# Towards Bridging the Gap between Multilingual Q&A Communities with Deep Learning

**—An Empirical Study of Russian Stack Overflow**

**Qiacheng Wang**

A subthesis submitted in partial fulfillment of the degree of

Bachelor of Science (Honours) at

The Department of Computer Science

Australian National University

October 2017

we strive, we dream, we inspire, always towards something greater. all the odds we defy, the risks we take, the challenges we endure only make us stronger.

# Acknowledgements

Prima facea, I wish to express my sicere thanks to My supervisor Dr. Zhenchang Xing, for teaching me all the necessary knowledge for the research.

I am also grateful to Dr. Chunyang Chen, who is doing research in NTU. I am extremly thankful and indebted to him for sharing expertise, and sincere and valuable guidance and encouragement extended to me.

I also thank my parents for the unceasin encouragement, support and attention.

Last but not the least, thanks for that young man who never give up.

# Abstract

The language policy of Q&A websites has been under the hot debate for a long time and there is still no convincing conclusion. Many of these communities are holding an English-Only Policy but operating some foreign language variants at the same time. Recognizing whether it is meaningful to develop multilingual flagship sites is a fundamental research question, and an advanced method towards bridging the gap between multilingual communities with deep learning can be very attractive to solve this problem. Traditional research on the Q&A websites focus on the same language or the accuracy of language processing method. However, these methods are not robust to deal with the lexical gap between different languages(e.g. English and Russian). In this paper, we formulate the problem of predicting similarity of question pairs in different languages and solve the problem by deep learning concept. Except for the neural language model we adopt, we mining large amounts of duplicate question pairs from Stack Exchange data dump of September 2017[1] to train deep semantic similarity model (DSSM). The results confirm us the probability and value of developing the cross-lingual deep learning methods, and also show us that this deep-learning approach significantly outperforms traditional methods in the cross-language fields.

# Contents

Chapter 1

# An Introduction to My Thesis

Q&A websites have been witnessing the unprecedented developemnt in the recent years. Although English is as close a global langauge, many Q&A websites debates on the "English Only" language policy have never come to an end. According to the Wikipedia [2], English is the third place in the rank of first language speakers and second place for the first and the second language speakers. It means that there are nearly three-quarters of people in this world speak other languages. Some website like Quora insisted the English only policy, but the Spanish version of Quora was open on 19th Oct 2016. Another example is Stack Overflow, who has launched many other languages variants for users who are native speakers of other languages including Russian, Spanish and Japanese.

**Table 1.1**: User amount and post amount of all Stack Overflow sites(By 1st Sep. 2017)

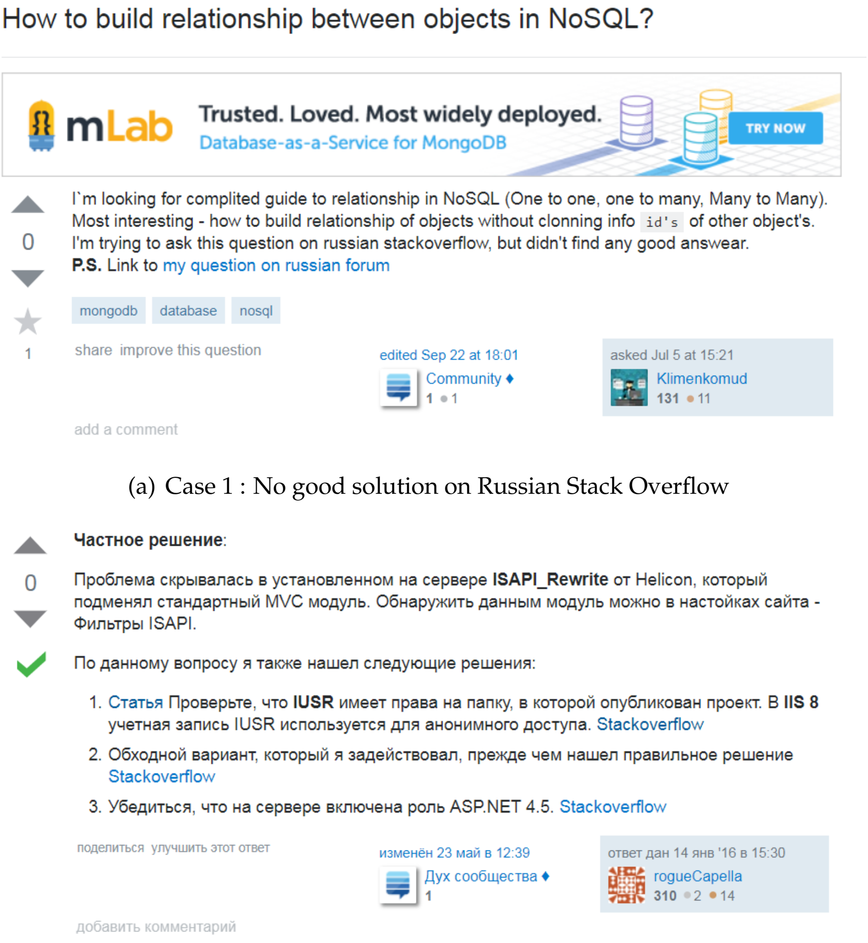
| Site | #User | #Post | #Comment |
|---|---:|---:|---:|
| Main site | 7,617,191 | 37,215,528 | 60,098,125 |
| Russian | 86,826 | 366,485 | 678,556 |
| Portuguese | 61,280 | 180,877 | 322,241 |
| Spanish | 49,033 | 76,278 | 123,662 |
| Japanese | 13,457 | 28,589 | 27,297 |

Stack Overflow is a platform with millions of users and hundreds of millions of posts and answers. Table 1.1 presents the statistics of users, posts and comment of Stack Overflow main site and other language sites. As we can see, there exists a huge differences in user amount and post amount between the Stack Overflow main site and all other language variants. Therefore, it is meaningful if a clear answer can be found that whether Q&A websites and communities should keep developing their multilingual variants, and more contributions can be made to bridging the lingual gap for all users who are not a native English speaker. It is highly believed that humans are accustomed to a more familiar environment [24], especially language environment. Although some people learn a second language, the native language is absolutely more attractive. However, the lingual gap between different language communities is hard to bridge.

The active level of a community is determined by many factors, and the most important one is the amount of users, that is more people more interactions. Comparing the statistics in Tabel 1.1, it is clear that the number of user in Stack Overflow main site is 88 times that of Russian Stack Overflow, which means the gap of resources is enormous. There always exists a doubt that whether developers need to build multilingual communities. In this empirical study, our main research object is Russian Stack Overflow, which supports more users than other variants. Compared with the Stack Overflow main site, we will analyze how the activity of users from different language environment can influence communities and explore the importance of the multilingual Q&A communities.

There are several interesting cases on the sites of Stack Overflow, which are presented in Figure 1.1. The sub-figure a) shows a very common phenomenon that users from the other language subsite of Stack Overflow like the Russian variant often cannot find satisfying and high-quality answers on other language Stack Overflow sites, which need to be defined as **sub-sites**, for some reasons, while the main site can alway assists them with the solutions. The sub-figure b) is a answer posted on Russian Stack Overflow, which means, in English, there are some previous solutions on the Stack Overflow English main site for this specific Russian question.

Imagine that a Russian speaker who cannot speak English fail to find a good answer to his or her question. Meanwhile, the reply to his post is quite slow because the number of users on Russian Stackoverflow is small. For this problem, our cross-lingual deep learning model can efficiently deal with the lingual gap problem and rec-



How to build relationship between objects in NoSQL?

mLab
Trusted. Loved. Most widely deployed.
Database-as-a-Service for MongoDB
TRY NOW
0
1

I`m looking for complited guide to relationship in NoSQL (One to one, one to many, Many to Many). Most interesting - how to build relationship of objects without clonning info `id's` of other object's. I'm trying to ask this question on russian stackoverflow, but didn't find any good answear.
**P.S.** Link to my question on russian forum

(a) Case 1 : No good solution on Russian Stack Overflow

**Частное решение**:

Проблема скрывалась в установленном на сервере **ISAPI_Rewrite** от Helicon, который подменял стандартный MVC модуль. Обнаружить данным модуль можно в настойках сайта - Фильтры ISAPI.

