In Favor of or Against Multi-Lingual Q&A Sites? Exploring the Evidences from User and Knowledge Perspectives

ANONYMOUS AUTHOR(S)

Many Q&A sites initially run only in English, and then gradually release their multi-lingual variants to serve users who speak other languages. The launch of such multi-lingual sites always leads to intense dispute about the pros and cons of multi-lingual sites. Although all arguments and concerns sound reasonable, people can rarely provide solid evidence to convince each other. In this paper, from users' comments about the launch of several non-English Stack Overflow sites, we first identify three major concerns including community split, knowledge needs and interests in other languages, and knowledge fragmentation and duplication. To validate these three concerns, we conduct an evidence-based data analysis and comparison of user characteristics, tag usage and cross-site links between the Russian Stack Overflow and the English Stack Overflow on these three concerns. Our study sheds the light on the existence value and risks of multi-lingual Q&A sites.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing*;

Additional Key Words and Phrases: Multi-lingual Q&A Site; Stack Overflow; Community Split; Knowledge Needs and Interests; Knowledge Fragmentation and Duplication; Cross-Lingual Retrieval

ACM Reference Format:
Anonymous Author(s). 2010. In Favor of or Against Multi-Lingual Q&A Sites? Exploring the Evidences from User and Knowledge Perspectives. *Proc. ACM Hum.-Comput. Interact.* 9, 4, Article 39 (March 2010), 19 pages. https://doi.org/0000001.0000001

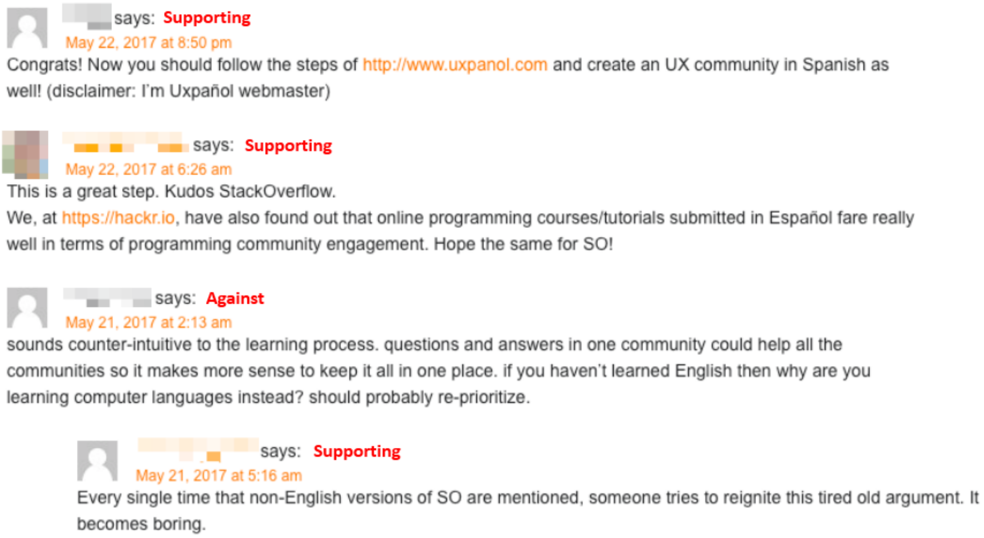
1 INTRODUCTION

A community question and answer (Q&A) website provides a collaborative information seeking platform for users to ask and answer questions. The accumulated questions and answers over time create a crowdsourced knowledge base which can benefit many people (in addition to the question askers themselves) who have similar questions. Recent years have witnessed the unprecedented development of Q&A websites such as Stack Exchange and Quora.

Originally, most of the Q&A sites run in one language (mostly English). However, for those who do not speak English or do not use English as their working language, they cannot directly access such English Q&A sites due to the language gap. To make the site available for non-English population, many Q&A sites begin to release their multi-lingual variants to serve users who speak other languages. Note that a multi-lingual variant of a Q&A site is neither the translation version of the original English site, nor the original site with language localisation features. Instead, it is a totally new site for a non-English language but adopts the same policies, practices and supports the same look and feel and user experience as the original English site. For example, there are Russian, Japanese, Portuguese, Spanish Stack Overflow[1], and French, Spanish, Japanese, German, Italian Quora[2]. Despite these benefits, the decisions of launching some multi-lingual variants of the

[1] https://stackoverflow.blog
[2] https://blog.quora.com

2573-0142/2010/3-ART39 $15.00
https://doi.org/0000001.0000001



[redacted] says: Supporting
May 22, 2017 at 8:50 pm
Congrats! Now you should follow the steps of http://www.uxpanol.com and create an UX community in Spanish as well! (disclaimer: I'm Uxpañol webmaster)

[redacted] says: Supporting
May 22, 2017 at 6:26 am
This is a great step. Kudos StackOverflow.
We, at https://hackr.io, have also found out that online programming courses/tutorials submitted in Español fare really well in terms of programming community engagement. Hope the same for SO!

[redacted] says: Against
May 21, 2017 at 2:13 am
sounds counter-intuitive to the learning process. questions and answers in one community could help all the communities so it makes more sense to keep it all in one place. if you haven't learned English then why are you learning computer languages instead? should probably re-prioritize.

[redacted] says: Supporting
May 21, 2017 at 5:16 am
Every single time that non-English versions of SO are mentioned, someone tries to reignite this tired old argument. It becomes boring.

Fig. 1. Example user comments and discussions on the launching of Spanish Stack Overflow site [14]

original English site always cause intense dispute about the pros and cons of multi-lingual variant sites among users. This is evident from the supporting- or against-comments that users left on the announcement of the new multi-lingual sites [6, 9, 11, 12, 14]. Users also created some posts on Meta Stack Exchange and Meta Stack Overflow to discuss the points in favour of or against multi-lingual sites [2, 4, 5, 7, 8, 10, 13].

Fig. 1 presents one typical example of such dispute about the multi-lingual Stack Overflow sites. We conduct a formative study in which we crawl all the user comments on the launch of multi-lingual Stack Overflow sites. Three major concerns emerge from our analysis: *community split*, *knowledge needs and interests in English and other languages*, and *knowledge fragmentation and duplication*. On the one hand, users who are against the multi-lingual sites are worried about: this would cause the community split across different sites, whether there are enough needs and interests in computer programming knowledge in other languages, and knowledge fragmentation and duplication across different sites would become a serious issue to deal with. On the other hand, users who support the multi-lingual sites argue that: the multi-lingual sites could involve more non-English-speaking users in the community, there are different knowledge needs and even unique knowledge interests for non-English-speaking users, and knowledge duplication may not be a bad thing as long as good questions and answers in one language could be referenced or translated in another language.

People use different examples, reasons, arguments to defend their opinions, but they can rarely provide solid evidences to convince others, and thus the same dispute repeats once another multi-lingual site is launched. Some users point out the needs for more evidence-based analysis of multi-lingual sites, rather than just conjecturing what would or would not happen. For example, this user suggests in his post [4] that "*We need to see charts of activity by country to gauge how many people this would bleed away from the main site. My guess is, it's a lot more than you'd expect*". With the launch of multi-lingual sites and the data accumulated on these sites over time, it becomes feasible for evidence-based analysis and comparison between the multi-lingual sites and the original

English site, which will give us insights into the value of multi-lingual sites, their impacts on the original English sites, and the necessary tools to support the multi-lingual sites.

In this paper, we conduct such an evidence-based data analysis and comparison between the Russian Stack Overflow (RSO) and the English Stack Overflow (ESO). We choose Stack Overflow as the subject site because it is the most successful and popular Q&A site for computer programming and it provides up-to-date data dump of the entire site information for public use and research. We choose Russian Stack Overflow because it is the largest multi-lingual Stack Overflow variant with sufficient users and posts for the comparative study. Inspired by the three major concerns we identify in our formative study of the user comments on the launch of multi-lingual Stack Overflow sites, we focus our analysis on three points. First, we study the risk of community split from the user perspective by analyzing user registration, user reputation score, and user posting behavior on RSO and ESO. Second, we study the knowledge needs and interests in English and Russian from the tag perspective by analyzing tag frequency correlation and tag uniqueness between RSO and ESO. Third, we study the knowledge fragmentation and duplication issue from the cross-site linking perspective by analyzing existing manual cross-site links by users and potential cross-site links that users are not aware of.

