

A Large-Scale Community Questions Classification Accounting for Category Similarity: An Exploratory Study^{*}

Galina Lezina¹ and Pavel Braslavski²

¹ Ural Federal University

galina.lezina@gmail.com

² Ural Federal University/Kontur Labs pb@kontur.ru

Abstract. The paper reports on a large-scale topical categorization of questions from a Russian community question answering (CQA) service Otvet@Mail.Ru. We used a data set containing all the questions (more than 11 millions) asked by Otvet@Mail.Ru users in 2012. This is the first study on question categorization dealing with non-English data of this size. The study focuses on adjusting category structure in order to get more robust classification results. We investigate several approaches to measure similarity between categories: the share of identical questions, language models, and user activity. The results show that the proposed approach is promising.

Keywords: Question topic categorization, community question answering, question retrieval, large-scale classification

1 Introduction

Community question answering (CQA) sites allow users to ask questions almost on every topic and get timely answers from other community members. Examples of general-purpose CQA platforms are Yahoo! Answers and its Russian counterpart Otvet@Mail.Ru (*otvet* means *answers* in Russian). Another popular CQA resource StackOverflow has a narrower scope – users ask there questions exclusively about software programming. Such services became a good complement of major web search engines such as Google and Bing. Users resort to their peers, when they have low search engine proficiency, encounter a complex search problem, or just want a more social search experience. CQA services have collected a vast amount of data and attract quite a big audience of users. Yahoo! Answers claimed reaching one billion answers in May 2010³; Otvet@Mail.Ru

^{*} This work is partially supported by the Russian Foundation for Basic Research, project #14-07-00589 “Data Analysis and User Modelling in Narrow-Domain Social Media”.

³ <http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served>

has accumulated almost 80 million questions and more than 400 million answers by August 2012⁴.

Topical classification of questions is an area of active research. Question classification can be helpful in several ways. First, category prompt for arriving questions makes question asking process easier for the user, maintains topical consistency within categories, and increases utility of categories for potential answerers (which again benefits questioners). Second, CQA archives contain a vast amount of topically labeled questions. Though partly noisy, these data can be still a valuable resource for question classification in external question answering tasks.

We describe an experiment on topical classification of a large data set of Russian questions originated from Otvet@Mail.Ru. The main purpose of this experiment is to learn to recommend appropriate category for the new arrived question. When posting a new question the user has to assign it to a category using drop-down lists; currently no hints are provided. By choosing the topically correct category for the posted question the user increases her chance of getting a good answer in the nearest future. In this paper we show that most users are not familiar with original category structure and rely on the experienced users is impractical.

In addition to inexperienced users problem we explore that Otvet@Mail.Ru categories structure has some drawbacks. This leads to a further category structure violations. Some categories are ambiguous to the user and overlaps with others. Again this leads question assignment to incorrect category.

The idea is to find similar categories and connect them together. These new categories can be accounted in question classification task. To do that we propose three different methods to calculate similarity of categories using the following features: sharing of identical questions, similarity distributions of words, and user activity.

Finally our contributions are threefold: 1) we describe a yearly non-English data set of questions that has not been previously used in research, 2) we perform a classification on a large data set that significantly exceeds in size data sets reported in the literature, 3) we investigate several approaches to category similarity, including users activity that can be helpful for category alignment in case of unbalanced and noisy label information.

The paper is organized as follows. The next section survey papers on question categorization within CQA context. Section 3 describes Otvet@Mail.Ru platform and the data set used in the study, including category structure, distribution of questions over categories, user activity throughout the year. The approaches to quantify closeness of categories are proposed in Section 4. Section 5 discuss classification methods and reports overall performance including our approach.

⁴ <http://otvet.mail.ru/news/#hbd2012> – accessed in July 2013.

2 Related Work

CQA data and tasks attract numerous researchers. Various methods for finding similar questions, search over large collections of questions and answers, experts search, etc. are proposed in the literature. Recent works made an attempt to organize (classify) CQA questions into an existing category hierarchy.

