Question Answering Evaluation Survey

L. Gillard, P. Bellot, M. El-Bèze

Laboratoire d'Informatique d'Avignon (LIA) – Université d'Avignon
339 ch. des Meinajaries, BP 1228 ; F-84911 Avignon Cedex 9 (France)
{laurent.gillard, patrice.bellot, marc.elbeze}@univ-avignon.fr

Abstract

Evaluating Question Answering (QA) Systems is a very complex task: state-of-the-art systems involve processing whose influences and contributions on the final result are not clear and need to be studied. We present some key points on different aspects of the QA Systems (QAS) evaluation: mainly, as performed during large-scale campaigns, but also with clues on the evaluation of QAS typical software components; the last part of this paper, is devoted to a brief presentation of the French QA campaign EQueR and presents two issues: inter-annotator agreement during campaign and the reuse of reference patterns.

1. Introduction

Question Answering (QA) has been widely studied since the first TREC Question Answering Track in 1999. Question Answering System (QAS) has evolved but still remain mapped on a typical architecture involving many steps such as question analysis, document and/or passage retrieval and, lastly, final answer selection strategies. Contribution of each of this steps need to be evaluated separately in order to understand their impact on final performance of the QAS, and one should do it from a component point of view.

In this paper, we first present a survey of the main evaluation campaigns available to QA community, but also the measures used to evaluate each kind of question considered in these campaigns. Second part discusses about evaluation of main QAS components and propose key points on how to do. Last part is devoted to a brief presentation of the French QA campaign EQueR and introduces two issues: inter-annotator agreement during the campaign and the reuse of reference patterns.

2. Evaluating QA

2.1. Overview of the main evaluation campaigns

This section develops a survey of different QA tasks as they had been defined during the large-scale monolingual campaigns such as TREC, NTCIR (QAC1 and QAC2) and multilingual campaigns such as CLEF. Metrics used for these tasks are presented and discussed: Precision and Recall, Mean Reciprocal Rank, Confidence Weighted Score, an others. Some evaluation metrics best suited for the List or Definition Questions such as Jimmy Lin's Pourpre are also under examination.

Since the first evaluation of Question Answering Systems (QAS), organized during TREC-8 (Voorhees, 1999), QAS are considered as a whole and evaluation performed only on final answers.

2.1.1. The TREC Question Answering Campaigns

The main purpose of QAS was (and still is) to move from Document Retrieval (DR) to Information Retrieval (IR) by extracting relevant and concise answers to open domain questions in a large collection of documents. As a precursor, TREC-QA successive tracks (Voorhees and Harman, 2005) evolved over the years to explore new and more difficult QA issues.

In TREC-8, the challenge was to obtain 50 or 250-bytes document chunks containing answers to some given fact-based questions.

For TREC-9, tracks remained the same, but both questions and documents collection were larger. Questions were more realistic and derived from real users factoid questions (rather than built specifically for the QA task from answers mined from the collection). Also, questions set contained few syntactic variants of an original one.

For the Main track of TREC-2001, required chunk size was reduced to 50 bytes and questions were no more guaranteed to have an answer in the collection (NIL answer). This campaign also introduced two new tracks: List questions whose answers were scattered across multiple documents *(What are 9 novels written by John Updike?)*; and Context which contains series of related questions (to test tracking discourse objects ability).

Since TREC-2002, systems had to provide only one Exact answer (instead of five candidates in previous TREC campaigns). For this campaign, List track was continued and Context abandoned. Another novelty was that answers set must be ordered by confidence (most confident supposed answers should be ranked first, and the least confident should be ranked last).

Passage track, consisting of 250-bytes answers, was again in TREC-2003. While Main track added to previously Factoid and List new type of Definition questions (such as *Who is Colin Powell?* or *What is the disease SARS?*) and ordering all answers by confidence was given up.

TREC-2004 merged all TREC-2003 tracks in one and evaluated series of questions. Each series was related to the same object (called a "definition target"; such as persons, *Jean Harlow*, organization, *Harlem globe trotters*, or things *Cassini space probe*). It contained "several factoid questions, zero to two list questions, and exactly one Other question" which should be understood as "Tell me other interesting things about this target I don't know enough to ask directly".