Our results show that some special knowledge needs and interests in non-English languages warrant the value of multi-lingual Q&A sites, and multi-lingual Q&A sites do not result in community split. However, knowledge fragmentation and duplication across multi-lingual sites is a valid concern. Although manual cross-site linking can partially deal with this issue, the extent of knowledge fragmentation and duplication calls for automated cross-site linking and cross-site translation support.

The main contributions of this paper are listed below:

- This work is the first quantitative study of the phenomena and effects of multi-lingual Q&A sites. It provides evidences to validate the existence value of multi-lingual Q&A sites, the risk of community split, and the issue of knowledge fragmentation and duplication.
- Our study identifies a new problem for the development of multi-lingual Q&A sites, i.e., manual cross-site linking is not sufficient for bridging the knowledge across multi-lingual sites. We provide a cross-site retrieval method which has potential to tackle this problem.
- Inspired by our findings, we envision cross-site retrieval and cross-site translation techniques to better support multi-lingual Q&A sites, and identify key challenges in developing such techniques.

2 RELATED WORK

There are many crowd-source knowledge-sharing websites in the world such as Q&A websites (like Quora and Stack Overflow) and Wikipedia. Although most of them begin with only English, the support of other languages has beed added later to make the site more accessible to non-English speakers. There is a growing field of research on multilingual Wikipedia. The general consensus on multilingual Wikipedia is that while English (the largest language edition) has a content advantage, each language edition has a great deal of unique information that is not available in other language editions [23, 24]. In contrast, the launch of Q&A sites in different languages has caused intense dispute, no matter for Stack Overflow [4, 5] or Quora [13]. Albeit bitter debate, it does not attract much attention from researchers to investigating multi-lingual Q&A sites. Our work attempts to fill in such a gap by carrying out a quantitative empirical study of multi-lingual Stack Overflow sites to explore the evidences in favor of or against multi-lingual Q&A sites.

As Wikipedia now owns more than 200 different editions for different languages, it encourages users to translate articles from the English Wikipedia into other languages with official guideline [17]. However, as sorely manual translation is time-consuming and labor-extensive, many researchers propose methods to help cross-lingual linking [33, 34, 36], or aid humans in propagating information from one Wikipedia (typically English) to another (typically a language edition with a small number of articles) by using machine translation [18, 19, 27].

There are also some works supporting multi-lingual retrieval specifically in Stack Overflow [21, 37]. To find related questions in Stack Overflow, Xu et al. [37] translated the Chinese query into English with online translation service, while Chen et al. [21] mapped both Chinese queries and English questions into the same vector space for cross-lingual retrieval. Different from their works, we mainly focus on the validation of users' concerns of multi-lingual Q&A sites including community split, knowledge interests and needs in other languages, and also knowledge fragmentation and duplication. We discuss the potential of extending our method in Section 7.1 for linking posts across multi-lingual sites, and also the feasibility of customising the state-of-the-art machine translation to help translate domain-specific posts in Section 7.2.

3 FORMATIVE STUDY OF THE CONCERNS ABOUT MULTI-LINGUAL STACK OVERFLOW SITES

In Stack Overflow, questions, answers and comments are required to be written in English [3]. As not all developers can ask or answer questions in English, the founders of English Stack Overflow would like to explore the feasibility to launch multi-lingual Stack Overflow, i.e., Stack Overflow in other languages. After the beta development process[3], several non-English Stack Overflow sites have been launched including Russian, Portuguese, Spanish, and Japanese one. Each time a new non-English Stack Overflow site has been launched, there was a public announcement in the official blog. For example, the announcement of the launch of the Spanish Stack Overflow site can be found at https://stackoverflow.blog/2017/05/20/stack-overflow-en-espanol-graduated/. Following the announcement, Stack Overflow users expressed their opinions about the decision in the blog comments (Fig. 1). People often disagreed with each other, which led to further dispute within some comment threads.

Although such arguments rarely come to an agreement in the end, they leave a rich dataset for us to identify the Stack Overflow users' concerns about multi-lingual sites. We crawl all comments of the blog articles that announced the launch of a new non-English Stack Overflow site. From the four announcement blogs, we crawl in total 348 comments including single comments, and comment threads i.e., comments following another comment.

61 comments are in non-English languages. We use Google Translate service [15] to translate such non-English comments into English. Then, one author reads one comment at a time and classifies it into one of the four categories: *against*, *support*, *neutral*, and *unrelated. unrelated* means that a comment is not related to multi-lingual dispute such as "*So, what other languages are available?*", "*A nice post, very informative*", or comments as response to points in other comments like "*@Rickster In Soviet Russia, websites access people.*" . According to our observation, some comments are very long with detailed reasons in favor of or against building multi-lingual Stack Overflow, while some of them are rather short like *"Great!!!!", "Pretty bad idea"*. So for each against/support/neutral comment, we further separate it as either *simple* or *verbose* by counting its word number. If a comment has less than 4 words, we count it as *simple*, otherwise as *verbose*.

Table 1 shows the analysis results. Overall, about 41.4% of comments support the launch of the non-English site, while 24.7% comments are against such multi-lingual sites. Note that 16%

[3] https://stackoverflow.com/help/whats-beta

Table 1. Sentiment analysis of comments on the launch of multi-lingual Stack Overflow sites, and the number of simple comments is in the square bracket.

Sites	**#Against**	**#Support**	**#Neutral**	**#Unrelated**
Japanese	23 [1]	30	11	33
Portuguese	44 [2]	66 [10]	5	45
Russian	14	37 [13]	4	12
Spanish	5	11	7	1
Total	86 (24.7%)	144 (41.4%)	27 (7.8%)	91 (26.1%)

Table 2. Example supporting and against comments on multi-lingual Stack Overflow sites

Supporting:
Someone who is truly struggling with english will have a better experience on a site in their native language.
Congrats, Make stackoverflow to all
Well done! Other languages cannot be ignored, so this is a step ahead.
Why can't those haters of Russian SO just continue to use English SO? You don't like Russian SO, you don't use it, as simple as that.
I wanted to write a huge comment, but English is not my first language (nor strongest), I'll refrain! Localization is awesome!
I think the philosophy of SE is that if a site can support itself and there is demand for it, then it has every right to exist.
Against:
It risks fragmenting our knowledgebase over multiple languages and segregating our users. (Concern #1 & Concern #3)
I actually think the multiple versions could bring the demise of StackOverflow (Concern #1)
This will lead to many bad habits ... English is critical to any developer / architect software. (Concern #2)
People should stop being lazy and learn English instead, as it's the de facto language for programming anyway. (Concern #2)
More importantly, the separation will duplicate work and will certainly hide good answer from each other. (Concern #3)
Just learn English. I am against any localized version of stackoverflow because it leads to fragmentation of knowledge. (Concern #3)

supporting comments are simple. In contrast, most of the against comments (96.5%) are verbose. That is, many users just simply say that they like the multi-lingual sites but do not give detailed reasons, while most users who are against the multi-lingual sites provide the detailed reasons. As such, although the number of the verbose supporting comments (83) is actually less than that of the verbose against comments (121), the difference is not so much. We also note that a large portion of comments (26.1%) are unrelated to the discussion of multi-lingual sites. This is because many comments are responses to some points in the prior comments which is unrelated to our analysis like "*OMG! They killed Hashcode!*", or just opinions about the blog article like "*A nice post, very informative*".

Next, the author further examines the verbose supporting and against comments and uses open coding method to code the major concerns in these comments. Three major concerns emerge from our open coding process: *community split (concern #1)*, *knowledge needs and interests in English and other languages (concern #2)*, and *knowledge fragmentation and duplication (concern #3).* Fig. 1 and Table 2 show some examples of the supporting and against comments on these three concerns. As these examples shown, users who are against the multi-lingual sites are worried about: if the multi-lingual sites would cause the community split across different sites, whether there are enough needs and interests in computer programming knowledge in other languages, and knowledge fragmentation and duplication across different sites would become a serious issue to deal with. Users who support the multi-lingual sites argue that: the multi-lingual sites could involve more non-English-speaking users in the community, there are different knowledge needs and even unique knowledge interests for non-English-speaking users, and if knowledge fragmentation and duplication do exist, they could be addressed by cross-site linking and/or machine translation.

