“Who was Pietro Badoglio?” Towards a QA system for Italian History

Stefano Menini[1-2], Rachele Sprugnoli[1-2], Antonio Uva[2]
[1]Fondazione Bruno Kessler, [2]Università di Trento
[1]Via Sommarive 18, 38123 Povo (TN), Italy [2]Via Sommarive 9, 38123 Povo (TN), Italy
{menini;sprugnoli}@fbk.eu; antonio.uva@unitn.it

Abstract

This paper presents QUANDHO (*QUestion ANswering Data for italian HistOry*), an Italian question answering dataset created to cover a specific domain, i.e. the history of Italy in the first half of the XX century. The dataset includes questions manually classified and annotated with Lexical Answer Types, and a set of question-answer pairs. This resource, freely available for research purposes, has been used to retrain a domain independent question answering system so to improve its performances in the domain of interest. Ongoing experiments on the development of a question classifier and an automatic tagger of Lexical Answer Types are also presented.

Keywords: question answering, dataset creation, annotation

1. Introduction

Question Answering (QA) systems provide users, who ask questions in natural language, with a short passage containing the answer and some context to validate it (Hirschman and Gaizauskas, 2001). The development of such systems requires a multidisciplinary approach combining techniques of different fields: it is not by chance that the literature reports various studies dealing with QA systems from different perspectives, e.g. Information Retrieval (Kolomiyets and Moens, 2011) and Semantic Web (Lopez et al., 2011).

In the last few years, QA has received a lot of attention thanks to the success of the IBM’s Watson system in the Jeopardy! game (Ferrucci et al., 2010) and to its application to the clinical domain (Ferrucci et al., 2013). In addition, the organization of international evaluation campaigns such as TREC in the USA (Voorhees et al., 2005)[1], CLEF in Europe (Mothe et al., 2015)[2] and NTCIR in Japan[3] fostered the extension of QA research to languages other than English. Nevertheless only two systems, both domain independent, are currently available for Italian, namely Wikiedi[4] and the Italian Minimal Structural Reranking Pipeline (henceforth, It-MSRP) (Uva and Moschitti, 2015).

In this paper we describe QUANDHO (*QUestion ANswering Data for italian HistOry*), an Italian question answering dataset that includes questions manually classified and annotated with lexical answer types, and a set of question-answer pairs. This dataset, freely available to the research community[5], focuses on a specific domain, i.e. the Italian history of the first half of the XX century. In addition, we present the evaluation of Wikiedi and It-MSRP on this domain and we report on the retraining of the latter using our dataset so to improve its precision and make it an effective tool to be incorporated in a real application. The final aim is to integrate an interactive version of the QA system in ALCIDE (Moretti et al., 2014), a web-based platform for historical content analysis. ALCIDE contains different corpora related to the historical domain among which the complete corpus of Alcide De Gasperi’s[6] writings is the largest one[7]: about 3,000,000 tokens covering the history of Italy between 1901 and 1954. A QA system focused on this specialized area would help readers to find additional information related to the historical documents they are reading.

The paper is structured as follows. In the first part we present the creation of QUANDHO giving details on how we defined, classified and annotated our set of questions (Section 2.), and how we associated each question with a pool of candidate answers (Section 3.). In the second part of the paper, i.e. Section 4. and Section 5., we focus on the evaluation of the two Italian QA sytems on our dataset, on the adaptation of It-MSRP to the historical domain and on some ongoing experiments. Conclusions are drawn in Section 6.

2. Question Set Creation

Given that no copyright-free list of historical questions for Italian is available (all school textbooks are copyright protected), we chose to create our own pool of questions starting from the Italian Wikipedia[8].

Since our work was focused on the Italian history in the first half of the 20th century, we found a suitable starting point in the Wikipedia page “Storia d’Italia (1861-oggi)”[9], the main page about the history of the unified Italy, containing links to the events and leading figures of this historical period. We crawled 2 levels of links from this seed page collecting a total number of 3,060 pages. Then, we

[1] `http://trec.nist.gov/`

[2] `http://www.clef-initiative.eu/`

[3] `http://research.nii.ac.jp/ntcir/index-en.html`

[4] `http://www.wikiedi.it/`

[5] `https://dh.fbk.eu/technologies/quandho`

[6] Alcide De Gasperi was one the founding fathers of the Italian Republic and of the European Union.