По данному вопросу я также нашел следующие решения:

1. Статья Проверьте, что **IUSR** имеет права на папку, в которой опубликован проект. В **IIS 8** учетная запись IUSR используется для анонимного доступа. Stackoverflow
2. Обходной вариант, который я задействовал, прежде чем нашел правильное решение Stackoverflow
3. Убедиться, что на сервере включена роль ASP.NET 4.5. Stackoverflow

(b) Case 2 : Good solutions exist on Stack Overflow main site

**Figure 1.1**: Examples of cross-lingual information needs

ommend useful resources across different languages. A key benefit of this approach is the non-English speakers can efficiently utilize the huge knowledge base from the dominant English community. The main contributions of this paper are listed below:

- A deep analysis of the value of multilingual communities and explore the user activity features in technical communities.
- Mining duplicate post pairs that can be used for deep learning model for Stack Overflow duplicate question detection.
- Build deep learning model for Stack Overflow duplicate question detection with

high accuracy.

* Build deep learning model for semantic similarity (DSSM) for comparing cross-lingual post similarity.

* Qualitative analysis of our experimental results and the benefits of our approach.

Chapter 2

# EmpiricalStudy

We carry out the empirical study between Stack Overflow and its corresponding Russian site in two perspectives, i.e., the users between them and the content within each site.

## 2.1 Users

In this part, we mainly analysis the whether it is necessary to build multilingual Stack Overflow by deeply compare the user activity, the knowledge base status and the areas of focus between Stack Overflow main site and Russian Stack Overflow site. The combination of all research results can lead to a solid conclusion for the meaning of multilingual Q&A community development.

It has been highly believed that the user is highly important for the analysis of a community because the whole knowledge base is the achievement of all users during a long time of accumulation. Users in the intersection of Stack Overflow main site user set and Russian Stack Overflow user set, who own both accounts should be considered as the main research object. According to the Stack Overflow policy, a user has a different user id on the specific site, while a unique account id, which is the way to link these two users on the entire Stack Exchange Network [4]. According to Table 1.1, By September 1st, 2017, there are 7,617,191 users on Stack Overflow main site and 86,826 users on Russian Stack Overflow site. We can easily get that there are 45,764 users owns both accounts at the same time, which is about 52.7% of the total number of the user amount on Russian Stack Overflow.

### 2.1.1 Creation Date

Focusing on this intersection user base, the User Migration is calculated by comparing their account creation dates of each site. In this set, 16,155 users sign up their Russian Stack Overflow account first and then sign up for a Stack Overflow account, while the other 29,607 users are in a diametrical way. On one hand, the user base is segregated by the new sub-site and the number of users leaving the main site to their native language sub-site is growing with the time goes by. On the other hand, if we define a user migrate from one site to another by his or her account's creation date, when 35 out of 100 users in the users who currently own both accounts migrate to Stack Overflow from Russian Stack Overflow, the other 65 of those 100 people migrate to Russian Stack Overflow from Stack Overflow, which means the Russian version offers a number of users as feedback to Stack Overflow main site.

To illustrate the user migration in details, Figure 2.1 presents the statics of the users who sign up a Russian Stack Overflow account from Jan 1st, 2014 to Sep. 1st, 2017. As we can see in the graph, the new users who migrated from Stack Overflow main site occupy a high proportion, and also the Russian site attracts and feedbacks a growing number of users to the main site with the time going by. In other words, the main site and the multilingual version both have a positive effect on the other one, which means both of them can assist the development of the other. This phenomenon means the development of multilingual version is helpful for the development of the main site.

### 2.1.2 Reputation

Reputation is a mark [5], which represents how much contribution a user has made, how much the community trusts you and how much the peer users think about your work. Of course, the more reputation a user earned, the higher level of activity the user is. So it is very easy to measure the high-level user activity by calculating this variable. Figure 2.3 shows the statics of reputation of users in Russian Stack Overflow

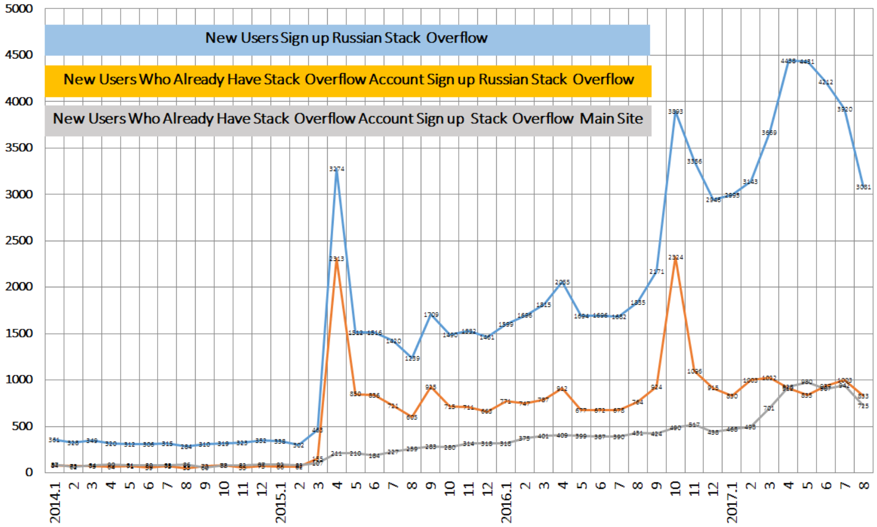


**Figure 2.1**: New User Sign Up for Russian Stack Overflow

by segregating all 86,826 users into 3 groups.

Local users who do not own an account of the main site are 41,062, and the other two labels have been introduced are 29,609 and 16,155. The compare of user amount of these three groups shows in Figure 2.2.

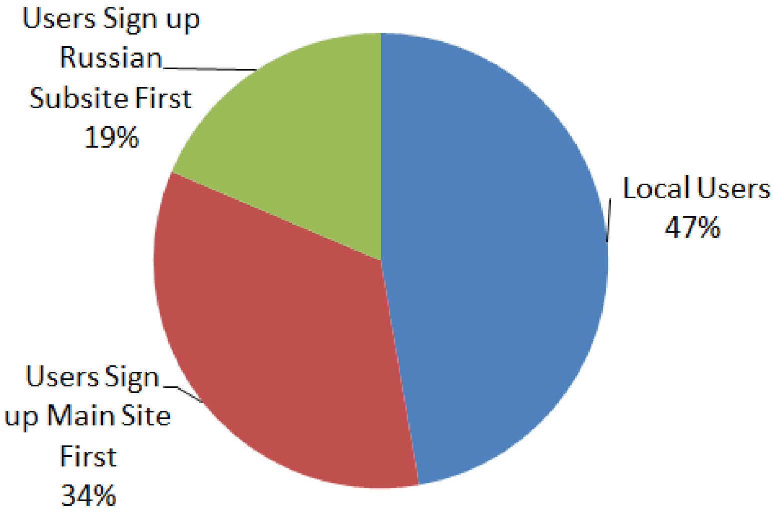


**Figure 2.2**: User Amount Compare for Russian Stack Overflow Users

Considering the Reputation Bonus Policy that a starting plus 100 reputation bonus to users who already have a 200+ account of any site belongs to Stack Exchange [5], the 29,609 users who migrated from the main site reasonably have a large number of the reputation level of 100 to 200. However, despite the local user, the remains are the user who owns both account of the main site and Russian version. Under the circumstance that the number of migrants is almost 2 times of that of the 16,155 users who sign up Russian version account first, the later contributes somehow an equal number of active users.