Last, TREC-2004 campaign proposed two tracks. The first one, Main was similar to for TREC-2003 but added event targets (*Russian submarine Kursk sinks*), and more difficult questions with temporal constraints, more anaphors and even reference to previous answers. Another track was Relationship. It stated that QAS users do not have a clearly defined information need and therefore cannot be associated with a single semantic answer type. An example of Relational Questions is *The analyst is interested in Cuba's modern role in Angola. Despite the end of combat support, is Cuba still a significant presence in Angola?*.

2.1.2. The CLEF Question Answering Campaigns

Cross-Language Evaluation Forum (CLEF) introduced its first QA evaluation as a pilot track in 2003 (Magnini et al. 2003). It was centered on European Languages and proposed both monolingual (Dutch, Italian, Spanish) and multilingual (actually mining an English corpus from Italian, Spanish, Dutch, French and German questions) exercises. Answer strings could be 50-bytes long or Exact answers. At most three responses were allowed per questions and NIL answers were possible.

For CLEF 2004 (Magnini et al. 2004), Main evaluation shifted to nine sources languages and seven targets to constitute fifty different mono and bilingual tasks. Questions sets were augmented with How and Definition questions, only one Exact answer per question was allowed. In parallel, a new pilot task was added: List questions but with conjunctives and disjunctives series; temporal restrictions in questions (before, during, or after a date, a period, or an event). It studied QAS self-confidence ability.

CLEF 2005 (Vallin et al. 2005) continued as CLEF 2004's Main, the increased number of covered languages and added subtypes of factoid temporally restricted questions. Due to the number of tasks, and difficulties to assess them, How and Object questions were withdrawn.

2.1.3. The NTCIR QA campaigns

The NTCIR QA campaigns were the TREC and CLEF equivalents for Asian languages. Question Answering Challenges (QAC) were large-scale monolingual, while last CLQA (Sazaki, 2005) was a Cross-Language QA (Questions to Documents: English to Chinese, English to Japanese and their opposites; but included also Chinese to Chinese track).

QAC1 (Fukumoto et al., 2004) was subdivided in three subtasks: first, was formulated like TREC-9 but answer string were limited to nouns, noun-phrases and values (rather than chunks of document) and didn't need to be extracted from the collection. Second subtask was an equivalent of TREC List series. Last subtask dealt with follow-up questions with ellipses or pronominalized elements (as it frequently occurs in Japanese sentences).

The QAC2 campaign was similar to QAC1 but each answer strings needed to be extracted from a document contained in the collection (Fukumoto et al., 2004). Also, Subtask 3 (Kato et al., 2004). added more than one follow-up questions of different types to measures abilities of dialogue.

Last QAC3 (Kato, 2005) was "limited" to Subtask 3 which was called Information Access Dialogue (IAD) and still simulated an interactive use of a QAS.

From this survey, one can see that QA has shifted from single, simple, fact-based questions to more complex information access dialogue situation (Burger et al. 2001; Strzalkowski, 2005). Current QAS has to deal with: tracking discourse object inside questions, but also inside documents collection; restrictions such as temporal one; gathering many types of information (not only nouns, noun phrases or Entities) with their interrelations; abilities to synthesized the information mined; and even being able to "push" useful information to the user (as for *Others* questions of recent). All this enumerative should be preferably done in a multilingual context. It worth notice that other evaluation campaigns already propose platform to evaluate some of these challenging aspects. An example is The Document Understanding Campaign (DUC) where some topics are only expressed with questions, sometimes nested, while the final result to provide is the corresponding summary.

Indeed, for being really usable, QAS will also need to be able to present all their conclusions (rather than *results* due to the amount of knowledge involved) in clean user interface (probably *browsable* to point out new exploration directions) and to cope with some interactions (to precise some needs when they occur). These conditions stated, QAS should be satisfactory from an user point of view.