[7] A demo of ALCIDE, not containing this corpus, is available at the following URL: `http://celct.fbk.eu:8080/Alcide_Demo/`

[8] `https://it.wikipedia.org/`

[9] `https://it.wikipedia.org/wiki/Storia_d\%27Italia_(1861-oggi)`

removed all the non-relevant pages by applying a set of filters. For example, the content of infoboxes was exploited to filter out events and people chronologically placed out of our period of interest (e.g. people who died before 1900 or were born after 1954), whereas the Wikipedia system of categories was used to detect pages out of our domain (e.g. pages belonging to the rock music or cycling portals). After this filtering 274 pages were retained covering 6 categories: people (e.g. *Mussolini*), political parties (e.g. *Democrazia Cristiana / Christian Democracy*), events in both domestic and foreign policy (e.g. battles, promulgation of laws), ideologies and concepts (e.g. *antifascismo / anti-fascism*), places (e.g. *Colonia Eritrea / Italian Eritrea*), organizations (e.g. *Società delle Nazioni / League of Nations*), and other (e.g. *Dirigibile Norge / Norge Airship*).

We used the 274 selected web pages to create a set of text snippets by splitting the text into paragraphs (one snippet for each paragraph), and then by cleaning it out removing all the HTML tags . The result was a set of 10,200 plain text snippets. Starting from these text snippets, we created 627 questions whose answer was contained in a snippet[10]. For the creation of these questions we followed rules inspired by the guidelines provided in the QA tasks[11] at CLEF:

- all questions must be constructed on the basis of a snippet containing the relevant information; in other words, relying on world knowledge alone is not permitted because each question must be guaranteed to have at least one snippet containing the answer;
- questions must be well-formed and grammatically correct: it is important to prefer simple and precise wording thus avoiding ambiguous and nonsense words;
- in case of questions having a list as answer (e.g. *Quali sono state le colonie italiane tra il 1912 ed il 1939? / What were the Italian colonies between 1912 and 1939?*), all the requested items of the list must be present in the snippet;
- questions must not contain anaphoric links to entities not mentioned in the snippet;
- term overlap between the question and the snippet is allowed but should be reduced to a minimum by using different types of lexical and syntactic variations such as synonyms (Example 1), morphological derivations (Example 2), from active to passive voice conversion (Example 3). Questions can also introduce simple inferences (Example 4).

(1) Snippet: *Nel dopoguerra Amendola dichiarò inoltre di aver scelto personalmente il Polizeiregiment Bozen come* ***obiettivo****[...] / During the postwar Amendola declared to personally have chosen the Polizeiregiment Bozen as the objective...*
Question: *Perché Amendola scelse il Polizeiregiment Bozen come* ***bersaglio*** *per l'attentato di via Rasella? / Why Amendola chose the Polizeiregiment Regiment Bozen as target for the attack in via Rasella?*

(2) Snippet: *L'****occupazione*** *italiana del Regno di Albania ebbe luogo tra il 1939 al 1943. [...]/ The Italian occupation of the Albanian Kingdom took place between 1939-1943.*
Question: *In quali anni l'Italia* ***occupò*** *il Regno d'Albania? / What years did Italy occupy the Albanian Kingdom?*

(3) Snippet: *Il 22 giugno la Germania, rompendo il patto di non aggressione del 1939,* ***invadeva*** *la Russia (operazione Barbarossa). [...] / On June 22, Germany, breaking the non-aggression pact of 1939, invaded Russia (Operation Barbarossa).*
Question: *Che paese* ***fu invaso*** *con l'operazione Barbarossa? / What country was invaded with the Operation Barbarossa?*