The task of determining CQA question topic has two goals. First is to facilitate browsing questions in CQA resources [1,5,11]. The category structure used in these papers resembles Yahoo!Answers in many ways, including user interface, rules, and incentives. In [1] authors proposed a kernalized framework to classify questions over hierarchical structure. Target category structure is a part of Yahoo! Answers structure: 6 top categories that includes the most popular and least popular categories. Totally they classified 11,354 questions from 127 leaf categories. In [5] authors randomly chosen 2057 Yahoo! Answers questions from 5 academic disciplines categories. Thus classified questions have less noise because they was asked in more formal categories. In [11] 3,900 questions from Yahoo! Webscope data set classified over 1,096 leaf categories. Authors compared different classification approaches using this data set. In [9] authors experimented with large-scale data set. They used more than 2 millions of questions for classification over Yahoo! Answers categories structure that includes 26 top-level and 1262 leaf level categories.

The second goal of CQA classification task is one of the question retrieval (QR) [3,7,8] problems.

The [3] proposes a category-based framework for search in CQA archives. Work conducts experiments with a data set that has 3,116,147 training and 127,202 test questions obtained from Yahoo! Answers. Authors build a classification model to classify a query question over structure that has 26 top-level and 1263 leaf level categories. In [7] authors determined question topic in question search task. Data set obtained from Yahoo! Answers includes 525,401 items from two categories which has 378 leaf categories. They used Yahoo! Answers taxonomy to get the specificity of topic terms. In [8] authors also exploited category information for improving performance of question retrieval. Experimental dataset includes 3,116,147 questions and 26 top-level and 1263 leaf level Yahoo! Answers categories.

The first problem solution would significantly improve user experience while the second makes possible to offer to the user similar questions from CQA archives and possibly avoid the user from posting the question. In our work we address to the first problem.

All those papers use Yahoo! Answers categories hierarchy as a target structure. Our data set differs in many ways from Yahoo! Answers. The most remarkable difference is the total number of categories. We describe in details our category structure in Section 3.1. Moreover some papers use not full Yahoo!Answers categories structure what probably should overestimate classification performance. Finally we have very large amount of source questions organized into much smaller number of categories.

All these papers deals only with data processing and classification method configuration and do not explore original category structure disadvantages. An adjacent to this is the problem named category hierarchy maintenance. Paper [12] propose a new approach to modify a given category hierarchy by placing documents into more topically suitable categories. Authors experiment with Yahoo!Answers, AnswerBag⁵ and Open Directory Project⁶ hierarchies. This work is built on the assumption that new topics arrive with a new documents and that semantics of the existing topics may change over time.

In our work we address to the problems that are not considered in previous works. First we address to the problem of target category structure disadvantages exploration. We explore its drawbacks through user experience. We do not try to discover new topics and to find their location in the category hierarchy but we try to find the most confusable to user categories and use this information about structure violations in the process of category prediction.

And second to our knowledge this is the first work that highlight the problem of classification of large-scale Russian-language questions by CQA categories hierarchy.

3 Otvet@Mail.Ru Data Collection

In this section we overview Otvet@Mail.Ru service structure and present the data collection that we use in our experiments.

3.1 Categories

In Otvet@Mail.Ru all questions are organized in categories hierarchy that has 28 top-level nodes and 186 leaf nodes.⁷ Figure 1 shows part of the Otvet@Mail.Ru categories hierarchy.

Some categories are fine grained in the subcategory level: the largest categories are “Food, cooking” and “Legal advice” since they have 14 subcategories which encompass a wide range of sub-topics. The smallest are “Science, Technology, Languages” and “Style & Fashion” as they have 4 subcategories which are quite coarse. Also some top-level categories such as “Humor”, “Adult” and “Other” do not branch.

Generally topics of the categories represent the interests of the community. Common quite understandable “seasonal fluctuations” on some topics could be traced over time. Figure 2 shows an examples of user activity in four subcategories. In the “Education – Homework” (“edu_homework”) category the number of question decreases in summer starting from June till September. Percentage of questions about homework per month range from 0.26% to 3.8% throughout the year. The maximum of asked questions in the “Travel, Tourism – Holidays

⁵ <http://www.answerbag.com>

⁶ <http://www.dmoz.org>

⁷ See <http://otvet.mail.ru/categories> for a full list of categories.

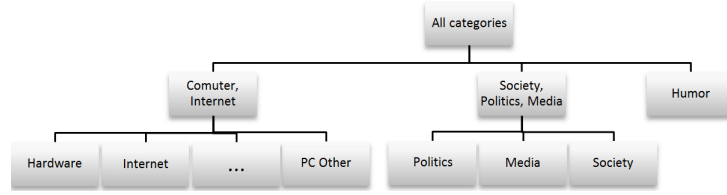


Fig. 1. Part of Otvet@Mail.Ru categories hierarchy.

