UDC 004.93; 004.8
IRSTI 20.53.19

DEVELOPMENT OF A SYSTEM OF QUESTIONS AND ANSWERS FOR THE KAZAKH LANGUAGE BASED ON RULE-BASED AND HMM

D. Rakhimova[1,4], N. Khairova[2], D. Kassymova[1,3]. K.U. Janibekovich[4]

[1]Institute Information and Computational Technologies CS MES RK, Almaty, Kazakhstan
[2]National Technical University Kharkiv Polytechnic Institutedisabled, Kharkiv, Ukraine
[3]Academy of Logistics and Transport, Almaty, Kazakhstan
[4]Al Farabi Kazakh National University, Almaty, Kazakhstan
https://orcid.org/0000-0003-1427-198X , https://orcid.org/0000-0001-6152-8317
di.diva@mail.ru, dikakassymova@gmail.com, uk02059@gmail.com

Abstract. This study describes and evaluates the methods developed for the module analysis of problems of the closed subject system of answers to questions for the Kazakh language (QA). Analyzing questions in order to obtain the necessary information to determine what is asked and how to respond to them, analyzing questions is one of the most important components of a quality control system. Therefore, we propose new methods for analyzing questions based on two main problems, in particular, on the basis of rules (Rule-based) and on the Hidden Markov model (HMM), based on the integration of the system classification approach, focus extraction and question classifiers, both of which use the relationship of dependence between the words in the question. A comparison of these solutions with basic models is also given. This study also provides a manual summary and annotated data of the gold standard for further research in this area.

Key words: Rule-base, Hidden Markov Model (HMM), Class extraction.

1 Introduction

The question answer system (QA) aims to obtain automatically generated answers to questions generated in natural languages. Over the past decade, radical improvements in natural language processing (NLP) and information retrieval (IR) methods have led to the development of well-known quality control systems, some of which are available for public use, such as AnswerMachine and WolframAlpha. It was even possible to develop a quality control system in which a person could compete with his rivals in a TV show [8]. However, it makes sense to create a fully operational system of quality control, mainly problem analysis (including preliminary processing and classification of questions), information retrieval, cross-linguistics and response generation (including obtaining and formulating answers), as well as rewriting. This is complicated by the many complex transmissions that need to be addressed, such as some lower-level transmissions, such as implications or link resolution. In addition, the architecture of the quality control system, as well as the methods used, usually depend on factors such as subject area and language. Many researchers have considered individual issues related to such systems. Although some of them are considered solved, most of the problems are still open for further research [9,1]. In this study, an attempt was made to analyze the first problem of the quality control system - the problem. In the general system you can ask questions in the Kazakh language. For the system to be truly useful, in addition to accuracy, each module must be carefully designed with a common architecture that is thoroughly analyzed and evaluated individually. In this study, we propose the development and evaluation of the first module, in particular, the analysis of issues for use in the prototype of the geography subject area on the assembly line of our system. The main task of the analysis of questions is to obtain useful information from the given question, which is used in the following modules to form the final correct answer. In particular, information that reflects a particular type or central property of a given object, in addition to classifying the problem into predefined classes in the subject area, can significantly reduce the amount of work space in

subsequent stages of the system, such as information retrieval or candidate generation. In the following example, we are asked for information that shows the name of the plane called the focus and the form of classification. ENTITY.CITY helps us to orient ourselves around these concepts on the basis of knowledge, looking for answers.

"What is the name of the largest city in Kazakhstan?"

To get the focus, we developed a rule-based model, as well as a statistical model based on the hidden stamp model (HMM). We study the accuracy of these two factors when getting the focus. In addition, to classify the problems, we show that the rule-based model is more successful in finding rough classes than the tf-idf-based bag of words base model, which uses the frequency of words in a question. It is especially difficult for the Kazakh language to develop such a module of problem analysis, not to mention the QA system, as it is an agglutinative language with a rich morphology and derivation structure. Therefore, we process preliminary questions by performing morphological analysis and unambiguous elimination, as well as dependency analysis using the NLP conveyor [16,6,15]. Morphological analysis and elimination of ambiguity leads to the acquisition of root forms of words and signs of their speech (POS). Dependency analysis creates a dependency relationship between the words in this sentence. The tags used by the dependency analyzer are defined in the Kazakh bank of the dependency tree, which is the subject, object, proposition, modifier, classifier, owner, etc. b. tags such as [6,7]. We propose a new approach to classifying issues and defining the focus based on the integration of a rule-based method with a HMM-based chain classification method for a closed quality control system. In addition, we offer a first-hand collection and annotation of gold standard analysis data for the Kazakh language. Data on the Kazakh issue of input codes and the gold standard will be made publicly available for reproduction and further study.