2.2. Measures for QA evaluation

Before surveying main measures to evaluate QAS, one must define how answers and correctness are defined, at least in a QA campaign point of view. So, an answer is judged Correct if it is responsive to a question and extracted from a document which stated it. If document does not support a correct answer string, it is Unsupported. If some bits are missing or added, it is judged Inexact (in contrast, many QA campaign ask for only Exact answers). And, if none of the previous conditions is realized it is Incorrect. Sometimes, a question does not have any known correct answers in the document collection, this is called a NIL answer and it will be judged correct if it was answered as expected.

To provide some automatic evaluation facilities, for each question, pairs of [correct answers patterns, supporting document identifier] are often derived from the *Known* sets of correct answers (which we will call *R*-set, and it is built from union of all correct answers provided by all participant QAS and assessors). When using only patterns, evaluation is Lenient (and in a way, does not "ensure" that document stated the answer) while Strict in the other case.

Measures commonly used in QA are presented below:

Recall and Precision. They have been long used in Information Retrieval as soon as a relevant set of objects can be related to a need (as relevant documents to a user query). Precision is a measure of the accuracy whereas Recall is a measure of its exhaustivity. They may be both combined in a weighted harmonic mean called F-Measure, which could then be averaged to a Mean F-Measure.

Actually evaluating Recall is a difficult task in QA as it supposes to be able to consistently enumerate all the correct answers available in a collection for each question! Most of times, it is approximated by the *R*-set.

Mean Reciprocal Rank: Mean Reciprocal Rank (MRR) measures the ability of a QAS to answer a set (factoid) questions (Q). Score of an individual question is its Reciprocal Rank (RR), defined as the inverse of the order of the first correct answer or zero is no correct answer is given (five ordered candidate answers were commonly allowed per question). The whole set is then scored with the mean of reciprocal ranks as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} RR(q_i)$$

$$RR(q_i) = \begin{cases} \dfrac{1}{\text{rank of first Correct answer for } q_i \text{, up to N}} \\ \text{or 0, if no Correct answer for } q_i \text{ in N first} \end{cases}$$

MRR was first used during TREC-8 (N=5). As stated by (Voorhees, 1999) advantages of RR are: its closeness to average precision measure used to evaluate Document Retrieval systems; its inclusive bounds between 0 and 1; its good average capacity; and "not retrieving any correct answer for a question is penalizing but not unduly so". Drawbacks are: the finite values a score can take (for up to five answers: 0, .2, .25, .33, .5 and 1) and the fact that no

credit is provided "for retrieving multiple (different) correct answers" or "realizing it did not know" it.

"Confidence Weighted" Score, (CWS, also known as Average Precision) was defined for TREC-11 and is inspired from Document Retrieval's uninterpolated average precision. Assuming exactly one possible answer by question (from a *Q* set), and that all these answers must be ordered by system's confidence in its response, CWS rewards correct answers placed earlier more than later one. It is defined as:

$$CWS = \frac{1}{|Q|}\sum_{i=1}^{|Q|}(\text{number correct in first } i \text{ ranks})$$

CWS ranges between 0 (no correct at all) and 1 (perfect score). It was abandoned because it was inappropriate due to changes in the TREC-12 track definition, but also because, it could reward QAS ordering answers strategies rather than their correctness to answer.

K-Measure was designed at CLEF 2004's Pilot Task (Herrera et al., 2004) for rewarding systems "that return as many different correct answers as possible to each question, but at the same time, punishing incorrect answers". It needs as a prerequisite a "self" confidence score, between 0 and 1, for each answer candidates. K-Measure is defined as follows:

$$K = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{\sum_{a \in sysAnswers_i} confidence(a) \cdot judgement(a)}{\max\{|R_i|, |sysAnswers_i|\}}$$

where, for a question *i*: R_i is the set of correct answers; $sysAnswers_i$ is the set of answer candidates (*a*); *confidence(a)* is the confidence score given by QAS to an answer; and *judgement(a)* is the judgment given by an human assessor valued between {1, 0, -1} respectively for an answer judged correct, for an answer already judged, and for an incorrect one. K-Measure is ranged between 0 (no evidence of correctness) and 1 (certainty of correctness). As authors pointed out, the main difficulty is still determining an exhaustive R_i set for each question. So, they propose an alternative, called K1-Measure, when only one single answer is admitted per question. It assumes R_i can be approximated by the *Known* set of correct answers. K1 is defined like K-measure without the recall part (the "maximum" denominator). Also, to have better clues on correlation between self-scoring confidence and assessors judgment correctness, they defined a correlation coefficient called *r*.