It seems that all arguments are reasonable and understandable, but everyone lacks solid evidences to support their arguments. In the remainder of the paper, we use English Stack Overflow (ESO) and Russian Stack Overflow (RSO) as the subject Q&A sites, and explore the evidences from the three perspectives (i.e., users, tags, and cross-site links) to answer the three major concerns that users expressed in their comments on multi-lingual Stack Overflow sites.

4 KNOWLEDGE NEEDS AND INTERESTS? EVIDENCES OF TAG UNIQUENESS AND TAG RANK CORRELATION

In English Stack Overflow and Russian Stack Overflow, each question can have up to five tags. A tag is a word or phrase that describes the topic of the question. As a whole, tags used in a Stack Overflow can represent the overall technology landscape that the posts of the site are about [20]. In this section, we compare the tag usage in ESO and RSO to understand whether and how knowledge needs and interests differ across multi-lingual sites.

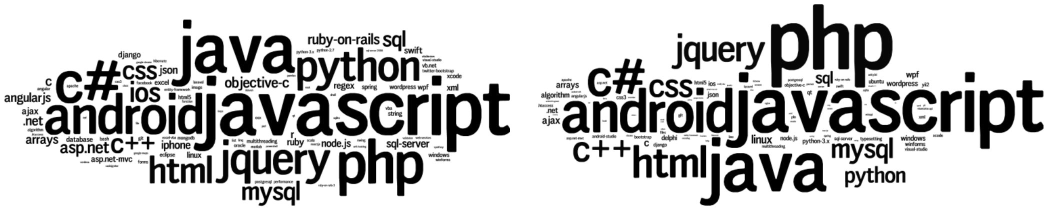
4.1 Tag Rank Correlation

There are 50,812 tags in English Stack Overflow and 3957 tags in Russian Stack Overflow. Among these 3957 tags in Russian Stack Overflow, 821 tags contain Russian characters and the rest 3136 tags are in English. We use Google Translate service [15] to translate 821 tags with Russian characters into English.

The majority (2560, 64.7%) of the 3,957 tags in RSO also exist in ESO. This shows that users of the two sites share much common interest. Fig. 2 shows the word cloud of the top-100 most frequently-used tags in ESO and RSO, respectively. The font size reflects a tag's relative frequency in its own site, i.e., the larger a tag is, the more frequent it is used in the site. Among these top 100 frequent tags, 56 of them appear in both sites such as *javascript*, *java*, *C#*, *android*, etc. Furthermore, 8 of the top 10 most frequently used tags of RSO are also in the top 10 most frequently used tags of ESO. To further understand the interest correlation between two sites, we perform a Spearman Rank Correlation analysis [38] to 2560 common tags that appear in both sites. The Spearman rank correlation coefficient between them is 0.69 ($p-value < 0.005$). It shows that an implicit increasing monotonic trend between the tag frequency ranks in English Stack Overflow and Russian Stack Overflow.

We then further analyze tags whose relative frequency in one site is rather different from the other site. We first normalize a tag's frequency to a value in (0, 1] through dividing each tag's post number in a site by the total post number of the site. Fig. 3 shows the comparative ratio of common tags' relative frequency in Russian Stack Overflow to their relative frequency in English Stack Overflow. The higher (lower) this ratio is, the relatively more (less) frequent a common tag is in Russian Stack Overflow than in English Stack Overflow. We consider a common tag's relative frequency in the two sites are close if the comparative ratio is within $0.5 < r \leq 2$, i.e., the relative frequencies differ less than twice.

According to our results, the relative frequencies of 1172 (45.8%) of 2560 common tags in the two sites are close. This shows that the level of interests in the technologies these tags represent are not significantly different. However, the relative frequencies of the rest 1388 common tags differ more than twice. There are 262 (10.2%) common tags whose relative frequencies differ more than 10 times, such as *xnet*, *symfony3*, *phpixie*, *vds*, *amazon-web-services*, *android-emulator*, *itext*, *cocoapods*, etc. This suggests that the level of interests in these technologies is very different in the two sites. This interest difference could be caused by many reasons, such as different coding environment, certain technique or tool's popularity in different regions. For example, the comparative ratio of the tag *amazon-web-services* is around 0.0063. This could be due to the much higher popularity of Amazon web services in English speaking countries than Russia.



(a) English Stack Overflow

(b) Russian Stack Overflow

Fig. 2. Top-100 most frequently used tags in ESO and RSO

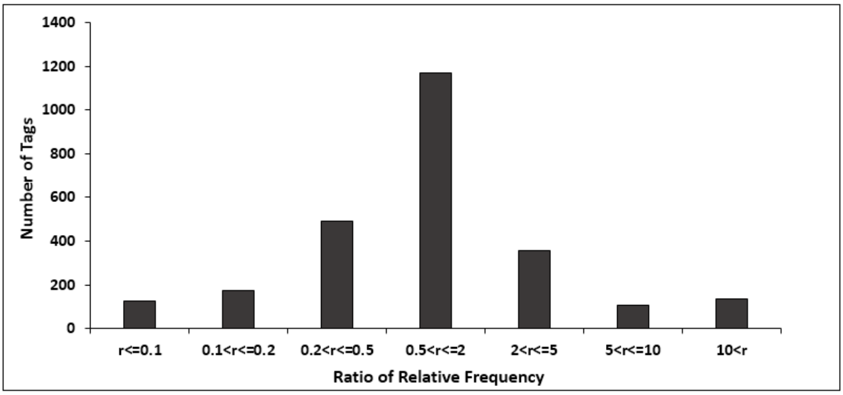
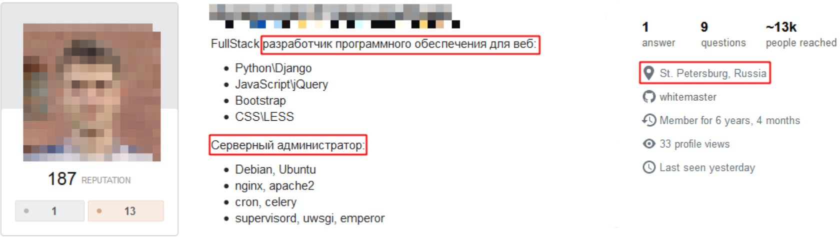


Fig. 3. Comparative ratios (r) distribution of the common tags' relative frequency in RSO to its relative frequency in ESO

4.2 Tag Uniqueness

1397 of the 3957 tags in Russian Stack Overflow do not exist in English Stack Overflow. That is, about 35.3% technical terms are of interest by only RSO users such as **Yandx** (the largest Russian search engine in Russia like Google), **VKontakte** (biggest social network company in Russia like Facebook), **Denwer** (a popular local server in Russia), **bitrix** (a popular CMS in Russia written in php), etc. As these techniques or services that these tags represent are Russian-specific, they are rarely used outside of Russia. Therefore, ESO does not have questions about these techniques or services. However, Russian Stack Overflow provides a perfect place to discuss such Russian-specific techniques or services. These site-specific tags have been used to annotate 61,662 posts in RSO which account for 15.4% of all 400,185 posts in RSO.

Summary: According to the tag usage analysis, the majority of tags in RSO are also in ESO. However, the level of usage frequency of the common tags often differs between the two sites, which reveals different level of interest in the corresponding technology. Furthermore, RSO also has its unique tags whose related posts make an non-negligible proportion of all posts in RSO. This shows that non-English speaking users do have unique knowledge needs and interests other than the knowledge in English.



[redacted]

FullStack разработчик программного обеспечения для веб:

- Python\Django
- JavaScript\jQuery
- Bootstrap
- CSS\LESS

Серверный администратор:

- Debian, Ubuntu
- nginx, apache2
- cron, celery
- supervisord, uwsgi, emperor

187 REPUTATION

● 1 ● 13

1	9	~13k
answer	questions	people reached

St. Petersburg, Russia

whitemaster

Member for 6 years, 4 months

33 profile views

Last seen yesterday

Fig. 4. One user portfolio in English Stack Overflow. This user may speak Russian.