Abroad” (“travel_abroad”) category asked in the July - usually the holiday season - and the minimum is in the December. Percentage of questions varies from 0.18 % to 0.45 %. Questions about holidays are asked 2.5 times more often in July than in December.

Subcategories “Food, Cooking – Other Cooking” (“food_other”) and “pc_other” have no such fluctuations. “food_other” and “pc_other” subcategories has small changes throughout the year - from 0.225% to 0.258% and from 2.0% to 2.4% respectively.

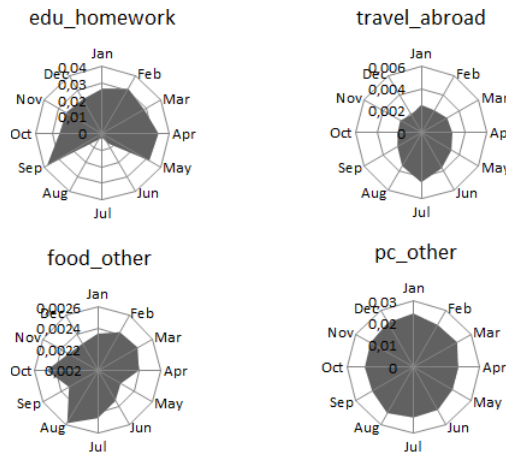


Fig. 2. Distribution questioners activity in categories “edu_homework”, “travel_abroad”, “food_other” and “pc_other” in 2012.

Almost every category has “other” subcategory (like “PC other” in the “IT” category) which itself are noisy because they contain all questions that possibly have no suitable subcategory or could be assigned to more than one subcategory. This drawback heavily violates categories structure, makes them coarse and indistinguishable between each other.

Another major problem is that people often ask at Otvet@Mail.Ru exactly the same and very similar questions in different subcategories, so categories and

subcategories overlaps. All this make categories structure hard to use for both questioners and answerers.

On the one hand user may be confused at the level of subcategories. In the example 1 user asks question about graphics card in “Computer, Internet – Other Computer” (“pc_other”) subcategory while the similar question is asked in “Computer, Internet – Hardware” (“hardware”) category (example 2).

*Example 1. “What graphics card is better? GTX 560 or GT 630”*⁸

*Example 2. “What graphic card is better?”*⁹

On the other hand user may confuse top-level categories. For example the question from 3 is asked in the “Animals and Plants – Wildlife”¹⁰, “Animals and Plants – Houseplants”¹¹, “Family, Home, Kids – Housekeeping”¹² and “Animals and Plants – Gardening”¹³ categories. This question is related to different top-level categories “Animals and Plants” and “Family, Home, Kids”.

Example 3. “what is the name of the flower on the picture?”

For some sort of questions user assumes some categories to be synonymous. In the current Otvet@mail.ru categories structure some questions could be assigned to more than one category.

This violates categories structure and makes user experience with the CQA service much worse. The classifier trained on this data set will probably confuse the categories that confuses the user. Our goal is to find similar subcategories to modify original structure. The approaches of categories structure modifications are described in details in Section 4.

3.2 Experimental Data Collection

To modify categories structure and train classifier we use all questions asked in 2012. The data set was obtained through Otvet@mail.ru API¹⁴. This data set contains 11,170,398 questions from different categories and subcategories some of which are not used in the service anymore. Examples of such useless categories are “Beauty and Health – Doctor” and “Newcomers”. So we do not use this categories in our predictions.

Category named “Golden” is useless because it is not topical. According to formal definition the “golden” category is a special one and it includes selected questions about some facts which may be of interest to a wide range of users. We also do not use this category in our experiments.

⁸ <http://otvet.mail.ru/question/167517346>

⁹ <http://otvet.mail.ru/question/83696264>

¹⁰ <http://otvet.mail.ru/question/69108691>

¹¹ <http://otvet.mail.ru/question/69166385>

¹² <http://otvet.mail.ru/question/69656908>

¹³ <http://otvet.mail.ru/question/69709407>

¹⁴ <http://otvet.mail.ru/api/v2/question?qid=24141950>

We removed all questions asked in these three categories and finally we get 10,739,727 questions asked in 186 categories for experiments.