List Questions or measuring many answers at the same time: Expected answers (instances) to List questions constituted a closed set usually assembled from many documents. It can be evaluated using **Accuracy**, i.e. the ratio of distinct correct answers to target number of instances; or even with Recall, Precision and F-Measure since the *R*-set is actually exhaustive for these questions. Last possibility is to adapt the Non-Interpolated Average Precision (NIAP) measure borrowed to Document Retrieval community and defined as:

$$NIAP = \frac{1}{|R_i|}\sum_{a \in R_i}\frac{\left|\{a' \mid a' \in R_i, rank(a') \le rank(a)\}\right|}{rank(a)}$$

where, for a question *i*: R_i is the set of correct answers; and *rank(a)* is the rank of an answer in the (ordered) answer candidates to evaluate.

Measures for Definition Questions. Such questions pose a problem: they accept many possible correct answers, but all will not be accurate or interesting, from a user's point of view (profiled as "an adult, a native speaker of English, and an average readers of 'US' newspapers"). Therefore, since recent evaluations, assessors built a by pipelining many processing list of vital nuggets of information, a QAS should return to a Definition question (Voorhees, 2003). Using this list, questions can be *manually* evaluated at a conceptual level (to abstract vocabulary, syntactic and others paraphrases mismatches). Recall, Precision (depending on an allowance length score to penalize verbosity) and F-measures are then computed.

(Lin and Demner-Fushman, 2005) and recently (Marton, 2006) proposed automatic metrics to replace this final manual evaluation. First authors, presented a measure, traced from ROUGE measure - (Chin-Yew and Hovy, 2003) used in automatic document summarization - which made the same assumption that the term co-occurrence statistics (unigrams) can replace the manual semantic matching process. They showed that their metrics highly correlate with TREC assessors' evaluation. Marton, echoed and investigated further to better approximate assessment score. His method used binary classifiers, based on up to trigrams, using function of an *idf*-based weight, an informativeness score and a recall threshold. He showed many improvements such as interpretable score, a confidence interval, exactly reproducing already done human assessments, better accuracy through additional annotation and support for using judgment data.

3. Evaluation of QAS components

The majority of QAS can be seen as "black boxes" producing an answer to a question, by pipelining many processing steps. Nevertheless, evaluating each step is challenging in order to study loss and gain but also to detect any components weakness. Ideally, each components should be evaluated separately, and whenever possible compared by using widely spread and recognized measurements. This section dissects a typical QA architecture and surveys different evaluation methodologies for their components.

Typical QA architecture relies on four main steps, which are pipelined most often:

- *Question Analysis*, to extract a semantic type of the expected answer;
- *Document Retrieval* (DR) to restrict the amount of processed data by further components;
- *Passage Retrieval* (which can also be merged with the previous step) to choose the best answer passages from documents;
- *Final Answer Extraction Strategies* to determine best answer candidate(s) from the selected passages.

Question analysis: Question analysis extracts some clues from the question such as (Ferret, et al. 2001): expected answer type(s), question category, interesting terms, focus (which corresponds to a noun or a noun phrase that is likely to be present in the answer, like *king* from *Which king signed the Magna Carta?*), focus head modifiers, semantic supporting words, syntactic relations, and main verb. From this analysis, some extraction patterns or any other adapted answering strategies can be used in the following steps. It can also build specific DR queries.

Another interesting issue which can be associated with this step is the prediction of the question difficulty (Grivola et al., 2005; Sitbon et al., 2006) in order to, for example, adapt the processing pipeline or, as already done (though not based on this difficulty prediction but only on not retrieving *enough* documents in the DR step), relax constraint on keywords used for Document Retrieval.

Concerning questions category, the QA Typology from ISI (Hovy et al., 2002) worth notice and contains 94 nodes

(47 leaf nodes) from 17,384 questions analyzed (provided by Answers.com).

