Building an Evaluation Corpus for German Question Answering by Harvesting Wikipedia

Irene Cramer, Jochen L. Leidner, Dietrich Klakow

LSV – Spoken Language Systems, Building C7 1
Saarland University, 66123 Saarbrücken, Germany.
Irene.Cramer@lsv.uni-saarland.de, Jochen.Leidner@lsv.uni-saarland.de, Dietrich.Klakow@lsv.uni-saarland.de

Abstract

The growing interest in open-domain question answering is limited by the lack of evaluation and training resources. To overcome this resource bottleneck for German, we propose a novel methodology to acquire new question-answer pairs for system evaluation that relies on volunteer collaboration over the Internet. Utilizing Wikipedia, a popular free online encyclopedia available in several languages, we show that the data acquisition problem can be cast as a Web experiment. We present a Web-based annotation tool and carry out a distributed data collection experiment. The data gathered from the mostly anonymous contributors is compared to a similar dataset produced in-house by domain experts on the one hand, and the German questions from the from the CLEF QA 2004 effort on the other hand. Our analysis of the datasets suggests that using our novel method a medium-scale evaluation resource can be built at very small cost in a short period of time. The technique and software developed here is readily applicable to other languages where free online encyclopedias are available, and our resulting corpus is likewise freely available.

1. Introduction

There is currently a growing interest in open-domain question answering, as evidenced by the large number of research teams participating in international system evaluations such as CLEF (mono-/cross-lingual), TREC (English monolingual) or NTCIR (Japanese monolingual). These evaluations provide new test questions each year, and research groups typically resort to the previous years' test sets for development, training and evaluation. However, after a system has once been exposed to a particular corpus, system performance it is typically better than for "fresh" (unseen) data, making the corpus less useful as resources to work with (over-training). Progress in the field is thus negatively affected by the limited availability of evaluation and training resources. Likewise, to date for the evaluation of German question answering (Q&A) systems, the only available data described in the literature is the CLEF series of evaluation test-sets (Magnini et al., 2004). While these datasets are very useful, they are small–typically just 200 questions and answer keys for German per year–and have the disadvantage that they have been heavily used for system development and past evaluations. However, thorough system evaluations and even more so experiments with supervised machine learning methods require a large number of ideally unseen instances to work with. In the context of the German SMARTWEB project (Wahlster, 2004; SmartWeb Consortium, 2006), our aim was to find an affordable way to create an ideally large dataset of German question-answer pairs that could not only be used to evaluate our prototypes, but also to train statistical models. For methodological reasons, this requires different subsets to be used for development ("train"), development testing ("devtest"), and final evaluation ("test").
To overcome the resource bottleneck, we propose a new methodology to acquire new question-answer pairs for system evaluation which relies on volunteer collaboration over the Internet. Utilizing WIKIPEDIA, a popular free online encyclopedia available in several languages, we show that the data acquisition problem can be cast as a Web experiment, in which individuals mark text spans in encyclopedia articles that are answers to self-constructed questions and "donate" those corresponding questions in the same way that the WIKIPEDIA articles were initially donated by volunteers. To this end, we have developed a Web-based annotation tool and carried out a distributed data collection experiment. The data gathered from the mostly anonymous contributors is compared to a similar dataset produced inhouse by domain experts.
The contributions of this paper are as follows:

- a novel methodology to harvest question-answer pairs for system evaluation from online encyclopedia entries using a volunteer community;
- a Web-based annotation tool that implements the method for German, but which may easily be deployed for further languages;
- the resulting evaluation corpus for German open-domain Q&A; and
- a comparative analysis between datasets created inhouse and another one created via a Web experiment on the one hand side, and between this Web dataset and an existing 200 question corpus, namely the CLEF QA 2004 German monolingual questions, on the other hand (Magnini et al., 2004).