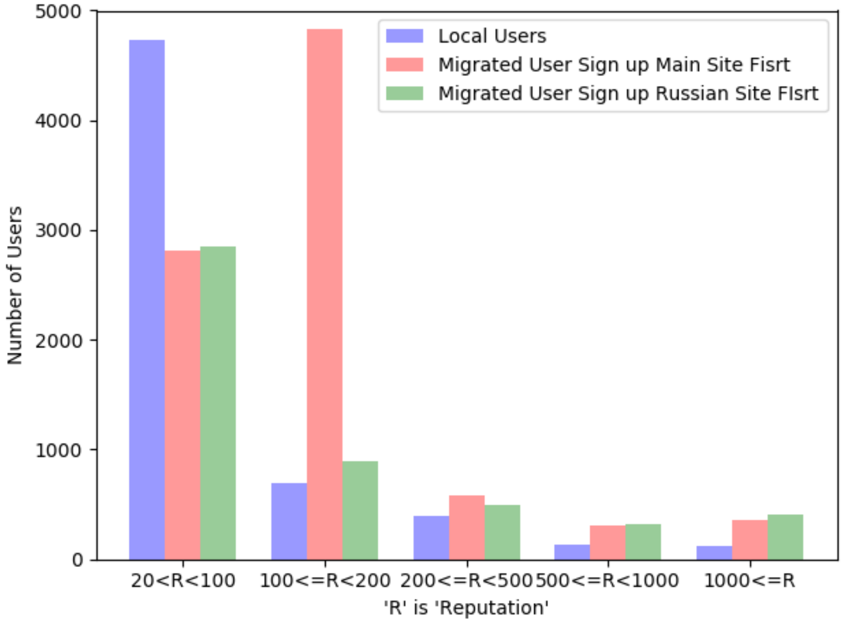


**Figure 2.3**: Statics of Reputation

To some degree, the reputation reveals that multilingual communities are not dominated by a large number of external users even the later has an overwhelming advantage on the user amount, which means that multilingual communities are relatively independent in the respect of user contribution in the community development process.

## 2.2 Post

Although every user has the right to post a question and add comments, the composition of the knowledge base ought to obey the Pareto Law, which means 20% of the users make 80% of the contribution. In this section, we analyse the behaviours of those 20% users.

Imagine that a post is a single unit, and millions of posts consist of the knowledge base, and each post includes maybe a lot of answers and comments in the tree structure. From Table 1 we know that there are totally 366,485 posts for the Russian Stack Overflow, and using the unique account id to calculate the amount of post for each user on Russian Stack Overflow, we have the results that present in Table 2.1.

**Table 2.1**: Post Statics

| User Post count | Post Sum | Post Sum/ Post Amount | User Sum | User Sum/User Amount |
|---|---|---|---|---|
| P_C>=2 | 340,903 | 93.4% | 33,669 | 26.1% |
| P_C>=4 | 316,768 | 86.8% | 12,305 | 14.2% |
| P_C>=6 | 301,225 | 82.5% | 8,784 | 10.1% |
| P_C>=8 | 289,100 | 79.2% | 6,901 | 7.9% |
| P_C>=10 | 278,688 | 76.4% | 5,665 | 6.5% |

From the table, we can easily know the static distribution. As mentioned above, with the number of a single user post sum growing, the user proportion decrease. Notice that 5,665 people from the 86,826 Russian Stack Overflow users who have 10 or more posts contribute 76.4% of the post amount. Here we need to set a research object called Ćore User.́ We need to assume these 5,665 people are the core users who have the highest level of contribution. Again, in these 5,665 core users, 1,685 people are local users while 3,980 people own both main site account and Russian sub-site account at the same time. In these 3,980 users in the intersection, 1,676 people are migrants from the Stack Overflow main site. Next, we will compare the level of the contribution between the migrants and the others in the Core User group. The user composition is showed in Figure 2.4.

For the 1,676 of Core Users who sign up the Stack Overflow account first, which also known as the migrant from the main site, they post 94,478 posts on Russian sub-site that count 25.9% of the post amount on Russian sub-site. For the 2,304 of Core Users who sign up the Russian sub-site account first, which also known as the migrant

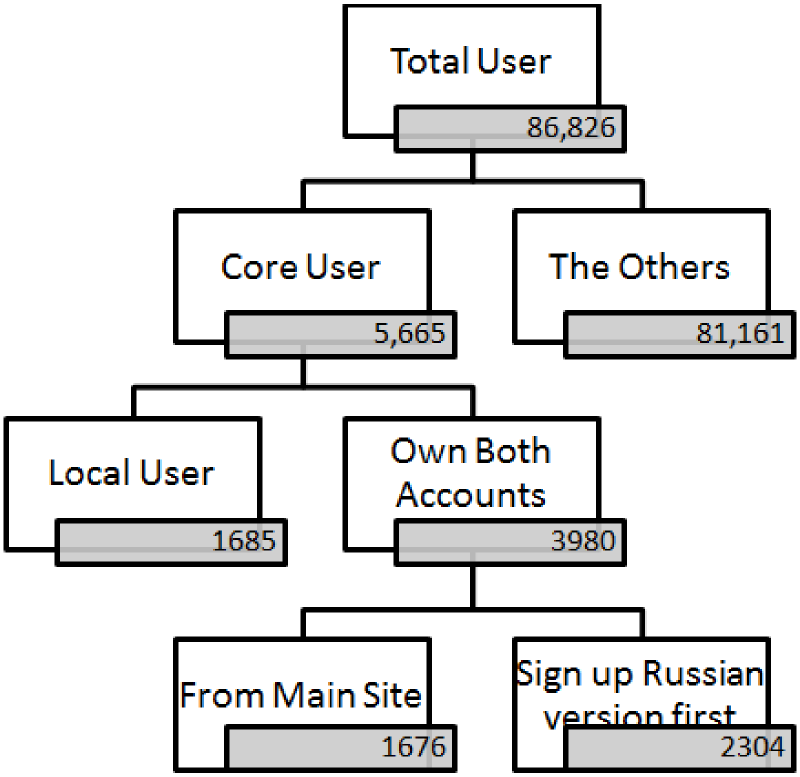


**Figure 2.4**: Statics of components on Russian Stack Overflow

from the main site, they post 125,891 posts on Russian sub-site that count 34.5% of the post amount on Russian sub-site. According to the average level, the external users do not dominate the post area.

Moreover, for the 1,676 of Core Users who sign up the Stack Overflow account first, according to their frequency of post activity, we can also conclude that they are still spending more time and concentrate on the Stack Overflow main site. The formula of judging a migrant is Formula (2.1). If the f value is greater than 1, the user transferred his or her main active area to the new site.

$$f = \frac{Russian_{post}SumAfterMigration/TimeLengthAfterMigration}{Mainsite_{post}CountBeforeMigration/TimeLengthBeforeMigration} \tag{2.1}$$

Assume that a reasonable range for the unchanged active level that is 0.8 to 1.2 for f, the results in Table 2.2 reveal that more than half of those active users from the main site tend to be more active in a multilingual subsite. Although they are from another site, they make a lot of contribution and seem to be like using the new sub-site regularly. The conclusion of the research on post aspect is the multilingual community has a group of users as backbone so that the community is not dominated by the users from the Stack Overflow main site, but the migrants from the main site are also willing to contribute in the new community.

**Table 2.2:** Statics for the 1,676 Users who post more than 10 and from Stack Overflow main site

| User Post count | User Amount | f < 0.8 | 0.8<=f<=1.2 | f > 1.2 |
|---|---|---|---|---|
| P_C>=10 | 1,676 | 637 | 106 | 933 |
| P_C>=20 | 400 | 187 | 35 | 178 |
| P_C>=50 | 207 | 123 | 15 | 69 |
| P_C>=100 | 132 | 93 | 10 | 29 |

## 2.3 Content

### 2.3.1 Tag

A tag is a word or phrase that describes the topic of the question. The tag is a means of connecting experts with questions, and they will be able to answer the question by sorting questions into specific and well-designed categories. Research on the tags is able to reveal a number of most hot fields. There are 50,000 tags on Stack Overflow main site and 3,779 tags including 688 Russian character tags and 3091 non-Russian character tags on Russian sub-site. Ranking the top 10 frequent tags in several sets to show what are the most popular fields in both sites.

