Overview of the INEX 2010 Question Answering Track (QA@INEX)

Eric SanJuan[1], Patrice Bellot[1], Véronique Moriceau[2], and Xavier Tannier[2]

[1] LIA, Université d'Avignon et des Pays de Vaucluse (France)
{patrice.bellot,eric.sanjuan}@univ-avignon.fr
[2] LIMSI-CNRS, University Paris-Sud 11 (France)
{moriceau,xtannier}@limsi.fr

Abstract. The INEX Question Answering track (QA@INEX) aims to evaluate a complex question-answering task using the Wikipedia. The set of questions is composed of factoid, precise questions that expect short answers, as well as more complex questions that can be answered by several sentences or by an aggregation of texts from different documents.

Long answers have been evaluated based on Kullback Leibler (KL) divergence between n-gram distributions. This allowed summarization systems to participate. Most of them generated a readable extract of sentences from top ranked documents by a state-of-the-art document retrieval engine. Participants also tested several methods of question disambiguation.

Evaluation has been carried out on a pool of real questions from OverBlog and Yahoo! Answers. Results tend to show that the baseline-restricted focused IR system minimizes KL divergence but misses readability meanwhile summarization systems tend to use longer and standalone sentences thus improving readability but increasing KL divergence.

1 Introduction

The INEX QA 2009-2010 track [1] aimed to compare the performance of QA, XML/passage retrieval and automatic summarization systems on special XML enriched dumps of the Wikipedia : the 2008 annotated Wikipedia [2] used in the INEX ad-hoc track in 2009 and 2010.

Two types of questions were considered. The first type was factual questions which require a single precise answer to be found in the corpus if it exists. The second type consisted of more complex questions whose answers required a multi-document aggregation of passages with a maximum of 500 words exclusively.

Like for the *2010 ad-hoc restricted focus task*, systems had to make a selection of the most relevant information, the maximal length of the abstract being fixed. Therefore focused IR systems could just return their top ranked passages meanwhile automatic summarization systems need to be combined with a document IR engine. The main difference between the QA long type answer task and the *ad-hoc restricted focus* one is that in QA, readability of answers[3] is as important as the informative content. Both need to be evaluated. Therefore answers

S. Geva et al. (Eds.): INEX 2010, LNCS 6932, pp. 269–281, 2011.

cannot be any passage of the corpus, but at least well formed sentences. As a consequence, informative content of answers cannot be evaluated using standard IR measures since QA and automatic summarization systems do not try to find all relevant passages, but to select those that could provide a comprehensive answer. Several metrics have been defined and experimented with at DUC [4] and TAC workshops [5]. Among them, Kullback-Leibler (KL) and Jenssen-Shanon (JS) divergences have been used [6] to evaluate the informativeness of short summaries based on a bunch of highly relevant documents. In this edition we used the KL one to evaluate the informative content of the long answers by comparing their n-gram distributions with those from 4 highly relevant Wikipedia pages.

In 2009 a set of encyclopedic questions about ad-hoc topics was released[1]. The idea was that informativeness of answers of encyclopedic questions could be evaluated based on the ad-hoc qrels[7]. This year, a set of "real" questions from *Over-Blog*[1] website logs not necessarily meant for the Wikipedia was proposed. A state of the art IR engine powered by Indri was also made available to participants. It allowed the participation of seven summarization systems for the first time at INEX. These systems only considered long type answers and have been evaluated on the 2010 subset. Only two standard QA systems participated to the factual question sub-track. Therefore most of QA@INEX 2010 results are about summarization systems versus a state of the art restricted focused IR system.

The rest of the paper is organized as follows. First, the focused IR system used as baseline is introduced in Section 2. It is described in (§2.1) and evaluated in (§2.2). Section 3 details the collection of questions (§3.1-3.3) and available reference texts for assessments (§3.5). Section 4 explains the final choice of metrics used for the evaluation of the informativeness of long answers after several experiments. Results are reported in Section 5. Finally, Section 6 discusses our findings and draws perspectives for next year edition.