We argue that through remote collaboration of non-expert volunteers a medium-scale evaluation resource can be built at very small cost in a very short period of time. Our statistical analysis further suggests that the data gathered this way is not very different in nature from question-answer pairs gathered by domain experts. The technique and software developed here is readily applicable to other languages where free online encyclopedias are available.
Paper plan. The remainder of this paper is structured as follows: Section 2. presents related work. Section 3. describes our new method for distributed data collection of

question-answer pairs using WIKIPEDIA and outlines the implementation of our annotation tool. Section 4. presents our resulting dataset, and compares its characteristics to a similar dataset produced by domain experts in-house, and to CLEF QA 2004 (German monolingual). Section 5. discusses some challenges for our approach, and Section 6. summarizes and concludes the paper.

2. Related Work

To the best of our knowledge, to date there exists no other attempt to harvest open-domain question-answer pairs from online encyclopedias, neither for German nor for other languages. Even for English, which receives much more research attention than German, the resource situation is scarce. In this section we review some related efforts.

CLEF and TREC. The Cross Language Evaluation Forum (CLEF) has been carrying out international mono-lingual and cross-lingual question answering evaluations for many languages, including English and German, and the organizers have curated evaluation datasets, DISEQuA, Multi-Six, Multi-8/Multieight-04, Multi9-05 (cf. (Magnini et al., 2004), for instance) on an annual basis. It was inspired by the Text Retrieval Conference (TREC), which has been offering a monolingual question answering track for English since 1999 (e.g. (Voorhees, 2003)). Both TREC and CLEF use regular expressions to describe sets of answers in a compact fashion.

Previous Use of Online Encyclopedias for Q&A. (Lita et al., 2004) analyzed the potential contribution of various evidence sources to open-domain question answering. They found that 35.81 % of TREC 8-12 answers could be found in the English edition of WIKIPEDIA. It would be interesting to carry out a similar study for German. While we have used WIKIPEDIA to create a dataset for Q&A evaluation and training of classifiers, its use in question answering itself is not new (Kupiec, 1993).

Further Work into Q&A Evaluation. Clifton and Teahan propose a semi-automatic technique for evaluating open-domain question answering systems that creates evaluation questions from documents in the text collections (Clifton and Teahan, 2005): in a first pass, using automatic named entity tagging and human-devised pattern rules a list of candidate questions is created without human intervention. These may contain two types of errors: firstly, some questions might be too generic; secondly, some questions may be ill-formed due to the crude nature of the patterns. In a human post-processing phase, over-generic questions are discarded and ill-formed questions are hand-corrected. Clifton and Teahan argue that the method is still more fine-grained and ultimately more appropriate than a TREC-style approach. Their experiments are using English examples, but the method as such would be applicable to other languages. (Leidner et al., 2003) present a reading comprehension corpus for Canadian English. A set of Web pages containing CBC newswire stories modified to be suitable for a teenage audience was annotated with remedial questions and answers, and the resulting corpus was labeled with a large number of strata comprising linguistic information (POS, parse trees etc.). The aim of this *Annotated CBC4Kids Corpus* is to provide a corpus for English text understanding that can be re-distributed and that incorporates multi-layer annotation into the corpus itself, so that experimenters are spared from producing their own idiosyncratic tool pipeline and can instead simply select from already existing annotations required by their methods. However, the resource is not available for German.

Web Experiments. (Reips, 2002) proposes some standards for Web-based experiments. WEXTOR (Reips and Neuhaus, 2002) and WebExp2 (Mayo et al., 2005) are environments to assist the computer-supported creation of Internet-based experiments, and they are mostly used by experimental psychologists and psycho-linguists, respectively.

3. Method

3.1. Design

In experiments, subjects asked to come up with questions typically find it hard to do so spontaneously. To counter this "empty backboard effect", we decided to present randomly chosen encyclopedia articles to each subject. Subjects were then asked to mark a span of characters with the mouse that contained an interesting fact and enter the question that they would use to find out this answer. Figure 1 shows an example entry that was created this way.

`Mirjam Müntefering (∗ 29. Januar 1969` im `Sauerland) ist eine deutsche Schriftstellerin.`
↓ (mark text span with mouse)
A: `Sauerland`
(Sauerland)
↓ (keyboard entry of corresponding question)
Q: `Wo wurde Mirjam Müntefering geboren?`
(Where was Mirjam Müntefering born?)