(4) Snippet: *[...] il successivo ponte aereo, organizzato dal mondo occidentale* ***per assicurare la sopravvivenza della popolazione di Berlino Ovest****, è entrato nella storia. / [...] the next airlift organized by the Western world to ensure the survival of the West Berlin population, has entered into history.*
Question: *Quale era lo* ***scopo*** *del ponte aereo di Berlino? / What was the purpose of the Berlin airlift?*

2.1. Question Set Classification

Once the questions were collected, we classified them following the question taxonomy proposed in (Li and Roth, 2002). This taxonomy has six classes and fifty subclasses and is briefly described below:

- ABBREVIATION has two subclasses, one for abbreviated expressions (e.g. *Cosa significa l'acronimo TLT? / What does the acronym TLT stand for?*) and one for acronyms (e.g. *Con quale sigla veniva ufficialmente chiamato l'Impero Coloniale Italiano? / What abbreviation was used to call the Italian Empire?*);
- ENTITY includes 22 subclasses among which *term* (e.g. *Come si chiama l'area tra trincee contrapposte? / What do you call the area between opposing trenches?*) and *event* (e.g. *A quale secessione prese parte Gronchi? / Which secession did Gronchi take part?*):
- DESCRIPTION has four subclasses covering, for example, definitions (e.g. *Cosa è stata la linea Gotica? / What was the Gothic Line?*) and reasons (*Perché l'Italia decise di espandersi verso l'Africa? / Why did Italy decide to expand to Africa?*);
- HUMAN includes four subclasses such as *individual* (*Chi definì Mussolini "uomo della Provvidenza"? / Who called Mussolini "the man sent by Providence"?*) and *group* (*A quale partito aderì Mariano Rumor? / Which party Mariano Rumor joined?*);

[10] Question creation and annotation were performed by two of the authors who are PhD candidate in NLP, have a background in humanities studies and experience in linguistic annotation. In the rest of the paper, the expression "annotators" is used to refer to them.

[11] http://nlp.uned.es/clef-qa/repository/

- LOCATION has five subclasses used for geographic references of different types such as cities (*Dove venne firmato l'armistizio corto? / Where was the Short Armistice signed?*) and countries (*Quali paesi firmarono il Patto Tripartito? / Which countries signed the Tripartite Pact?*);

- NUMERIC comprises thirteen subclasses, among which *date* (*In che anno avvenne la marcia su Roma? / In which year the March on Rome happened?*) and *count* (*Quante perdite ci furono nella battaglia del Don? / How many losses were incurred in the battle of the Don River?*).

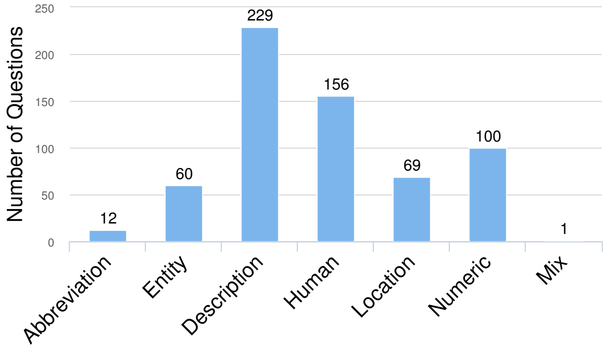
The *Mixed* class was added to the aforementioned 6 classes to cover one question requiring two types of answers: *Dove e quando vennero fondati i Fasci di combattimento? / Where and when were Italian Fasci of Combat founded?*
The classification was performed manually by two annotators. Given the lack of comprehensive guidelines, annotators relied their choices on the concise definitions of question classes and on the analysis of the labeled TREC training and test sets, both available online.[12] At the beginning of the classification process, a subset of fifty randomly chosen questions was used to discuss critical issues. During this phase was decided, for example, to include ideologies (e.g. fascism) in the *other* subclass of the ENTITY class and to extend the *country* subtype of the LOCATION class to cover nations not existing anymore but frequent in our snippets, for example Yugoslavia or Soviet Union. Later on, the annotators independently classified the rest of the question set and at the end they reconciled discrepancies. The inter-annotator agreement (IAA) (Artstein and Poesio, 2008) on classes before reconciliation was 92,7% (581/627) with a kappa statistic of 0.90. For subclass attribution, limited to the questions having a perfect agreement in class assignment, the agreement was 80.4% (504/627), with a kappa of 0.85.