Most of the removed questions was asked in the “Newcomers” category. Figure 3 shows 10 most popular categories in 2012. Almost 10% of the total number of questions in the data set was asked in the “Humor” category. These top 10 subcategories comprise 40% of all questions and the other 60% are asked in the rest 176 (!) subcategories.

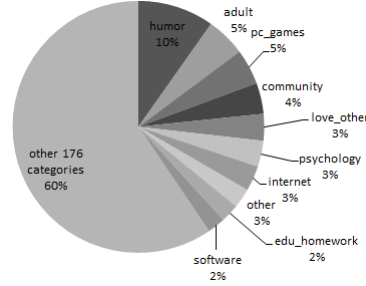


Fig. 3. The most popular categories in 2012.

This percentages changes slightly from month to month, but top categories remain the same. We used questions asked in 11 month to find similar categories and to train classifier classifier. Questions asked in the December of 2012 are used to evaluate classification results. Originally December data set had 989,521 questions but after removing redundant categories we get 939,472 questions.

We did lexical pre-processing of questions before experiments. We perform data pre-processing in three steps:

1. Remove punctuation and lowercase questions.
2. Lemmatize words using AOT¹⁵. AOT is a software for automatic text processing and is intended mainly for the analysis of the Russian language.
3. Remove stopwords.

3.3 Users

In 2012 at Otvety@Mail.Ru 2,287,417 unique users asked at least one question. More than half (1,406,132) of all active users asked question in the service only once. Figure 4 (a) shows dependence of number of questions on the number of users who asked this number of questions.

More frequently users ask questions in one or two categories. The figure 4 (b) shows that 236,670 users ask more than two questions in one category (but possibly different subcategories). In this figure we do not take into account the users who ask only one question. Most frequently users ask questions in two

¹⁵ <http://www.aot.ru/>

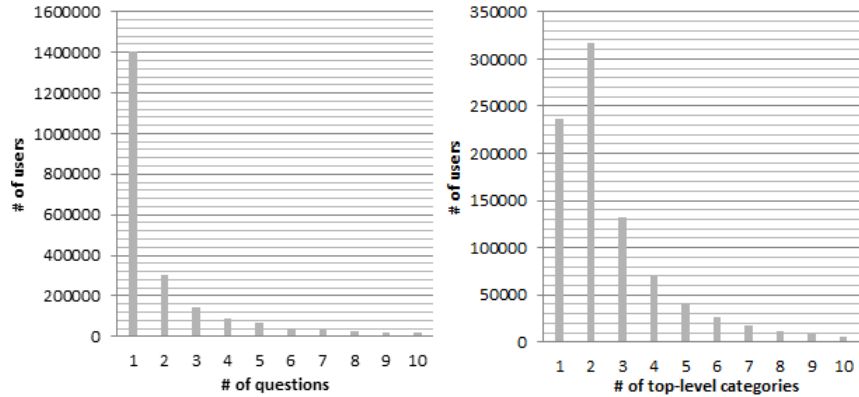


Fig. 4. Distribution of the number of questions depending on the number of users who ask this number of questions (a). Number of categories depending on the number of users who ask questions only in this number of categories (b)

different categories and only one user asked questions in each of 28 categories¹⁶. The same situation is typical for subcategories where user ask questions in a limited number of subcategories - mostly in two subcategories. This limited set of categories of the user possibly is an area of her interests but there is another way to explain this behaviour of the user.

For users who ask questions only in two subcategories we assume that they might not be sure what subcategory best suits the question. Some users might post one question in two different but topically similar subcategories. We check this assumption in the section 4.3.

4 Finding Similar Categories

In this paper we learning to predict the most probable category for the question. Section 5.1 describes baseline classification category prediction approach. For clarity, presenting classification results we also provide examples of categories which confuses the baseline classifier.

Our goal is to improve classification performance by finding categories that often confuses users. Regardless of the baseline classification results we try to find these confusable categories. Further in the paper we show that our approach allows to connect categories that often confuses baseline classifier. We assume that categories that confuses users probably will confuse the classifier too so we modify categories structure to make it more unambiguous and less confusable to the classifier and as a consequence to the user.

We find semantically similar subcategories and connect them so these connections form the new categories from the old one so ambiguous questions could be assigned to them. We use three similarity measures to find similar subcategories.