Figure 1: Example of the Question-Answer Pair Harvesting Process.

Subjects were asked to repeat the cycle up to three times for a given encyclopedia entry, based on their own interest, in order to introduce a notion of practical relevance to the data gathering effort. Then the next random article was displayed. WIKIPEDIA is a free online encyclopedia based on the Wiki principle[1], i.e. anyone can freely create new articles or modify existing ones. The ease of participation and a sense of community spirit has quickly led to a sizeable encyclopedia (Figure 2). By casting our data gathering effort as a Web experiment, we aimed to utilize this community spirit to enhance question answering research. Figure 3 shows the architecture of the software developed for the experiment described here. An offline program downloads a desired number of WIKIPEDIA pages in advance (we chose 20,000 articles) and pre-processes them (for the most part, this involves a conversion to plain text). A Web client first

[1] The German WIKIPEDIA is available at http://de.wikipedia.org/.

EN	1 million	DE	363,000	FR	248,000	PL	217,000
JA	187,000	DU	141,000	IT	141,000	SV	141,000
PT	118,000	ES	98,000	RU	62,000	FI	52,000

Figure 2: Number of entries in WIKIPEDIA (by language, as of March 2006).

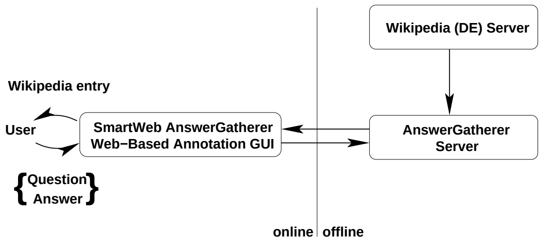


Figure 3: SmartWeb AnswerGatherer Architecture.

explains the experiment, then enquires information about the subjects, and then allows for the annotation of an arbitrary number of randomly chosen WIKIPEDIA entries.

3.2. Software Implementation

The annotation tool was developed in JavaScript/ECMAscript (client) and Python 2.3 (server), using standard the Common Gateway Interface (CGI). To speed up development, interaction with the complex local IT infrastructure was minimized by providing an embedded Web server that allows for easy testing and log file inspection. The resulting GUI is shown in Figure 4. At the top, a title is displayed in order to provide orientation and avoid confusion between our experimental setup and the "real" WIKIPEDIA site. Below, a pane displays the WIKIPEDIA entry. At the bottom, input fields allow for the entry of between one to three questions and comfortable markup of corresponding answers. A help page can be opened that summarizes the task, and subjects can contact the authors by email to ask questions. Finally, the data gathered can be submitted to the server. In addition, a button allows for entries to be skipped if subjects do not understand the entry, or do not find it interesting enough. This helped keep the motivation up and thus ensured continued participation.

3.3. Deployment

The Web based annotation tool was set up on our departmental Web server running the Apache Web server.[2] In order to reach a wide audience, we advertised our Web experiment using the following channels:

1. **Personal e-mails.** The authors sent personal emails to 40-50 recipients who were personal friends and colleagues and German speakers.

2. **USENET.** A call for volunteer participation was sent to the USENET group `de.sprache.etc.misc`.

3. **Web Experiment Portals.** As there is growing interest in Web experiments in psychology and psycholinguistics, there exist several portal sites where researchers can register their own online experiments. We selected two portals whose target audience appeared to include the largest number of German speakers, namely *Language Experiments* in Edinburgh[3] and the *Web Experiment Psychology Lab* in Zurich[4], and asked to have our experiment hyper-linked from there to allow volunteer subjects to find the experiment.

	CLEF	EXPERTS	WEB
Number of Q&A pairs	200	652	1,454
Number of topics	n/a	218	626
Avg. question length	7	6	6
Avg. answer length	n/a	6	6

Figure 5: Comparison Between Q&A Evaluation Datasets.

Subjects who left their details were eligible to win two book vouchers in a price draw. The setup described here has been publically available to subjects since October 1, 2005. We analyze a snapshot of the data (1 October - 1 December). Meanwhile, we continue hosting the experiment to collect more data.

