Show Me How to Tie a Tie: Evaluation of Cross-Lingual Video Retrieval

Pavel Braslavski[1], Suzan Verberne[2], and Ruslan Talipov[1]

[1] Ural Federal University pbras@yandex.ru, roosh90@mail.ru
[2] Radboud University Nijmegen s.verberne@cs.ru.nl

Abstract. In this study we investigate the potential of cross-lingual video retrieval for *how-to* questions. *How-to* questions are the most frequent among *wh*-questions and constitute almost 1% of the entire query stream. At the same time, *how-to* videos are popular on video sharing services. We analyzed a dataset of 500M+ Russian *how-to* questions. First, we carried out manual labelling of 1,000 queries that shows that about two thirds of all *how-to* question queries are potentially suitable for answers in the form of video in a language other than the language of the query. Then, we evaluated video retrieval quality for original and machine translated queries on a crowdsourcing platform. The evaluation reveals that machine translated questions yield video search quality comparable to the quality for original questions. Cross-lingual video search for *how-to* queries can improve recall and diversity of search results, as well as compensate the shortage of original content in emerging markets.

Keywords: how-to questions, video retrieval, question answering, cross-lingual information retrieval, machine translation, query translation, evaluation

1 Introduction

Several studies reported an increase in the share of question-like queries in search engine logs in recent years [11,19]. This phenomenon can be explained by different trends: users' desire for a more natural interface, users' laziness or low search proficiency, a large amount of Web content in the form of questions and answers that can be found through search engines, as well as proliferation of voice search. *How-to* questions are the most frequent question type in community question answering (CQA) [17] and Web search [11,19].[3] The substantial share of question queries and prevalence of *how-to* questions mark a significant evolution in user behavior in the last decade. For example, in the late 1990s, queries in question form comprised less than 1% of the entire search engine query stream and the most common question type was `[where can i find...]` for general information on a topic [16].

Another modern trend is the proliferation of multimedia content on the Web, video in particular. Pew Research Center found in 2013 that 72% of adult Internet users use video sharing services and that *how-to* videos are among the

[3] https://www.google.com/trends/2014/story/top-questions.html

top interests, watched by 56% of online adults.[4] Video is a natural medium for answering many *how-to* questions [3,18]. One aspect that distinguishes video from text is that videos can often be understood with visual information only. This means that even if the textual information accompanying the video is in a language that is not well-understood by the user who asked the question, they might still consider their question answered by the video. As a result, retrieving videos in a different language than the user's query language can potentially give a higher recall for the user's question. In the countries with emerging Web and growing Internet access, content creation in the users' first language lags behind the demand, and cross-language video retrieval can enrich the supply. Further, most *how-to* questions are longer and more coherent than the average search query, which makes them more suitable for machine translation (MT).

In this study, we investigate the potential of cross-lingual video retrieval for *how-to* questions. The main goals of the study are to

1. get a better understanding of *how-to* question queries, their properties, structure, and topics; as well as their potential for video results in a language other than the language of the original query;
2. evaluate Russian→English machine translation quality for *how-to* questions;
3. evaluate the complete pipeline for cross-lingual *how-to* video retrieval.

We used a large log of Russian *how-to* questions worth 500M+ queries submitted throughout a year, which constitutes almost 1% of the entire query stream. We performed a thorough automatic analysis of the data and manually labeled a considerable subset of queries. We retrieved videos through the YouTube API both for original Russian queries and their machine translations. After that we performed search results evaluation on a crowdsourcing platform.

The contributions of this paper compared to previous work are two-fold: First, we show an in-depth analysis of Russian *how-to* queries in order to estimate the proportion of queries for which cross-lingual video retrieval would be valuable. Second, we evaluate the pipeline for cross-lingual video retrieval step by step on a sample of queries from a leading Russian search engine, using crowd judgments for assessing video relevance.

2 Related work

Cross-language question answering (QA) was investigated before in the context of the well-known IR evaluation campaigns CLEF [5] and NTCIR [9]. The study [13] explores a translingual QA scenario, where search results are translated into the language of the query. The authors discuss implementation strategies and specific MT errors that are critical to solving the problem.

A rich set of features for ranking answers from CQA archives in response to *how-to* questions is investigated in [17]. Weber et al. [20] first extract queries with *how-to* intent from a search log, then answer them with "tips" from CQA.