2 Baseline System: Restricted Focused IR System

Several on-line resources have been made available to facilitate participation and experiment the metrics. These resources available *via* a unique web interface at *http://termwatch.es/Term2IR* included:

1. a document index powered by Indri,
2. a sentence and Part of Speech tagger powered by the TreeTagger,
3. a summarization and Multi-Word Term extractor powered by TermWatch,
4. a tool for automatic evaluation of summary informativeness powered by FRESA,
5. links to document source on the TopX web interface.

2.1 Features

The system allows to test on the INEX 2009 ad-hoc corpus the combination of a simple IR passage retrieval system (Indri Language Model) with a baseline summarization system (a fast approximation of Lexrank).

[1] http://www.over-blog.com/

Different outputs are available. The default is a selection of relevant sentences with a link towards the source document in TopX. Sentences have been selected following approximated LexRank scores among the 20 top ranked passages returned by Indri using a Language Model over INEX 2008 corpus. Multiword terms extracted by shallow parsing are also highlighted.

A second possible output gives a baseline summary with less than 500 words, made of the top ranked sentences. The Kullback-Leibler divergence between distributions of n-grams in the summary and in the passages retrieved by Indri are also shown. They are computed using the FRESA package. It is also possible to test any summary against this baseline.

Finally, the passages retrieved by Indri are available, in several formats: raw results in native INEX XML format, raw text, POS tagged text with TreeTagger.

Questions and queries can be submitted in plain text or in Indri language. The following XML tags have been indexed and can be used in the query: *b, bdy, category, causal_agent, country, entry, group, image, it, list, location, p, person, physical_entity, sec, software, table, title*. These are examples of correct queries:

- Who is Charlie in the chocolate factory?
- #1(Miles davis) #1(Charles Mingus) collaboration
- #1(Charles Mingus).p, #combine[p](Charles Mingus)

2.2 Evaluation on the 2010 *Restricted Focus Ad-Hoc Task*

Let us first give some details on this restricted focus system.

As stated before it starts by retrieving n documents using an Indri language model. These sentences are then segmented into sentences using shallow parsing. Finally sentences are ranked using a fast approximation of LexRank. Basically, we only consider sentences that are at distance two from the query in the intersection graph of sentences. These are sentences that share at least one term with the query, or with another sentence that shares it. The selected sentences are then ranked by entropy.

We evaluated this baseline system on the Ad-hoc restricted focused task, by setting $n = 100$. We then retrieve for each sentence all passages in which the same word sequence appears, with possible insertions. We return the first 1000 characters.

The precision/recall function of this system starts high compared to other participant runs. It gets among automatic runs, the third char precision (0.3434) and the best iP[0.01] with a value of 0.15 (0.1822 for the best manual run).

3 Sets of Questions and References

A total set of 345 questions has been made available. There are four categories of questions:

1. factual and related to 2009 ad-hoc topics (151),
2. complex and related to 2009 ad-hoc topics (85),

3. factual from Over-Blog logs (44),
4. complex from Over-Blog logs (70)

This year evaluation has been carried out on the fourth category. Answers for the first category are available. A run is also available for categories 1 and 3. Informativeness of answers to questions in category 3 can be partially evaluated based on qrel from ad-hoc 2009 INEX track.

3.1 Encyclopedic *vs.* General Questions

236 questions are related to 2009 INEX ad-hoc topics. Most of the remaining questions come from a sample of the log files from the search engine on Over-Blog. These are real questions submitted to their website by visitors looking for answers among the blogs hosted on their website. We have selected a subset of these questions such that there exists at least a partial answer in the Wikipedia 2008. Then we have mixed these questions with others from Yahoo! Answers website[2].

We considered three different types of questions: short_single, short_multiple and long.

3.2 Short Type Questions

Those labeled short_single or short_multiple are 195 and both require short answers, *i.e.* passages of a maximum of 50 words (strings of alphanumeric characters without spaces or punctuations) together with an offset indicating the position of the answer.

Short_single questions should have a single correct answer, *e.g.* question 216: *Who is the last olympic champion in sabre?* whereas multiple type questions will admit multiple answers (question 209: *What are the main cloud computing service providers?*).

For both short types, participants had to give their results as a ranked list of maximum 10 passages from the corpus together with an offset indicating the position of the answer. Passages had to be self-contained to decide if the answer is correct or not.