Even, if there didn't exist any common/reference hierarchy (it is strongly dependent on other components like Name Entity detection or specific processing), from an evaluation point of view, it can be seen as a categorization task against a hierarchy.

Document Retrieval (DR) finds relevant documents to a user query extracted from a document collection. In QAS, this query is a subset of question words or a result from the question analysis step. This crucial step may lead to a "not found" (probably not satisfactory from a user point of view) or, worst, to an incorrect answer if no documents containing a correct answer can be found.

In order to evaluate DR one should rely on an evaluation platform such as TREC ad-hoc tasks for reference corpora. Measures commonly used are Recall and Precision but many other alternative measures are also described in (Baeza-Yates and Ribeiro-Neto, 1999).

This evaluation need has also been expressed during the last TREC (2005) since participants were asked to provide an ordered list of documents for each question, so that DR step and its effects could be further studied.

From a campaign evaluation perspective, (Nomoto et al. 2004) identified at least two features related to Document Retrieval which affected the performance of the QAS evaluated during QAC campaign.

Passage retrieval narrows the search to smaller chunks of text from two points of view. From an "arbitrary" one, by using fixed size windows in terms of sentences or words around candidate answers. If so, evaluation is done as for the Document Retrieval step. The second point of view is to keep some coherence inside passage containing a candidate answer. In this case, the evaluation methodology is similar to the one used in topic segmentation. Measures commonly used are Recall and Precision but also Beeferman measure (Beeferman et al., 1997) and WindowsDiff (Pevzner and Hearst, 2002). See (Sitbon and Bellot, 2006) for a discussion on these measures and an introduction on how segmentation reference corpora can be built.

Final answer(s) strategies: at the end of the processing pipeline, QAS has to choose answer(s) to provide from a set of passages containing answers. These are selected by using extraction patterns, mapping between a fine grained Named Entity and expected answer types, or any other syntaxic or semantic overlapping. Then, final choice utilizes a simple cut-off or some computations (like question words density metrics, redundancy metrics, etc.).

This quick description, shows some of the difficulties to evaluate final answer production from a QAS. QA campaigns do final answers evaluation but take for granted that performances from other components are already known. From a strictly numerical point of view, when judges only score the first five answers: they miss the fact that questions might have a correct answer at the sixth ranks; or that the sixth answer had the same score as the previous five and that they appears in arbitrary order. It might also happen that the correct answer was provided many times at different positions (maybe is it a clue that QAS failed or missed some tiling?). All these issues are lost in QA campaigns.

From a strategy point of view, extraction patterns can be evaluated like Information Extraction tasks.

Any score will probably need its own evaluation framework. For an example for evaluating a density metric, one should isolate some characteristic passages and study influence of adding some useful or non-useful words to the passage. And there still be the problem of designing a reference or a baseline to compare with.

To evaluate the ordering scheme, an intermediate component could be added to do permutation inside the sets of (final or passages) answers to see it effects on final ordering strategy and the final score.

Named Entity Tagging: many QAS system dealing with factoid questions are based upon an expected answer types hierarchy mapped to a (more or less fine grained) Named Entities (NE) hierarchy. Answering a question is then pairing the most appropriate question type to a detected NE sharing some context with the question.

NE tagging can be done anywhere in the pipeline, but it is generally done during Question Analysis and Passage Retrieval.

There is no exhaustive Named Entity hierarchy which is probably too task dependent and, when restricted to open domain as covered by factoid QA, almost any object can become an "extended" entity. Nevertheless, some like the one proposed in (Sekine and Nobata, 2004) are promising (it proposes 200 named entity categories) and could be used as a reference.

Another problem is the lack of annotated reference corpora with sufficiently fined grained named entities as, for example, in the context of geographic references annotation (Clough and Sanderson, 2004).

To develop a base system, one could use the data available from the Message Understanding Conference (MUC-6, MUC-7 proposed NE Recognition between 7 categories but also two other tasks: co-reference resolution, and template relation) Information Retrieval and Extraction Exercise (IREX, in Japanese language with 8 categories) and Automatic Content Extraction (in English, 5 categories). There are no equivalent corpora in all languages or multilingual contexts although (Poibeau et al., 2003) and (Bering et al., 2003) proposed interesting approaches. The latter propose to reuse corpora intended to other communities. For example, (Favre et al., 2005) derived a NE Recognition System from the *French ESTER Rich Transcription program on Broadcast News* data.