4. Results

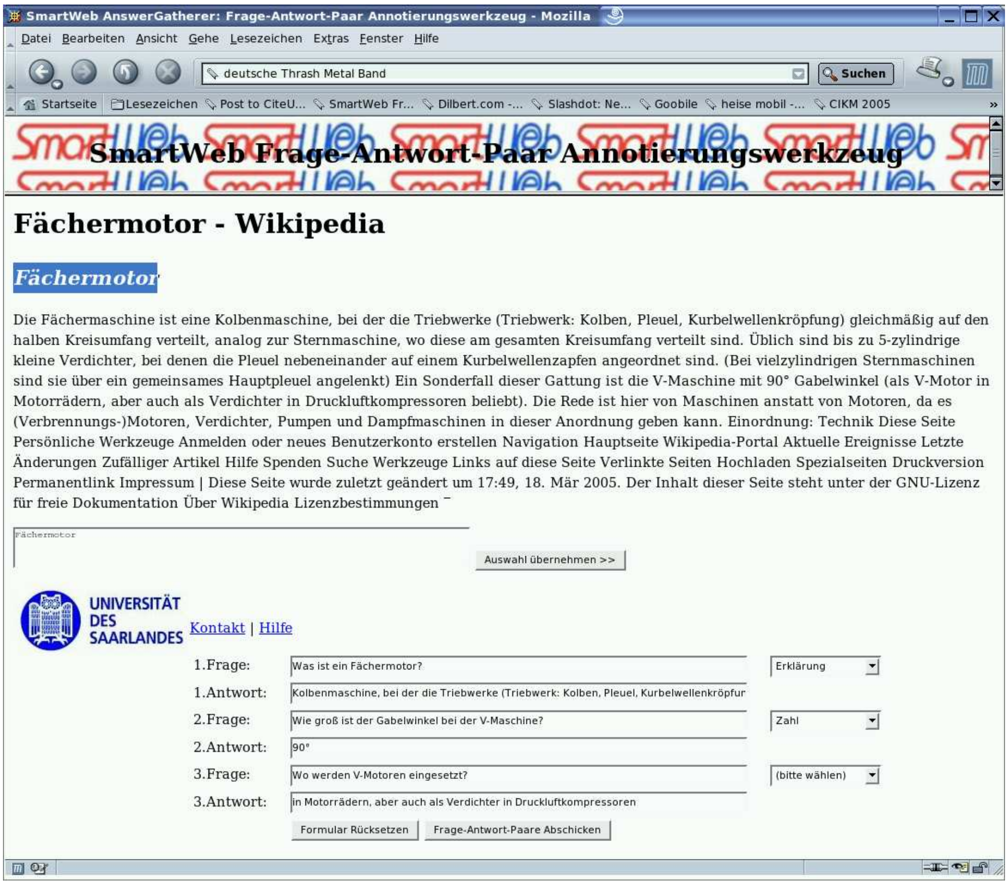
In the two-month time window that we discuss here, between 57 and 107 subjects visited our experiment online[5] and contributed 1,454 question-answer pairs in total. At the same time, another dataset was created by in-house experts, amounting to 652 question answer pairs. Figure 6 shows the CLEF QA 2004 (Magnini et al., 2004) German monolingual corpus ("CLEF" for short) next to the two datasets produced in this study, one created by experts—including the authors of this paper ("EXPERT")—and the other being a dataset produced by volunteers on the Web ("WEB"). As can be seen, the question length distributions for both the in-house and the Web experiments are approximately normally distributed, and so are the CLEF QA 2004 German questions. They are very similar; however, the same does not hold for the answer length: its distribution does not comply with a Gaussian (Figure 7). We believe the former is due to the syntactic constraints (as subjects were asked to form grammatical questions), whereas the latter is caused by the absence of such grammatical requirements (we asked subjects to choose the minimal answer span that they considered necessary). Figure 8 shows the distribution of initial words of the questions (an indicator of the question type) for the CLEF QA 2004, AnswerGatherer EXPERT, and AnswerGatherer WEB datasets, respectively. For eight out of ten question type classes, we observe strong similarities between all three datasets; only for questions with *wogegen/wozu/wohin* ("against what/what...for/where..to") and imperatives with *nenne* ("name...") can we observe minor differences. As can bee seen in Figure 6, we also observed a small number of one-word and two-word questions due