Besides, we collected manually answers for the first category (factual and related to 2009 ad-hoc topics) in the Wikipedia INEX collection. These answers will be made available as a development set for 2011 campaign. Moreover, a sample run is available for all factual questions. This run has been produced by FIDJI, an open-domain QA system for French and English [8]. This system combines syntactic information with traditional QA techniques such as named entity recognition and term weighting in order to validate answers through different documents. Question analysis in FIDJI turns the question into a declarative sentence. It also aims to identify the syntactic dependencies, the expected type(s) of the answer (named entity type) and the question type (factoid, definition, complex, list questions).

[2] http://answers.yahoo.com/

3.3 Long Type Questions

Long type questions require long answers up to 500 words that must be self-contained summaries made of passages extracted from the INEX 2009 corpus. Are considered as words any sequence of letters and digits. An example of a long type question is (#196): *What sort of health benefit has olive oil?* There can be questions of both short and long types, for example a question like *Who was Alfred Nobel?* can be answered by "a chemist" or by a short biography. However, most of the selected long type questions are not associated with obvious name entities and require at least one sentence to be answered.

3.4 Format Submission

The submission format has been simplified to follow INEX TREC ad-hoc format:

`<qid> Q0 <file> <rank> <rsv> <run_id> <column_7> <column_8> <column_9>`
where:

- the first column is the topic number.
- the second column currently unused and should always be Q0.
- the third column is the file name (without .xml) from which a result is retrieved, which is identical to the ¡id¿ of the Wikipedia document.
- the fourth column is the rank the result is retrieved, and fifth column shows the score (integer or floating point) that generated the ranking.
- the sixth column is called the "run tag" and should be a unique identifier for the group AND for the method used.

The remaining three columns depend on the question type (short or long) and on the chosen format (text passage or offset).

For textual content, raw text is given without XML tags and without formatting characters. The resulting word sequence has to appear in the file indicated in the third field. This is an example of such output:

```
1 Q0 3005204 1 0.9999 I10Run1 The Alfred [...] societies.
1 Q0 3005204 2 0.9998 I10Run1 The prize [...] Engineers.
1 Q0 3005204 3 0.9997 I10Run1 It has [...] similar spellings.
```

An Offset Length format (FOL) can also be used. In this format, passages are given as offset and length calculated in characters with respect to the textual content (ignoring all tags) of the XML file. File offsets start counting from 0 (zero). Previous example would be the following in FOL format:

```
1 Q0 3005204 1 0.9999 I10Run1 256 230
1 Q0 3005204 2 0.9998 I10Run1 488 118
1 Q0 3005204 3 0.9997 I10Run1 609 109
```

This would mean that results are from article 3005204. The first passage starts at the 256th character (so 257 characters beyond the first character), and has a length of 230 characters.

In the case of short type questions, we use an extra field that indicates the position of the answer in the passage. This position is given by counting the number of words before the detected answer.

3.5 Reference Texts

For each question we have selected four highly relevant Wikipedia pages from which we have extracted the most relevant sections. Questions for which there were too few relevant passages were not submitted to participants. These passages that were not publicly available have been then used as reference text to evaluate long type answers using KL divergence.

4 Evaluation of Long Answers

Only long answer evaluation is presented. As short answer runs come from organizers' systems, we decided not to evaluate them but they will be made available for future participants.

The informative content of the long type answers are evaluated by comparing the several n-gram distributions in participant extracts and in a set of relevant passages selected manually by organizers. We followed the experiment in [6] done on TAC 2008 automatic summarization evaluation data. This allows to evaluate directly summaries based on a selection of relevant passages.

Given a set R of relevant passages and a text T, let us denote by $p_X(w)$ the probability of finding an n-gram w from the Wikipedia in $X \in \{R, T\}$. We use standard Dirichlet smoothing with default $\mu = 2500$ to estimate these probabilities over the whole corpus. Word distributions are usually compared using one of these functions:

– Kullback Leibler (KL) divergence:

$$KL(p_T, p_R) = \sum_{w \in R \cup T} p_T(w) \times \log_2 \frac{p_T(w)}{p_R(w)}$$

– Jensen Shannon (JS) divergence:

$$JS(p_T, p_R) = \frac{1}{2}(KL(p_T, p_{T \cup R}) + KL(p_R, p_{T \cup R}))$$

In [6], the metric that obtained best correlation scores with ROUGE semi-automatic evaluations of abstracts used in DUC and TAC was JS. However, we have observed that JS is too sensitive to abstract size; therefore we finally used KL divergence to evaluate informative content reference texts or passages.

