Information Processing and Management 47 (2011) 856–869

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Learning to select the correct answer in multi-stream question answering

Alberto Téllez-Valero [a], Manuel Montes-y-Gómez [a,*], Luis Villaseñor-Pineda [a], Anselmo Peñas Padilla [b]

[a] Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Sta. María Tonantzintla, Pue. 72840, Mexico
[b] Depto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Juan del Rosal, 16, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 1 April 2009
Received in revised form 26 February 2010
Accepted 21 March 2010
Available online 13 May 2010

Keywords:
Data fusion
Multi-stream QA
Textual entailment
Answer validation

ABSTRACT

Question answering (QA) is the task of automatically answering a question posed in natural language. Currently, there exists several QA approaches, and, according to recent evaluation results, most of them are complementary. That is, different systems are relevant for different kinds of questions. Somehow, this fact indicates that a pertinent combination of various systems should allow to improve the individual results. This paper focuses on this problem, namely, the selection of the correct answer from a given set of responses corresponding to different QA systems. In particular, it proposes a supervised multi-stream approach that decides about the correctness of answers based on a set of features that describe: (i) the compatibility between question and answer types, (ii) the redundancy of answers across streams, as well as (iii) the overlap and non-overlap information between the question–answer pair and the support text. Experimental results are encouraging; evaluated over a set of 190 questions in Spanish and using answers from 17 different QA systems, our multi-stream QA approach could reach an estimated QA performance of 0.74, significantly outperforming the estimated performance from the best individual system (0.53) as well as the result from best traditional multi-stream QA approach (0.60).

1. Introduction

The great amount of available information has motivated the development of different tools for searching and browsing large document collections. The major examples of these tools are information retrieval (IR) systems, which focus on identifying relevant documents for general user queries. Search engines such as Google[1] and Yahoo[2] are a special kind of IR systems which allow to retrieve information from the Web.

It is clear that IR systems have made possible the processing of large volumes of textual information; however, they present serious problems for answering specific questions formulated by users. In fact, they can not properly tackle this task: once the user obtained a list of relevant documents for her question, she still has to review all documents in order to find the desired information. This limitation, along with a growing need for improved information access mechanisms, triggered the emergence of *question answering* (QA) systems. These systems aim at identifying the exact answer to a question from a given document collection. In other words, given an user query in the form of natural language question (e.g., `Which country did Iraq invade in 1990?`), a QA system must detect the text fragment that respond the question (e.g., `Kuwait`) instead of returning a list of documents related to the question words.

* Corresponding author. Tel.: +52 222 2 663 100x8218; fax: +52 222 2 663 152.
E-mail addresses: albertotellezv@ccc.inaoep.mx (A. Téllez-Valero), mmontesg@ccc.inaoep.mx (M. Montes-y-Gómez), villasen@ccc.inaoep.mx (L. Villaseñor-Pineda), anselmo@lsi.uned.es (A.P. Padilla).
[1] http://www.google.com.
[2] http://www.yahoo.com.

0306-4573/$ - see front matter © 2010 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2010.03.007

Recent research in QA has been mainly fostered by the TREC,[3] CLEF,[4] and NTCIR[5] conferences. These conferences consider different languages (English, European languages and Asian languages respectively) and contemplate different kinds of questions, such as factual questions (e.g., `Where is the Taj Mahal?`), definition questions (e.g., `What is the Quinoa?`), and list questions (e.g., `Who were the members of The Beatles?`). The results from these conferences have shown some interesting facts. On the one hand, they have shown that there are several QA systems that, in spite of their general poor performance, are highly accurate to respond certain kinds of questions. For instance, in the Portuguese QA track at CLEF 2008 (Forner et al., 2008), the system tagged as "diue081" correctly responded 89% of the definition questions, whereas, it only could respond to 35% of the factual questions. On the other hand, these results have also indicated that there is a high complementarity among different QA systems. Just as an example, the combination of the correct answers from all participating systems of the Portuguese QA track at CLEF 2008 (nine systems) could outperform by 49% (reaching a 82% of accuracy) the best individual result for factual questions (55%).

Based on these facts, some advanced approaches known as *multi-stream QA systems* attempt to improve the individual results by taking advantage of the complementarity from existing QA systems. Therefore, the major challenge of this kind of systems is to *select* the correct answer for a given question by combining the evidence from different input systems (or streams). Moreover, considering that for some questions no system is able to extract the correct answer, this challenge also involves determining the cases that require a *nil* response.

Traditional approaches for multi-stream QA rely on measuring the confidence of streams or the redundancy of answers across them (Burger et al., 2002; Clarke et al., 2002; de Chalendar et al., 2002; Jijkoun & de Rijke, 2004; Rotaru & Litman, 2005; Roussinov, Chau, Filatova, & Robles-Flores, 2005). On the other hand, recent approaches consider the application of *textual entailment recognition* (RTE)[6] techniques, which decide about the correctness of answers based on information from a given support text[7] (Glöckner, Sven, & Johannes, 2007; Harabagiu & Hickl, 2006). Motivated by all these previous efforts, in this paper we propose a new kind of *hybrid approach*, which is based on a supervised learning method that combines features from the traditional and textual-entailment approaches. In particular, it considers a set of features that describe the redundancy of answers across streams, the compatibility between question and answer types, as well as the overlap and non-overlap information between the question–answer pair and the support text.

Our experimental results in a set of 190 questions in Spanish language, considering answers from 17 different QA systems,[8] demonstrate the appropriateness of the proposed method. It reached an estimated QA performance of 0.74, significantly outperforming the estimated performance from the best individual system (0.53) as well as the result from best traditional multi-stream QA approach (0.60).

The rest of the paper is organized as follows. Section 2 describes the previous work in multi-stream QA. Section 3 presents our supervised multi-stream QA method, given special attention to the description of the used features. Section 4 shows the evaluation results in a collection of 190 Spanish questions, and compares these results against those from other traditional multi-stream approaches. Finally, Section 5 exposes our conclusions and outlines some future work directions.

2. Related work

A typical QA system consists of three main processes that are carried out in sequence: question analysis, document/passage retrieval, and answer selection. One problem with this kind of architecture is that it is highly affected by cascade errors (Rodrigo, Peñas, & Verdejo, 2008a). In order to reduce this problem new alternative approaches known as *meta QA systems* and *multi-stream QA systems* has been proposed.

On the one hand, meta QA systems internally combine several components or techniques at each QA process. For example, Pizzato and Molla-Aliod (2005) describe a QA architecture that uses several document retrieval methods, and Chu-carroll, Czuba, Prager, and Ittycheriah (2003) present a QA system that applies two different components at each process.

On the other hand, multi-stream QA approaches go a step forward by achieving a superficial combination of several QA systems. Most of these approaches are adaptations of multi-stream techniques from document retrieval (Belkin, Kantor, Fox, & Shaw, 1995; Lee, 1997); nevertheless, in this case, they are mainly focused on *selecting the correct answer* for a given question rather than on ranking all candidate answers. Following, we introduce the traditional approaches for multi-stream QA. In particular, we organize them in five categories taking into consideration some ideas proposed elsewhere (Diamond, 1998; Vogt & Cottrell, 1999). It is important to notice that the first two categories denote *system-centered* approaches, which generate their decisions using information about the system's confidences. Whereas, in contrast, the third and fourth categories correspond to *answer-centered* approaches, which select the final answer exclusively based on its frequency of occurrence across streams.