5 COMMUNITY SPLIT? EVIDENCES OF USER REGISTRATION, POSTING BEHAVIOR AND REPUTATION SCORE

Users, especially the core users are crucial to the success of Q&A sites [28], as they play an important role in asking and answering questions, and in curating post content and quality. Will the multi-lingual sites result in the split of user community by the languages? What kinds of users contribute the most to the multi-lingual sites? Will multi-lingual sites affect the posting behavior of users who have accounts in both English and non-English sites? We answer these questions by analyzing user registration, user reputation score, and user posting behavior.

5.1 User Registration

By December 8th, 2017, there are 8,123,937 users in the English Stack Overflow (ESO) site and 98,125 users in the Russian Stack Overflow (RSO) site. As both sites belong to Stack Exchange networks [16], we can identify the intersection of users who own both ESO and RSO accounts by their Stack Exchange ID. We find that 51,415 users own both accounts at the same time, accounting for 52.4% of the total number of the users in Russian Stack Overflow.

We further check the account registration time of these 51,415 users. Among these 51,415 users, 32,288 (62.8%) sign up their ESO account first and then sign up RSO account. Those users account for 32.9% of the 98,125 registered users in RSO which shows the importance of the ESO site to the development of the RSO site. However, compared with the total 8.1 million users in ESO, 32,288 users account for only 0.4%. Therefore, even all of these 32,288 users stop using ESO and contribute only to RSO, it will not result in a significant loss for ESO. Let alone many uses who own both accounts do not stop contributing to ESO after signing up RSO (see Section 5.2). On the other hand, 19,125 (37.2%) users sign up their RSO account first, and then sign up ESO account. It means that to some extent, the launch of RSO may help to attract users to ESO. This could further reduce the potential loss of users in the English site.

Next, we would like to investigate the impact of building RSO site to the specific group of users in the original ESO site who can speak Russian. To that end, we try to identify users in ESO who are from Russia or can speak Russian. In Stack Overflow, users can create their own profiles by filling in their names, location, AboutMe (self introduction), GitHub link, etc. Fig. 4 shows an example user profile who can speak Russian. We first check if the user's location is in Russia. If the location contains any Russian letters, we regard the place as in the Russia. If the location is written in English, we determine whether the location is in Russia by checking the Russian cities list on Encyclopaedia Britannica[4]. We assume that the users whose location is in Russia can speak Russian. Then for other users who do not have location information, we further check if their AboutMe contains any Russian letters, and if so we also regard such users as ones who can speak Russian.

[4] https://www.britannica.com/topic/list-of-cities-in-Russia-2040243

Table 3. Statics of user posting behavior in Russian Stack Overflow

#Post by User	#Post	% of All Posts	#User	% of All Users
>=2	372,046	92.9%	24,908	25.4%
>=4	345,543	86.3%	13,526	13.8%
>=6	328,398	82.1%	9,644	9.8%
>=8	314,833	78.7%	7,537	7.7%
>=10	303,604	75.9%	6,201	6.3%

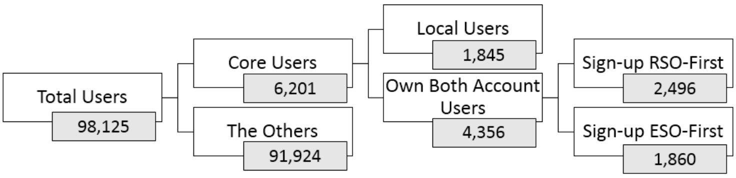


Fig. 5. User composition in Russian Stack Overflow

Using the above two heuristics, we find in total 34,705 users in ESO who may speak Russian. Among these users, 10,699 (30.8%) have registered RSO. Within these 10,699 users, 6,537 (61.1%) users sign up ESO first, and the rest 4,162 (38.9%) users sign up RSO first. While for the other users in ESO (8,089,232), only 40,716 (0.5%) have RSO accounts. The results show that, compared with users who do not speak Russian, RSO does attract many potentially Russian-speaking users in ESO. However, this does not mean a unidirectional loss of users from ESO to RSO, because RSO also helps attract users to ESO.

5.2 User Posting Behavior

In a Q&A website, the main user interaction is to ask and answer questions. We regard both questions and answers as posts. In RSO, there are 400,185 posts as of December 8th, 2017. We count the number of posts created by different users. The results can be seen in Table 3. We can see that 6,201 (6.3%) users who have created more than 10 posts have contributed to 75.9% posts in RSO. So, we regard these 6,201 users as the core users in RSO.

As shown in Fig. 5, 4,356 of the 6,201 core users own both ESO and RSO accounts. 1,860 users sign up ESO first, but they created 139,275 posts in RSO which account for 34.8% of all posts in RSO. 1,845 local users and 2,496 users who sign up RSO first created 260,910 posts in RSO which accounts for 65.2% of all posts. We can see that a large proportion of core users comes from ESO. These core users make significant contributions to Russian Stack Overflow, but the local core users and the core users who sign up RSO first make even relatively more contributions.

If a core user in RSO who signs up ESO first and then RSO contributes comparatively less to ESO after signing up RSO, we say that RSO negatively influence this user's contribution to ESO. To identify such users, we define the following formula:

$$f = \frac{PostCountonESOAfterSignupRSO/DaysAfterSignupRSO}{PostCountonESOBeforeSignupRSO/DaysBeforeSignupRSO} \tag{1}$$

For a user who signss up ESO first and then RSO, this equation compares the user's average post number per day in ESO before and after the user signs up RSO. The smaller value indicates that the user makes comparatively less contribution to ESO after signing up RSO. Let's take one user as an

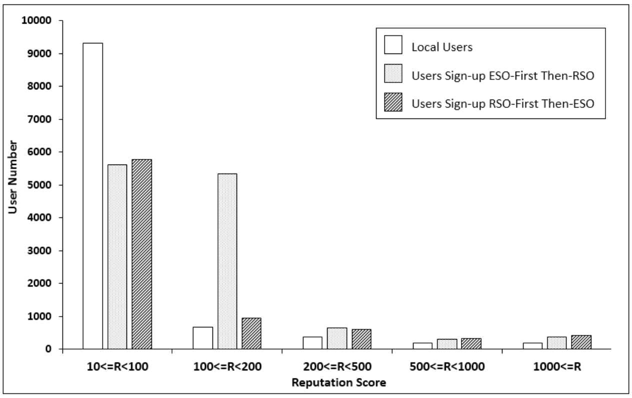


Fig. 6. The reputation distribution of three-group users in RSO

example whose id is *Vlad* in English Stack Overflow[5] and *VladD* in the Russian one[6]. He joined the ESO first on February 2010, and then joined the RSO on November 2012. Before joining the RSO, he has answered 669 questions in ESO in about 2 years, but only answered 189 questions in ESO for 5 years after his join into RSO. Instead, he has answered 3,798 questions in RSO and earned 173,213 reputation score so far, while his current reputation score in ESO is only 28,781. This user can be regarded as a user migrated from ESO to RSO.

We consider $0.8 \leq f \leq 1.2$ as a reasonable range for relative stable level of contribution to ESO after the user signs up RSO. That is, the user's posting behavior in ESO does not change significantly after signing up RSO if $0.8 \leq f \leq 1.2$. Among 1860 core users in RSO who sign up ESO first, 120 (6.5%) are with the f value in range of 0.8 to 1.2, i.e., their contribution is relatively stable in ESO after signing up RSO. 1033 (55.5%) users are even more active in ESO with $f > 1.2$. Only 707 (38%) users become less active in ESO. It indicates that although many users from ESO contribute a lot to RSO, they do not stop making consistent contributions in ESO. It further demonstrates that the launch of RSO will not cause severe community split of users between ESO and RSO.