¹⁶ <http://otvet.mail.ru/profile/id9112629>

4.1 Connecting Subcategories Using Similar Questions

At Otvet@Mail.Ru some similar questions belong to different categories because sometimes it is hard for the user to determine which category is more topically appropriate to the question. We assume that subcategories are similar if they share many similar questions. We denote this method of finding similar question by $QSim$.

Question similarity $QSim$ is calculated as follows:

$$QSim(S_i, S_j) = \sum_{q \in Q_{ij}} \frac{\min(S_i(q), S_j(q))}{S_i(q) + S_j(q)}, \quad (1)$$

here Q_{ij} is the set of questions that is assigned to the S_i and S_j ; $S_i(q)$ and $S_j(q)$ are numbers of times question q was assigned to the S_i and S_j respectively.

To find similar and duplicate questions we use simhash [2] algorithm. Simhash is based on comparison of bags of words and gives the same hash values for the same and similar questions. In our application questions are similar if they have the same vocabulary but may have different set of particles and stopwords. Questions from examples 4 and 5 in Russian language have the same meaning but they differs lexically. Simhash can handle this case because we remove stopwords and particles before calculating hash values.

Example 4. “Who has any plans for today?”¹⁷

Example 5. “What are your plans for today?”¹⁸

As an example this measure gives a strong connection between “pc_other”, “Computers, Internet – Software” (“software”), “Computers, Internet – Internet” (“internet”) and “hardware” subcategories. These 4 subcategories share common top-level “Computer, Internet” category in original Otvet@Mail.Ru hierarchy. $QSim$ also connects “pc_other” subcategory with the subcategories “Science, Technology, Languages – Technology” (“technics”) and “Goods and Services – Mobile devices” (“mobiles”) from different top-level categories. Indeed in “technics” subcategory users ask many questions about computers and hardware like in the example 6.

Example 6. “Hp laptop speakers are hissing, what I should I do?”¹⁹

According to $QSim$ the subcategories “technics” and “mobiles” has weak connection but they are connected too. Example 7 shows the question that is more suitable to the “mobiles” category but was asked in the “technics” subcategory.

Example 7. “What is better to buy HTC One Mini Silver or Iphone 4s 8 GB”²⁰

¹⁷ <http://otvet.mail.ru/question/76074787>

¹⁸ <http://otvet.mail.ru/question/75570807>

¹⁹ <http://otvet.mail.ru/question/167836262>

²⁰ <http://otvet.mail.ru/question/167848364>

4.2 Connecting Subcategories Using Vocabulary

Another approach is to find similar subcategories using vocabularies. We assume that similar subcategories have similar set of words because users ask similar questions. The Kullback-Leibler Divergence (KL-divergence) is a good measure to find subcategories that are lexically similar.

KL-divergence can be calculated as follows:

$$D_{kl} = \sum_{w \in W_{ij}} \log \left(\frac{P_{S_i}(w)}{P_{S_j}(w)} \right) P_{S_i}(w), \quad (2)$$

here $P_{S_i}(w)$ is the probability that word w occurs in the S_i subcategory; W_{ij} is the set of words that occur both in S_i and S_j subcategories.

KL-divergence is an asymmetric measure: $D_{kl}(S_i||S_j) \neq D_{kl}(S_j||S_i)$ so to calculate distance between two subcategories we use the sum of these measures. We denote this measure by $KLSim$ and calculate it as follows:

$$KLSim(S_i, S_j) = D_{kl}(S_i||S_j) + D_{kl}(S_j||S_i) \quad (3)$$

$$KLSim(S_i, S_j) = \sum_{w \in W_{ij}} (P_{S_i}(w) - P_{S_j}(w)) \log \left(\frac{P_{S_i}(w)}{P_{S_j}(w)} \right) \quad (4)$$

According to equation 2 the KL-divergence operates with an intersection of vocabularies W_{ij} of two subcategories S_i and S_j whence the KL-divergence cannot be computed if this intersection is small or empty. To overcome this drawback we use smoothing that was proposed in [4]. Instead of $P_{S_i}(w)$ probability we use smoothed $D_{S_i}(w)$:

$$D_{S_i}(w) = \begin{cases} \gamma P_{S_i}(w) & \text{if } w \in W_i \\ \beta & \text{otherwise} \end{cases}, \quad (5)$$

here W_i is the set of words occurring in S_i subcategory; the parameter β is a positive number smaller than the minimum word probability occurring in either S_i or S_j subcategories and γ is a normalization coefficient and it is based on the requirement:

$$\sum_{w \in W_i} \gamma P_{S_i}(w) + \sum_{w \in W_i, w \notin W_j} \beta = 1 \quad (6)$$

The parameter γ is calculated as follows:

$$\gamma = 1 - \sum_{w \in W_i, w \notin W_j} \beta \quad (7)$$

The parameters γ and β are calculated for each pair of subcategories independently.