We used the FRESA package[3] to compute both KL and JS divergences between n-grams ($1 \leq n \leq 4$). This package also allows to consider skip n-grams.

Evaluating informative content without evaluating readability does not make sense. It clearly appears that if readability is not considered then the best summarizer would be the random summarizer on n-grams which certainly minimizes KL divergence but produces incomprehensible texts.

The readability and coherence are evaluated according to "the last point of interest" in the answer which is the counterpart of the "best entry point" in

[3] http://lia.univ-avignon.fr/fileadmin/axes/TALNE/Ressources.html

Table 1. Cumulative KL divergence for best runs per participant

ID	Specificity	unigrams	bigrams	4 skip grams	average	readability
98	Focused IR	1599.29	2207.56	2212.49	2006.45	1/5
92	MWT expansion	1617.35	2226.61	2231.56	2025.17	2/5
860	System combination	1617.6	2227.37	2232.43	2025.8	3/5
857	Question reformulation	1621.14	2234.57	2239.85	2031.85	3/5
855	Semantic expansion	1625.76	2235.21	2240.35	2033.77	3/5
943	Long sentences	1642.93	2252.28	2257.22	2050.81	4/5
557	JS minimization	1631.29	2237.61	2242.83	2037.24	3/5

INEX ad-hoc task. It requires a human evaluation by organizers and participants where the assessor indicates where he misses the point of the answers because of highly incoherent grammatical structures, unsolved anaphora, or redundant passages.

5 Results

We received runs for long type questions from seven participants. All of these participants generate summaries by sentence extraction. This helps readability even if it does not ensure general coherence. Extracts made of long sentences without anaphora are often more coherent but have higher KL scores. To retrieve documents, all participants used the IR engine powered by Indri, available at track resources webpage[4].

Table 1 shows results based on KL divergence on long-type questions from OverBlog logs. The cumulative divergence is the sum of KL scores between participant extracts and selected pages.

As expected, baseline-restricted focused IR system (98) minimizes KL divergence but the resulting readability is poor. Meanwhile the system (943) having best readability favors long sentences and gets highest divergence figures. The most sophisticated summary approach is the Cortex system (860) which reaches a compromise between KL divergence and readability.

But query formulation to retrieve documents looks also important, the approach based on query enrichment with related MultiWord Terms (92) automatically extracted from top ranked documents, gets similar divergence scores. Meanwhile this is a system slightly adapted from the focused IR system used in previous INEX 2008 and 2009 ad-hoc track [9,10].

Surprisingly sentence JS minimization (557) does not seem to minimize overall KL divergence. This system ranks sentences in decreasing order according to their JS divergence with the query and the retrieved documents.

Only score differences between the baseline and the other systems are significant, as shown in Table 2.

[4] http://qa.termwatch.es/

Table 2. Probabilities of signed t-tests over KL divergence scores

ID	92	860	857	855	943	557
98	* 0.0400	* 0.0149	** 0.0028	* 0.0172	**** 0.0005	*** 0.0000
92		0.3777	0.1037	0.2304	0.0821	0.0529
860			0.1722	0.4442	0.1497	0.1104
857				0.2794	0.5047	0.1788
943					0.1798	0.1013
557						0.1857

The standard deviation among systems KL divergences varies. The ten question minimizing standard deviation and, therefore, getting most similar answers among systems are:

2010044 What happened to the president of Rwanda death?
2010107 What are the symptoms of a tick bite?
2010096 How to make rebellious teenager obey you?
2010066 How much sadness is normal breakup?
2010062 How much is a typical sushi meal in japan?
2010083 What are the Refugee Camps in DRC?
2010046 How to get Aljazera sports?
2010047 How to be a chef consultant?
2010005 Why did Ronaldinho signed for Barcelona?
2010049 Where can I find gold sequined Christain Louboutin shoes?

All these questions contain at least one named entity that refers to a Wikipedia page. Therefore, the systems mostly built their answer based on the textual content of this page and KL divergence is not accurate enough to discriminate among them.