5.3 User Reputation Score

In Stack Overflow, users can earn reputation score if their questions, answers or comments get upvotes. They can also get bonus if they commit to editing others' posts, creating tag wiki which help to maintain the quality of the site. So, apart from the number of posts, reputation score is another good indicator of how much contribution a user has made to the community and how much the community trusts his work. According to the privilege policy of Stack Overflow[7], users who earn more than 10 reputation score can remove the new user restriction. As new users do not contribute much to the site, we only consider 31,029 users with more than 10 reputation score in RSO in this analysis. These 31,029 users in RSO belong to three groups: 10,702 (34.5%) local users who only register RSO, 12,292 (39.6%) users who sign up ESO first and then sign up RSO, and 8,035 (25.9%) users who sign up RSO first and then sign up ESO.

Fig. 6 shows the distribution of reputation score of these three groups of users in five different intervals. We can see that the distributions of reputation score in the first two low-reputation intervals are different from those in the three high-reputation intervals. This phenomenon is due to

[5] https://stackoverflow.com/users/276994
[6] https://ru.stackoverflow.com/users/10105
[7] https://stackoverflow.com/help/privileges

the Association Bonus Policy of Stack Exchange [1]. According to this policy, a user will earn 100 bonus reputation score in RSO when the user signs up a RSO account, if the user has already owns an ESO account with 200+ reputation score. Therefore, compared with local users and bi-account users who sign up RSO first, there is a relatively less proportion of bi-account users who sign up ESO first in the reputation interval $10 \leq R < 100$. In contrast, a large proportion of bi-account users who sign up ESO first have reputation score within $100 \leq R \leq 200$. This results in that the number of bi-account users who sign up ESO first is more than the sum of local users and bi-account users who sign up RSO first in the interval $100 \leq R \leq 200$.

But in the other three intervals (i.e., $R \geq 200$), we can see that the number of local users plus the number of users who sign up RSO first then ESO is more than the number of users who sign up ESO first. Furthermore, in the three intervals $R \geq 200$, there are roughly the same number of users in the group of users sign up ESO first and in the group of users sign up RSO first. However, considering that there are 12,292 users in total who sign up ESO first and 8,035 users who sign up RSO first, the later group has much higher percentage of high-reputation users than the former group.

This analysis suggests that the community of RSO has its own core contributors. However, its community is not completely independent of that of ESO, because a large number of users from ESO (especially high-reputation users from ESO) makes similar contribution to RSO, as those local users and those who sign up RSO first.

Summary: In term of post contribution and reputation score, Russian Stack Overflow benefits a lot from English Stack Overflow as many users in RSO come from ESO, especially those who can speak Russian. But there is no significant community split between the two sites, as those from-ESO users only account for a very small proportion of the user base in ESO and the majority of them still make stable contributions to ESO even after they sign up RSO. In addition, RSO also helps to attract new users to ESO.

6 KNOWLEDGE FRAGMENTATION AND DUPLICATION? EVIDENCES OF CROSS-SITE LINKING

ESO and RSO are both programming-related websites and our tag usage analysis shows that users of the two sites share much common interests. Therefore, it is reasonable that some new questions asked in the Russian Stack Overflow may have been already answered or have highly related questions in English Stack Overflow. This could create knowledge fragmentation and duplication across multi-lingual sites. But the supporters of multi-lingual sites argue that it is likely that users on non-English Stack Overflow may cross-reference posts to English Stack Overflow, which would solve the knowledge fragmentation and duplication issue. Below is a user's comment about the concern of knowledge fragmentation and duplication and the suggestion that cross-site linking could solve the issue.

> *"This is the world wide web and it is built on a thing called hypertext. Multiple languages? Cool, that's the "world" part. But don't forget about hypertext. Please, make sure that questions and answers can be linked easily across the different language sites. If someone asks a question in Japanese that is essentially a duplicate of an existing English question, linking the two questions will allow Japanese speakers (who might understand *some* English, or at least be able to read code snippets) to get immediate value from past answers, without waiting for new answers... "*
>
> — one comment on the announcement of Japanese Stack Overflow [9]

Table 4. Links statics

Source	#Total links	#Links to ESO	#Links to RSO
ESO Posts	13,636,973	1,581,888 (11.6%)	88
ESO Comments	5,959,110	1,835,405 (30.8%)	164
RSO Posts	131,964	5,674 (4.3%)	5,014 (3.8%)
RSO Comments	55,471	4,548 (8.2%)	5,658 (10.2%)

In this section, we investigate existing cross-site links made by the users on Russian Stack Overflow and potential cross-site links that the users are not aware of. This will provide evidences whether knowledge fragmentation and duplication is a serious issue for multi-lingual sites, whether manual cross-site linking is sufficient to battle this issue, and what kind of tools could better support cross-site linking.

6.1 Explicit Cross-Site Links by Users

In Stack Overflow, users can add hyperlinks in their posts (both questions and answers) and comments to reference to other resources. We collect all hyperlinks in ESO and RSO and analyse the link direction within and across the two sites. The statistics of hyperlinks in the two sites are summarised in Table 4. We can see that a large proportion of hyperlinks in ESO posts and comments references to its own content, while only a few hyperlinks in ESO posts and comments reference to the RSO content. This is unsurprising because English speaking users would not have much interest in Russian content. In contrast, comparable proportion of hyperlinks in RSO reference to its own content and the ESO content, respectively. This phenomenon indicates that there could be certain level of knowledge fragmentation and duplication between the two sites and RSO users attempt to address the issue by cross-site linking.

To further confirm this phenomenon, we manually examine the relationships between a RSO post and an ESO post that is referenced by a hyperlink in the RSO post. We collect in total 5674 hyperlinks in RSO that reference to ESO. For each hyperlink, we collect its residing RSO post and the ESO post being referenced. As we are most familiar with *Python* related questions, we randomly select 50 such pairs of RSO-ESO posts tagged with *Python*. Examining these 50 pairs of RSO-ESO posts reveal three kinds of relationships between a RSO post and its hyperlinked ESO post: duplicate questions (14 pairs), related and useful for problem solving (30 pairs), and related but not useful for problem solving (6 pairs).

First, hyperlinks in 14 (28%) RSO posts reference to some duplicate questions in ESO. For example, the Russian question "как запустить скрипт на pypy?"(postid=722948, “how to run a script on pypy”) has a hyperlink to the English question “How to use PyPy on Windows?” (postid=9893317). Both the Russian question and the English question are about how to run a script of pypy on Windows.

Second, hyperlinks in 30 (60%) RSO posts reference to some related but not duplicate English posts in ESO that are useful for solving the problem in the Russian questions. For example, the Russian question "Обновление Label из цикла в tkinter"(postid=581331, “updating the label of the cycle in tkinter”) has a hyperlink to an ESO post “Update Tkinter Label from variable” (postid=2603169). The Russian question asks how to use the *update* function on Tkinter Label in a loop structure. The referenced ESO question does not solve this specific problem, but it discusses how to use the *update* function, which is relevant.

Third, sometimes RSO users add hyperlinks to the ESO posts in their questions to tell that they have examined some related ESO posts, but these ESO posts cannot solve their problems. This kind of RSO posts accounts for 6 (12%) of the 50 examined samples. Fo example, one Russian question

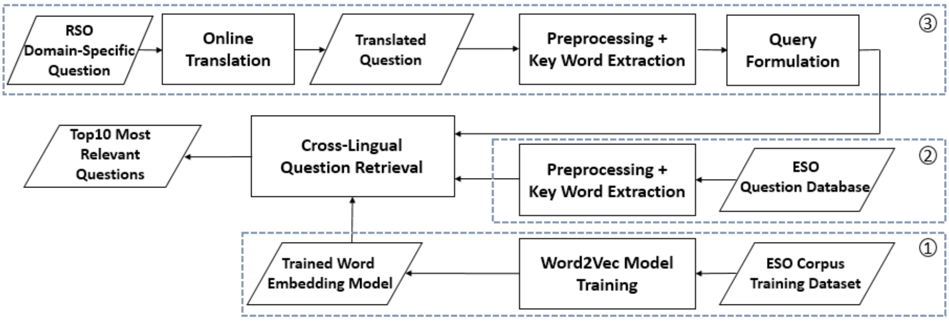

RSO Domain-Specific Question
Online Translation
Translated Question
Preprocessing + Key Word Extraction
Query Formulation
③
Top10 Most Relevant Questions
Cross-Lingual Question Retrieval
Preprocessing + Key Word Extraction
ESO Question Database
②
Trained Word Embedding Model
Word2Vec Model Training
ESO Corpus Training Dataset
①

Fig. 7. Overall framework for cross-site retrieval

with the title "Почему не работает унаследованная форма?" (postid=318618 'Why does not the inherited form work?") has a hyperlink to an ESO question "why does not work in the form of validation?" (postid=23534615). The Russian question states that the answers to this ESO question do not work for his problem.