4. Evaluations and issues of EQueR

In July 2004, the first French QA campaign Technolangue EQueR (French acronym of *Évalaution en Question Réponse*) has been organized (Ayache et al., 2005 & 2006). Participating systems had to deal with Factoid, Definition, List questions but also *Yes/No* questions which introduced for the first time during a QA evaluation campaign.

EQueR Evaluation Platform purposed two tracks:

- A General track covered open domain QA from French speaking newspapers articles and French laws' text and is comparable to other QA campaigns.
- Another track, aimed at restricted QA in context of medical issues from a specialized documents and questions set (for more information, see Ayache et al., 2005).

For these tracks, questions were broken down in subtasks for: Factoid, Definition, List and Yes/No questions. Two types of answer strings were allowed: Exact answers, and Passage answers (i.e. 250-bytes chunk of text containing an answer) both accompanied by a document ID from the collection (to regard Supported and Unsupported issues). Exact and Passage answers had to be judged independently

one from the other but in the context of document provided. The only exception was for Yes/No questions where the passage must explain why the final *yes* or *no* answer

Factoid, Definition and Yes/No questions were evaluated by using MRR (with up to five candidates answers). In comparison, as previously seen, for last TREC, Definition questions had to receive all possible interesting answers. Yes/No questions were of course allowed only one answer, thus MRR was actually just correctness (0 or 1). List questions were evaluated by using NIAP. All final score were normalized in one final score. Results of participating system are presented in (Ayache et al., 2005 & 2006).

Inter-annotator agreements insight. For EqueR, two assessors score all results provided by the QAS. The results were not "pooled" (like it was during TREC evaluation: the same answer provided by two different systems was presented more than once to judge, thus inducing some inconsistency. Few answers pairs ([Exact answer, coming from a support document]) were assigned opposite judgments: 4 out of 594 for the first assessor, and 18 out of 702 for the second which is quite negligible.

On the other hand, this raises the problem of how to derive a reliable reference from assessed answers: should these inconsistencies be discarded or is there some information that can be extracted from these? Are hard to evaluate questions also hard to process by QAS?

Disagreements and errors still remaining are currently being studied further, even if calculation of the Kappa coefficient (Di Eugenio and Glass, 2004) reveals a good strength in agreement with a value of 0.887. This Kappa value is obtained by admitting a binary judgment: an answer is or is not Correct (merging Inexact, Unsupported, and Incorrect judgements). By the way, these differences in assessment might be modeled by a weighted agreement coefficient as those proposed by Krippendorff (Alpha coefficient) and Cohen (weighted-Kappa).

Reusability of a reference. After the campaign, in order to automatically evaluate our QAS, we derived a set of reference patterns accompanied by their document ID (to perform Strict and Lenient evaluation). For each Factoid and Definition question, patterns were built manually from the Correct answers provided by all participants. We then validated these reference patterns by checking their coverage on all the Exact answers of all the QAS (19708 answers) involved in EQueR, and match 97,5% of assessors' judgments.

After modifications in our Document Retrieval step, passages utilized by our QAS were totally changed and we notice a significant loss of final performances (even when applying Lenient conditions i.e. not checking if a document is supporting an answer). Therefore this run was manually evaluated. This evaluation was done in a absolute similar way to the EQueR campaign. It showed that the loss was actually due to a gap between our references patterns (and yet derived from the campaign) and the "new" discovered answers. Indeed, many "new" answers were Correct but were never encountered before and could not be matched. It has already been pointed out by (Lin and Katz, 2005) and can be explained as follows: a non participating QAS (or a heavily modified one) can not be reliably post-evaluated by using a reference built from other effective participants unless the number of assessed answers is far more greater. To address this, they proposed to use a verified and controlled subset of questions and answers.

5. Conclusion

In this paper we have surveyed the main Question Answering campaigns but also the measures used to evaluate Question Answering Systems (QAS).