As a result the set of connected using $KLSim$ categories pairs is very similar to the set of pairs obtained using $QSim$ described in the previous section. We give a short comparison of these measures in the Section 4.4.

4.3 Connecting Subcategories Using User Activity

Recall that users who ask more than one question in Otvet@Mail.Ru more often assign them only two different subcategories. We motivated by the assumption that users who are confused between two semantically similar categories ask question in two similar categories. We use this assumption to compute categories similarity. We call this measure User similarity and denote it by $USim$. User similarity is calculated as follows:

$$USim(S_i, S_j) = \frac{U_{ij}}{U_i + U_j}, \quad (8)$$

here U_{ij} is the number of users who asks questions both in S_i and S_j ; U_i and U_j are the total number of users who ask questions in the S_i and S_j subcategories respectively.

As the result $USim$ measure connects subcategories from one common category of the original Otvet@Mail.Ru categories hierarchy. It gives only two pairs of connected subcategories which subcategories is assigned to different categories in the original structure. These pairs of connected subcategories are “music”/“drama” and “drama”/“internet”.

4.4 Similarity Thresholds

Original 186 Otvet@Mail.Ru subcategories produce 17,205 pairs and it makes no sense to connect all subcategories so we have to choose the most similar subcategories. We select thresholds for all three measures independently and if pair similarity value does not pass the threshold’s value we connect them in new one. In the section 5.3 figure 6 shows the performance of classifier depending on the selected threshold of similarities for all measures. Empirically selected threshold values corresponds to the moment where classifier performance begins to sharply increase.

Finally we take 106 pairs of similar subcategories for $QSim$ similarity result, 78 pairs for $KLSim$, and 40 pairs for $USim$.

Table 1 lists the number of connected subcategories pairs for each measure.

Table 1. Connected with $QSim$, $KLSim$ and $USim$ pairs.

Similarity measure	# of connected pairs	# of connected subcategories from different categories	Total # of categories in the new structure
$QSim$	107	62	218
$KLSim$	78	41	217
$USim$	40	2	202

Modified structures built using $QSim$ and $KLSim$ measures are very similar to each other because they connect 56 same subcategories. $USim$ and $KLSim$ share only 22 pairs, while $USim$ and $QSim$ have 31 pairs in common.

Generally *USim* connects subcategories from one common category. This means that users who ask only two questions is interested in one common topic and address one question within one top-level category. *QSim* and *KLSim* connect subcategories both from same and different top-level categories.

5 Classification

In this section we describe classification approach and evaluation methods. Table 2 presents evaluation results for different classification tasks and in section 5.2 we describe all methods with its notations.

5.1 Baseline Approach

Our baseline is a standard approach for text classification tasks - support vector machine with bag of words features vector. It classifies questions by original Otvety@Mail.Ru hierarchy.

Figure 5 shows the Hinton diagrams of baseline classifier’s confusion matrices for flat top-level and lower-level classification. Figure 5 (b) shows the part of confusion matrix obtained for flat top-level categories classification. It is interesting that generally humor is the most confusable category and it is less often confused with technical categories than with non-tech (more frequently it is confused with “society”, “philosophy”, “love” and “adult” categories).

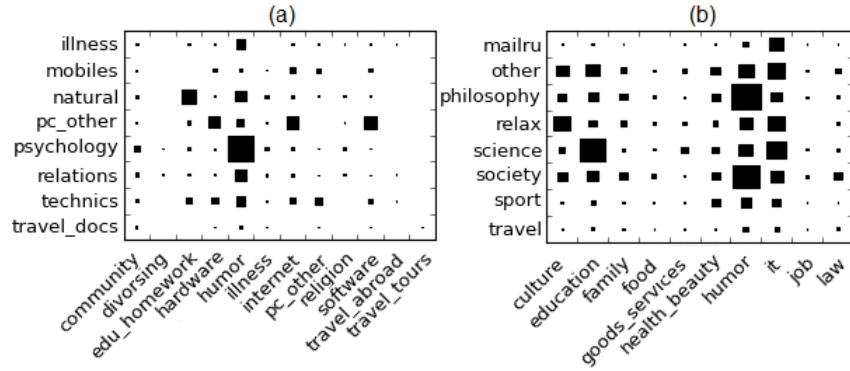


Fig. 5. Part of confusion matrices of baseline classifier for lower-level (a) and top-level (b) categories