[3] Text REtrieval Conference; http://trec.nist.gov/.
[4] Cross Language Evaluation Forum; http://www.clef-campaign.org/.
[5] NTCIR Project; http://research.nii.ac.jp/ntcir/.
[6] RTE is determine whether, given two text fragments, the meaning of one text could be reasonably inferred, or textual entailed, from the meaning of the other text (Dagan, Magnini, & Glickman, 2005).
[7] The text fragment from which the answer was extracted.
[8] All these systems were evaluated at the Spanish QA track at CLEF-2006.

- *Ordered Skimming approach*: in this case the streams (individual QA systems) are ordered by their general confidence, and the final answer is selected from the stream with the highest one. Some systems based on this approach are described by Clarke et al. (2002) and Jijkoun and de Rijke (2004).
- *Dark Horse approach*: it can be considered an extension of the previous approach because it also considers the confidence of streams; however, in this case, these confidences are calculated for each type of question. Using this approach each stream has different confidences associated to factual and definition questions. Jijkoun and de Rijke (2004) described a multi-stream QA system based on this approach.
- *Answer Chorus approach*: it relies on the answer redundancies; it selects as the final respond the answer with the highest frequency across streams. Some systems based on this approach are described in de Chalendar et al. (2002), Burger et al. (2002), Jijkoun and de Rijke (2004), Roussinov et al. (2005), and Rotaru and Litman (2005).
- *Web Chorus approach*: it uses information from the Web to evaluate the relevance of candidate answers. It selects the answer with the greatest number of Web pages containing the answer terms along with the question terms. It was proposed by Magnini, Negri, Prevete, and Tanev (2001), and subsequently it was evaluated in Jijkoun and de Rijke (2004).
- *Hybrid approach*: it considers the combination of criteria from the system and answer-centered approaches. The method described in Jijkoun and de Rijke (2004) is an example of this kind of approach. It uses the system's confidences to differentiate between answers having the same frequency of occurrence. Its evaluation results indicated that this combination could outperform the results obtained by other multi-stream QA systems based on one single strategy.

In addition to the traditional approaches adopted from IR, more recently emerged a new type of multi-stream QA approach based on the application of *textual entailment recognition* (RTE) techniques (Dagan et al., 2005). The idea behind this approach is to decide about the correctness of answers based on their textual entailment with a given support text (Peñas, Rodrigo, Sama, & Verdejo, 2008), and, therefore, using these decisions to identify the more appropriate answer for the question at hand. Two systems based on this new approach are the ones reported by Glöckner et al. (2007) and Harabagiu and Hickl (2006), for Dutch and English respectively.

The method proposed in this paper is a new kind of *hybrid approach*, which is based on a supervised learning method that combines features from the answer-chorus and textual-entailment approaches. It mainly differs from previous multi-stream methods in the following concerns:

First, different from other hybrid methods, it does not consider any information about the confidence of the input QA systems. Instead, it uses several features that attempt to evaluate the textual entailment of the answers with a given support text. This difference makes our method independent from the used QA systems, and, consequently, makes it easily adaptable to work with different systems.

Second, different from other answer-chorus methods, our proposal does not consider the answer frequencies as the main or unique criterion for answer selection. This characteristic allows our method to be very appropriate for dealing with poor performance QA systems (something common in most non-English languages), where correct answers tend to show low frequencies.

Third, different from previous textual-entailment approaches (and also varying from supervised answer-validation methods (e.g., Jijkoun & de Rijke, 2006; Kozareva, Vázquez, & Montoyo, 2007)), our proposal not only considers features that indicate the overlap between the question–answer pair and the support text, but also includes some features that evaluates the non-overlapped information, allowing, in this way, to correctly analyze the situations where exists a high overlap but not necessarily an entailment relation between these two elements. In addition, our method is only based on a lexical–syntactic analysis of texts, avoiding the use of deep semantic analysis as well as the consult of diverse external knowledge sources. We consider this last characteristic to be very important for constructing multi-stream QA applications in different languages.

The following section describes in detail the proposed approach.

3. System description

Fig. 1 shows the general scheme of the proposed multi-stream QA approach. It consists of two main stages. In the first stage, called *QA stage*, several QA systems extract—in parallel—a candidate answer and its corresponding support text for a given question. Then, in the second stage, called *selection stage*, a classifier evaluates all candidate answers and assigns to each of them a category (correct or incorrect) as well as a confidence value (ranging from 0 to 1). At the end, the correct answer having the greatest confidence value is selected as the final response. In the case that all answers were classified as incorrect, the system returns a *nil* response.

The following subsection describes in detail the answer selection method. It is important to comment that the current method is an extension of our previous work evaluated in the answer-validation exercise at CLEF (Téllez-Valero, y Gómez, & Pineda, 2007). In particular, it considers a novel set of features that indicate the non-overlapped information between the question–answer pair and the support text, and therefore, allow to significantly improve our previous results in multi-stream QA (Téllez-Valero, y Gómez, Pineda, & Peñas, 2008).

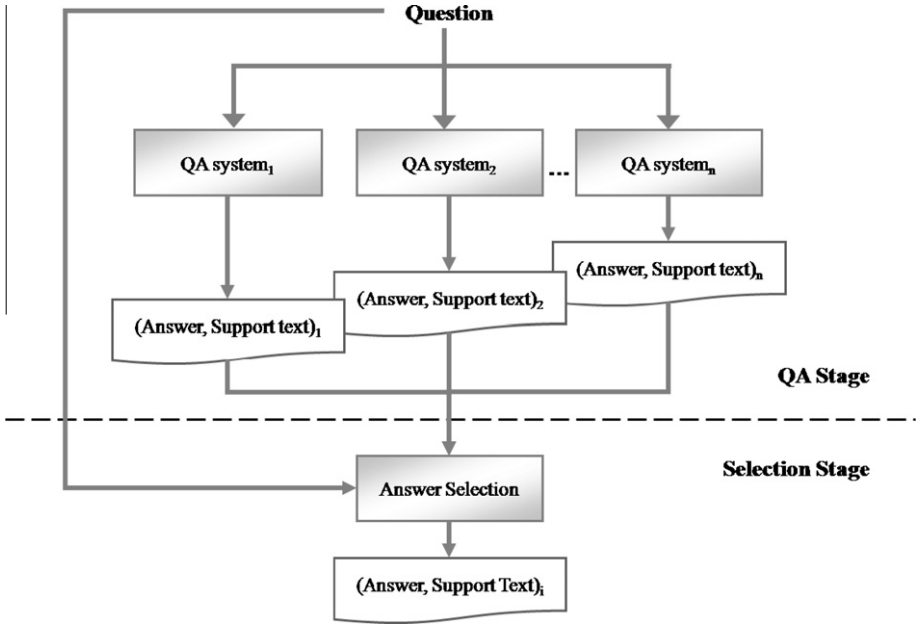

Fig. 1. General scheme of our multi-stream QA system.