According to the result shown in Table 2.3, it is clear that the popular areas of the two sites are similar. Considering the difference in the size of different communities, the hit counts of the same tag on different site are totally not on the same level. While

**Table 2.3**: Tag Ranking Statics

| Main site tags | f | Russian site tags | f | Russian site Russian tags | Translation | f |
|---|---|---|---|---|---|---|
| javascript | 0.339‰ | php | 0.651‰ | массивы | arrays | 0.066‰ |
| java | 0.304‰ | javascript | 0.583‰ | база-данных | database | 0.062‰ |
| c# | 0.263‰ | java | 0.531‰ | регулярные-выражения | regular expressions | 0.056‰ |
| php | 0.259‰ | android | 0.443‰ | алгоритм | algorithm | 0.056‰ |
| android | 0.238‰ | c# | 0.389‰ | веб-программирование | web programming | 0.042‰ |
| jquery | 0.201‰ | html | 0.321‰ | вёрстка | coding | 0.035‰ |
| python | 0.187‰ | jquery | 0.282‰ | многопоточность | multithreading | 0.030‰ |
| html | 0.159‰ | c++ | 0.275‰ | ооп | oop | 0.028‰ |
| c++ | 0.123‰ | css | 0.247‰ | строки | lines | 0.026‰ |
| ios | 0.122‰ | mysql | 0.208‰ | файлы | files | 0.026‰ |

for Russian Stack Overflow site, after translating the top 10 popular tags, it is clear that most of the Russian character tags are also common on the Stack Overflow main site. This fact indicates that for a series of website, all the tag sets of different language subsites are pretty similar, although there must be some distinction resulted by culture divergence.

### 2.3.2 Links

Link is a good tool to recommend and refer the existing works to other users. Not only links can refer the knowledge base in the same site, but also across different sites. This part of work is based on the statics of links on Stack Overflow main site and Russian Stack Overflow. People use links in their posts and comments, and connect their evidence and reference by links. The link amount and the direction reveal the similarity and difference between these two sites.

Basically, in our study, links can be divided into two sets. One is the group of posts and comments refer the existing posts on Stack Overflow main site, and the other is those who refer the existing posts on Russian Stack Overflow. Some statistics are presented in the Table 2.4.

As we can see, the number of links that Russian users used to refer posts on Stack

**Table 2.4**: Links Statics

| | Link Amount | Links referring posts of Stack Overflow | Links referring posts of Russian Stack Overflow |
|---|---|---|---|
| Stack Overflow Posts | 13,187,126 | 1,512,118 | 80 |
| Stack Overflow Comments | 5,737,401 | 1,759,284 | 150 |
| Russian Stack Overflow Posts | 120,860 | 5,425 | 4,324 |
| Russian Stack Overflow Comments | 50,178 | 4,160 | 5,054 |

Overflow main site is smaller than the number of links that they used to refer posts on the Russian sub-site. Apparently, the results of the links show that Russian Stack Overflow sub-site is relatively independent because it owns a considerable proportion of links referring the existing post in its own knowledge base. The Russian sub-site does not completely depend on Stack Overflow main site, as the Russian users are not always referring the existing posts on main site. However, in some ways the Russian Stack Overflow sub-site has a non-negligible demand of knowledge reference from the main site. So the content of these two sites are not totally same and not totally different, especially the Russian subsite has its necessity of existence.

## 2.4 Conclusion

The above empirical comparison between two sites shows that Russian Stack Overflow owns many unique features which are not included by the main Stack Overflow, no matter its users or Russian-specific content. Such difference demonstrate that the existence of the multi-lingual Stack Overflow is meaningful and useful to some specific users. Furthermore, the multi-lingual deviation does not significantly undermine the knowledge accumulation or user participation of the main site.

Despite the uniqueness of the Russian site, there are a lot of interaction between two sites, e.g., many main-site posts are quoted in the Russian one. In addition, as a relatively new site, questions asked in the Russian site may have already been posted in the main Stack Overflow site. To assist site interaction and avoid potentially duplicated questions, a tool is needed to help retrieve related English posts in the main Stack Overflow when Russian users are asking or answering questions in the Russian site.

Chapter 3

# approach

This section is the main part of this paper. It illustrates how we formulate the problem we want to solve and what technical skill we use.

## 3.1 Dataset

The main data source of our research is from the Stack Exchange data dump of 1st Sep. 2017 [1]. As all the duplicate pairs have been manually marked by experts on Stack Overflow, we can collect 0.31 million duplicate question pairs. In addition, we can generate as many non-duplicate question pairs as we want.

We randomly divide the duplicate question pairs into three groups that 80% for training, 10% for validation and 10% for testing. The 80% subset is mix up with 0.15 million non-duplicate pairs for training the DSSM model, the first 10% subset is used for validation, and the last 10% subset is used for the comparison process between our approach and the baselines.

## 3.2 Overall Architecture

We consider the problem as a binary prediction one. Considering an input unit including a well-written title and a list of well-chosen tags, our approach uses the tags to filtering candidates and uses the title to predict and rank those candidates. The main architecture of the DSSM is illustrated in Figure 3.1.

First of all, what we need to do in Step 1 is mining the duplicate question pairs

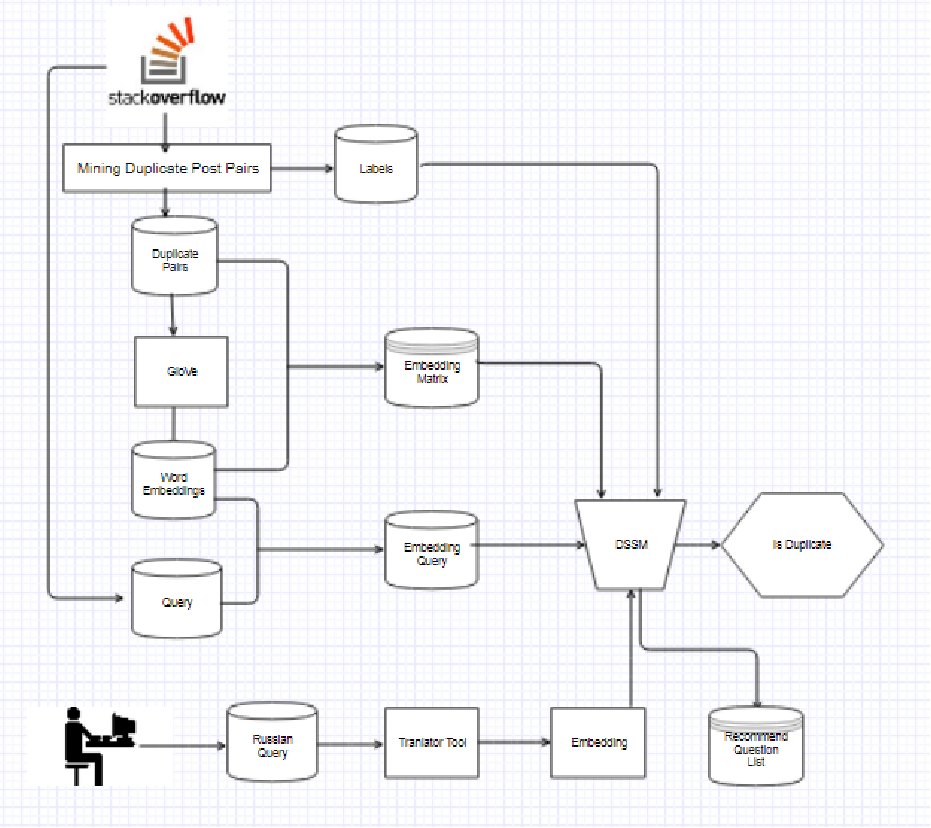


**Figure 3.1**: Overview of the Main architecture

from the raw data of Stack Overflow [1], using the duplicate mark in the Post History file can generate about 310,000 pairs of duplicate questions. We extract the titles and tags of these posts. Step 2, use GloVe in the process of word embedding and represent all question sentences in vectors after tokenizing them. Step 3, assembling the embedding matrix, which is used for calculating the Weight. Step 4, building the deep learning model. Concatenating the duplicate question pair as input and get a binary variable *is_duplicate* as output. In the dense layer. Step 5, training the model and record the best performance from the callbacks. Finally, we get the model that can be used in the recommendation system.