[2] http://www.lsv.uni-saarland.de/answergatherer/

[3] http://www.language-experiments.org/

[4] http://www.psychologie.unizh.ch/

[5] The lower bound is the number of subjects that revealed their e-mail addresses; an upper bound is given by counting distinct IP addresses.


SmartWeb AnswerGatherer: Frage-Antwort-Paar Annotierungswerkzeug - Mozilla
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Fenster Hilfe
deutsche Thrash Metal Band
Suchen
Startseite
Lesezeichen
Post to CiteU...
SmartWeb Fr...
Dilbert.com -...
Slashdot: Ne...
Goobile
heise mobil -...
CIKM 2005
SmartWeb Frage-Antwort-Paar Annotierungswerkzeug
Auswahl übernehmen >>
UNIVERSITÄT DES SAARLANDES
Kontakt | Hilfe
Formular Rücksetzen
Frage-Antwort-Paare Abschicken

Fächermotor - Wikipedia

Fächermotor

Die Fächermaschine ist eine Kolbenmaschine, bei der die Triebwerke (Triebwerk: Kolben, Pleuel, Kurbelwellenkröpfung) gleichmäßig auf den halben Kreisumfang verteilt, analog zur Sternmaschine, wo diese am gesamten Kreisumfang verteilt sind. Üblich sind bis zu 5-zylindrige kleine Verdichter, bei denen die Pleuel nebeneinander auf einem Kurbelwellenzapfen angeordnet sind. (Bei vielzylindrigen Sternmaschinen sind sie über ein gemeinsames Hauptpleuel angelenkt) Ein Sonderfall dieser Gattung ist die V-Maschine mit 90° Gabelwinkel (als V-Motor in Motorrädern, aber auch als Verdichter in Druckluftkompressoren beliebt). Die Rede ist hier von Maschinen anstatt von Motoren, da es (Verbrennungs-)Motoren, Verdichter, Pumpen und Dampfmaschinen in dieser Anordnung geben kann. Einordnung: Technik Diese Seite Persönliche Werkzeuge Anmelden oder neues Benutzerkonto erstellen Navigation Hauptseite Wikipedia-Portal Aktuelle Ereignisse Letzte Änderungen Zufälliger Artikel Hilfe Spenden Suche Werkzeuge Links auf diese Seite Verlinkte Seiten Hochladen Spezialseiten Druckversion Permanentlink Impressum | Diese Seite wurde zuletzt geändert um 17:49, 18. Mär 2005. Der Inhalt dieser Seite steht unter der GNU-Lizenz für freie Dokumentation Über Wikipedia Lizenzbestimmungen ¯

Fächermotor

1.Frage:	Was ist ein Fächermotor?	Erklärung
1.Antwort:	Kolbenmaschine, bei der die Triebwerke (Triebwerk: Kolben, Pleuel, Kurbelwellenkröpfur	
2.Frage:	Wie groß ist der Gabelwinkel bei der V-Maschine?	Zahl
2.Antwort:	90°	
3.Frage:	Wo werden V-Motoren eingesetzt?	(bitte wählen)
3.Antwort:	in Motorrädern, aber auch als Verdichter in Druckluftkompressoren	

Figure 4: SmartWeb AnswerGatherer Web Interface.