[4] http://www.pewinternet.org/2013/10/10/online-video-2013/

Research in multimedia QA started over a decade ago with mono-lingual video retrieval. The early work focuses on factoid questions and uses a speech interface, transcribing the queries and the videos using automatic speech recognition [22,2]. The first work addressing cross-language video QA aims at answering questions in English using a corpus of videos in Chinese [21]. The authors use OCR to extract text from videos, and MT to translate the video text to English. Their QA module takes a classic approach, applying question analysis, passage retrieval and answer selection.

In the late 2000s, it was found that non-factoid questions are more frequent than factoid questions on the web, and that *how-to* questions constitute a large proportion of *wh*-questions [17]. Thus, the scope of research in multimedia QA was broadened from factoid questions to non-factoids such as *how-to* questions. It was argued that video retrieval is especially relevant for *how-to* questions [3]. The approach taken by Chua et al. [3] consists of two steps: (1) finding similar questions on Yahoo! Answers with which the terminology from the original question is expanded, and retrieving videos from YouTube for those expanded queries; and (2) re-ranking the retrieved videos based on their relevance to the original question. The authors evaluate their system on 24 *how-to* queries from Yahoo!Answers that have corresponding video answers in YouTube. In two follow-up papers [7,8], Li et al. improve the analysis of *how-to* questions in order to extract better key phrases from the question. They also improve video re-ranking using visual features, user comments and video redundancy.

In their overview paper [6], Hong et al. identify three directions for research in multimedia QA: 1) the creation of large corpora for evaluation, especially for definition and *how-to* QA; 2) the development of better techniques for concept detection and multimedia event detection; and 3) extension of the existing approaches to general domains. The latter of these three goals is addressed by [10].

In the current paper, we address the problem of answering *how-to*-questions with videos. We use Russian queries as a source, and retrieve videos for both the original query and its English translation. Our main contribution compared to previous work is that we show the large potential of *cross-lingual* video retrieval for answering *how-to* questions.

3 Data

Our study uses a subset of question-like Russian queries submitted to a major Russian search engine. The initial dataset comprises of all queries for the year 2012 containing question words and their variants. Under the agreement with the search engine, we only have access to the queries containing question words for research purposes; we have no access to the other queries issued by the same users or to the search results. The nearly 2 billion initially acquired questions form about 3–4% of the actual query log.

The initial data underwent a multi-step cleaning to keep only queries that represent actual question-asking information needs. First, spam and bot users were removed from the log based on total number of submitted queries, unnat-

Table 1. Top-10 most frequent queries.

Query	Count
[how to download music from vkontakte]*	1,048,845
[how to kiss in a right way]	880,326
[how to remove page in odnoklassniki]*	717,639
[how to make a slime toy]	691,358
[how to do it in a right way]	545,554
[how to download video from youtube]	542,397
[how to make a magic wand]	396,700
[how to earn money]	345,653
[how to quit smoking]	297,358
[how to build a website]	286,657

*popular Russian social network sites

ural query 'bursts', very long queries, and long sequences of almost identical questions. Second, only queries with a question word in the first position were retained. Further, we filtered out queries matching Wikipedia titles, crossword puzzles formulations and questions from TV game shows. Finally, we filtered out those question queries that contain only one word after we removed stop-words and question words. The cleaning removed more than half of the originally sampled questions; the remaining dataset contains about 915 million question queries from about 145 million users. This represents up to 2% of the entire query stream. A detailed description of log cleaning can be found in [19].

For the current study we extracted all queries starting with *как (how)* from the cleaned log, which resulted in $573,129,599$ total queries ($237,846,014$ unique queries). *How-to* questions comprise about 63% of all question queries and up to 1% of the entire search engine query stream.

Table 1 cites the top-10 *how-to* questions along with their log frequencies (here and in subsequent tables the Russian queries and distinct terms have been translated for the reader's convenience). Table 2 summarizes the most frequent last words of the questions (suffixes) that are a good indicator of query intent. The list supports our hypotheses that many *how-to* questions seek for easy-to-perform instructions (*at home* in different formulations and *DIY*) and visual information (*video, photo*). *How-to* questions reflect also the ubiquity and popularity of social networks (*vkontakte, odnoklassniki*). 14.2% of the queries contain Latin characters. The presence of the Latin characters can be seen as an indirect signal that it might be useful to translate the query because the topic is potentially non-local. Table 3 lists the top-10 words in Latin script. The table shows that the most frequent words in Latin script are foreign words related to computer software, games and mobile devices. Presence of *youtube* in the list again supports our assumption that many *how-to* questions seek for visual content.