Figure 5 (a) show the part of confusion matrix of original lower-level categories classification result. In this figure we can see confusions between subcategories at leaf level. Here we can see that classifier frequently confuses subcategories from one common category like “hardware” and “pc_other”, “religion” and “psychology”, etc.

5.2 Methods

We evaluate question classification performance over original Otvety@Mail.Ru categories hierarchy and three modified categories structures. Evaluation methods of classification over modified categories structures in table 2 is denoted similarly with its measures: *QSim*, *KLSim* and *USim* respectively. We independently classify questions over top-level (TLC) categories and over lower-level (LLC) subcategories of original Otvety@Mail.Ru hierarchy.

Otvety@Mail.Ru categories hierarchy is useful resource not only for internal question category recommendation task. It also can be used to determine topic of the question from external resource - a search engine query subject for example. Recall that query topic identification is actively used in the question retrieval task. Some categories from original structure is not useful for topic prediction task. These categories are “humor”, “other” and “about Mail.Ru project”. “humor” and “other” is not objective while “about Mail.Ru project” is meaningful only for Otvety@Mail.Ru users. The category “other” has questions on many topics as well as “humor”. We exclude these three categories from original Otvety@Mail.Ru categories structure and evaluate classification performance over top-level (TLC*) categories and lower-level (LLC*) subcategories.

Hierarchical classification is an effective approach in hierarchical classification task. We denote it by TLC/LLC. In hierarchical classification approach we build one classifier to predict top-level category and classifiers for every top-level category to predict subcategory.

Recall that we have 10,739,727 questions asked in 2012 and test data set includes 939,472 questions. After removing questions from useless categories we have 8,456,252 questions for training and 815,170 for testing.

All methods use the same baseline classifier. They differs only in evaluation approach.

5.3 Classification performance

Table 2 presents evaluation results in terms of accuracy.

Accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \quad (9)$$

here T means True, F is False, P is Positive and N is Negative.

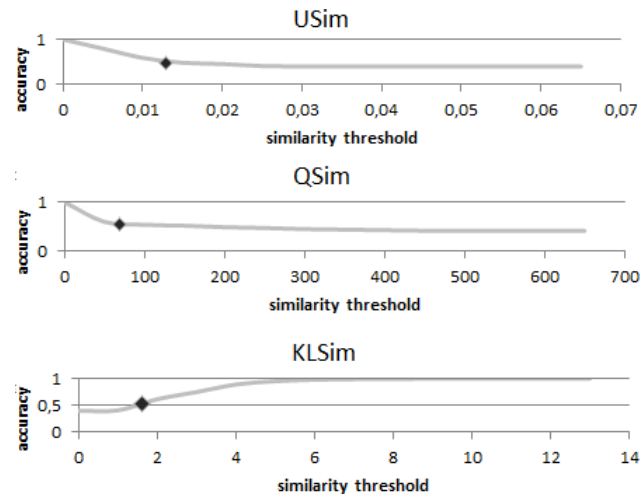
In the case of top 3 evaluation we take three most probable categories predicted by the classifier and see whether correct category are in the predicted. So we give the user an opportunity to choose between recommended categories.

Classifier needs relatively small amount of data for training. Recall that we have about 10 million of questions for training but the accuracy stops growing after 500 thousands of training samples. Figure 7 shows accuracy values depending on the size of training data set for different classification tasks.

We evaluating classification by modified structures for *QSim*, *KLSim* and *USim* measure. Evaluation on structure built with *USim* measure gives us the

Table 2. Classification results

Evaluation method	# of classes	Accuracy	
		Top 1	Top 3
TLC	28	0.56	0.79
LLC	186	0.40	0.63
Baseline TLC*	25	0.61	0.83
LLC*	171	0.42	0.65
TLC/LLC	183	0.66	0.91
QSim	218	0.57	0.80
KLSim	217	0.52	0.76
USim	202	0.49	0.70

**Fig. 6.** Classifier accuracy depending on the selected measure threshold

lowest accuracy. Generally *USim* connects subcategories belonging to the one common category in the original Otvet@Mail.Ru hierarchy while *QSim* and *KLSim* connect subcategories from different categories that users often confuse. *USim* possibly reflect an areas of user's interest and not subcategories that users often confuse within one common category because they do not know which subcategory is more appropriate for a given question.