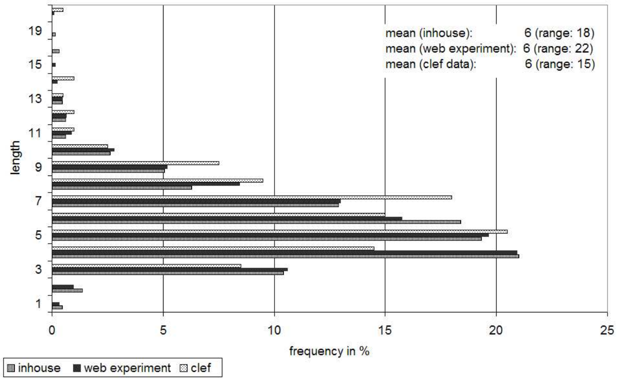


Figure 6: Distribution of Question Length.

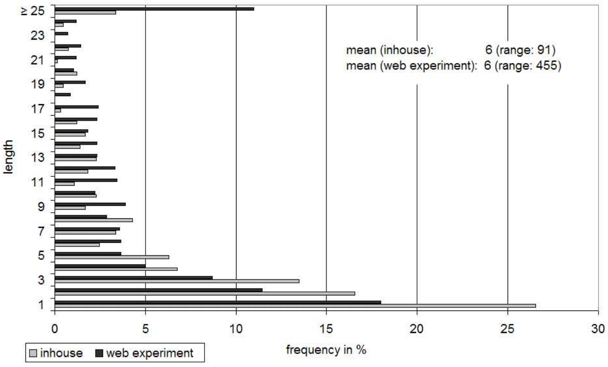


Figure 7: Distribution of Answer Length.

to elliptical constructions. In addition, our resulting corpus has a lot of occurrences of anaphoric questions, since our guidelines did not explicitly exclude them Figure 9.

Figure 10 shows the growth of our harvested collection during the period when the Web experiment was carried out. The large jumps (e.g. on October 11th and 20th, respectively) indicate activities on our part to attract more volunteers (such as emails or postings to mailing lists). However, it is also clear that–while effective in principle–without steady "marketing" activities the participation quickly levels off (November 15th). Manual inspection of our results revealed one case of "e-vandalism", i.e. a sequence of meaningless letter sequences entered as questions and answers. However, we assume that this was accidental, probably caused by a non-speaker of German who was curious about our GUI, but did not know how to use it (our experiment targeted only German speakers, so all instructions were also in German). We removed the pairs concerned. We also discovered some minor cases where non-native speakers "contributed" to our experiment. It is difficult to filter out such cases.

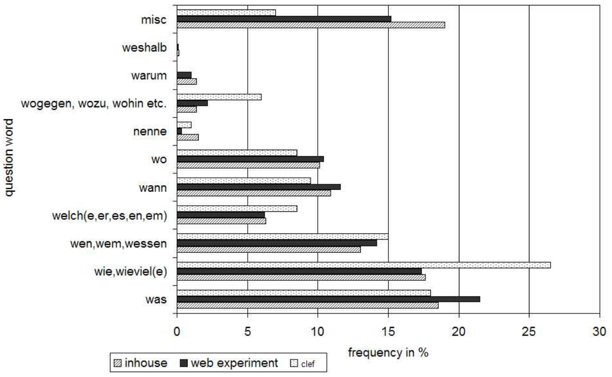


Figure 8: Approximate Question Type Distribution in Inhouse and Web Experiments, Respectively.

Q_{1a}:	`Definiere die freie Enthalpie.` *Define the free enthalpy.*
Q_{1b}:	`Wie wird sie noch genannt?` *How is it also called?*
Q_{2a}:	`Wann veröffentlichte Milán Füst seine ersten Gedichte?` *When did Milan Fust publish his first poems?*
Q_{2b}:	`Und worin?` *And where?*

Figure 9: Examples of Elliptical and Anaphoric Usage.