## 3.3 Learning Word Representation

Mapping all the words into a dense low-dimensional vector space, words that appear in the similar sentence have a short distance in the embedding space, and each dimension represents a latent semantic [10,11]. The assumption is that words appeared in the similar context have similar meanings, so their mapped vectors in the embedding vector space should be relatively close. Word embedding only needs a large amount of text to learn as it is an unsupervised method.

In this paper, we use the Keras tokenizer to tokenize all the questions in the training data set that we mined from the raw data. We also adopt GloVe, which is a popular word embedding tool. For the embedding space, each dimension represents a latent semantic feature of a word. We get a dictionary of all the word that appeared in the text, and finally, there are 0.21 million dimensions in the embedding space.

For example, given an English duplicate question pairs, the word embedding system will convert it into embedding space by mapping each word with the dictionary of words after preprocessing. Then, concatenating the two vectors as a unit, it will be the input of the deep learning model.

## 3.4 Model Training

Figure 3.2 presents the architecture of our deep semantic similarity model(DSSM), which is based on the Stanford Natural Language Inference [12]. This model takes a vector pair of embedded questions as input. Firstly, tokenize the two questions into word sequence then process the word embedding. We need to add padded dimensions to keep all embedding sentences in the same shape (assume each sentence vector has 25 dimensions). Also, we use the GloVe the form a Weight matrix Rw t*Dw that t is the total words number and Dw is the max dimension number of the embedding (assume Dw is 300).

we choose the $relu$ activation function because of its performance better than other activation functions in the one hot vector calculation. Also, we determine to use sig-

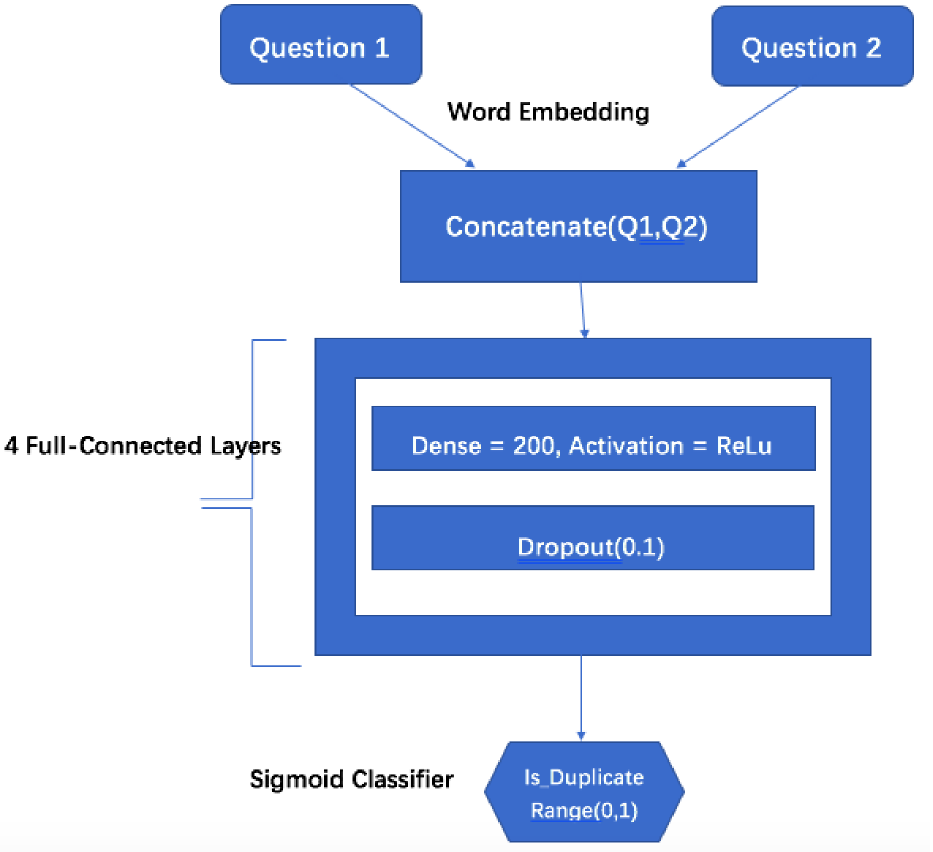


**Figure 3.2**: Model architecture

moid as the activation function in the bottom layer because the sigmoid activation function is better for the binary classification problem. Our approach also adopts the Batch normalization algorithm[13] that can faster the learning speed and higher overall accuracy. There are four fully connected layers which dense is 200 and use dropout to overcome the overfitting problem.

For the last layer, the DSSM should output a parameter called is_duplicate that range from 0 to 1, which indicates the similarity of the input question pair. We use the sigmoid activation function and binary cross entropy for this binary classification problem.

Chapter 4

# Experiments

In this section, several experiments to evaluate the effectiveness of our approach are presented. Selecting some approaches as baselines, our approach needs to be compared with these traditional ones and advanced ones to show its efficiency.

## 4.1 Dataset

The data prepared for baselines is also from the Stack Exchange data dump of 1st Sep. 2017 [1]. We collect 0.31 million duplicate question pairs and divide the duplicate question pairs into three groups that 80% for training, 10% for validation and 10% for testing. The 80% subset is mix up with 0.15 million non-duplicate pairs for training the CLSM model, the first 10% subset is used for validation, and the last 10% subset is used for the comparison process between our approach and the baselines.

## 4.2 Baseline Building

Two baselines are built below, which are Term Frequency-Inverse Document Frequency (TF-IDF) and The Word-n-Grams Letter-Trigram(CLSM).

### 4.2.1 TF-IDF

Term frequency-inverse document frequency(TF-IDF) is a well-known approach that predicts textual similarity with cosine distance [14]. This method is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

Term Frequency measures how frequently a term occurs in a document. And Inverse Document Frequency measures how important a term is. The equation of the weight for term in a document is showed in Equation (4.1).

$$Tf - Idf_{t,d} = Tf_{t,d} \cdot \log \frac{N}{df_t} \tag{4.1}$$

For this baseline development, we use the same tokenizer as we have used in our approach to tokenize the same data, which is the training data for our DSSM. By calculating the term frequency and inverse document frequency for each word in the total word set, we can easily get the cosine similarity for a question pair.

### 4.2.2 Word-n-Grams and Letter-Trigram CLSM

From the literacy review, we found some excellent work that has been done in the more bottom level than the semantic level, such as the word level and letter level research [15]. Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil used the word-n-gram and letter-trigram model to build up a latent semantic model with pretty good results. The difference between the letter-trigram embedding with traditional word embedding is that every single word in the corpus needs to generate some letter trigrams. The letter trigrams represent all the sentences in the low dimensional vector space and then calculate the similarity between them. This method works very well in the letter and word levels, but the difficulty is that there is no well-developed letter trigram base. Nevertheless, this approach is also implemented. We trained this approach and compare the results with those of our DSSM approach in the experiment section. We generate a letter-trigram base from the corpus that we mined from Stack Overflow. Although the amount of letter trigram is not as great as we thought at first, we still apply the trigram base for letter embedding.