Figure 1 shows cumulative query frequency distribution. Unique queries comprise 36.9% of the whole query mass; only 55 queries have frequencies larger than 10^5. Figure 2 shows length distribution of *how-to* questions (note that two-word questions were removed on the previous log cleaning stage). Question queries are

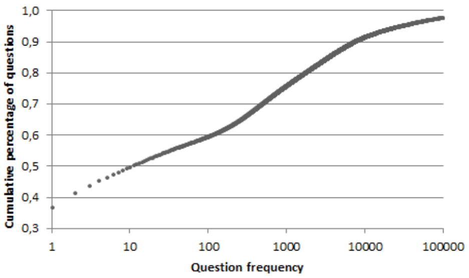

Fig. 1. Question frequency distribution.

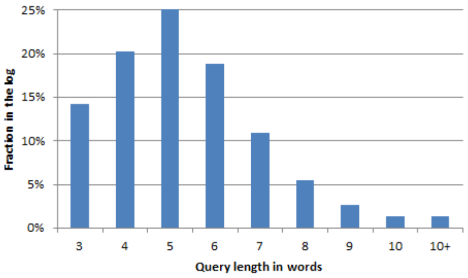

Fig. 2. Question length distribution.

Table 2. Most frequent words in the last position of the question.

Suffix	Count	%
at home	9,342,893	1.63
video	8,409,924	1.47
[*windows*] 7	3,421,005	0.60
minecraft	3,415,061	0.60
vkontakte	3,321,133	0.58
DIY	2,794,189	0.49
photo	2,715,022	0.47
free	2,441,100	0.43
odniklassniki	2,409,490	0.42
at home	2,148,758	0.37

Table 3. Most frequent Latin character words in Russian question queries.

Query term	Count	%
windows	5, 699, 513	0.99
minecraft	5, 418, 118	0.95
iphone	2, 229, 155	0.39
wifi	1, 758, 564	0.31
samsung	1, 486, 654	0.26
xp	1, 346, 204	0.23
cs (counter strike)	1, 304, 039	0.23
nokia	1, 236, 462	0.22
ipad	1, 165, 871	0.20
youtube	1, 158, 642	0.20

longer than average web queries: the most frequently occurring query length in our data sample is five words, constituting one fourth of all *how-to* questions.

4 Results

In this section, we will first present the results of the in-depth manual analysis of a small query sample (Section 4.1), then evaluate the automatic translation of Russian queries (Section 4.2) and finally present the results of video retrieval evaluation (Section 4.3).

4.1 Manual Analysis of 1,000 Queries

To get a better understanding of *how-to* queries and the potential of answer them with videos in a different language, we randomly sampled 1,000 unique queries with frequencies 100 and higher from the dataset and analyzed them manually. This sample corresponds to 564, 866 queries submitted by users. Despite the fact that special attention was paid to cleaning the initial data from non-interrogative queries that look like questions (see previous section), two queries of this sort – a song title and a TV series title – were found in the sample.

Table 4. Query labeling results, in percentages of the unique 1,000 queries (proportions accounting for log frequencies are similar).

yes	**possibly**	**no**	**hard to answer**
Are video results potentially useful?			
74.5	12.8	11.1	1.7
Are English results potentially useful?			
83.6	1.3	14.8	0.4

The thousand queries were labeled by two authors in regard to three facets: (1) whether or not video would be a good answer medium for this question; (2) whether or not results in English (regardless – text or video) would be useful; (3) what the question's topical category is. The questions were first labeled by two authors independently; then the labels were discussed and discrepancies were reconciled. In cases where the authors could not interpret the information need behind the question, the question was labeled as *hard to answer*. Some opposed decisions (*yes/no* for facets 1 and 2) resulted in averaging to the label *possibly* upon negotiation.