3.1. Answer selection

As we previously mentioned, this stage focuses on the selection for the final response by combining evidence from the answers of different input QA systems. In particular, the classification of the candidate answers (in correct and incorrect categories) as well as the estimation of their confidence values is carried out by means of a supervised learning approach that considers three main processes: preprocessing, feature extraction and answer classification.

3.1.1. Preprocessing

The objective of this process is to extract the main content elements from the question, answer and support text, which will be subsequently used for deciding about the correctness of the answer. This process considers two basic tasks: on the one hand, the identification of the main constituents from the question–answer pair, and, on the other hand, the detection of the core fragment of the support text as well as the consequent elimination of the unnecessary information.

3.1.1.1. Constituent identification. We detect three basic constituents from the questions: its main action, the action actors, and if exist, the action restriction. As an example, consider the question in Table 1. In this case, the action is represented by the verb `invade`, its actors are the syntagms `Which country` and `Iraq`, and the action restriction is described by the propositional syntagma `in 1990`.

In order to detect the question constituents we firstly apply a shallow parsing to the given question.[9] Then, from the resulting syntactic tree (*Qparsed*), we construct a new representation of the question (called Q') by detecting and tagging the following elements:

1. *The action constituent*: It corresponds to the syntagm in Q_{parsed} that includes the main verb.
2. *The restriction constituent*: It is represented by the prepositional syntagm in Q_{parsed} having at least one explicit time expression (e.g., `in 1990`), or including a preposition such as `after` or `before`.
3. *The actor constituents*: These constituents are formed by the rest of the elements in Q_{parsed}. It is commonly divided in two parts. The first one, henceforth called *hidden actor constituent*, corresponds to the syntagm that includes the interrogative word and it is generally located at the left of the action constituent. The second part, which we call the *visible actor constituent*, is formed by the rest of the syntagms, generally located at the right of the action constituent.

Finally, we also consider an *answer constituent*, which is simply the lemmatized candidate answer (denoted by A').

3.1.1.2. Support text's core fragment detection. Commonly, the support text is a short paragraph—of maximum 700 bytes according to CLEF evaluations—which provides the necessary context to support the correctness of a given answer. However,

[9] For all language processes, namely, lemmatization, part of speech tagging, named entity recognition and classification, and shallow parsing, we employed Freeling (Atserias et al., 2006).

Table 1
Example of a question, its answer and its support text.

Question	Which country did Iraq invade in 1990?
Candidate answer	Kuwait
Support text	Kuwait was a close ally of Iraq during the Iraq—Iran war and functioned as the country's major port once Basra was shut down by the fighting. However, after the war ended, the friendly relations between the two neighboring Arab countries turned sour due to several economic and diplomatic reasons which finally culminated in an Iraqi invasion of Kuwait

in many cases, it contains more information than required, damaging the performance of RTE methods that are based on lexical–syntactic overlaps. For instance, the example of Table 1 shows that only a part of the last sentence in the support text (i.e., `Iraqi invasion of Kuwait`) is useful for validating the given answer; whereas, the rest of the text only contributes to produce an irrelevant overlap (e.g., `Kuwait was a close ally of Iraq`). In order to reduce the support text to the minimum useful text fragment according to the candidate answer, we proceed as follows:

- First, we apply a shallow parsing to the support text, obtaining the syntactic tree (S_{parsed}).
- Second, we match the content terms (nouns, verbs, adjectives and adverbs) from the question constituents against the terms from *Sparsed*. In order to avoid some minimal writing differences of the same concept not solved by the morphological analysis (e.g., `Iraq` against `Irak` or `Iraqi`). We compare the terms using the Levenshtein edition distance.[10] Mainly, we consider that two different words are equal if their distance value is less than 0.4.
- Third, based on the number of matched terms, we align the question constituents with the syntagms from the support text.
- Forth, we match the answer constituent against the syntactic tree (S_{parsed}). The idea is to find all occurrences of the answer in the given support text.
- Fifth, we determine the minimum context of the answer in the support text that contains all matched syntagms. This minimum context (represented by a sequence of words around the answer) is what we call the *core fragment* (denoted by T'). In the case that the support text includes several occurrences of the answer, we select the one with the smallest context.

Applying the procedure described above we determine that the core fragment of the support text showed at Table 1 is `in an Iraqi invasion of Kuwait`.

3.1.2. Feature extraction

This stage gathers a set of processes that allow extracting several features from the question, answer and support text. These features can be categorized in two groups: those that indicate characteristics of the question and the answer as well as their relation, and those that measure the entailment relation between the question–answer pair and the support text. The following sections describe both kinds of features and explain the way they are calculated from Q', A' and T'.

3.1.2.1. Features about the question and answer.

– Question characteristics

We consider four different features from the question: the question word (what, how, where, etc.), the question category (factual or definition), the expected answer type (date, quantity, name, or other), and the type of question restriction (date, period, event, or none).

The question word, question category, and the expected answer type are determined using a set of simple lexical patterns. Some of these patterns are showed below. It can be observed that each of them includes information about the question category and the expected answer type. It is important to mention that actually exists several machine learning approaches that have been successfully applied to question classification (e.g., Zhang & Lee, 2003, Blunsom, Kocik, & Curran, 2006). However, because the proposed question's classes are very general, we prefer using the classic approach based on hand-crafted rules (Lee, Oh, Huang, Kim, & Choi, 2000; Pasca & Harabagiu, 2001).

`(WHAT` *OR* `WHO) is` [*whatever*] → DEFINITION – OTHER
`HOW (many` *OR* `old)` [*whatever*] → FACTUAL – QUANTITY
`WHEN` [*whatever*] → FACTUAL – DATE
`WHERE` [*whatever*] → FACTUAL – NAME
`WHAT` is the name [*whatever*] → FACTUAL – NAME

On the other hand, the value of the question restriction (date, period, event or none) depends on the form of the restriction constituent. If this constituent contains only one time expression, then this value is set to "date". In the case the restric-

[10] The Levenshtein edition distance has been previously used in other works related to answer validation in Spanish (e.g., Rodrigo, Peñas, Herrera, & Verdejo, 2007).

tion constituent includes two time expressions, it is set to "period". If the restriction constituent does not include any time expression, then the question restriction is defined as "event". Finally, when the question does not have any restriction constituent, the value of the question restriction is set to "none". Following we show some examples of questions having different kinds of question restrictions.

`Where did Freud move to live` *in 1939*`?` → DATE
`What is the name of the collection of pictures painted by Goya` *from 1819 to 1823*`?` → PERIOD
`Where did Francis Cammaerts work with the Maquis` *during World War II*`?` → EVENT
`Who was Alexander Graham Bell?` → NONE

– Question–answer compatibility

This feature indicates if the question and answer types are compatible. The idea of this feature is to capture the situation where the semantic class of the evaluated answer does not correspond to the expected answer type. For instance, having the answer `yesterday` for the question `How many inhabitants are there in Longyearbyen?`

This is a binary feature: it is equal to 1 when the answer corresponds to the expected answer type, and it is equal to 0 if this correspondence does not exist.

– Answer redundancy

Taking into account the idea of "considering candidates as allies rather than competitors" (Dalmas & Webber, 2007), we decided to include a feature related to the occurrence of answers across streams.