The main architecture of the Word-n-Grams and Letter-Trigram convolutional latent similarity model is in Figure 4.1. It contains (1) the word-n-gram layer uses a contextual sliding window to get the word trigram sequence of the input sentence, (2)

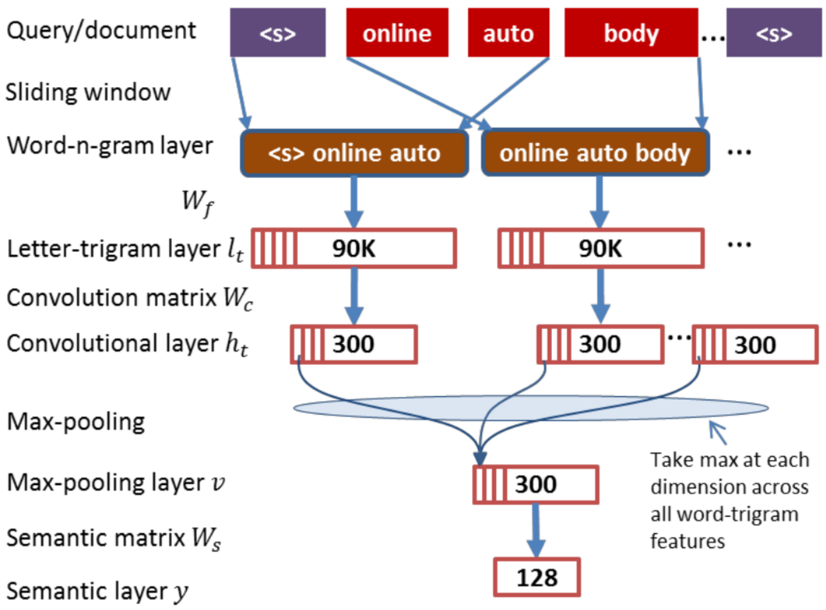


**Figure 4.1**: Model architecture of the Word-n-Grams and Letter-Trigram CLSM[15]

the letter-trigram layer embedding each word trigram into the embedding space, (3) the convolutional layer that combines the features for each word and both its neighbors. (4) the max-pooling layer transforms the word trigram features into a sentence level vector and (5) the final semantic layer that represents the semantic level vector of an input sentence.

The input unit of this model includes a query, a positive sentence, and N negative sentences. And the output unit is a list of similarity, the sum of this list is 1 and represents the comparison of similarity between each sentence with the query. This baseline is used to evaluate the recommendating functionality of our model.

## 4.3 Evaluation Metrics

### 4.3.1 Semantic Evaluation Metrics

According to some other excellent research [16, 17, 18, 19], some good evaluation metrics like accuracy, precision, recall rate and F1-score are the most widely used criterions in their experiments that can be applied to comparing the efficiency and accuracy of our approach and the baselines. Before illustrate how to calculate these criterions, there is a table for introducing some fundamental parameters.

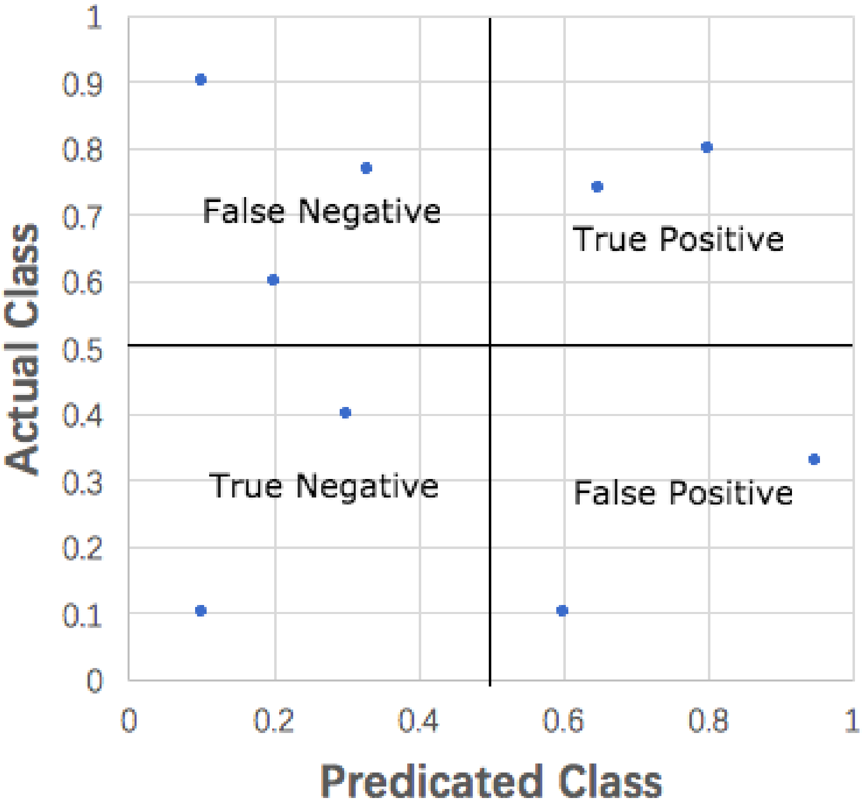


**Figure 4.2**: Illustration: Relationship Between Predicted Class and Actual Class

Assume the method predict a question pair and get the similarity as a number in the range of 0 to 1, and from the label, we know the correct class is 0 or 1. In Figure 4.2, we divide 4 areas for different kinds of prediction. True Positives (TP) is predicted positive values and in the correct range according to the actual class. True Negatives (TN is predicted negative values and in the correct range according to the actual class. False Positives (FP) is predicted positive values and in the incorrect range according to the actual class. False Negatives (FN) is predicted negative values and in the incorrect range according to the actual class.

Accuracy is simply a ratio of correctly predicted observation to the total observations, which is shown in the equation Equation (4.2).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.2}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, which is shown in the equation Equation (4.3).

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class, which is shown in the equation Equation (4.4).

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account, which is shown in the equation Equation (4.5).

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.5}$$

### 4.3.2 Retrieval Evaluation Metrics

According to the existing research [20], we can use the Precision@k (pre@k) and Mean Average Precision (MAP) as evaluation metrics. Given a question pair as input, we have a query and its Ground Truth (GT). Simply evaluate the performance of our approach by calculating the metrics above. Precision@k determined by the most k relevant question of the query and the position of the Ground Truth. Mean Average

Precision is the average value of all the testing queries.

## 4.4 Evaluation

Some research problems related to the efficiency and functionality of our approach will be discussed and answered in this section.

### 4.4.1 Semantic Similarity Evaluation

**Question1: How much advantages the new approach have compared with the traditional approaches in the semantic similarity field?**

| | | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Baseline1 | TF-IDF | 0.4070 | 0.9049 | 0.1555 | 0.2654 |
| Our Approach | DSSM | 0.8026 | 1.0 | 0.8026 | 0.8904 |

**Table 4.1**: Overall comparison

Our semantic similarity model is based on the deep learning technique and uses word embedding to calculate the semantic similarity between two question titles, which is totally different with the traditional TF-IDF approach. And the comparison between these two approaches is in Table 4.1. As we can see, our approach has achieved a tremendous progress on the accuracy of predicting the semantic similarity. Respectively, our deep semantic similarity mode outperforms the baseline in several important criterions.

### 4.4.2 Retrieval Evaluation

**Question2: How much advantages the new approach have compared with the current deep learning approaches in the information retrieval field?** The Table 4.2 shows the comparison between the Word-n-Gram and Letter-Trigram CLSM baseline and our approach. All the Pre@K metrics of our approach are higher than those of the baseline2, But our approach does not far more better than the baseline. Because the data we used to train our deep semantic similarity model is the duplicate pairs man-

| | | Pre@1 | Pre@5 | Pre@10 | MAP |
|---|---|---|---|---|---|
| Baseline2 | TF-IDF | 0.25 | 0.36 | 0.42 | 0.31 |
| Our Approach | DSSM | 0.30 | 0.46 | 0.59 | 0.38 |

**Table 4.2**: Overall comparison

ually marked by experts of Stack Overflow community, there are many other highly similar questions are still not marked.