Results for facet 1. Potentially any kind of content can be represented by video: text, music, still images, and video proper. Text can be rehearsed, presented as running lines or a sequence of textual fragments (and such videos can be found on the Web in plenty). When labeling queries in this aspect, we tried to assess to which extent a video answer would be appropriate and helpful. To be marked with *yes* the query must relate to a real-life, tangible problem. The topic of the question might be abstract, as long as the answer can be shown on video. For example, the query `[how to solve absolute value inequalities]` relates to a rather abstract mathematical problem, but was labelled as allowing a video answer, taking into account proliferation of MOOCs and supported by plenty of relevant video results for this query. `[how to calculate profitability]` is an example of a 'non-video' query.

Results for facet 2. Some queries are local – relate to national mobile network operators (MNO) `[how to remove a number from the blacklist megafon]`, locally used software `[how to work in 1C 7.7]`, taxation or legislation `[how to pay the vehicle tax]`, and customs `[how to dress on 1 september]`. This kind of questions was tagged as inappropriate for translation.

Table 4 shows the outcome of the manual labelling for the first two facets. The target subset of the queries for this study – for which both video results and results in other languages are potentially useful – constitutes 68.9% of unique queries in our sample and 66.4% in the corresponding query stream.

Results for facet 3. As a starting point for query categorization, we took the YouTube channel topics.[5] The list consists of 16 items and is not an ideal flat taxonomy – in our case, almost all the queries could have been assigned to the *How-to* category, so we tried to choose the most specific category. In

[5] `https://www.youtube.com/channels?gl=US`

the course of labeling, we slightly modified the list of categories: 1) *Cooking & Health* was divided into two separate categories; 2) *Legal & Finance* and *Adult* were introduced, 3) *Tech* was renamed into *Computers, Internet and Cell phones.* The latter category became expectantly the largest one, but we did not divide it further, because many questions correspond simultaneously to several related concepts, e.g. `[how to download a photo from iphone to computer]` or `[how to setup Internet on MTS]` (MTS is a national MNO). Table 5 shows the breakdown of categories in our question set along with fractions of queries, for which video results and results in a language other than Russian are potentially useful in each category. It can be seen from the table that *Legal & Finance* category is the least suitable both for video and non-local results. Low figures in *Science & Education* are due to questions like `[how to translate...]` and `[how to spell...]`. Many questions from *Lifestyle & Social* are about relationships, dating, etc. that can be well *illustrated*, but hardly *answered* with video.

Manual investigation of questions allowed us for making several additional observations: About 2% of questions contain spelling errors; 8% contain slang (e.g. *бесплатка* for a free text service 'please call me') or transliterated names of software, computer games, or services (e.g. *фотошоп* for *photoshop*). Spelling errors can be seen as a lesser problem, since search engines are good at correcting misspelled queries; slang and transliteration can potentially harm translation quality to a larger extent.

In the sample, we saw that question queries starting with *how* followed by a verb can be divided in two groups: 1) a large group of question queries starting with *how + infinitive* that have a clear practical intent (equivalent to the English *how to*), e.g. `[how to cook stuffed peppers]`, and 2) a much smaller group of queries starting with *how + finite verb form* that have a more abstract curiosity intent, e.g. `[how was the clock invented]`. The ratio of infinite/finite verbs following *как* is 16:1. Also note that not all Russian questions starting with *как* correspond to English *how* counterparts. For example, *как зовут X?*-questions – literally *how is X called* – correspond to *what is X's name?.*

How-to questions are about actions that are usually described by verbs. Therefore it is interesting to note that some categories are very characteristic for their verb use in question, though most generic verb *do/make* is presented in all categories in different proportions. The most distinguished category is *Cooking*, which presents the entire range of culinary manipulations – *cook* (by a large margin), *bake*, *soak, pickle, cut, jerk, marinate*, etc. Another example – *Computers, Internet and Cell phones*, where leading *do/make* is followed by more specific *create, install, configure, download*, and *remove.*

These insights can be valuable for automatic analysis and categorization of *how-to* question queries.