Next, from a typical QA architecture, we have discussed the key points of a component-based evaluation. Even if, few recent QAS are already able to evaluate themselves by proofing their own answers like (Harabagiu et al., 2003) such a component evaluation is necessary in order to further investigate the impact of each of them on the quality of the system.

Lastly, we briefly presented the first French QA campaign EQueR. Good agreement between the two assessors of the campaign was revealed from a Kappa coefficient calculation. We also echoed on the difficulty to reuse a reference patterns derived from a QA campaign for a post-evaluation.

Acknowledgement. The authors would like to thank Professor Guy Lapalme for its helpful ideas and suggestions.

6. References

Ayache, C., Choukri, K., Grau, B. (2005). Campagne EVALDA/EQueR Evaluation en Question-Réponse. http://www.technolangue.net/IMG/pdf/rapport_EQueR_1.2.pdf

Ayache, C., Grau, B., Vilnat, A. (2006). EQueR: the French Evaluation campaign of Question Answering Systems. In *this volume.*

Baeza-Yates, R., Ribeiro-Neto, B. (1999). *Modern information retrieval.* ACM Press books. Addison-Wesley.
ISBN-0-201-39829-X

Beeferman, D., Berger, A., and Lafferty, J. (1997). Text segmentation using exponential models. In *Proceedings of the 2nd conf. on Empirical Methods in Natural Language Processing*, USA.

Bering, C., Drozdzynski, W., Erbach, G., Guasch, C., Homola, P., Lehmann, S., Li, H., Krieger, H., Piskorski, J., Schäfer, U., Shimada, A., Siegel, M., Xu, F., Ziegler-Eisele, D. (2003). Corpora and evaluation tools for multilingual named entity grammar development. In Proceedings of Multilingual Corpora Workshop at Corpus Linguistics 2003, Lancaster, pp. 42-52.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C., Maiorano, S., Miller, G., Moldovan1, D., Ogden1, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., Weishedel, R. (2001). Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). NIST. http://www-nlpir.nist.gov/projects/duc/roadmapping.html

Chin-Yew, L. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-Occurrences Statistics. In *The Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Edmonton, Alberta.

Clough, P., Sanderson, M. (2004). A proposal for comparative evaluation of automatic annotation for geo-referenced documents. In *The proceedings of Workshop on Geographic Information Retrieval SIGIR.*

Di Eugenio B. and Glass M. (2004). The Kappa statistic: a second look. Computational Linguistics, Volume 30, Issue 1, pp. 95-101.

Favre B., Bechet F., Nocera P. (2005). Robust Named Entity Extraction from Spoken Archives. In

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, Canada, pp. 491–498.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I., Vilnat, A. (2001). Finding An Answer Based on the Recognition of the Question Focus. In *Proceedings of the 10th Text Retrieval Conference*. Gaithersburg, Maryland, USA. NIST Special Publication 500-250, pp. 362-370.

Fukumoto, J., Kato, T., Masui, F. (2004). An Evaluation of Question Answering Challenge (QAC-1) at the NTCIR Workshop 3.

Fukumoto, J., Kato, T., Masui, F. (2004). An Evaluation of Question Answering Challenge (QAC-1) at the NTCIR Workshop 3. In The Proceedings of the SIGIR NTCIR Workshop.

Fukumoto, J., Kato, T., Masui, F (2004) Question Answering Challenge for Five ranked answers and List answers – Overview of NTCIR4 QAC2 Subtask 1 and 2. In The Working Notes of NTCIR-4. Tokyo, Japan.

Grivolla, J., Jourlin, P., de Mori, R., (2005) Automatic Classification of Queries by Expected Retrieval Performance. In *The Proceedings of SIGIR 2005, Predicting Query Difficulty Workshop*, Salvador, Brazil.

Harabagiu, S.M., Moldovan, D.I., Clark C., Bowden M., Williams J., Bensley J. (2003). Answer Mining by Combining Extraction Techniques with Abductive Reasoning. In *Proceedings of the 12th Text Retrieval Conference*, Gaithersburg, Maryland, USA pp. 375-382.

Herrera, J., Peñas, A., Verdejo, F. (2004) Question Answering Pilot Task at CLEF 2004. In *Proceedings of the CLEF 2004 Workshop*, Bath, United Kingdom.