	Abb	Ent	Des	Hum	Loc	Num	Mix
Abb	11	0	0	0	0	0	0
Ent	0	53	6	3	1	0	0
Des	1	14	209	2	1	0	0
Hum	0	0	10	141	0	0	0
Loc	0	2	2	4	66	0	0
Num	0	0	0	0	0	100	0
Mix	0	0	0	0	0	0	1

Table 1: Confusion matrix of manual class assignment before reconciliation

As shown by the confusion matrix on class assignment (Table 2.1.), NUMERIC proved to be the less ambiguous class while the most common disagreement was registered between ENTITY and DESCRIPTION. This disagreement was due to a misunderstanding about the classification of events. During the reconciliation, only named events (e.g. *First World War*) were classified as ENTITY of subclass *event*; all the others, such as *combing actions against partisans*, were assigned to the DESCRIPTION class. Moreover, the distinction between subclasses of the class DESCRIPTION (*definition*, *description*, *reason* and *manner*) was not always clear thus producing the 35% of the disagreement in subclass assignment.
Figure 1 shows the final classification, i.e. after reconciliation, of the 627 questions into the main seven classes. Although the questions are for the most factoid (i.e. the answer is a single word token or a short noun phrase) such as the ones under the NUMERIC class, there is also a significant portion of non factoid questions represented by the DESCRIPTION category (229 questions, 36.5% of the total), that requires a more articulated answer.

Figure 1: Bar chart representing the distribution of the questions into different classes



2.2. Lexical Answer Type Annotation

The two annotators that have classified the question set also performed the annotation of Lexical Answer Types (LATs). The LAT is a noun that, without belonging to a predefined category, describes the type of answer corresponding to the question (Ferrucci et al., 2010).
LAT annotation followed the same process as in the classification phase: after discussing together the annotation of fifty questions randomly chosen, the annotators worked independently, then IAA was calculated and disagreements were reconciled. The IAA before reconciliation was 94,1% (590/627): given that for five questions the annotators could not find any agreement, the opinion of a third annotator[13] was asked.
In the final annotation, 283 out of 627 questions show an explicit LAT, as in *In che* ***giorno*** *le truppe di Clark hanno liberato Roma? / On what* ***day*** *did the troops of Clark liberate Rome?*, all the others have an implicit LAT such as *Quando avvenne la battaglia di Caporetto? / When did the battle of Caporetto happen?* Both explicit and implicit LATs have been annotated: for example in the latter question the noun marked as LAT was *battaglia / battle*. Moreover, we found 318 unique LATs, only 11 of them having a

[12] http://cogcomp.cs.illinois.edu/Data/QA/QC/

[13] The third annotator was the third author of the paper.

frequency above 10[14] and covering the 23% of the question set.

3. Creation of the Question-Answer Pairs

Once the set of questions was defined, we associated each question with a pool of candidate answers thus creating question-answer pairs following two main steps.

In the first step, for each of the 627 questions, we used Lucene[15] to extract up to 20 candidate answers from the 10,200 Wikipedia snippets collected as explained in Section 2. In this way we gathered 12,474 question-answer pairs. Each of these pairs has been manually marked as *true* if the answer was correct with respect to the question, and *false* otherwise. Answers marked as *true* included the ones we used to define the questions. This process resulted in a set of 1,230 correct answers as well as 11,244 pertinent, and thus more challenging, wrong answers.