Table 5. Question category breakdown. To several categories (*Music, Comedy, Film & Entertainment, From TV, Animation, Causes & Non-profits, News & Politics*) only zero to two queries were assigned, so we do not cite figures for them in the Table. For reader's convenience, the table shows data for unique queries only. Category breakdown accounting for query frequencies is roughly the same with minor deviations. "Comp, Int & Phon" refers to the category "Computers, Internet and Cell phones"

Category	Share (%)	Video? (%)	Translate? (%)	Example
Gaming	5.9	89.2	95.4	`[how to save Mordin]`
Beauty & Fashion	7.0	98.7	98.7	`[how to get rid of freckles]`
Automotive	2.7	89.7	79.3	`[how to improve sound in car]`
Sports	1.9	100.0	100.0	`[how to jump on a skateboard]`
How-to & DIY	18.7	94.6	95.1	`[how to fix hooklink]`
Comp, Int & Phon	26.6	86.3	82.1	`[how to change local disk icon]`
Science & Education	7.6	45.8	72.3	`[how is beeswax made]`
Cooking	6.3	98.6	97.1	`[how to make puff pastry]`
Health	5.0	70.9	100.0	`[how to treat wound]`
Lifestyle & Social	4.9	13.0	87.0	`[how to attract a guy]`
Legal & Finance	5.0	3.6	20.0	`[how to calculate income tax]`
Adult	2.7	93.1	100.0	`[how to give erotic massage]`
Other	4.6	8.0	48.0	`[how to call to Thailand]`

4.2 Machine Translation

The 1,000 queries were machine translated using three free online services – Yandex[6], Google[7], and Bing[8] – to avoid bias in the gold standard translation: a professional translator post-edited randomly picked translations from the three MT engines' outputs. As a byproduct we obtained a comparative evaluation of three services.

We calculated two popular MT quality measures widely adopted by the MT community: BLEU (Bilingual Evaluation Understudy, [12]) and TER (Translation Error Rate, [15]) – using the post-edited translation as reference. Table 6 summarizes automatic translation quality scores for the three MT engines. While BLEU indicates the proportion of common n-grams in reference and machine translations (larger scores mean better translations), TER measures the number of edits required to change a system output into the reference (the lower the better). Both measures rank the three systems equivalently. The obtained BLEU scores are significantly higher than those by the best performing systems for Russian-English pair in the WMT'2015 evaluation campaign (around 0.29) [1]. This is expected, since reference translations were obtained as a result of post-editing, not as 'from scratch' translations.

For our retrieval evaluation experiments we took the output of Google Translate (the lowest score in our list) to be not overoptimistic.

[6] `https://translate.yandex.com/`

[7] `https://translate.google.com/`

[8] `http://www.bing.com/translator`

Table 6. Quality evaluation of the machine translation engines for query translation (1,000 unique *how-to* queries), using a professional post-edited translation as reference.

MT engine	BLEU	TER
Yandex	0.52	30.95
Google	0.44	37.18
Bing	0.49	34.17

4.3 Video Retrieval Evaluation

We used the YouTube search API[9] to retrieve videos for 100 queries. We sampled these queries from our *target* subset, i.e. queries for which we marked that both video answers and answers in a different language could be useful. Thus, the evaluation results can be considered an upper limit of cross-lingual video retrieval, when no query analysis is performed. We retrieved 10 videos from YouTube for each original Russian query, and 10 videos for the query's English translation by Google Translate. We had each of the videos assessed by three workers on the crowd sourcing platform Amazon Mechanical Turk[10] — resulting in a total of 6,000 $(100 * (10 + 10) * 3)$ HITs.

Unfortunately, Russian native speakers are marginally presented on Mechanical Turk [14]. Therefore, we modelled cross-language video retrieval in a reverse direction: workers supposedly proficient in English – we set locale to US as qualification for the HITs – were presented with the reference query translation and had to evaluate videos retrieved both for original Russian query and its English machine translation. Note that this reverse setting is more rigorous than the true Ru→En direction: we can imagine that even an average Russian-speaking user possesses some elementary knowledge of English, whereas the odds are much lower that MTurk workers with US locale are proficient in Russian.

We set up an annotation interface in which we showed a question together with one retrieved video and the question "How well does the video answer the question?" Relevance labeling was done on a four-point scale: (3) Excellent answer; (2) Good answer; (1) May be good; (0) Not relevant. In order to validate that the MTurk labels by speakers of English are a good approximation for the assessments by Russian speakers, we recruited Russian volunteers through online social networks. The volunteers labeled the search results for 20 original Russian queries and their translations (10 results per query) in the same interface. Each question–video pair was assessed by two volunteers; 45 volunteers took part in the labeling.