Different from other redundancy methods (like the one present by Roussinov et al. (2005)) that directly use the frequency of occurrence of the answers, the proposed feature indicates the normalized sum of the inverse of the edition distances between the actual evaluated answer to each one of the other candidate answers. The edition distance strategy allows dealing with the great language variability and also with the presence of some typing errors. In this way, an answer `X` contributes to the redundancy rate of another answer `Y` and vice versa, even though `X` and `Y` are not exactly the same (e.g. `spacial telescope` and `Hubble telescope`).

3.1.2.2. Features related to the textual entailment recognition. The features of this category are of two main types: (i) features that measure the overlap between the support text and the hypothesis (an affirmative sentence formed by combining the question and the answer); and (ii) attributes that denote the differences (non-overlap) between these two components.

It is important to explain that, different from other RTE methods, we do not use the complete support text; instead, we only use its core fragment T'. In addition, we neither need to construct a hypothesis text, instead we use as hypothesis the set of question–answer constituents (that is, the union of Q' and A', which result is H').

– Overlap characteristics

These features express the degree of overlap—in number of words—between T' and H'. In particular, we compute an overlap feature for each one of the four types of content terms (nouns, verbs, adjectives and adverbs) as well as for each one of the six types of named entities (names of persons, places, organizations, and other things, as well as dates and quantities). We generate these ten different overlap features for each one of the five constituents in H' (the action constituent, the restriction constituent, the hidden actor constituent, the visible actor constituent, and the answer constituent). This way, we get a total of fifty features that represent the overlap characteristics, which take values between 0 and 1 indicating the degree of overlap. It is important to comment that the absence of some type of content term or named entity from both components (T' and H') is considered as a coincidence, and, therefore, its related overlap feature is set to 1.

Similar to the calculation of the answer redundancy, in this case we also apply the edition distance to evaluate the overlap between the terms from H' and T'. Besides, we also apply a special procedure to evaluate the temporal restriction fulfillment. For instance, we consider that the restriction constituent `in August 3, 1996` is satisfied by `in August 1996`, and that the restriction constituent `between 1990 and 1998` is overlapped by the text `in 1996`.

In order to illustrate the computation of these characteristics we consider the question `Which country did Iraq unjustly invade in 1990?`, the answer `ONU`, and the core fragment `ONU against Iraq since invade Kuwait in 1990`. In this case, the ten features corresponding to the action constituent are all set to 1 (because they have the same verb and the other type of content terms are not present) except for the one related to adverbs, which has value of 0 since the term `unjustly` from the action constituent of H' does not occur in the corresponding constituent of T'.

– Non-overlap characteristics

These features indicate the number of non-overlapped terms from the core fragment of the support text, that is, they indicate the number of terms from T' that are not present in any of the detected constituents. Mainly, we compute this value by counting the non-overlapped words from the text between the answer constituent and each one of the other constituents

Table 2
Resume of the proposed features used for answer classification.

Number/kind of features	Description
4 Question characteristics	Nominal features representing the question word, the question category, the expected answer type, and the type of question restriction
1 Question–answer compatibility	Boolean feature indicating the correspondence between the answer and the expected answer type
1 Answer redundancy	Numerical feature that indicates the normalized sum of the inverse of the edition distances between the given candidate answer and the rest of the answers
50 Overlap characteristics	Numerical features which describe the degree of overlap between T' and H'. These features are computed for each one of the four types of content terms and for each one of the six types of named entities. In addition, these 10 different overlap features are computed for each one of the five constituents from H'
40 Non-overlap characteristics	Numerical features that indicate the number of terms from T' that are not overlapped by H', and that occur between the answer constituent and each one of the four questions constituents. These four different non-overlap features are calculated for each one of the four types of content terms as well as for each one of the six types of named entities

(i.e., from the pairs answer-action, answer-restriction, answer-hidden actor, and answer-visible actor). We do that for each type of content term as well as for each type of named entity, generating a total of forty different non-overlap features.

Continuing with the previous example, the features extracted from the text between the answer constituent and the restriction constituent (i.e., `...against Iraq since invade Kuwait...`) are all set to 0 except for the one corresponding to place named entities, which has value of 1 because the term `Kuwait` from the core fragment does not occur in H'.

3.1.3. Answer classification

This process determines if each candidate answer is correct or incorrect, and also estimates a classification confidence (β) for each of them. The confidence values help to select the final response in situations where several candidate answers are classified as correct; in such cases the answer with the greatest confidence is selected as the final response.

The classification of candidate answers is carried out by means of a supervised learning approach, which determined the category (correct or incorrect) of each answer based on the 96 features described in Table 2. Using this approach, the confidence value for each answer corresponds to its classification probability distribution. In particular, we used the ADTree algorithm that is a combination of decision tress with boosting, and which generates classification rules that are small and easy to interpret (Freund & Mason, 1999). We employed the algorithm implementation by Weka (Witten & Frank, 1999), and configured it to apply 10 iterations and does not consider any prune method.

It is important to comment that during the development process we also carried out some experiments considering other learning algorithms available at Weka, such as Support Vector Machines, Naïve Bayes, C4.5 decision tree, *K*-Nearest Neighborhood as well as their ensembles using boosting and bagging approaches. Despite the classification results achieved by all approaches were comparable[11] (applying a 10 Fold Cross Validation over the train set described in Section 4.1.1), we decided using the ADTree algorithm because it allows obtaining more disperse confidence values, which make easy the selection of the final response for questions having several "correct" candidate answers.

4. Experimental results

4.1. Experimental setup

4.1.1. Training and test data

As we described in Section 3, the core component of the proposed multi-stream QA system is the answer selection module, which relies on a supervised learning approach.

In order to train this module we used the SPARTE corpus (Peñas, Rodrigo, & Verdejo, 2006). This corpus was build from the Spanish corpora used at the CLEF conferences for evaluating QA systems from 2003 to 2005. It contains 2962 training instances represented by the tuple $(\langle\text{question}\rangle, \langle\text{answer}\rangle, \langle\text{support-text}\rangle, \langle\text{answer-category}\rangle)$, where $\langle\text{answer-category}\rangle$ is a binary variable that indicates whether the answer is correct or incorrect following the CLEF evaluation criteria (i.e., an answer is correct if it is justified by its support text). One important fact about this corpus is that it is very unbalanced: 77% of the training instances are negative (their answer category is `incorrect`), whereas, just 695 instances (the other 23%) correspond to positive examples (answers labeled as `correct`). In some way, this distribution reflects the current performance of Spanish QA systems.

On the other hand, we used a set of 190 questions and the answers from 17 different QA systems[12] for evaluating the proposed approach. Table 3 describes some implementation details from these systems (streams). In total, our evaluation corpus

[11] We compared the classification accuracy of the ADTree against the accuracy from the other learning schemas, taking as null hypothesis (H_0) that *pA* is equal to *pB* where *pA* is the ADTree accuracy and *pB* is the accuracy to be compared. After applying a statistical significance test to evaluate the difference of two proportions (*pA* and *pB*) as is described in Dietterich (1998), we found that, with a critical region $|z| > Z_{\alpha=0.05}$, it was not possible reject the null hypothesis.