6 Conclusion

Top-level classification by the structure without general categories like “humor” and “other” and specific “about Mail.Ru Project” category exceeds the classification by original categories structure results by 5%. “humor” and “other”

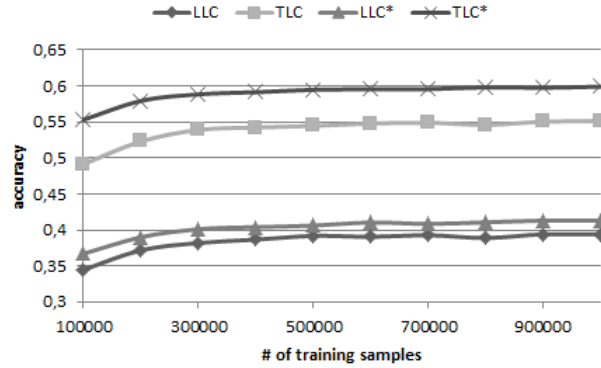


Fig. 7. Accuracy over the number of training samples.

categories itself are noisy because they have questions from all possible categories. In the “humor” category users can post jokes of any topic. Determining jokes is another scientific problem and it is not addressed in this paper. Performance of classification by subcategories (LLC*) without categories “humor” and “other” does not differ from classification by subcategories (LLC) of original structure. Classifier is often confused between subcategories of different top-level categories. The same is relevant to the users.

Hierarchical classification is an effective approach in such problems. In 91% cases a correct category is in top 3 predicted categories. In this case we even do not take into account the similar categories.

Another CQA question classification problem is that questions itself are short sparse texts while sparse text classification is another well known problem. For example this problem is described in [6]. We do not handle question text sparseness in our paper.

An open-ended question is how to choose similarity measures’ thresholds. Recall that we selected it empirically and we do not provide clear guidelines how to choose it. Anyway the more similar pairs coincided with the most confusable by baseline classifier categories. But it could just be the feature of our data set.

References

1. Chan, W., Yang, W., Tang, J., Du, J., Zhou, X., Wang, W.: Community question topic categorization via hierarchical kernelized classification. In Proceedings of the 22nd ACM international conference on Conference on information and knowledge management, 959-968 (2013)
2. Charikar, M. S.: Similarity estimation techniques from rounding algorithms. In Proceedings of the thirty-fourth annual ACM symposium on Theory of computing, 380-388 (2002)
3. Cao, X., Cong, G., Cui, B., Jensen, C. S., Zhang, C.: The use of categorization information in language models for question retrieval. In Proceedings of the 18th ACM conference on Information and knowledge management, 265-274 (2009)

4. Bigi, B: Using Kullback-Leibler distance for text categorization, pp. 305-319 (2003)
5. Blooma, M. J., Coh, D. H.-L, Chua, A. Y.: Question classification in social media. In *International Journal of Information Studies*, 101-109 (2009)
6. Li, B., King, I., Lyu, M. R.: Question routing in community question answering: putting category in its place. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2041-2044 (2011)
7. Duan, H., Cao, Y., Lin, C. Y., Yu, Y.: Searching Questions by Identifying Question Topic and Question Focus. In *ACL*, 156-164 (2008)
8. Cao, X., Cong, G., Cui, B., Jensen, C. S.: A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of the 19th international conference on World wide web*, 201-210 (2010)
9. Cai, L., Zhou, G., Liu, K., Zhao, J.: Large-scale question classification in cQA by leveraging Wikipedia semantic knowledge. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 1321-1330 (2011)
10. Broder, A. Z., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., Zhang, T.: Robust classification of rare queries using web knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 231-238 (2007)
11. Qu, B., Cong, G., Li, C., Sun, A., Chen, H.: An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology*, 889-903 (2012)
12. Yuan, Q., Cong, G., Sun, A., Lin, C. Y., Thalmann, N. M.: Category hierarchy maintenance: a data-driven approach. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 791-800 (2012)