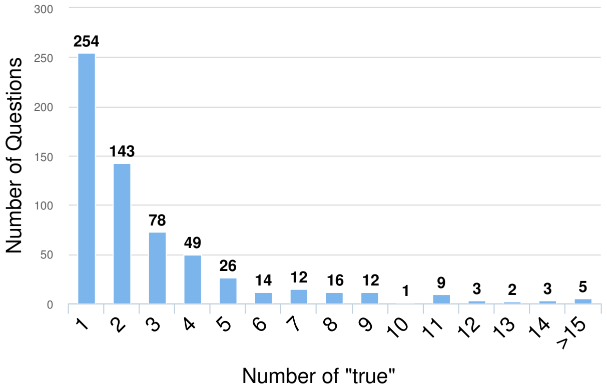
In the second step, we extended the pool of question-answer pairs by introducing answers less related to the domain of our questions and not necessarily from Wikipedia pages covering historical topics. To this end, we relied again on Lucene to extract up to 40 candidate answers from the whole Italian Wikipedia dump, creating 25,080 additional question-answer pairs. Given that the manual annotation of all these pairs would have required a very high effort in terms of time, we defined and followed a three-stage procedure: *i)* for each question we defined a set of keywords representative of the correct answer (e.g. Question: *Chi proclamò la Reggenza Italiana del Carnaro? / Who proclaimed the Italian Regency of Carnaro?*, Keywords: *Gabriele D'Annunzio - D'Annunzio*; *ii)* we used these keywords to find good answer candidates, automatically marking as *true* each snippet containing them; *iii)* we removed duplicates and, to avoid false positives, we made a manual check of the pairs automatically marked as *true*, so to understand if the presence of the keywords actually corresponded to a correct answer. The snippets marked as *true* in the second stage of the procedure were 1,153: after cleaning up duplicates and false positives we had them reduced to 572. The remaining 22,584 answers, not containing the keywords were marked as *false*.

The resulting set of pairs, released in the QUANDHO resource, is composed by a total of 35,630 question-answer pairs, with 1,802 answers marked as *true* and 33,828 marked as *false*. As shown in Figure 2, each question is associated with at least one correct answer.

4. From Open-Domain to Close-Domain QA

This Section presents the evaluation of two existing domain independent QA systems for Italian on the historical domain and the retraining process of one of these systems so to improve its precision. Evaluation and retraining were both based on the QUANDHO dataset.

Figure 2: Distribution of the answers marked as *true* over the 627 questions



4.1. Evaluation of Open-Domain QA Systems

We evaluated the performances of Wikiedi and It-MSRP on a subset of our question set. Wikiedi is a Web application built on top of the QuestionCube framework (Molino and Basile, 2012) while It-MSRP is a system, adapted from the model proposed by Severyn et al. (2013b), that reranks answer passages for factoid questions in Italian. Both systems deal with unstructured textual sources and are domain independent. Moreover, they both use Wikipedia to retrieve candidate passages, a search engine scoring function based on the BM25 model (Robertson and Zaragoza, 2009) and a set of Natural Language Processing modules (e.g. Part-Of-Speech tagger, Named Entity Recognizer) to analyze questions and candidate answers. The main difference between the two systems is that It-MSRP applies Support Vectors Machines algorithms using tree kernels to rank answer passages but does not include any question classification component. On the contrary, Wikiedi classifies the questions on the basis of Machine Learning techniques and hand-written rules. Moreover, It-MSRP has been trained and tested on TREC data translated to Italian, as described in Uva and Moschitti (2015). The performances of It-MSRP on the TREC dataset are reported in Table 2.

Model	P@1	MRR	MAP
BM25	15.22	23.11	0.18
It-MSRP	22.29	30.74	0.25

Table 2: Performance of the It-MSRP on the TREC dataset

To test Wikiedi we queried the system using the web interface while for It-MRSP we run the pipeline on a local machine. In both cases we used 209 questions as test set (i.e. one third of the whole question set) and we checked the correctness of the first given answer so to calculate the Precision at rank 1 (P@1), i.e. the percentage of relevant answers ranked at position 1. The results of this evaluation are reported in Table 3 where details are given for each class of questions together with the overall P@1[16]. It-MSRP P@1

[14] *anno/year, ruolo/role, guerra/war, nome/name, battaglia/battle, trattato/treaty, paese/country, partito/party, legge/law, patto/pact, scopo/goal*

[15] `https://lucene.apache.org/core/`

[16] The question belonging to the MIX class is not part of this test set.

is 21.06 points lower than Wikiedi P@1: the only class for which It-MSRP outperforms Wikiedi is *Location* (31.82% versus 27.27%) while the class for which the two systems have the biggest gap is *Description* (8.57% versus 45.71%).