On the contrary, the 10 following questions are the top ten that maximized standard deviation and have the greatest impact in the ranking of the systems:

2010093 Why is strategy so important today?
2010114 What is epigenetics and how does it affect the DNA/genes in all of our cells?
2010009 What does ruddy complexion mean?
2010066 What do nice breasts look like?
2010022 How to get over soul shock?
2010092 How to have better sex with your partner?
2010080 How to be physically attractive and classy?
2010014 Why is it so difficult to move an mpeg into imovie?
2010010 What do male plants look like?
2010075 WHAT IS A DUAL XD ENGINE?

Clearly, these questions are not encyclopedic ones and do not refer to particular Wikipedia pages. Meanwhile partial answers exist in the Wikipedia but they are spread among several articles.

Figure 1 shows the KL divergence values for 4-skip n-grams over the most discriminative questions. It can be observed that cumulative KL divergence varies along questions. These variations reveal the gaps between reference documents and textual content extracted by participant systems.

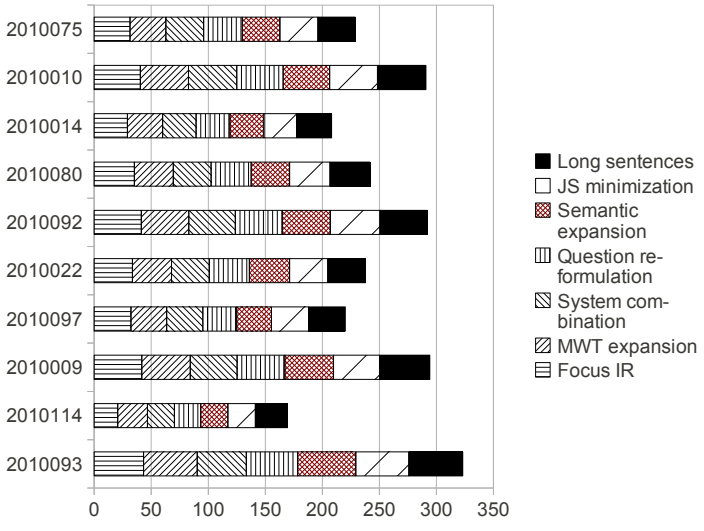

Fig. 1. KL divergence per system over the most discriminative questions

Figure 2 shows the same values but scaled in order to reveal main differences between systems.

6 Plans for Next QA Track 2011 Edition

In 2011, we will keep on addressing real-world focused information needs formulated as natural language questions using special XML annotated dumps of the Wikipedia, but we will mix questions requiring long answers and those not.

6.1 Merging Short and Long Answer Task

Contrary to long answer task, we had too few participant teams interested in short answer evaluation in 2010. We think that this is mainly due to the fact that this task required too many modifications to traditionalINEX participant systems.

To avoid this problem, we plan to make the task more manageable for questions where short answers are expected.

Consequently, potential short answers will be tagged in the collection and considered as traditional XML retrieval units. The provided collection will thus be enriched with named entity annotations. These entities are generally the most