[12] The test set gathers the outputs from all QA systems that participated in the QA track of CLEF 2006 (Magnini et al., 2007), and it was employed at the first Spanish Answer Validation Exercise (Peñas, Rodrigo, Sama, & Verdejo, 2007).

Table 3
Implementation details of the QA systems used as streams in our multi-stream approach.

Stream	(i) Question analysis, (ii) Document/passage retrieval, (iii) Answer extraction
1	(i) Classification based on syntactic patterns (24 classes), it takes as input questions in English; (ii) passage retrieval applying the IR-n system; (iii) answer selection using syntactic patterns (Ferrández et al., 2007)
2	The same as stream 1, but in this case input questions are in Spanish
3	(i) Classification based on machine learning (10–15 classes); (ii) passage retrieval by means of the Indri system; (iii) answer selection using the relevance of passages and the frequency of answers (Tomás & González, 2007)
4	The same as stream 3, but in this case the retrieved passages are re-ranked using Latent Semantic Analysis
5	(i) Do not apply, it takes as input questions in English; (ii) it retrieves documents from the Web using Google, and passages from the document test collection using Lucene; (iii) answer selection based on the redundancy of answers (Whittaker et al., 2007)
6	The same as stream 5, but in this case input questions are in Spanish
7	The same as stream 5, but in this case input questions are in French
8	(i) Classification based on lexical patterns (3 categories); (ii) passage retrieval applying the vector space model (VSM) and *n*-gram comparison; (iii) answer selection using lexical patterns as well as a classifier based on lexical features (Juárez-González et al., 2007)
9	(i) Classification based on syntactic parsing (the number of classes are not specified); (ii) passage retrieval by means of the VSM; (iii) answer selection based on the frequency of answers (Bowden et al., 2007)
10	(i) Classification based on lexical patterns (4 classes); (ii) passage retrieval applying the Xapian system, in this case some multi-words terms are indexed (iii) answer selection using the frequency of answers (de Pablo-Sánchez et al., 2007)
11	The same as stream 10, but using only single words in the search index
12	(i) Classification based on contextual rules at a semantic level (86 categories); (ii) passage retrieval applying VSM and named entities comparison; (iii) answer selection using the contextual rules and the frequency of answers (Cassan et al., 2007)
13	The same as stream 12, but in this case input questions are in Portuguese
14	(i) Classification based on machine learning (6 classes); (ii) passage retrieval applying the IR-n system and named entities comparison; (iii) answer selection using lexical patterns, the relevance of the passages and the frequency of answers (García-Cumbreras et al., 2006)
15	(i) Classification based on lexical patterns (17 classes); (ii) passage retrieval using the JIRS system (*n*-gram comparison); (iii) answer selection using lexical patterns, the relevance of the passages and the frequency of answers (Buscaldi et al., 2007)
16	The same as stream 15, but in this case the passages are retrieved using Lucene
17	(i) Classification based on lexical patterns (3 categories); (ii) passage retrieval applying VSM and *n*-gram comparison; (iii) answer selection analyzing co-occurrences in a lexical–syntactic level and using lexical patterns (Pérez-Coutiño et al., 2007)

considered 2369 candidate answers with their corresponding support texts. Because this corpus only contains 671 correct answers (28% of the total), it represents a great challenge for the answer selection module, which must to detect these answers in order to allow a successfully multi-stream QA performance.

4.1.2. Evaluation measure

The evaluation measure most commonly used in QA is the *accuracy*, i.e., the percentage of correctly answered questions. Following the CLEF evaluation criteria, this measure is calculated as the fraction of correct answers plus correct *nil*-answers[13] with respect to the total number of questions (refer to Formula (1)).

$$\text{accuracy} = \frac{\#\_\text{correct\_answered\_questions} + \#\_\text{correct\_unanswered\_}nil\_\text{questions}}{\#\_\text{questions}} \tag{1}$$

In particular, in our experiments we used an evaluation measure called *estimated QA performance* (Rodrigo et al., 2008a), which was proposed at the 2008 Answer Validation Exercise (Rodrigo, Peñas, & Verdejo, 2008b). This evaluation measure is specially suited to compare the results from multi-stream QA systems against those from the input streams. Based on the fact that multi-stream QA systems do not have access to the target document collection, and, therefore, that their performance is always restricted by the answers extracted from the input streams, this measure (see formula (2)) not only considers the traditional accuracy but also a *reject accuracy*, which indicates the ability of a multi-stream QA system to correctly discard all candidate answers when all of them are incorrect.

$$\text{estimated\_QA\_performance} = \text{accuracy} + \text{reject\_accuracy} * \text{accuracy} \tag{2}$$

$$\text{reject\_accuracy} = \frac{\#\_\text{correct\_rejected\_questions}}{\#\_\text{questions}} \tag{3}$$

4.2. Experiments

4.2.1. First experiment: our multi-stream approach and the single streams

The objective of any multi-stream QA method is to combine the responses from different QA systems (streams) in order to increase the number of correct answers. In other words, its goal is to obtain a better performance than that from the best input stream.

[13] *Nil* questions do not have an answer in the target document collection, or even worst, they do not have any possible answer. As an example considers the question `What is the capital of Neverland?`; for this kind of questions given no answer is considered a correct response.

In a first experiment, we attempted to evaluate the achievement of this objective. We mainly compared the estimated QA performance of our method against that from the input streams. Table 4 shows the results of this experiment. This table also includes the result corresponding to a *perfect selection* of the correct answers from all streams (0.946). This value indicates the maximum reachable estimated QA performance for any multi-stream approach in this data set.

From this table, it is noticeable that our multi-stream approach completely satisfied its basic purpose given that its result significantly outperformed the results from all streams.[14] On the other hand, concerning the perfect combination result, our approach could correctly answer and reject 151 questions from a total of 190 (i.e., approximately a 80%); what is more, it could correctly answered 84% of the questions having a correct response.

4.2.2. Second experiment: our multi-stream approach against other approaches

As a second experiment we compared our method against the traditional multi-stream QA approaches (refer to Section 2). In particular, we implemented some methods from these approaches based on the following criteria:

- *The Ordered-Skimming Method*: It selects the answers in accordance to the global stream accuracies. In other words, it aims to select the answer from the best available (not *nil*) option. Taking into account the values from Table 4, it prefers returning answers from stream eight.
- *The Dark-Horse Method*: It selects the responses based on the stream accuracies for answer type. In particular, it prefers to extract the answers for factual questions from stream twelve, and the answers for definition questions from stream eight.
- *The Answer-Chorus Method*: It selects the answers based on their repetitions across different streams. That is, for each question it selects the most frequent answer. In the case that two or more answers have the same frequency, it randomly selects the final response.
- *The Web Chorus Method*: It selects the answers based on the number of Web pages, retrieved by Google, that contain the terms from the question (without the question word) along with the terms of the answer. Similar to the previous method, in the case that two or more answers obtained the same qualification, the final response is selected randomly.
- *TheAnswer-Chorus–Dark-Horse Method*. It selects the answers based on their repetitions across different streams (Answers Chorus Method), but in the case that several answers obtain the same maximal frequency, it applies the Dark Horse criteria to select the final response.