5. Discussion

5.1. Advantages

To sum up the results observed in the previous section, both the comparison of the question type distributions and the length distributions provide evidence to support our claim that gathering question answer pairs from Web subjects using WIKIPEDIA in a scenario as proposed here is a feasible way to create Q&A evaluation datasets, and results in datasets not too different from questions designed by experts and/or taken from search logs as in TREC/CLEF. We believe the selection of topics covered in encyclopedias like WIKIPEDIA for a language is not universal, but reflects the salience attributed to themes in a particular culture that speaks the language. Our approach thus benefits from the availability of the German WIKIPEDIA, in that topics covered will perhaps be more likely to be of interest to a German-speaking audience than translated topics from other languages/cultures. When building a Q&A system, the use of encyclopedia articles allows the use of context of the entry to be utilized, i.e. for statistical learning, word sense disambiguation, and the resolution of pronouns; for instance, the referent of *er* in *Wann wurde er geboren?"* ("When was he born?") is likely to be the person that the article is about. Our method also gives subjects the freedom to choose interesting questions/answers. They choose how to formulate question, and what would best satisfy their information need (answer), for example regarding the length of the selected text passage. Again, this means that preferences can be induced from a reasonably sized dataset.

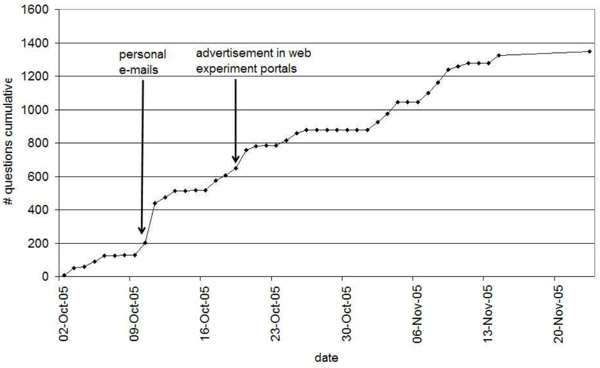


Figure 10: Participation in the Web Experiment as a Function of Time (cumulative).

5.2. Challenges

Answer Granularity. Consider the following example:

Q: Wann wurde Saarbrücken erstmals urkundlich erwähnt?
(When was Saarbrucken first mentioned?)
A_1: 999 (in 999)
A_2: im Jahre 999 (in the year 999)
A_3: Saarbrücken wird in einer Schenkungsurkunde Kaisers Otto III. im Jahre 999 erstmals als Königsburg "castellum Sarabrucca" erwähnt, die dem Bistum Metz geschenkt wird.
(Saarbrucken gets mentioned in a donation bull by emperor Otto III. in the year 999 as royal castle "castellum Sarabrucca" (Lat.) for the first time.)

Three questions become immediately apparent: (a) What is the best unit size for mark-up? (b) What is the most useful answer for the user? (c) What is the most useful unit for system evaluation? We have not focused to find the answer to these questions in this paper, but rather specified the shortest unit that can be the answer as the "correct" text span that subjects were to annotate in our guidelines.

Stylistic Alignment. Consider another example:

A: Saarbrücken wird in einer Schenkungsurkunde Kaisers Otto III. im Jahre 999 erstmals als Königsburg "castellum Sarabrucca" erwähnt, die dem Bistum Metz geschenkt wird.
(Saarbrucken gets mentioned in a donation bull by emperor Otto III. in the year 999 as royal castle "castellum Sarabrucca" (Lat.) for the first time.)
Q_1: Wann wurde Saarbrücken erstmals urkundlich erwähnt?
(When was Saarbrucken first mentioned in a document?)
Q_2: Wie alt ist Saarbrücken?
(How old is Saarbrucken?)

We have seen many cases in our data where subjects formulated questions in the style of Q_1 rather than the more natural style of Q_2. This can be taken as evidence that people get influenced stylistically by the documents that they harvest the answers from in the sense that they take over lexical, syntactic and idiomatic choices of the "encyclopedic register" when formulating the question. This is an instance of *alignment* (Levelt and Kelter, 1982), and we have observed it despite instructions in our guidelines asking for questions formulated in the way subjects would naturally ask them. (Branigan et al., 1999) found that the persistence of alignment in writing dissipates quickly when a single sentence intervened between prime and target. However, we do not know of any controlled studies that are dedicated to how to best break alignment in online experiments. An intuitive way to break alignment would be to detract subjects from the formulations of the WIKIPEDIA article by displaying intermediate material between marking the answer and formulating the corresponding questions. This, however, is left for future research. Despite these two challenges, we have shown that harvesting of open content sources like WIKIPEDIA for Q&A can benefit research; but ultimately, instead of relying on volunteer contributions research should feed back into the pool of open knowledge and improve the state of the art in content access, to help the model sustain and flourish.