To sum up, the first category of cross-site links reveals knowledge duplication, and the second category may have the risk of knowledge fragmentation. However, the third category has nothing to do with knowledge fragmentation and duplication.

6.2 Implicit Cross-site Links that Users Are Not Aware of

In Table 4, we find that the ratio of hyperlinks in RSO posts and comments that reference to RSO and ESO posts is relatively lower than that of hyperlinks in ESO post and comments that reference to ESO and RSO posts. This makes us hypothesize that there could be cross-site links between RSO and ESO posts that users are not aware of, and thus have not been explicitly hyperlinked. To test this hypothesis, we carry out another empirical study. As ESO has millions of questions and answers, it is impossible to manually find if a RSO question has some related posts in ESO. Therefore, we develop a cross-site retrieval method to locate potentially related questions in ESO, given a RSO question. Then, we analyze the relationships between the given RSO question and the retrieved ESO questions to understand implicit cross-site links.

6.2.1 Method for cross-site question retrieval. Fig. 7 displays the flow chart of our cross-site retrieval method. The method can be divided into three main parts. In the first part, we pre-train a word embedding model for retrieving related question for a given query. Word embedding [30] is a method to convert the words into dense vectors based on the assumption that words appear in similar context tend to have similar meanings. We train a word embedding model using all posts collected from ESO including both question titles and descriptions. We adopt the Principle Component Analysis [25] to reduce the high-dimensional word embedding into 2-dimension for manual observation. Examples of some word embeddings are visualized in Fig. 8. We can see that similar words are close to each other in the vector space, such as "js" and "javascript", "image" and "picture", "bug" and "defect". In addition, word embedding model is also robust to word morphological transformation (like plural "images", "bugs") and misspellings (e.g., "javacript" and "jascript" for "javascript"). Compared with traditional keyword matching, word embedding model helps to bridge the lexical gap between different words which share similar meanings [21].

In the second part, we prepare the posts from English Stack Overflow as a question database. In this study, we build a database of 779,370 ESO questions tagged with *Python*. We pre-process each question by tokenzing the sentence and removing the stop words in both question title and

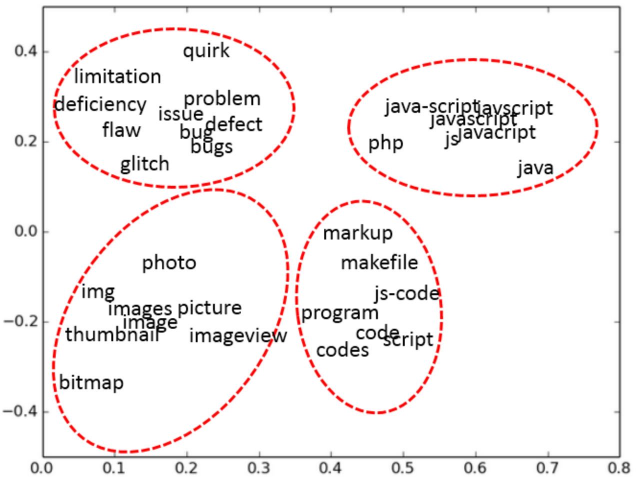


Fig. 8. Visualization of word embeddings by principle component analysis [25]

description. All words in the question title are considered as keywords. We extract keywords from the question description because the question description contains too many technical details that could negatively influence the retrieval accuracy. We use two different popular keywords extraction algorithms (TF-IDF and TextRank) to extract keywords from the question description. TF-IDF (term frequency-inverse document frequency) [31] is a numerical metric corresponding to the importance of a word in a corpus. The TF-IDF metric of a word is proportional to the frequency of the word appearing in one post and inverse proportional to the number of posts in which the word appears. TextRank [29] is a popular unsupervised approach for automatic keywords extraction. It performs random walk on a graph of word co-occurrences to determine the most important words in a post. These two algorithms yield two different but complementary sets of keywords. To reduce the bias of using a single algorithm, we choose the union of the two keyword sets to build our final set of keywords. If a keyword appears in both title and description, we consider it as a title keyword.

In the third part, we formulate a RSO question as a query to retrieve related ESO questions in the ESO question database. We first translate the RSO question's title and description into English with Google Translate [15]. We use the same pre-processing and keyword extraction steps to build a query from the given RSO question. As question title represents more compact and meaningful information than question description, we assign double weight to title keywords in the query than description keywords. Given the query of a Russian question (R) and an English question (E) in the ESO question database, We define the relevance between a query keyword (R_i) and a question keyword (E_j) as the cosine similarity between the word embeddings of R_i and E_j, i.e., $Rel(R_i, E_j) = CosineSim(v_{R_i}, E_j)$. For each query keyword R_i, we compute its relevance with each keyword of the English Question and find the one with the largest similarity:

$$Rel(R_i, E) = \max_{j \in E} Rel(R_i, E_j) \times R_i.Weight \tag{2}$$

Note that this relevance is weighted by the query keyword's weight $R_i.Weight$. We then define the relevance between the query of a Russian question (R) and an English question (E) in the ESO question database as:

$$Rel(R, E) = \sum_{i \in R} Rel(R_i, E) \tag{3}$$

Table 5. Examples from implicit cross-site links that users are not aware of

Translated Question	IDE for python (#464)	the slow implement of the code, a coin toss (#432934)	books and educational resources for python (#420125)
Result #1	[duplicated] Searching for a Python lightweight IDE (or text-editor) (#6011678)	[related] python code for coin toss issues (#6486877)	[related] Resource/book suggestion to effectively writing software for python/c++ beginner (#9694065)
Result #2	[related] What's a good IDE for Python on Mac OS X? (#893162)	[related] python coin toss (#14882530)	[related] good resource for learning advanced or obscure python concepts (#14756145)
Result #3	[related] IDE for Python: test a script (#6023377)	[related] Simulate multipe coin toss streak (#15761259)	[related] what are good online resources to learn using Pyhton to get and post data in a webpage? (#41241481)

We compute the relevance of the query and all questions in the ESO question database, and return the top 10 questions with the highest relevance as the most relevant English questions to the given Russian question.

6.2.2 Manual validation of cross-site reference. We randomly sampled 50 RSO questions tagged with *Python* that do not have hyperlinks to ESO. We use these RSO questions as query questions and obtain the top-10 related ESO questions for each RSO question using our cross-site retrieval method. We then examine if the retrieved ESO questions are duplicate or otherwise related to the query RSO questions.