		P@1	
Question Classes	**#Questions**	**Wikiedi**	**It-MSRP**
Abbreviation	2	0.00%	0.00%
Description	68	45.71%	8.57%
Entity	19	27.78%	22.22%
Human	52	40.00%	24.00%
Location	21	27.27%	31.82%
Numeric	47	38.30%	17.02%
OVERALL	**209**	**38.76%**	**17.70%**

Table 3: Precision at rank 1 of Wikiedi and It-MSRP on 209 questions: performances over the six classes and overall result

4.2. Adaptation to the Historical Domain

As shown by It-MSRP performances in Tables 2 and 3, the P@1 of the system trained on TREC dataset has a drop of 4.59% when dealing with historical questions (22.29% versus 17.70%). This result highlights the need of improving its performances on the target domain.

To this end, we used QUANDHO to retrain It-MSRP on the historical domain by conducting a set of 3-fold cross validation experiments. In particular, we tried different configurations (i.e. SubTree kernel, SubSet Tree kernel, SubSet Tree kernel with Bag of Words and a Partial Tree kernel) to learn the model for the It-MSRP system. Table 4 reports the results of the best configuration which was obtained by training the system with the Partial Tree Kernel (Moschitti, 2006). Results, calculated on the 209 test questions, are given in terms of Precision at 1 (P@1), Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP).

Model	**P@1**	**MRR**	**MAP**
BM25	17.22	25.03	0.21
It-MSRP trained on TREC	17.70	27.03	0.24
It-MSRP trained on QUANDHO	27.75	31.09	0.28

Table 4: It-MSRP performance on 209 historical questions

Table 4 compares the performance of the It-MSRP system on the 209 test questions by using two different models. The first one is obtained by training the system on the TREC data (domain independent), while the second one is obtained by training the system on the 418 historical questions not used for the test. The baseline model corresponds to the Lucene score (BM25).

Details about the performance of It-MSRP after the retraining process are given in Table 5 where the results of P@1 are reported over six classes. We registered an overall improvement of 10.05% and a beneficial effect (equal or above 10 percentage points) on 4 out of 6 classes. These improvements are particularly encouraging given that the system is based only on an SVM classifier using tree kernels applied to syntactic trees; no hand-written features are implemented.

	P@1	
Question Classes	**Result**	**Difference**
Abbreviation	0.00%	+0.00%
Description	18.57%	+10.00%
Entity	33.33%	+11.11%
Human	38.00%	+14.00%
Location	31.82%	+0.00%
Numeric	27.66%	+10.64%
OVERALL	**27.75%**	**+10.05%**

Table 5: P@1 of It-MSRP after the retraining on the historical dataset: the absolute difference for each question class is calculated with respect to the results reported in Table 3

5. Ongoing Experiments

The results obtained after the retraining process are promising but there is still room for improvements. As suggested in Severyn et al. (2013a), QA systems may highly benefit from information on the category of questions. For this reason we are working to add a question classifier and an automatic tagger of LATs in the It-MSRP system. Both these modules are trained on the pool of questions we manually annotated. At the moment of writing we can only report on some preliminary results.

As for the automatic tagger of LATs, a first experiment using 10-fold cross-validation achieved an accuracy of 0.76. An accuracy of 0.73 was scored by the question classifier for which we employed a stratified 10-fold cross-validation. The stratified k-fold cross-validation approach was used for this experiment so to make sure there was the same percentage of data for each class (ABBREVIATION, HUMAN, etc.) in the training and test folds. Table 5. presents the confusion matrix of the automatic classification of questions obtained in this first experiment. As observed in the manual classification, the less ambiguous class is NUMERIC, with a precision of 82.00%. ABBREVIATION and ENTITY are, on the contrary, the most problematic classes with a precision of 16.67% and 35.00% respectively. A close look at the errors showed that challenging questions are those having *Cosa - Quale - Che/What - Which* as question stems because these stems can be associated with many different classes.