We calculated the inter-rater agreement between the Russian volunteers and English-speaking MTurk workers as an indication for the validity of the MTurk labels. We found the following agreement scores in terms of weighted Cohen's κ:[11]

[9] `https://developers.google.com/youtube/v3/`

[10] `https://www.mturk.com`

[11] Weighted κ is a variant of κ that takes into account that the labels are interval variables: 3 is closer to 2 than it is to 1.

Table 7. Evaluation of the video re for the original Russian how-questions and their English translations. Precision@10 is the proportion of relevant videos (assessed relevance ≥ 2) in the top-10 YouTube results. Success@10 is the proportion of questions that have at least one relevant video in the top-10. DCG uses averaged relevance scores by MTurk workers.

	Precision@10 (stderr)	Success@10	DCG@10
For original questions	0.643 (0.250)	0.98	3.45
For translations	0.638 (0.311)	0.96	3.39

the agreement among MTurkers was 0.448; the agreement among the Russian volunteers was 0.402 and the agreement between Russians and MTurkers was 0.414. Since the agreement between Russian volunteers and MTurk workers is not lower than the agreement among Russian volunteers, the MTurk annotations can be considered a good approximation for the assessments by Russian speakers.

Table 7 shows the results for the video retrieval evaluation in terms of Precision@10, Success@10, and DCG@10[12]. In case of DCG we averaged MTurk's scores for each query-video pair; for evaluation with binary relevance, we considered the scores 2 and 3 (good or excellent answer) as relevant and used the majority vote of the three workers as final relevance judgment. According to a paired t-test the difference between DCG scores for original queries and their translations is not significant ($P = .73$, $n = 100$), neither is the difference between the Precision@10 scores ($P = .98$).

If we consider the proportions of relevance labels in the labelled videos, we see that 17% of the assessed videos was labelled as (0) Not relevant; 19% was labelled as (1) May be good; 29% was labelled as (2) Good answer and 35% was labelled as (3) Excellent answer. Figure 3 illustrates the differences in Precision@10 scores between original Russian queries and their English translations. In total, 48 queries show improved precision@10 for the translated query, in 39 cases the translated query gives worse results, and for 13 queries the results tie. Manual investigation of harmed queries reveals that the drop in quality is mainly due to translation flaws – either wrong translation for polysemic words or poor processing of Russian writings of English names such as *майнкрафт* for *minecraft*. In some cases the translation is fair, but either the topic is unpopular (e.g. `[how to clean white mink fur at home]`) or the search results are not precise (e.g. `[how to turn on flash player in chrome browser]` results in videos *how to install flash player in chrome, how to fix crash from flash player in chrome*, etc.). English content (or English queries submitted through the API) may be moderated more strictly, e.g. there is a relevant result for original query `[how to delay ejaculation]`, but no relevant results for its correct translation.

[12] We opted for DCG, since it also reflects how many relevant results were retrieved, not only how well the retrieved results were ranked (as in case of nDCG).

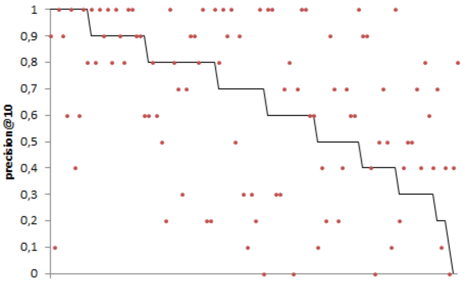

Fig. 3. Precision@10 for original queries (line) and corresponding translated queries (dots). The x-axis represents the individual queries, ordered by decreasing p@10 for original queries. All dots that occur above the line are queries that show improved p@10 when translated.

5 Conclusions

The main contribution of this study is that it shows the potential of cross-lingual video retrieval for *how-to* questions. We combined an in-depth manual analysis with evaluation by the crowd of the translation–retrieval pipeline. The result of the manual analysis was a set of questions for which we decided that videos in a different language are potentially useful answers. This claim was supported by our results: 98% of the Russian queries in our selection had at least one relevant video answer in the top-10 from YouTube. Precision@10 is 0.638, which implies that on average, 6 out of the first 10 video results are relevant ('Good answer' or 'Excellent answer') to the query. The results show that *how-to* queries translated to English by off-the-shelf systems give the same video retrieval performance as the original Russian queries. Cross-lingual video search for *how-to* queries can improve recall and diversity of search results, as well as to compensate the shortage of original content in emerging markets.