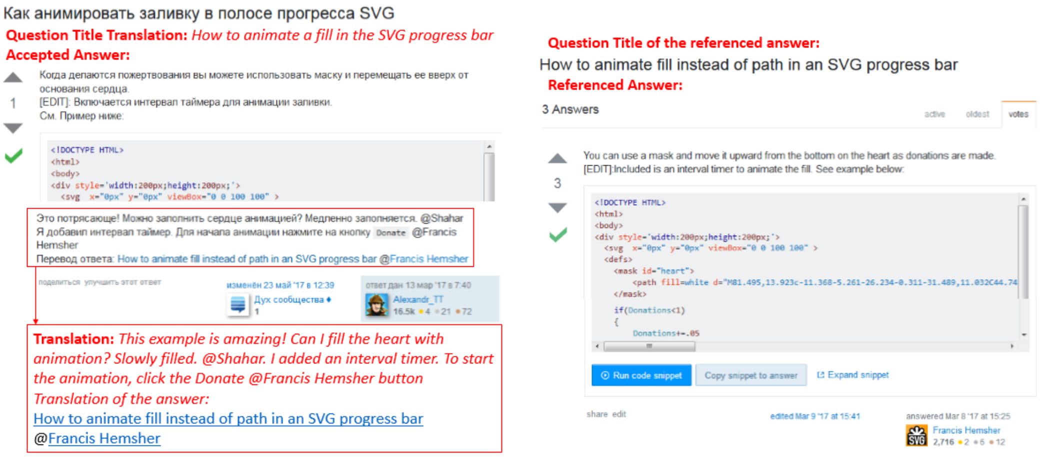
We find that for 6 (12%) RSO questions, we can locate duplicate questions in the top-10 ESO questions returned by our cross-site retrieval method. For the other 20 (40%) Russian questions, we can locate at least one related-but-not-duplicate ESO questions in the top-10 list. Table 5 presents three RSO questions (title translated into English) in our study. We only show the top 3 most relevant ESO questions due to the space limit. Among these three examples, our cross-site retrieval method finds a duplicate ESO question (postid=9694065, "Searching for a Python lightweight IDE (or text-editor)") to the RSO question (postid=464, "IDE for Python") and finds some related ESO questions to the RSO questions (postid=432934, "The slow implementation of the code a coin toss, and postid=420125, "Books and educational resources for Python").

For the rest 24 (48%) RSO questions, Our cross-site retrieval method does not find any related questions in the current ESO database. Note that it may be either due to the limitation of our method, or the fact that there are no duplicate or related questions in ESO. More future works are needed to confirm the detailed reasons.

Summary: There are certain level of knowledge fragmentation and duplication across ESO and RSO. RSO users recognize some duplicate and related ESO posts and reference them in RSO posts and comments. But there are many duplicate and related ESO questions that RSO users are not aware of and thus do not reference in RSO posts, which calls for tool support to better tackle knowledge fragmentation and duplication issue.

7 DISCUSSION

Our study not only provides the evidences to validate the users' concerns about multi-lingual Q&A sites, but also reveals some needs for better supporting multi-lingual Q&A sites. Fig. 9 shows an example in which a user asked a question in RSO but received no answer timely. And then, the user managed to find a duplicate question in English Stack Overflow (Fig. 9(b)) which can perfectly answer his question in Russian Stack Overflow. This user went back to Russian Stack Overflow and answered his own question (Fig. 9(a)) in which he referenced to the ESO question and translated the accepted answer of that ESO question in his post. Such cases further demonstrate the needs of cross-site retrieval and translation which are discussed in this section.



Как анимировать заливку в полосе прогресса SVG

Question Title Translation: *How to animate a fill in the SVG progress bar*

Accepted Answer:

1

Когда делаются пожертвования вы можете использовать маску и перемещать ее вверх от основания сердца.
[EDIT]: Включается интервал таймера для анимации заливки.
См. Пример ниже:

```
<!DOCTYPE HTML>
<html>
<body>
<div style='width:200px;height:200px;'>
  <svg  x="0px" y="0px" viewBox="0 0 100 100" >
```

Это потрясающе! Можно заполнить сердце анимацией? Медленно заполняется. @Shahar
Я добавил интервал таймер. Для начала анимации нажмите на кнопку `Donate` @Francis Hemsher
Перевод ответа: How to animate fill instead of path in an SVG progress bar @Francis Hemsher

изменён 23 май '17 в 12:39
Дух сообщества ♦
1

ответ дан 13 мар '17 в 7:40
Alexandr_TT
16.5k ● 4 ● 21 ● 72

Translation: *This example is amazing! Can I fill the heart with animation? Slowly filled. @Shahar. I added an interval timer. To start the animation, click the Donate @Francis Hemsher button*
Translation of the answer:
How to animate fill instead of path in an SVG progress bar
@Francis Hemsher

(a) #638741: A RSO post that references to and translates a duplicate ESO question

Question Title of the referenced answer:

How to animate fill instead of path in an SVG progress bar

Referenced Answer:

3 Answers

3

You can use a mask and move it upward from the bottom on the heart as donations are made.
[EDIT]:Included is an interval timer to animate the fill. See example below:

```
<!DOCTYPE HTML>
<html>
<body>
<div style='width:200px;height:200px;'>
  <svg  x="0px" y="0px" viewBox="0 0 100 100" >
  <defs>
    <mask id="heart">
        <path fill=white d="M81.495,13.923c-11.368-5.261-26.234-0.311-31.489,11.032C44.74
    </mask>

    if(Donations<1)
    {
        Donations+=.05
```

edited Mar 9 '17 at 15:41

answered Mar 8 '17 at 15:25
Francis Hemsher
2,716 ● 2 ● 6 ● 12

(b) #42672537: The referenced ESO question with the accepted answer

Fig. 9. An example of manual cross-site linking and translation

7.1 Cross-site Retrieval

Since its launch in 2007, English Stack Overflow has accumulated a large pool of computer programming knowledge, including 16 million questions and 24 million answers. In contrast, the non-English Stack Overflow sites have been launched in recent years. As the non-English Stack Overflow users and the English Stack Overflow users do share some common interests, it is very likely that some questions asked on a new non-English Stack Overflow site may have been already asked or even answered on the English Stack Overflow. If we can recommend relevant, high-quality questions and answers (e.g., accepted or highly-voted by the community) on English Stack Overflow to the users of the new non-English Stack Overflow site, it will benefit the new site.

There are two types of users in a multi-lingual Stack Overflow site: those speak only the native language, and those can at least read some English. For the users who can read some English, the recommendation will lead them to the relevant posts on English Stack Overflow which directly satisfy their information needs. For the users who speak only the native language, even though they cannot understand the English descriptions, they can at least understand some code snippets or APIs mentioned in the English posts, which are the real “meat” of computer programming knowledge. Therefore, these non-English speaking users can still benefit from the recommendation of relevant English posts.

However, it is not an easy task to retrieve duplicate or related posts across multi-lingual Stack Overflow sites for two reasons. First, there is a natural language gap between non-English Stack Overflow and English Stack Overflow. Second, there are too many posts in Stack Overflow to retrieve related posts, and these post themselves often have lexical gap, i.e., they discuss the same computer programming knowledge but use very different words [21]. So an effective cross-site retrieval tool is needed.

Our cross-site retrieval method in the last section has the potential to be adopted for this purpose, as it can bridge both the language gap and lexical gap to certain extent. To validate this potential, we use the 50 Russian *Python* questions that have manual hyperlinks to ESO posts in Section 6.1 as query questions to retrieve relevant ESO questions from the question database of 779,370 English

Python questions. For 29 (58%) of these 50 Russian *Python* questions, the retrieved top-10 ESO questions by our method contain the ESO posts that RSO users manually hyperlink in the 50 Russian *Python* questions. For the rest 21 Russian questions, we further analyze the human-hyperlinked ESO posts that our method fails to retrieve in the top-10 ESO questions. For example, the RSO question "Порядок удаления элементов списка в python"[8] ("the removal of elements of the list in python") has a hyperlink to an ESO question "Calling del on a list"[9]. In this example, "del" is a function name for deleting. However, our method currently cannot reliably capture such domain-specific information and sentence-level semantics.

Therefore, although promising, our cross-site retrieval method still need to be enhanced, for example, replacing the current word-level semantic modelling with more advanced sentence-level semantic methods [22, 26] for better performance. Furthermore, our current analysis is anecdotal. More formal evaluation should also be carried out to validate the effectiveness and efficiency of a cross-site retrieval method.

7.2 Cross-site Translation

Although cross-site retrieval across multi-lingual Q&A sites is useful for users, it is just the first step, not the ultimate solution. To assist non-English speaking users in fully exploiting the existing knowledge in English Stack Overflow, some English questions and answers especially those highly-viewed, highly-voted ones should be translated into other languages. It was also mentioned in some comments on the announcement of multi-lingual Stack Overflow sites, such as "*I think a better idea would be to let the community translate each question and answer–think of it as a kind of editing capability.*", "*...If someone finds a really good question on a site, with a good answer, they can translate it to other sites, post it on those sites (cite the original of course)...*".