	Abb	**Ent**	**Des**	**Hum**	**Loc**	**Num**	**Mix**
Abb	2	1	9	0	0	0	0
Ent	0	21	18	6	13	2	0
Des	2	11	187	16	10	3	0
Hum	0	9	14	125	7	1	0
Loc	0	8	12	6	41	2	0
Num	0	3	10	2	3	82	0
Mix	0	0	0	0	0	1	0

Table 6: Confusion matrix of automatic class assignment

Examples 5-8 show the high variability of *Quale* (both in its singular and plural form) in terms of classes:

(5) Question: ***Quali*** *sono i confini temporali del biennio rosso italiano? / What are the temporal boundaries of Italian red biennium?*
Class: NUMERIC.

(6) Question: ***Quale** fu il risultato dell'operazione Diadem? / Which was the result of Operation Diadem?*
Class: DESCRIPTION.

(7) Question: ***Quale** fazione era guidata da Gregor Strasser? / Which faction was led by Gregor Strasser?*
Class: HUMAN.

(8) Question: ***Quali** regioni divennero a statuto speciale tra il 1946 e il 1948? / What regions obtained a special status between 1946 and 1948?*
Class: LOCATION.

6. Conclusions and Future Work

In this paper we present a new Italian dataset for question answering in the domain of history. This resource, called QUANDHO, is made by: *i)* a set of questions manually classified in 7 main classes and several subclasses and annotated with LATs and *ii)* a set of question-answer pairs. QUANDHO is freely available for download and, to the best of our knowledge, is the first of this kind in the history domain. The dataset has been used to retrain a QA system and to design the implementation of a question classifiers and of an automatic tagger of LATs.

As for future works, the integration of more complex modules for linguistic analysis, e.g. a temporal expression recognizer, can be envisaged to improve system performances. We also intend to expand the resource with additional questions covering other periods of the Italian history.

The final aim is to incorporate the system in ALCIDE so to test its usability in a real scenario, i.e. with students reading De Gasperi's writings through the web platform.

7. Bibliographical References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., et al. (2010). Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79.

Ferrucci, D. A., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. (2013). Watson: beyond Jeopardy! *Artif. Intell.*, 199:93–105.

Hirschman, L. and Gaizauskas, R. (2001). Natural language question answering: the view from here. *natural language engineering*, 7(04):275–300.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Lopez, V., Uren, V., Sabou, M., and Motta, E. (2011). Is question answering fit for the Semantic Web? A survey. *Semantic Web*, 2(2):125–155.

Molino, P. and Basile, P. (2012). QuestionCube: a Framework for Question Answering. *IIR*, 835:167–178.

Moretti, G., Tonelli, S., Menini, S., and Sprugnoli, R. (2014). ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment. In *Atti della prima Conferenza Italiana di Linguistica Computazionale*.

Moschitti, A. (2006). Efficient convolution kernels for dependency and constituent syntactic trees. In *Machine Learning: ECML 2006*, pages 318–329. Springer.

Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, G., San Juan, E., Cappellato, L., and Ferro, N. (2015). *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9283. Springer.

Robertson, S. and Zaragoza, H. (2009). *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Severyn, A., Nicosia, M., and Moschitti, A. (2013a). Building structures from classifiers for passage reranking. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 969–978. ACM.

Severyn, A., Nicosia, M., and Moschitti, A. (2013b). Learning adaptable patterns for passage reranking. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 75–83.

Uva, A. and Moschitti, A. (2015). Automatic Feature Engineering for Italian Question Answering Systems. In *Proceedings of the Sixth Italian Information Retrieval Workshop*.

Voorhees, E. M., Harman, D. K., et al. (2005). *TREC: Experiment and evaluation in information retrieval*, volume 1. MIT press Cambridge.