Table 5 shows the results from this experiment. These results demonstrate the appropriateness of the proposed multi-stream QA method, which improved in 41% the result from the best input stream. In contrast, not any of the other approaches could increase in more than 10% the performance of the best input stream.[15] Moreover, some of the traditional approaches could not outperform the result from the best input stream; that is the case of the Ordered-Skimming and Dark-Horse Methods.

Taking into account that the estimated QA performance not only depends on the number of correctly answered questions, but also on the number of correctly rejected questions (refer to Formula (2)), we modified traditional multi-stream QA methods in order to allow them rejecting some answers. The idea was to incorporate some filtering conditions that obligate these methods to eliminate the less reliable answers. That is, in the case that no answer could satisfy these conditions, a *nil* response will be delivered. Following we describe the modifications incorporated to each one of the methods.

- *Ordered-Skimming Method**: It only considers answers from the best five streams (i.e., it only returns answers coming from the streams that have an estimated QA performance greater than 0.3 (see Table 4)).
- *Dark-Horse Method**: It only returns answers coming from the best five streams for each question type. In this case there were selected the best five streams for answering factual questions as well as the best five for answering definition questions.
- *Chorus Method**: It only considers answers recommended by two or more streams; therefore, it discards all infrequent answers.
- *Chorus–Dark Horse**: Firstly, it applies the Chorus Method*, but when many responses have the same maximal frequency it uses the Dark-Horse Method to select the final response.

Table 6 shows the results from this experiment. It is interesting to notice that all methods improved their estimated QA performance when they could reject some answers. The explanation of this behavior is that these modifications allowed all methods to correctly reject some questions. This experiment also helped to reveal another important characteristic of our method. It could correctly reject several answers without using any information about the confidence of streams and with-

[14] For this conclusion, we compared the estimated QA performance of our multi-stream approach (0.74) against the best stream result (0.53), taking as null hypothesis (H_0) that pA is equal to pB (where $pA = 0.74$ and $pB = 0.53$). Then, we applied a statistical significance test to evaluate the difference of the two proportions as is described in Dietterich (1998), and found that, with a critical region $|z| > Z_{\alpha=0.05}$ and $n = 190$, it was possible to get a $z = 4.25$ that allows us to reject H_0.

[15] The result of the proposed approach significantly outperformed the results from all traditional multi-stream QA methods. For reaching this conclusion we compared the estimated QA performance of our multi-stream approach (0.74) against the best traditional multi-stream result (0.58), taking as null hypothesis (H_0) that pA is equal to pB (where $pA = 0.74$ and $pB = 0.58$). Then, we applied a statistical significance test to evaluate the difference of the two proportions as is described in (Dietterich, 1998), and we found that with a critical region $|z| > Z_{\alpha=0.05}$ and $n = 190$ it was possible to get a $z = 3.29$ which allows us to reject H_0.

Table 4
Results from our multi-stream approach and the single streams.

Stream	# of correct			Estimated QA performance
	Answered questions	Unanswered *nil* questions	Rejected questions	
1	25	16	0	0.216
2	48	17	0	0.342
3	49	7	0	0.295
4	34	10	0	0.232
5	10	1	0	0.058
6	24	5	0	0.153
7	16	3	0	0.100
8	**88**	12	0	**0.526**
9	31	7	0	0.200
10	26	10	0	0.189
11	15	11	0	0.137
12	85	12	0	0.511
13	33	10	0	0.226
14	21	18	0	0.205
15	57	13	0	0.368
16	45	12	0	0.300
17	64	16	0	0.421
Our method	**123**	0	28	**0.743**
Perfect selection	146	0	44	0.946

Table 5
Comparing our approach against traditional approaches.

Method	# of correct			Estimated QA performance
	Answered questions	Unanswered *nil* questions	Rejected questions	
Ordered Skimming	98	0	0	0.516
Dark Horse	99	0	0	0.521
Chorus	101	0	0	0.532
Web Chorus	32	0	0	0.168
Chorus–Dark Horse	110	0	0	0.579
Our method	**123**	0	28	**0.743**
Best input stream	88	12	0	0.526
Perfect selection	146	0	44	0.946

Table 6
Results obtained after improving the rejection capability.

Method	# of correct			Estimated QA performance
	Answered questions	Unanswered *nil* questions	Rejected questions	
Ordered Skimming*	98	0	7	0.535
Dark Horse*	99	0	6	0.538
Chorus*	98	0	15	0.557
Chorus-Dark Horse*	106	0	15	0.602
Our method	**123**	0	28	**0.743**
Best input stream	88	12	0	0.526
Perfect selection	146	0	44	0.946

out considering any restriction on the answer frequencies. In particular, our method correctly rejects a 64% of the unanswered questions. In contrast, in the best case other multi-stream QA methods could only rejected a 34% of these questions.

4.2.3. Results analysis

The results from the first experiment showed us that our multi-stream QA method could satisfactorily reach its main objective. That is, its estimated QA performance (0.743) enhanced the best individual result (0.526) by a relative improvement of 41%. This improvement was more evident in the case of factual questions (refer to Table 7), where our method outperformed the best stream result by 55% (from 0.459 to 0.710). In the case of definition questions (see Table 8) our approach (0.861) could only obtained a slightly gain of 4% in relation to the best stream (0.833). In this case, however, it is important to observe the low average accuracy of the input streams.

Table 7
Results corresponding to the set of 148 factual questions.

Method	# of correct			Estimated QA performance
	Answered questions	Unanswered *nil* questions	Rejected questions	
Our method	92	0	21	0.710
Best stream	60	8	0	0.459
Streams average	30	6	0	0.238
Perfect selection	111	0	37	0.938

Table 8
Results corresponding to the set of 42 definition questions.

Method	# of correct			Estimated QA performance
	Answered questions	Unanswered *nil* questions	Rejected questions	
Our method	31	0	7	0.861
Best stream	28	7	0	0.833
Streams average	10	5	0	0.352
Perfect selection	35	0	7	0.972

Concerning the results from the traditional multi-stream QA methods, we may observe the following:

First, the methods that rank answers based on the stream confidences, namely, the Ordered-Skimming Method and the Dark-Horse Method, also obtained relevant results. Nevertheless, it is necessary to mention that—in our implementations—these methods made use of a *perfect estimation* of these confidences[16]. For that reason, and given that in a real scenario it is practically impossible to obtain these perfect estimations, we consider that our proposal is more robust than these two methods.

Second, the results also give evidence that the presence of several deficient streams (which generate a lot of incorrect answers) seriously affects the performance of the Chorus Method, which normally is reported as one of the best multi-stream approaches.

It is also important to comment that we attribute the poor results achieved by the Web Chorus Method to the amount of online information for Spanish (which it is considerably less than that for English). Therefore, in order to obtain better results, it is necessary to apply some question/answer expansions using for instance synonyms and hyperonyms.

Finally, Jijkoun and de Rijke (2004) describe a multi-stream QA architecture that combines the Answers Chorus and the Dark-Horse Methods. Its evaluation results indicated that this combination outperformed the results obtained by other systems based on one single traditional multi-stream strategy. Our experimental results confirmed the conclusions of Jijkoun and de Rijke, but most important, they demonstrated the competence of our method since it could outperform the result from this hybrid approach.