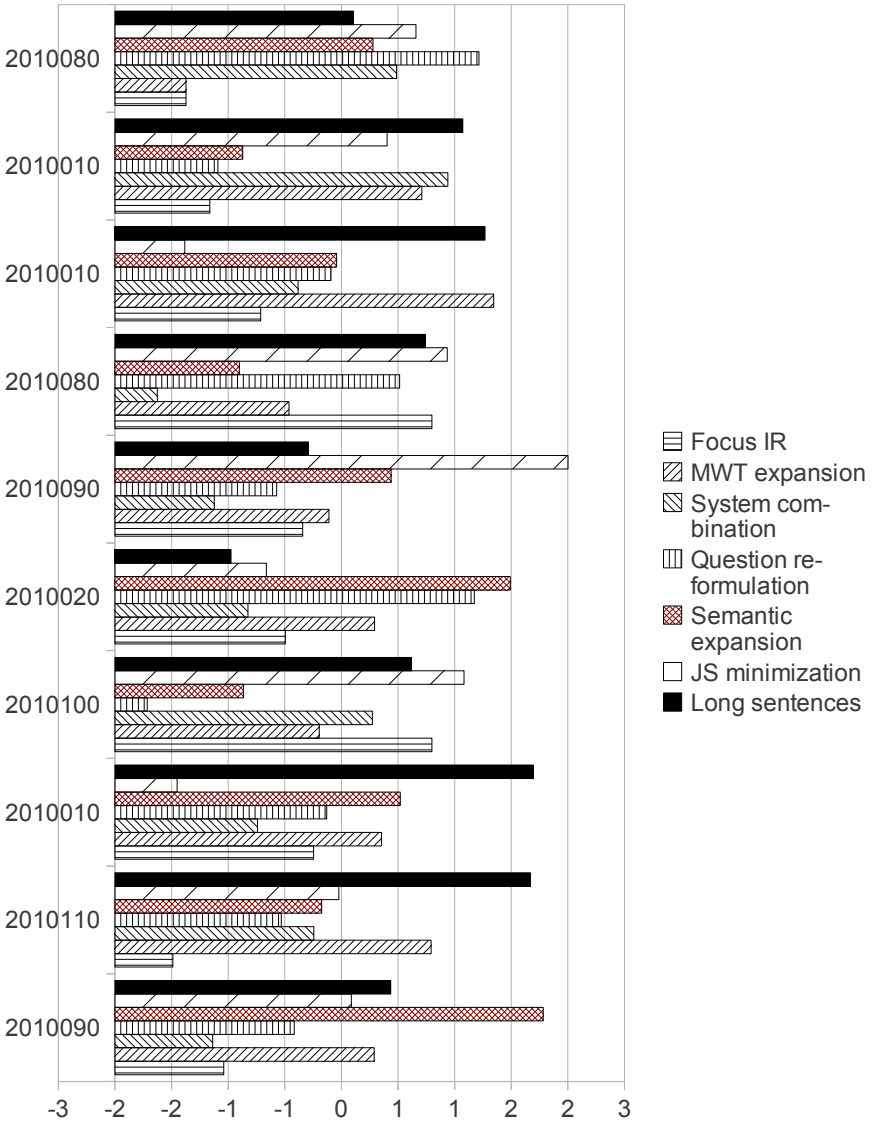

Fig. 2. scaled KL divergence per system over the most discriminative questions

relevant for short answers. However, the type of entities (persons, locations, organizations, etc.) will not be provided.

Moreover, the information concerning the short/long type of expected answers will be removed from the input question format. Given a question, the decision whether the answer should be rather short or long is then left to the system.

Concerning the evaluation, many different answers can be considered as relevant, and a manual assessment is necessary. As a short answer is now an XML tag, the global methodology should not differ from long answer task.

Finally, by restricting output to XML elements and not any passage, we shall improve readability. Indeed, we have observed that a significant number of readability issues were due to incorrect sentence segmentation or mixed document structure elements like lists, tables and titles.

6.2 Towards Contextualization of Questions

In 2011, we shall build and make available a new dump of the Wikipedia adapted for QA where only document structure, sentences and named entities will be annotated. For each question, participants will have to guess the context category (for example, "expert", "pupil", "journalist"...), provide a context as a query oriented readable summary and a list of possible answers. Summary and answers will be made of XML elements from Wikipedia dump. Real questions will be selected from OverBlog [5] website logs and Twitter.[6] Participants system will have to fisrt detect the research context. Then, they shall adapt their strategy to select document nodes that match the query, by computing, for example, complexity or readability measures, expertise level scores or genre.

The assumption is that context could improve the performances of Information Retrieval Systems. Modelling context in IR is considered as a long-term challenge by the community [11]. It is defined as the combination of retrieval technologies and knowledge on the context of the query and the user in a single model to provide the most suitable response compared to an information need. The contextual aspect refers to tacit or explicit knowledge concerning the intentions of users, the environment of users and the system itself. Modelling context is not an end in itself. The system must be able to decide the most adequate technologies compared to a given context, i.e.: to adapt the searching methods to the context. Of course, the user does not provide the system with the knowledge on the context of the required search. Classical IR approaches suppose a common objective which is topic relevance of top ranked documents. Opposite to this, approaches investigated by [12] allow to integrate scores orthogonal to the relevance score such as complexity. Document complexity must be matched to both users capability and level of specialisation in the target knowledge area. Different measures of difficulty can be incorporated to the relevance scores.