The obtained results suggest the following directions of research in the future. First, it would be interesting to study the topics of *how-to* question queries that benefit most from cross-lingual video retrieval. A recent study [19] demonstrates that even rare questions can be categorized in topical categories with acceptable quality (recall that about one third of *how-to* questions is unique in a yearly log). Categorization of questions could be paired with video categorization based on metadata and user comments [4]. Second, a user study must be carried out to analyze the users' experience with video retrieval results in a foreign language.

The annotated data is available for research purposes.[13]

Acknowledgments

We thank Yandex for preparing the dataset and granting access. Pavel Braslavski's work was supported by RFBR, research project #14-07-00589-a. This publica-

[13] `http://kansas.ru/howto-video/`

tion was partially supported by the Dutch national program COMMIT (project P7 SWELL). We also thank all volunteers, who took part in the evaluation.

References

1. Bojar, O., et al.: Findings of the 2015 workshop on statistical machine translation. In: WMT. (2015)
2. Cao, J., Nunamaker, J.F.: Question answering on lecture videos: a multifaceted approach. In: JCDL. (2004)
3. Chua, T.S., Hong, R., Li, G., Tang, J.: From text question-answering to multimedia qa on web-scale media resources. In: LS-MMRM Workshop. (2009)
4. Filippova, K., Hall, K.B.: Improved video categorization from text metadata and user comments. In: SIGIR. (2011)
5. Giampiccolo, D., et al.: Overview of the clef 2007 multilingual question answering track. In: Advances in Multilingual and Multimodal Information Retrieval. (2008)
6. Hong, R., Wang, M., Li, G., Nie, L., Zha, Z.J., Chua, T.S.: Multimedia question answering. IEEE Trans. Multimedia **19**(4) (2012)
7. Li, G., et al.: Video reference: question answering on youtube. In: MM. (2009)
8. Li, G., Li, H., Ming, Z., Hong, R., Tang, S., Chua, T.S.: Question answering over community-contributed web videos. IEEE Trans. Multimedia **17**(4) (2010)
9. Mitamura, T., et al.: Overview of the ntcir-7 aclia tasks: Advanced cross-lingual information access. In: NTCIR-7 Workshop. (2008)
10. Nie, L., Wang, M., Gao, Y., et al.: Beyond text qa: Multimedia answer generation by harvesting web information. IEEE Trans. Multimedia **15**(2) (2013)
11. Pang, B., Kumar, R.: Search in the lost sense of query: Question formulation in web search queries and its temporal changes. In: ACL, Vol. 2. (2011)
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. (2002)
13. Parton, K., McKeown, K.R., Allan, J., Henestroza, E.: Simultaneous multilingual search for translingual information retrieval. In: CIKM. (2008)
14. Pavlick, E., Post, M., Irvine, A., Kachaev, D., Callison-Burch, C.: The language demographics of amazon mechanical turk. TACL **2** (2014)
15. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: AMTA. (2006)
16. Spink, A., Ozmultu, H.C.: Characteristics of question format web queries: An exploratory study. Information Processing & Management **38**(4) (2002)
17. Surdeanu, M., Ciaramita, M., Zaragoza, H.: Learning to rank answers to non-factoid questions from web collections. Computational Linguistics **37**(2) (2011)
18. Torrey, C., Churchill, E.F., McDonald, D.W.: Learning how: the search for craft knowledge on the internet. In: CHI. (2009)
19. Völske, M., Braslavski, P., Hagen, M., Lezina, G., Stein, B.: What users ask a search engine: Analyzing one billion russian question queries. In: CIKM. (2015)
20. Weber, I., Ukkonen, A., Gionis, A.: Answers, not links: Extracting tips from yahoo! answers to address how-to web queries. In: WSDM. (2012)
21. Wu, Y.C., Chang, C.H., Lee, Y.S.: Clvq: Cross-language video question/answering system. In: IEEE MMSE. (2004)
22. Yang, H., Chaisorn, L., Zhao, Y., Neo, S.Y., Chua, T.S.: Videoqa: question answering on news video. In: MM. (2003)