Breaking the language barrier would make the entire world more knowledgeable, and even users who cannot speak English at all can fully benefit from the knowledge on English Stack Overflow. As translating could cost less effort than drafting an answer as good as the well-examined one on English Stack Overflow, it may save much effort of core users in the multi-lingual site who serve as the backbone for the relatively new site.

Recently, machine translation has made big progress by incorporating deep learning methods [32, 35]. To assist high-quality translation, we can adopt the power of advanced machine translation to first obtain an overall translation, and then let human users revise the translation. However, the general machine translation system may not work well for Q&A discussions, and it needs to be improved in two aspects. First, current machine translation models focus on sentence-level translation without contextual information. However, to translate an answer well in a Q&A discussion thread, the model not only needs to consider the current sentence in the answer, but also needs to take the corresponding question information into consideration. Second, Stack Overflow is a domain-specific site about computer programming, while the current machine translation system is trained for general purpose. Many domain-specific words will be out of vocabulary or have different meanings from daily life such as "bug", "port", etc. So the general machine translation model must be customised by incorporating the domain-specific knowledge.

In addition to technical aspect of cross-site translation, we also need to design some mechanisms to motivate users to contribute to the translation activity. We can set up a reward rule that once users translate one answer from English Stack Overflow to non-English site, they can earn some award points. Once other users vote their translated answer, they will earn more points and reputation

[8] https://ru.stackoverflow.com/questions/55464/
[9] https://stackoverflow.com/questions/8205102/

score. A translator badge can also be created and awarded to users whose translation score reaches to a certain level.

8 CONCLUSION AND FUTURE WORK

In this paper, we carry out an empirical study to explore the evidences in favor of or against multi-lingual Q&A sites. Our study examines not only the user perspective, including user registration, posting behavior and reputation score, but also the knowledge perspective, including tag uniqueness and tag correlation between multi-lingual sites, and explicit cross-site links created by users and implicit cross-site links that users are not aware of. Our study results help to resolve the arguments about the risk of community split and the knowledge needs and interests in non-English language. We also confirm the validity of the concern about knowledge fragmentation and duplication, and provide a promising cross-site retrieval method for mitigating this issue.

Our evidence-based study method provides a guideline to study the pros and cons of multi-lingual Q&A sites, and our findings shed the light on the development of multi-lingual Q&A sites. In the future, based on our findings, we will develop tools to help bridge the gap across multi-lingual Q&A sites, such as recommending duplicate or related posts on English Stack Overflow to Russian Stack Overflow questions, and developing domain-specific deep learning based machine translation model for Q&A discussions.

REFERENCES

[1] 2009. Cross-Site Account Associations. https://stackoverflow.blog/2009/07/08/cross-site-account-associations/. (2009). Accessed: 2018-03-25.

[2] 2009. Do posts have to be in English on Stack Exchange? https://meta.stackexchange.com/questions/13676/do-posts-have-to-be-in-english-on-stack-exchange/13684. (2009). Accessed: 2018-03-25.

[3] 2009. Non-English Question Policy. http://blog.stackoverflow.com/2009/07/non-english-question-policy/. (2009). Accessed: 2018-03-25.

[4] 2013. Do the benefits of having SO in multiple languages outweigh the risks involved? https://meta.stackexchange.com/questions/194959/do-the-benefits-of-having-so-in-multiple-languages-outweigh-the-risks-involved. (2013). Accessed: 2018-03-25.

[5] 2013. Should Spanish SO (and all the similar variants) be closed? https://meta.stackexchange.com/questions/187805/should-spanish-so-and-all-the-similar-variants-be-closed. (2013). Accessed: 2018-03-25.

[6] 2014. Announcing Stack Overflow in Portuguese. http://blog.stackoverflow.com/2014/01/ola-mundo-announcing-stack-overflow-in-portuguese/. (2014). Accessed: 2017-05-01.

[7] 2014. Can't We All be Reasonable and Speak English? https://stackoverflow.blog/2014/02/13/cant-we-all-be-reasonable-and-speak-english/. (2014). Accessed: 2018-03-25.

[8] 2014. Fluent in Spanish? We're hiring a Community Manager for a Spanish Stack Overflow. https://meta.stackoverflow.com/questions/271998/fluent-in-spanish-were-hiring-a-community-manager-for-a-spanish-stack-overflow. (2014). Accessed: 2018-03-25.

[9] 2014. Stack Overflow in Japanese. http://blog.stackoverflow.com/2014/12/stack-overflow-in-japanese/. (2014). Accessed: 2017-05-01.

[10] 2014. When will other sites follow Stack Overflow Portuguese. https://meta.stackexchange.com/questions/236890/when-will-other-sites-follow-stack-overflow-portuguese/236892. (2014). Accessed: 2018-03-25.

[11] 2015. Stack Overflow in Portugese: now with less beta! http://blog.stackoverflow.com/2015/06/stack-overflow-in-portuguese-now-with-less-beta/. (2015). Accessed: 2017-05-01.

[12] 2015. Welcome, Nicolas Chabanovsky and Stack Overflow in Russian! http://blog.stackoverflow.com/2015/06/welcome-nicolas-chabanovsky-and-stack-overflow-in-russian/. (2015). Accessed: 2017-05-01.

[13] 2016. Quora in Spanish is Now Open to the World. https://blog.quora.com/Quora-in-Spanish-is-Now-Open-to-the-World. (2016). Accessed: 2018-03-25.

[14] 2017. Stack Overflow en Español has Graduated! https://stackoverflow.blog/2017/05/20/stack-overflow-en-espanol-graduated/. (2017). Accessed: 2018-03-25.

[15] 2018. Google Translate. https://cloud.google.com/translate. (2018). Accessed: 2018-03-01.

[16] 2018. Stack Exchange Sites. https://stackexchange.com/sites. (2018). Accessed: 2018-03-25.

[17] 2018. Wikipedia:Translate us. https://en.wikipedia.org/wiki/Wikipedia:Translate_us. (2018). Accessed: 2018-04-15.

[18] Eytan Adar, Michael Skinner, and Daniel S Weld. 2009. Information arbitrage across multi-lingual Wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 94–103.

[19] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1075–1084.

[20] Chunyang Chen, Zhenchang Xing, and Lei Han. 2016. Techland: Assisting technology landscape inquiries with insights from stack overflow. In *Software Maintenance and Evolution (ICSME), 2016 IEEE International Conference on*. IEEE, 356–366.

[21] Guibin Chen, Chunyang Chen, Zhenchang Xing, and Bowen Xu. 2016. Learning a dual-language vector space for domain-specific cross-lingual question retrieval. In *Automated Software Engineering (ASE), 2016 31st IEEE/ACM International Conference on*. IEEE, 744–755.

[22] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. (2017).

[23] Elena Filatova. 2009. Multilingual wikipedia, summarization, and information trustworthiness. In *SIGIR workshop on information access in a multilingual world*, Vol. 3.

[24] Brent Hecht and Darren Gergle. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 291–300.

[25] Ian T Jolliffe. 1986. Principal component analysis and factor analysis. In *Principal component analysis*. Springer, 115–128.

[26] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. 3294–3302.

[27] A Kumaran, Naren Datha, B Ashok, K Saravanan, Anil Ande, Ashwani Sharma, Sridhar Vedantham, Vidya Natampally, Vikram Dendi, and Sandor Maurice. 2010. WikiBABEL: A system for multilingual Wikipedia Content. In *American Machine Translation Association (AMTA) Workshop*.

[28] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2857–2866.

[29] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[31] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21.

[32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[33] Yu-Chun Wang, Chun-Kai Wu, and Richard Tzong-Han Tsai. 2014. Cross-language and Cross-encyclopedia Article Linking Using Mixed-language Topic Model and Hypernym Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 586–591.

[34] Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 459–468.

[35] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[36] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 975–985.

[37] Bowen Xu, Zhenchang Xing, Xin Xia, David Lo, Qingye Wang, and Shanping Li. 2016. Domain-specific cross-language relevant question retrieval. In *Proceedings of the 13th International Conference on Mining Software Repositories*. ACM, 413–424.

[38] Jerrold H Zar. 1998. Spearman rank correlation. *Encyclopedia of Biostatistics* (1998).

Received February 2007; revised March 2009; accepted June 2009