6. Summary, Conclusions and Future Work

We have presented a method to utilize a community of volunteer subjects for creating an evaluation resource for open-domain question answering. To this end, we have implemented a tool that allows subjects to mark up passages containing facts of interest and to formulate questions asking for those facts. We have deployed the tool as a Web experiment and have quickly gathered a sizeable corpus of question-answer pairs that are useful to evaluate open-domain question answering systems. Our method has proved to be rapid to develop and cheap to deploy, but relied on steady activity to attract (and retain) volunteers. The resulting dataset was compared against an in-house dataset and against the CLEF QA 2004 set of German questions in order to study question type and length distributions in comparison. We also observed interesting phenomena in the resulting data, including ellipsis, anaphora, and stylistic alignment. Finally, the German Q&A corpus created by applying the method described in this paper can be freely obtained (please visit http://www.lsv.uni-saarland.de/resources/). The method proposed here and executed for German can without changes be applied to other languages for which localized WIKIPEDIAS are available in significant sizes (including English, Spanish, Polish etc).

Acknowledgments. This research was partly funded by the BMBF Project SmartWeb under Federal Ministry of Education and Research grant 01IM D01M.

We thank Tim Bartel from the Wikimedia Foundation for feedback. A big “thank you” goes to all volunteer subjects, without whom this would not have been possible. We also thank the two Web experiment portals for linking to our study and Holly Branigan for discussions about alignment.

7. References

H. P. Branigan, M. J. Pickering, and A. A. Cleland. 1999. Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, 6:635–640.

T. Clifton and W. J. Teahan. 2005. Semi-automated evaluation for question answering systems. Technical Report AIIA:05.5, University of Wales at Bangor, Bangor, Wales, UK.

Julian Kupiec. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA, USA, June 27 - July 1, 1993*, pages 181–190. ACM.

Jochen L. Leidner, Tiphaine Dalmas, Bonnie Webber, Johan Bos, and Claire Grover. 2003. Automatic multilayer corpus annotation for evaluating question answering methods: CBC4Kids. In *Proceedings of the Third Workshop on Linguistically Interpreted Corpora (LINC-3) held at the Tenth Annual Meeting of the European Chapter of the Association for Computational Linguistics 2003 (EACL’03)*, pages 39–46, Budapest, Hungary, April.

W. J. M. Levelt and S. Kelter. 1982. Surface form and memory in question answering. *Cognitive Psychology*, 14:78–106.

Lucian Vlad Lita, Warren A. Hunt, and Eric Nyberg. 2004. Resource analysis for question answering. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 162–165, Barcelona, Spain. Association for Computational Linguistics. Demonstration.

Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Ivanov Simov, and Richard F. E. Sutcliffe. 2004. Overview of the CLEF 2004 multilingual question answering track. In *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, pages 371–391.

Neil Mayo, Martin Corley, Frank Keller, and T. Florian Jaeger. 2005. WebExp2. [Online]. http://www.webexp.info/.

U.-D. Reips and C. Neuhaus. 2002. WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, and Computers*, 34:234–240.

Ulf-Dietrich Reips. 2002. Standards for Internet-based experimenting. *Experimental Psychology*, 49(4):243–256.

SmartWeb Consortium. 2006. Smartweb homepage. (Online). `http://smartweb.dfki.de/` (cited 2006-02-22).

Ellen M. Voorhees. 2003. Overview of the trec 2003 question answering track. In *TREC*, pages 54–68. U.S. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.

Wolfgang Wahlster. 2004. SmartWeb: Mobile applications of the Semantic Web. In Peter Dadam and Manfred Reichert, editors, *INFORMATIK 2004 – Informatik verbindet, Band 1. Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Ulm, 20.-24. September 2004*, volume 50 of *LNI*, pages 26–27, Heidelberg. Gesellschaft für Informatik (GI), Springer.