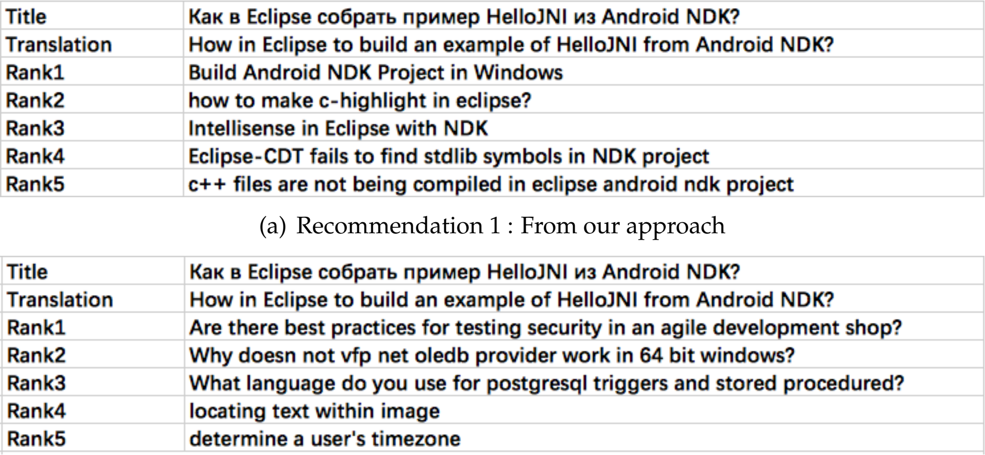
In addition, our mode is a binary classification model, so it basically divides all the questions into two types, which are duplicate and non-duplicate. However, some non-duplicate question pairs generated for training are not totally without relation, which means some questions that recommended by our approach should be very similar to the query, but they are not marked duplicate.

The results confirm us that our deep semantic similarity model is effective in the Information retrieval area. Generally, a user always needs to get a large number of relative posts by the recommendation system, and our approach is better than the baseline2 when they both recommending a large number of posts.

### 4.4.3 Cross-lingual Question Retrieval Comparison

**Question3: How does our approach performance in the cross-lingual question retrieval field?** To answer this question, we randomly take a thousand Russian Posts from Russian Stack Overflow, which own a lower number of answers than others. These Russian posts are considered as query input, and we adopt google translate here to bridge the lingual gap. A black-box testing has been made by a lot of people to find out the efficiency and accuracy of our approach in the cross-lingual area. We collect all the recommendation posts generated by our approach, all the recommendation posts generated by the baseline2 and the recommendation posts generated by the Stack Overflow Search Engine. These three groups of recommendation posts are evaluated manually by volunteers to determine how good our approach is. And the results are very heuristic for us to improve our approach.

Our approach is better in some cases. There are some examples in Figure 4.3.



| | |
|---|---|
| Title | Как в Eclipse собрать пример HelloJNI из Android NDK? |
| Translation | How in Eclipse to build an example of HelloJNI from Android NDK? |
| Rank1 | Build Android NDK Project in Windows |
| Rank2 | how to make c-highlight in eclipse? |
| Rank3 | Intellisense in Eclipse with NDK |
| Rank4 | Eclipse-CDT fails to find stdlib symbols in NDK project |
| Rank5 | c++ files are not being compiled in eclipse android ndk project |

(a) Recommendation 1 : From our approach

| | |
|---|---|
| Title | Как в Eclipse собрать пример HelloJNI из Android NDK? |
| Translation | How in Eclipse to build an example of HelloJNI from Android NDK? |
| Rank1 | Are there best practices for testing security in an agile development shop? |
| Rank2 | Why doesn not vfp net oledb provider work in 64 bit windows? |
| Rank3 | What language do you use for postgresql triggers and stored procedured? |
| Rank4 | locating text within image |
| Rank5 | determine a user's timezone |

(b) Recommendation 1 : From Baseline2

**Figure 4.3**: Examples of cross-lingual question retrieval

Chapter 5

# Related Work

This section introduces the related work. We mainly present the multi-lingual Q&A sites study and the multi-lingual retrieval, then a clear introduction to the deep learning techniques in software engineering.

## 5.1 Q&A Websites with Multiple Languages

Q&A websites have witnessed a booming development in recent years, especially the technique sites like Stack Exchange and Superuser. Some of the Q&A sites choose to start the multi-lingual sub-sites to attract non-native speakers. The debates on the language policy on Q&A sites have never come to an end [3 ,6, 7, 8, 9]. As we know, the language and cultural diversity can lead into a huge barrier. The development of Q&A websites is delayed by this problem. According to the increasing trend of the number and of users on these Q&A websites, it is worthy to evaluate the benefits and disadvantages of the language policy. The challenge we are facing is to bridge the language gap and the semantic gap.

Some insightful research like best answer detection [21], implications of technical mini-blogs Q&A exchange [22] and low-quality answer detection[23] present us many aspects of analysis on the Q&A websites. Also, the participation levels are different among different counties with different languages [24]. The existing research shows us the possibility to develop the cross-lingual approach to help people bridge the language gap or even the semantic gap.

## 5.2 Multi-lingual Retrieval

With the development of internet, Q&A websites have become vital knowledge base for users, especially the technical sites like Stack Overflow for the programmers. Information retrieval is a good approach to assist people utilizing the resource on these platforms. Researchers classify this topic into two main parts, which are document retrieval and code retrieval. For our study, we mainly focus on the document retrieval part. So far, a lot of heuristic research has been done in this particular area. Especially, some research on the dual-language information retrieval is very remarkable. A CNN cross-lingual domain specific model explored some doable methods to overcome the lingual gap [20]. They divide the relationship with two different questions into four types, which are **Duplicate**, **Direct Link**, **Indirect Link** and **isolated**, and formulate a multiclass classification approach, which is considered as a good future work for our approach.

Another excellent work has been carried out on cross-lingual issues in software engineering. The cross-lingual bug localization[25] and the domain-specific multi-classification approach [18] are both focusing on the cross-lingual problem between English and Chinese.

## 5.3 Deep Learning for Software Engineering

There are a lot of deep learning applications in the software engineering field and have made excellent results. Some effective tools like **word2vec** and **GloVe** [26]have been developed for learning words representation, which assists the researchers efficiently process the word embedding tasks. In our approach, we adopt GloVe to process the learning word representation tasks because it can achieve a high level of accuracy in a shorter time with using the negative samples.

The semantic similarity is always a most attractive topic in the neuro-linguistic programming area. A lot of good approaches have been developed on this topic. For example, the model for detecting the question similarity [27] and the model for detecting relevant tweets [28]. These existing well-designed models formulate the problem into a binary classification problem, which is duplicate and non-duplicate. It is necessary to note that there are a number of experts manually mark the highly similar questions as duplicate ones on the Stack Exchange Site. Considering the data structure that we can mine, these binary classification approaches are highly responsive to our scenario and worthy for us to learn.

A highly heuristic work has been carried out recently. As mentioned above, we generate 0.3 million of duplicate pairs for training the deep learning model. Somehow, the amount of training data might be small. Using post body as the corpus to pre-train an RCNN model for generating plenty of titles[Denoising bodies to titles]can solve the problem with insufficient training data. We consider this approach as a good supplement for our model.