5. Conclusions and future work

In this paper, we proposed a multi-stream QA method supported on a supervised learning approach. This method is founded on the idea of combining the output of different QA systems (streams) in order to obtain a better performance. Mainly, it is a new kind of *hybrid approach* that combines features from the answer-chorus and textual-entailment approaches.

The proposed method differs from previous multi-stream approaches in the following concerns: first, it does not consider any information about the confidence of the input QA systems; second, it does not consider the answer frequencies as the main or unique criterion for answer selection; and third, it includes some features that evaluates the non-overlapped information, allowing, in this way, to correctly analyze the situations where exists a high overlap but not necessarily an entailment relation between the question–answer pair and the support text. In addition, our method is only based on a lexical–syntactic analysis of texts, avoiding the use of deep semantic analysis as well as the consult of diverse external knowledge sources. All these features together make our method very appropriate for dealing with poor performance QA systems, which represent the current state for most non-English languages. In particular, we successfully evaluated our method in a Spanish test set, where current average answer accuracy is of 26% (please refer to Table 4).

Concerning the kind of used attributes, it is important to emphasize that, to our knowledge, our method is the first attempt to consider some features describing the *non-overlapped information* between the question–answer pair and the support text. Certainly, an evaluation of the proposed features during the development phase—using the information gain measure—showed us that the non-overlap and answer-redundancy attributes were the most discriminative.

[16] The confidences were calculated directly from the test set (refer to Table 4). We were obligated to do that because there is no correspondence between the systems that were used to generate the train and test sets.

From the evaluation results achieved on a test set of 190 Spanish questions gathered from the CLEF-2006 QA collection, we could observe the following:

- The proposed method significantly enhanced the estimated QA performance from the best individual stream. Its result (0.74) outperformed the best QA participating system (0.53) by a relative improvement of 41%. This relative increment was even better for factual questions, reaching a 55%.
- Although our method also takes advantage of the redundancy of answers across streams, it turned out to be less sensitive to their low frequency than other approaches. In particular, it outperformed the Answer-Chorus Method by 33%.
- Our method significantly outperformed the results from the Ordered-Skimming and Dark-Horse Methods, even though they used a perfect estimation of the system's confidences. This fact indicates that our method does not require any knowledge about the input streams, and, therefore, that it can be more easily adapted to different application scenarios.
- The proposed method allowed to significantly reduce the number of incorrect answers presented to the user. In relation to this point, our method was especially adequate to deal with questions having no correct answers. Particularly, it correctly rejected 64% of these questions, outperforming other multi-stream QA approaches in more than 85%.

Finally, it is clear that any improvement in the answer validation module will directly impact the performance of the proposed multi-stream method. Therefore, our future work will be focused on enhancing this module. In particular we plan to consider some new features in the entailment recognition process. On the one hand, we plan to include some additional discriminative features that allow describing with more detail the overlap between the question–answer pair with the support text. Mainly, we are considering using complex syntagms such as prepositional phrases and conjunctions/disjunctions. On the other hand, we plan to use Wordnet in order to consider synonyms and hyperonyms for computing the term and structure overlaps.

Acknowledgements

This work was done under partial support of CONACYT (Project Grant 43990 and scholarship 171610). We also thank the CLEF organizers.

References

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., & Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of the 5th international conference on language resources and evaluation (LREC'06)* (pp. 48–55).

Belkin, N. J., Kantor, P. B., Fox, E. A., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management, 31*(3), 431–448.

Blunsom, P., Kocik, K., & Curran, J. R. (2006). Question classification with log-linear models. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, & K. Järvelin (Eds.), *SIGIR* (pp. 615–616). ACM.

Bowden, M., Olteanu, M., Suriyentrakorn, P., Clark, J., & Moldovan, D. I. (2007). LCC's poweranswer at QA@CLEF 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 310–317). Springer.

Burger, J. D., Ferro, L., Greiff, W., Henderson, J., Mardis, S., Morgan, A., et al. (2002). MITRE's qanda at TREC-11. In *Text retrieval conference (TREC) TREC 2002 proceedings*. Department of Commerce, National Institute of Standards and Technology.

Buscaldi, D., Soriano, J. M. G., Rosso, P., & Sanchis, E. (2007). *N*-gram vs. keyword-based passage retrieval for question answering. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 377–384). Springer.

Cassan, A., Figueira, H., Martins, A. F. T., Mendes, A., Mendes, P., Pinto, C., et al. (2007). Priberam's question answering system in a cross-language environment. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 300–309). Springer.

Chu-carroll, J., Czuba, K., Prager, J., & Ittycheriah, A. (2003). In question answering, two heads are better than one. In *In Human language technology conference of the North American chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 24–31).

Clarke, C. L. A., Cormack, G. V., Kemkes, G., Laszlo, M., Lynam, T. R., Terra, E. L., et al. (2002). Statistical selection of exact answers (multitext experiments for TREC 2002). In *Text retrieval conference (TREC) TREC 2002 proceedings*. Department of Commerce, National Institute of Standards and Technology.

Dagan, I., Magnini, B., Glickman, O. (2005). The PASCAL recognising textual entailment challenge. In Proceedings of pascal challenge workshop on recognizing textual entailment (pp. 1–8). Southampton, UK.

Dalmas, T., & Webber, B. L. (2007). Answer comparison in automated question answering. *Journal of Applied Logic, 5*(1), 104–120.

de Chalendar, G., Dalmas, T., Elkateb-Gara, F., Ferret, O., Grau, B., Hurault-Plantet, M., et al. (2002). The question answering system QALC at LIMSI, experiments in using web and wordnet. In *Text retrieval conference (TREC) TREC 2002 proceedings*. Department of Commerce, National Institute of Standards and Technology.

de Pablo-Sánchez, C., González-Ledesma, A., Moreno-Sandoval, A., & Vicente-Díez, M. T. (2007). Miracle experiments in QA@CLEF 2006 in Spanish: Main task, real-time QA and exploratory QA using wikipedia (WiQA). In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 463–472). Springer.

Diamond, T. (1998). Information retrieval using dynamic evidence combination, unpublished Ph.D. Thesis Proposal, School of Information Studies, Syracuse University.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1924.

Ferrández, S., López-Moreno, P., Roger, S., Ferrández, A., Peral, J., Alvarado, X., et al. (2007). Monolingual and cross-lingual QA using aliqan and brili systems for CLEF 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information*

retrieval, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 450–453). Springer.

Forner, P., Peñas, A., Agirre, E., Alegria, I., Forascu, C., Moreau, N., et al. (2008). Overview of the clef 2008 multilingual question answering track. In *Working notes for the CLEF 2008 workshop*.

Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In *Machine learning: Proceedings of the sixteenth international conference* (pp. 124–133). Morgan Kaufmann.

García-Cumbreras, M. A., Ureña-López, L. A., Martínez-Santiago, F., & Perea-Ortega, J. M. (2006). Bruja system. The university of Jaén at the Spanish task of CLEFQA 2006. In *CLEF 2006 working notes*. Alicante, Spain.