Hovy, E.H., Hermjakob, U., Ravichandran, D. (2002). A Question/Answer Typology with Surface Text Patterns. In *Proceedings of the DARPA Human Language Technology conference (HLT).* San Diego, USA.

Kato, T., Fukumoto, J., Masui, F. (2004) Question Answering Challenge for Information Access Dialogue – Overview of NTCIR4 QAC2 Subtask 3. In The Working Notes of NTCIR-4. Tokyo, Japan.

Kato, T., Fukumoto, J., Masui, F. (2005) An oerview of NTCIR5 QAC3. In The Proceedings of the NTCIR-5 Workshop Meeting. Tokyo, Japan.

Lin, J., Demner-Fushman, D. (2005) Automatically Evaluating Answers to Definition Questions. Technical Report LAMP-TR-119/CS-TR-4695/UMIACS-TR-2005-04, University of Maryland, College Park.

Lin, J., Katz, B. (2005) Building a Reusable Test Collection for Question Answering. Journal of the American Society for Information Science and Technology, *forthcoming.*

Magnini, B., Romagnoli, S., Vallin, A., Jesús Herrera, J., Peñas, A., Peinado, V., Verdejo, F., de Rijke, M. (2003). The Multiple Language Question Answering Track at CLEF 2003.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R., (2004). Overview of the CLEF 2004 Multilingual Question Answering Track.

Marton, G., (2006). Nuggeteer: Automatic Nugget-Based Evaluation using Descriptions and Judgements. MIT CSAIL Work Product 1721.1/30604.

Nomoto, M., Fukushige, Y., Mitsuhiro, S., Suzuki, H. (2004). Are we making progress? An analysis of NTCIR QAC1 and 2. In *The Proceedings of NTCIR-4.* Tokyo, Japan.

Pevzner, L., Hearst, M.A. (2002). A critique and improvement of an evaluation metric for text segmentation. In *Computational Linguistics*, pp. 19–36.

Poibeau, T., Acoulon, A., Avaux, C., Beroff-Bénéat, L., Cadeau, A., Calberg, M., Delale, A., De Temmerman, L., Guenet, A.-L., Huis, D., Jamalpour, M., Krul, A., Marcus, A., Picoli, F., Plancq, C. (2003). The Multilingual Named Entity Recognition Framework. In *Proceedings of the European Association for Computational Linguistics Conference (EACL 2003).* Budapest, Hungary. pp. 155-158.

Sazaki, Y. (2005). Overview of the NTCIR-5 Cross Lingual Question Asnwering Track (CLQA1). In *The Proceedings of NTCIR-5 Workshop Meeting.* Tokyo, Japan.

Sekine, S., Nobata, C. (2004) Definition, dictionaries and tagger for Extended Named Entity Hierarchy. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, pp. 1977-1980.

Sitbon, L., Bellot, P. (2006) Tools and methods for topic segmentation of texts and contextual evaluation. In *this volume*.

Sitbon, L., Grivolla, G., Gillard, L., Bellot, P., Blache, P., (2006). Vers une prédiction automatique de la difficulté d'une question en langue naturelle, *forthcoming.*

Strzalkowski, T., Small, S., Hardy, H., Yamrom, B., Liu, T., Kantor, P., Ng, K.B., Wacholder, N. (2005). HITIQA: A Question Answering Analytical Tool. In The Proceedings of the International Conference on Intelligence Analysis, McLean, USA.

Vallin, A., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R. (2005) Overview of the CLEF 2005 Multilingual Question Answering Track.

Voorhees, E.M. (1999). The TREC-8 question answering track report. In *Proceedings of the 8th Text Retrieval Conference*, Gaithersburg, Maryland, USA, pp. 77-82.

Voorhees, E.M. (2003). Overview of the TREC 2003 Question Answering Track. In *Proceedings of the 12h Text Retrieval Conference*. Gaithersburg, Maryland, USA. pp. 54-68.

Voorhees, E.M., Harman, D. (2005) *TREC Experiment and Evaluation in Information Retrieval*. MIT Press 2005, chapter 10. pp. 233-257.