[5] http://en.over-blog.com

[6] This track relates to Contextual Information Retrieval and will be supported by the French National Research Agency ANR via the project CAAS (Contextual Analysis and Adaptive Search - ANR 2010 CORD 001 02).

The international community in information retrieval is indeed interested in research in contextual information retrieval since few years. The group IRiX (Information Retrieval in Context), proposed by Pr. Rijsbergen organizes an annual workshop on this topic since 2004. In the framework of the 6th PRCD (European commission), the network of excellence DELOS (Network of Excellence on Digital Libraries[7]) is interested in the access to information for users of various categories: individuals, specialists or population. Project PENG (Personalized News Content Programming[8]) is interested in defining an environment for programming and modelling the contents which makes it possible to offer interactive and personalized tools for multimedia information access to specialists or simple users.

7 Conclusion

In 2010 we provided a reusable resource for QA evaluation based on INEX adhoc 2009-2010 wikipedia document collection, including an original evaluation approach and software implementation.

In 2011 we will try to deal with two types of challenges. The first challenge is defining the query features that could help in predicting the query type and extracting it automatically from the query formulation or from the environment of the user. The second challenge is to be able to help the system to disambiguate the users expectation and use it for answering the queries.

References

1. Moriceau, V., SanJuan, E., Tannier, X., Bellot, P.: Overview of the 2009 qa track: Towards a common task for qa, focused ir and automatic summarization systems. In: [7], pp. 355–365 (2009)
2. Schenkel, R., Suchanek, F.M., Kasneci, G.: Yawn: A semantically annotated wikipedia xml corpus. In: Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C., Brochhaus, C. (eds.) BTW. LNI, vol. 103, pp. 277–291 GI (2007)
3. Pitler, E., Louis, A., Nenkova, A.: Automatic evaluation of linguistic quality in multi-document summarization. In: ACL, pp. 544–554 (2010)
4. Nenkova, A., Passonneau, R.: Evaluating content selection in summarization: The pyramid method. In: Proceedings of HLT-NAACL, vol. 2004 (2004)
5. Dang, H.: Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In: Proc. of the First Text Analysis Conference (2008)
6. Louis, A., Nenkova, A.: Performance confidence estimation for automatic summarization. In: EACL, The Association for Computer Linguistics, pp. 541–548 (2009)
7. Geva, S., Kamps, J., Trotman, A. (eds.): Focused Retrieval and Evaluation, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009, Brisbane, Australia, December 7-9, LNCS, vol. 6203. Springer, Heidelberg (2010) (Revised and selected papers)

[7] http://www.delos.info/
[8] http://www.peng-project.org/

8. Moriceau, V., Tannier, X.: FIDJI: Using Syntax for Validating Answers in Multiple Documents. Information Retrieval, Special Issue on Focused Information Retrieval 13, 507–533 (2010)
9. SanJuan, E., Ibekwe-Sanjuan, F.: Combining language models with nlp and interactive query expansion. In: [7], pp. 122–132
10. Ibekwe-Sanjuan, F., SanJuan, E.: Use of multiword terms and query expansion for interactive information retrieval. In: Geva, S., Kamps, J., Trotman, A. (eds.) INEX 2008. LNCS, vol. 5631, pp. 54–64. Springer, Heidelberg (2009)
11. Allan, J., Aslam, J., Belkin, N.J., Buckley, C., Callan, J.P., Croft, W.B., Dumais, S.T., Fuhr, N., Harman, D., Harper, D.J., Hiemstra, D., Hofmann, T., Hovy, E.H., Kraaij, W., Lafferty, J.D., Lavrenko, V., Lewis, D.D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J.M., Prager, J.M., Radev, D.R., Resnik, P., Robertson, S.E., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R.M., Singhal, A., Smeaton, A.F., Turtle, H.R., Voorhees, E.M., Weischedel, R.M., Xu, J., Zhai, C.: Challenges in Information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst. SIGIR Forum 37(1), 31–47 (2002)
12. Gao, J., Qi, H., Xia, X., Nie, J.Y.: Linear discriminant model for information retrieval, pp. 290–297. ACM, New York (2005)