Glöckner, I., Sven, H., & Johannes, L. (2007). Logical validation, answer merging and witness selection – A case study in multi-stream question answering. In *RIAO 2007 (Recherche d'Information Assistn++e par Ordinateur – Computer assisted information retrieval), Large-scale semantic access to content (text, image, video and sound)*. Pittsburgh, USA: Le Centre de Hautes Etudes Internationales d'informatique Documentaire – C.I.D.

Harabagiu, S., & Hickl, A. (2006). Methods for using textual entailment in open-domain question answering. In *ACL-44: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 905–912). Morristown, NJ, USA: Association for Computational Linguistics.

Jijkoun, V., & de Rijke, M. (2006). Recognizing textual entailment: Is lexical similarity enough? In I. Dagan, F. Dalche, J. Quinonero Candela, & B. Magnini (Eds.), *Evaluating predictive uncertainty, textual entailment and object recognition systems. LNAI* (Vol. 3944, pp. 449–460). Springer.

Jijkoun, V., & de Rijke, M. (2004). Answer selection in a multi-stream open domain question answering system. In S. McDonald & J. Tait (Eds.), *ECIR. Lecture notes in computer science* (Vol. 2997, pp. 99–111). Springer.

Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., y Gómez, M. M., & Pineda, L. V. (2007). Using machine learning and text mining in question answering. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 415–423). Springer.

Kozareva, Z., Vázquez, S., & Montoyo, A. (2007). University of Alicante at QA@CLEF2006: Answer validation exercise. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 522–525). Springer.

Lee, J. H. (1997). Analyses of multiple evidence combination. In *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval. Combination techniques* (pp. 267–276).

Lee, K.-S., Oh, J.-H., Huang, J., Kim, J.-H., & Choi, K.-S. (2000). TREC-9 experiments at KAIST: QA, CLIR and batch filtering. In: *Text retrieval conference (TREC) TREC-9 proceedings*. Department of Commerce, National Institute of Standards and Technology.

Magnini, B., Negri, M., Prevete, R., & Tanev, H. (2001). Is it the right answer? Exploiting web redundancy for answer validation. In *ACL '02: Proceedings of the 40th annual meeting on Association for Computational Linguistics* (pp. 425–432). Morristown, NJ, USA: Association for Computational Linguistics.

Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., et al. (2007). Overview of the CLEF 2006 multilingual question answering track. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 223–256). Springer.

Pasca, M., & Harabagiu, S. (2001). High performance question answering. In W. B. Croft, D. J. Harper, D. H. Kraft, & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR-01)* (pp. 366–374). New York: ACM Press.

Peñas, A., Rodrigo, Á., Sama, V., & Verdejo, F. (2007). Overview of the answer validation exercise 2006. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 257–264). Springer.

Peñas, A., Rodrigo, Á., Sama, V., & Verdejo, F. (2008). Testing the reasoning for question answering validation. *Journal of Logic and Computation, 18*(3), 459–474.

Peñas, A., Rodrigo, Á., & Verdejo, F. (2006). SPARTE, a test suite for recognising textual entailment in Spanish. In A. F. Gelbukh (Ed.), *CICLing. Lecture notes in computer science* (Vol. 3878, pp. 275–286). Springer.

Pérez-Coutiño, M. A., y Gómez, M. M., López-López, A., Pineda, L. V., & Pancardo-Rodríguez, A. (2007). Applying dependency trees and term density for answer selection reinforcement. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 424–431). Springer.

Pizzato, L. A. S., & Molla-Aliod, D. (2005). Extracting exact answers using a meta question answering system. In *Proceedings of the Australasian language technology workshop*. Sydney, Australia (pp. 105–112).

Rodrigo, Á., Peñas, A., Herrera, J., & Verdejo, F. (2007). The effect of entity recognition on answer validation. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 483–489). Springer.

Rodrigo, Á., Peñas, A., & Verdejo, F. (2008b). Evaluating answer validation in multi-stream question answering. In *Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: Information retrieval, question answering, and cross-lingual information access*.

Rodrigo, Á., Peñas, A., & Verdejo, F. (2008b). Overview of the answer validation exercise 2008. In *CLEF 2008 working notes*. Aarhus, Denmark.

Rotaru, M., & Litman, D. J. (2005). Improving question answering for reading comprehension tests by combining multiple systems. In *Proceedings of the American Association for Artificial Intelligence (AAAI) 2005 workshop on question answering in restricted domains*. Pittsburgh, PA.

Roussinov, D., Chau, M., Filatova, E., & Robles-Flores, J. A. (2005). Building on redundancy: Factoid question answering, robust retrieval and the "other". In *Proceedings of the thirteenth text retrieval conference (TREC 2005)* (pp. 15–18).

Téllez-Valero, A., y Gómez, M. M., & Pineda, L. V. (2007). A supervised learning approach to Spanish answer validation. In C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, & A. Peñas, et al. (Eds.), *CLEF. Lecture notes in computer science* (Vol. 5152, pp. 391–394). Springer.

Téllez-Valero, A., y Gómez, M. M., Pineda, L. V., & Peñas, A. (2008). Improving question answering by combining multiple systems via answer validation. In A. F. Gelbukh (Ed.), *CICLing. Lecture notes in computer science* (Vol. 4919, pp. 544–554). Springer.

Tomás, D., & González, J. L. V. (2007). Re-ranking passages with lsa in a question answering system. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 275–279). Springer.

Vogt, C. C., & Cottrell, G. W. (1999). Fusion via a linear combination of scores. *Information Retrieval, 1*(3), 151–173.

Whittaker, E. W. D., Novak, J. R., Chatain, P., Dixon, P. R., Heie, M. H., & Furui, S. (2007). CLEF2006 question answering experiments at Tokyo Institute of Technology. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval*, 7th workshop of the cross-language evaluation forum, CLEF 2006, Alicante, Spain, September 20–22, 2006. Revised selected papers. Lecture notes in computer science (Vol. 4730, pp. 351–361). Springer.

Witten, I. H., & Frank, E. (1999). *Data mining: Practical machine learning tools and techniques with java implementations*. Morgan Kaufmann.

Zhang, D., & Lee, W. S. (2003). Question classification using support vector machines. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 26–32). New York, NY, USA: ACM.

Alberto Téllez-Valero is a PhD student at the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. His areas of interest include question answering, answer validation, recognizing textual entailment, information extraction and retrieval, and text classification.

Manuel Montes-y-Gómez is a titular researcher at the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics (INAOE) in Puebla, Mexico. He obtained his PhD in Computer Science from the Computing Research Center of the National Polytechnic Institute of Mexico. His research interests include question answering, information extraction and retrieval, text classification, and text mining.

Luis Villaseñor-Pineda. He obtained his PhD in Computer Science from the Université Joseph Fourier, France. Since 2001, he is professor of the Department of Computational Sciences of the National Institute of Astrophysics, Optics and Electronics in Puebla, Mexico. His areas of interest include questions answering, text classification, and speech recognition and dialogue systems.

Anselmo Peñas Padilla. He is an associate professor at the Department of Languages and Computing Systems of the National University of Distance Education (UNED) in Madrid, Spain. He obtained his PhD in Computer Science from the UNED. His research interests include question answering evaluation, answer validation, recognizing textual entailment, and information retrieval for a multilingual information access.