Chapter 6

# Conclusion and Future Work

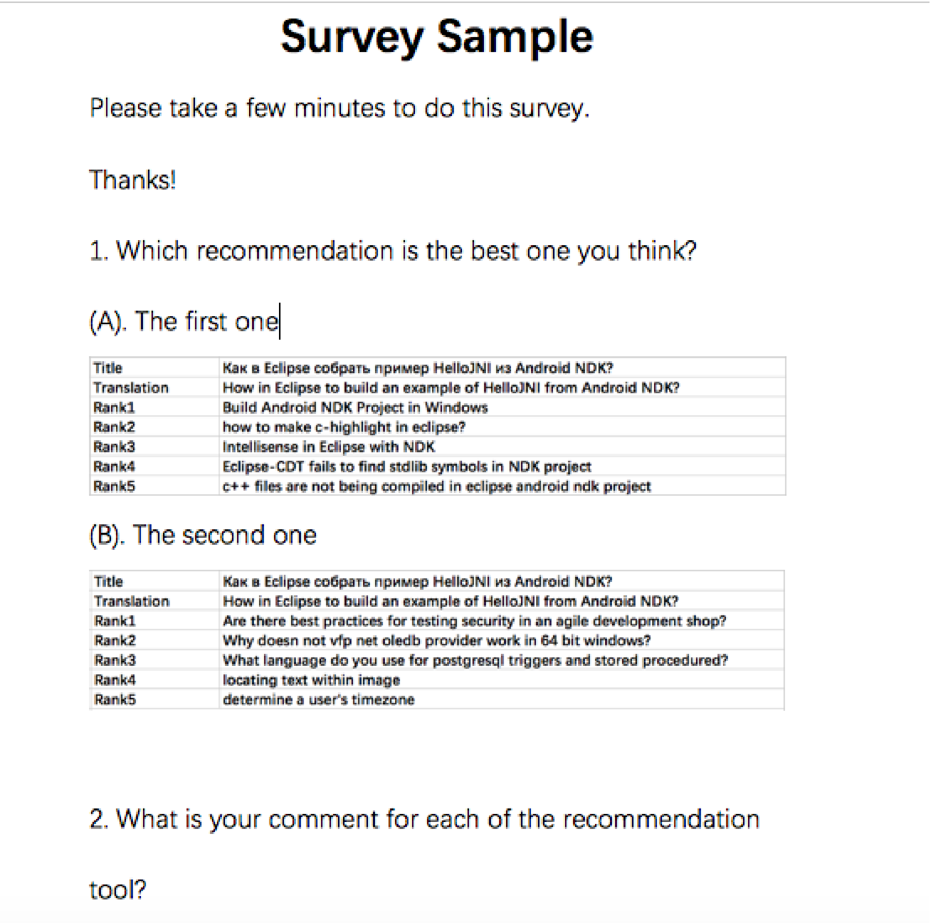
In this paper, we build a deep learning based semantic similarity model for recommending cross-lingual similar posts across Stack Overflow sites of different language. Our approach can predict semantic similarity between two different languages. At word level, our approach has adopted GloVe for word embedding to encode word semantics in dense low dimensional vector space. At document level, our approach has developed a deep semantic similarity model to learn the semantic relatedness between the two posts in a question pair. Our training data is mined from the Stack Exchange site [1], which are manually marked as duplicate pairs by experts on Stack Overflow site. The experiments that we have done confirm the robustness of our approach for overcoming the lingual gap problem in cross-lingual question retrieval and present the advantages of our approach for recommending the best K similar posts with the query, compared with the TF-IDF based baseline and the letter-trigram based deep learning model.

In the future, we will enhance our approach by using some advanced and efficient neural machine translation methods and some accurate feature filtering methods. Some techniques applied at word and letter levels like word-n-gram and letter-trigram will also be considered as the enhancement at the word level and letter level. From the aspect of the training data, according to the related research [29], the training data may be in a small number as we have mined from Stack Overflow, but we can pretrain the deep learning model by collecting a data set of millions of titles from the current post bodies so that it can largely enrich the training data for the deep learning model.

Last but not least, we will implement a new recommending tool that will be deployed on Stack Overflow or some other Q&A websites to assist non-English language speakers to utilize the knowledge in the English Q&A site better.

Appendix A

# Some Other Stuff



## Survey Sample

Please take a few minutes to do this survey.

Thanks!

1. Which recommendation is the best one you think?

(A). The first one

| | |
|---|---|
| Title | Как в Eclipse собрать пример HelloJNI из Android NDK? |
| Translation | How in Eclipse to build an example of HelloJNI from Android NDK? |
| Rank1 | Build Android NDK Project in Windows |
| Rank2 | how to make c-highlight in eclipse? |
| Rank3 | Intellisense in Eclipse with NDK |
| Rank4 | Eclipse-CDT fails to find stdlib symbols in NDK project |
| Rank5 | c++ files are not being compiled in eclipse android ndk project |

(B). The second one

| | |
|---|---|
| Title | Как в Eclipse собрать пример HelloJNI из Android NDK? |
| Translation | How in Eclipse to build an example of HelloJNI from Android NDK? |
| Rank1 | Are there best practices for testing security in an agile development shop? |
| Rank2 | Why doesn not vfp net oledb provider work in 64 bit windows? |
| Rank3 | What language do you use for postgresql triggers and stored procedured? |
| Rank4 | locating text within image |
| Rank5 | determine a user's timezone |

2. What is your comment for each of the recommendation tool?

**Figure A.1**: Survey Sample

# Bibliography

[1] http://archive.org/download/stackexchange

[2] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[3] https://meta.stackexchange.com/questions/13676/dopostshavetobeinenglishon-stackexchange/13684

[4] https://meta.stackoverflow.com/questions/332784/useridandaccountidwhatdo-theyreferto

[5] https://stackoverflow.com/help/whatsreputation

[6] https://meta.stackexchange.com/questions/194959/do-the-benefits-of-having-so-in-multiple-languages-outweigh-the-risks-involved

[7] https://meta.stackexchange.com/questions/52331/is-it-ok-to-have-non-english-question-and-answers-in-area-51

[8] https://meta.stackexchange.com/questions/187805/should-spanish-so-and-all-the-similar-variants-be-closed

[9] https://meta.stackexchange.com/questions/186000/the-language-of-stack-exchange

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781, 2013.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[12] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. ”A large annotated corpus for learning natural language inference,” in Proceed-

ings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), September 2015.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.

[14] C. Sun, D. Lo, S.-C. Khoo, and J. Jiang. Towards more accurate retrieval of duplicate bug reports. In Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering, pages 253–262. IEEE Computer Society, 2011.

[15] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In CIKM, 2014. 5

[16] X. Xia, D. Lo, E. Shihab, X. Wang, and B. Zhou. Automatic, high accuracy prediction of reopened bugs. Automated Software Engineering, 22(1):75–109, 2015.

[17] X. Xia, E. Shihab, Y. Kamei, D. Lo, and X. Wang. Predicting crashing releases of mobile applications.

[18] B. Xu, D. Lo, X. Xia, A. Sureka, and S. Li. Efspredictor: Predicting configuration bugs with ensemble feature selection. In 2015 Asia-Pacific Software Engineering Conference (APSEC), pages 206–213. IEEE, 2015.

[19] B. Zhou, X. Xia, D. Lo, C. Tian, and X. Wang. Towards more accurate content categorization of api discussions. In Proceedings of the 22nd International Conference on Program Comprehension, pages 95–105. ACM, 2014.

[20] C. Chen and Z. Xing, “Similartech: Automatically recommend analogical libraries across different programming languages,” in 31st IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE/ACM, 2016

[21] S. Kim, J. S. Oh, and S. Oh. Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective. presented at ASIST, 2007.

[22] Treude C, Barzilay O, Storey M (2011) How do programmers ask and answer questions on the web? In: Proceedings of the 33rd international conference on software

engineering, pp 804–807

[23] L. Ponzanelli, A. Mocci, A. Bacchelli, M. Lanza, D. Fullerton 2014. Improving Low Quality Stack Overflow Post Detection. ICSME 2014, 541-544.

[24] Oliveira, N., Andrade, N., Reinecke, K.: Participation differences in Q&A sites across countries: opportunities for cultural adaptation. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI). ACM Press (2016)

[25] X. Xia, D. Lo, X. Wang, C. Zhang, and X. Wang. Cross-language bug localization. In ICPC, pages 275–278, 2014.

[26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP 2014), October 2014.

[27] Zhang, Y., Lo, D., Xia, X. et al. J. Comput. Sci. Technol. (2015) 30: 981. https://doi.org/10.1007/s1139001515764

[28] A. Sharma, Y. Tian, and D. Lo. Nirmal:Automatic identification of software relevant tweets leveraging language model. In 2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER), pages 449 458, 2015

[29] Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi S. Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluıs Marquez. 2016. Semi-supervised ' question retrieval with gated convolutions. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, NAACL-HLT '16, pages 1279–1289.