ambientalista, livido]]*}, where the synsets are represented in square brackets for clarity. The intersection $\Im_1 = \Sigma \cap S_1$ is simple $\Im_1 = \{numero, verde, orari\}$ because these tokens have the highest Jaro-Winkler distance from the tokens in S_1. As consequence, $M = \{|\Im_1|\} = \{3\}$ and $m_2 = 3$.
In the third case, the support text is the tag set, so $P = \{informazioni,\ orari,\ numero,\ verde\}$ and $\Im = \{orari,\ numero,\ verde\}$. The m_3 value is simply $m_3 = |\Im| = 3$.
Finally, the m measure is computed adding the three calculated values, so $m = 3.95 + 3 + 3 = 9.95$ that represents the highest value among those computed for all FAQs in the dataset.

4 Evaluations

The dataset used for the evaluation was the one provided by the QA4FAQ task organizers; they released such a dataset as a collection of both questions and feedbacks that real customers provided to the AQP Risponde engine.
In particular, such dataset includes:

- a knowledge base of about 470 FAQs, each composed by the text fields we referred to;
- a set of query by customers;
- a set of pairs that allows organizers to evaluate the possible contestants. The organizers analyzed the feedbacks provided by real customers of AQP Risponde engine, and checked them for removing noise.

Training data were not provided: in fact AQP is interested in the development of unsupervised systems, like ChiLab4It is.
According to the guideline, we provided results in a text file purposely formatted, and for each query in the dataset we considered the first 25 answers. However, only the first FAQ is considered relevant for the scope of the task. ChiLab4It is ranked according to the *accuracy@1* (c@1), whose formulation is:

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

where n_R is the number of correct answers, n_U is the number of unanswered questions, and n is the total number of questions.
A participant could have provided two different runs, but in our case we considered only the best configuration of the system. In table 2 we show

Table 2: The final results for QA4FAQ task

TEAM	c@1
ChiLab4It	**0.4439**
baseline	*0.4076*
Team 1 run 1	0.3746
Team 1 run 2	0.3587
Team 2 run 1	0.2125
Team 2 run 2	0.0168

the final results with the ranks of all participants as provided by the organizers; our tool performed better than the other participants, and it was the only one ranked above the experimental baseline.

5 Discussion and Future Works

ChiLab4It has been presented in this work, that is a tool designed for participating to the QA4FAQ task in the EVALITA 2016 competition. ChiLab4It relies on QuASIt, a cognitive model for an artificial agent performing question answering in Italian, already presented by the authors. QuASIt is able to answer both multiple choice and essay questions using an ontology-based approach where the agents manages both domain and linguistic knowledge.

ChiLab4It uses the functions of QuASIt aimed at answering multiple choice questions using a support text to understand the query because a FAQ can be regarded exactly as a support text, that can be used to understand the query sentence and to provide the answer. Moreover our tool enhances the sentence similarity measure introduced in our reference cognitive model in two ways. First, three separate measures are computed for the three parts of a FAQ that is question text, answer text and tag set, and they are summed to provide the final similarity. Second, the synonyms of the query words are analyzed to match the query against each sentence of the answer text of the FAQ to achieve linguistic flexibility when searching for the query topic inside each text.

ChiLab4It was tested with the competition data, and it resulted to be the winner having a c@1 rank well above the fixed experimental baseline.

Future works are aimed at refining the development of the entire QuASIT system. Particular attention will be devoted in studying more refined versions of the similarity measure to take into account complex phrasal structures.

References

Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli. 2016. EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. 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).

Annalina Caputo, Marco de Gemmis, Pasquale Lops, Franco Lovecchio, and Vito Manzari. 2016. Overview of the EVALITA 2016 Question Answering for Frequently Asked Questions (QA4FAQ) Task. 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, Vincenzo Cannella, and Roberto Pirrone. 2014. I-ChatbIT: an Intelligent Chatbot for the Italian Language. In Roberto Basile, Alessandro Lenci, and Bernardo Magnini, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*. Pisa University Press.

Arianna Pipitone, Giuseppe Tirone, and Roberto Pirrone. 2016a. Chilab4IT: 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*. in press.

Luc Steels, 2011. *Introducing Fluid Construction Grammar*, chapter Design Patterns in Fluid Construction Grammar. John Benjamins.

William E. Winkler. 1990. String comparator metrics and enhanced decisionn rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.