2 Related works

The fundamental task of a quality control system is to determine the type of response, its properties and possible limitations. Given the demand in natural language, the quality control system often receives certain direct information, such as the class of the question (for example, who, when, etc.) on the basis of predefined types of answers [4]. The latest modern methods of classifying questions often include statistical methods [12,13]. In addition, some quality control systems are focused on general semantics and form a direct knowledge base from unprocessed texts of questions [10]. However, these systems only determine the type of problem. They no longer specify, for example, what type of object is requested, which significantly narrows the search space. One way to model the analysis of questions is to use general-purpose search engines. One of the earliest studies using such a strategy is the OpenBus quality control system. AnswerBus search engines use a word bag strategy that is evaluated based on the number of requests returned for each word. The total search engine score for a particular question is the amount of hits returned for each word of the question. Based on their total score, the best search engine is determined as the most suitable source of knowledge to answer the question. However, AnswerBus does not use any semantic information and does not receive any information to create a more Competent Response Strategy. The first successful Turkic-language facto quality control system used a hybrid approach (rule-based as well as statistical), but not for the analysis of questions, but for direct answers by comparing samples of questions and answers at the surface level [5]. It does not use explicit question analysis, except for pre-defined samples of questions and answers. Inspired by its significant achievements, our system adapts its strategies to analyze the problems used in one of the most powerful quality control systems IBM Watson [11]. To analyze this question, Watson first takes a piece of the key that is the answer link (focus); second, it takes terms that describe the type of value being asked (lexical answer type, LAT); third, the key class (QClass); and finally, some additional key elements (QSection) if it needs special processing. Lalli et al. Evaluate the importance of distilling such information to get the right answer. To obtain this information, Watson typically uses rules based on regular expressions in combination with statistical classifiers to assess the acquired reliability of the rules. On the other hand, to analyze a complete problem in

a closed quality control system, it is sufficient to extract only LAT and QClass, because in a complete question sentence Watson calls the focus, often the question word (for example, "what" in Section 1). Thus, the real focus of the problem, something we call focus, is actually called LAT according to Watson. In this regard, our focus definition:

"question terms that specify what type of object is being requested".

A more relevant study was conducted to analyze our problems [3], which are used together to obtain the focus of the question in an open quality control system based on rules and statistical methods. In this study, binary classification using reference vector (SVM) machines is performed for words in English problems analyzed by a constituency analyzer. In addition, experts with selected rules are used to define different functions, which are then applied in SVM. In contrast, our analysis is based on rules and uses statistical models individually to get the focus. It also performs a classification of questions for Kazakh questions to be analyzed with the help of a dependency analyzer. In addition, the classification of chains is performed using an algorithm based on the latent markov model (HMM), the results of which are combined with the results of experts based on the rules to obtain the final focus. Unfortunately, our study is inconsistent with this study. First, the definition of focus [3] is related to the analysis of voters and the decision-making of key decisions that are not currently available for the Kazakh language. Therefore, it is not possible to define alternative rules to the English data set, and the methods proposed in [3] cannot be applied to the Kazakh data set.

3 System structure

Although the main technical part of this study is the methodology (ie the combination of rules and statistical models), one of the principles of this work is to implement a quality control system based on this analysis module and create a starting point for the development of subsequent modules. Therefore, this section presents the general architecture of the system, as well as how to add a problem analysis module to it. The overall architecture of the system is based on the DeepQA technology presented in the work [8]. The basic principle in DeepQA is to have parallel blocks with several submodules, which produce different results for the candidates for each application, and then the trained models of machine learning are evaluated according to the evidence collected. The most likely candidate is then returned as the final answer. The focus obtained after the analysis of the question is used in the information retrieval module to obtain the appropriate 4 units of knowledge that can be cut and refined in QClass. These relevant units are then passed to a candidate response module with several different information retrieval algorithms to obtain all possible relevant response units. For each unit of candidate's answer, units of syntactic and semantic evidence are collected, which are then used to evaluate the candidate's answers, and those with low scores are cut. Finally, the strong candidates are synthesized into a final set of answers, in which the most likely answer is passed to the response generation module along with the other k top answers to provide optionality.

3.1 Question analysis module

The question analysis module consists of three parallel sub-modules shown in Figure 1, Distiller, HMM-glasses and ClassRules. The first two modules are for solving the focus of the question, and the third module is for determining whether the question is classified into a predefined class of classes (QClass).

Focus shows exactly what the question asks and what type it is. In the example in Section 1, the focus is on the sum of these parts: “city name” (name of a specific city) because the question asks for a name. In particular, it asks for the name of the city. Therefore, the phrase "city name" can be syntactically derived from the word "city name", because we have morphological roots in the question parts. Because "city" is a root, and "nyn" is a suffix meaning "city name". The QClass for this question is ENTITY (table 2). Note the following example - "Who is a sailor" and Qclass - this is HUMAN.INDIVIDUAL. The question is the basis of the trick that asks the person's name and it is known that the person is a sailor. We capture the distinctive properties of meaning in

question (for example, the first sailor), because at this point we are interested in the relationship "there" and "part", which indicates a particular type of object. The remaining properties are used by subsequent modules of the system to semantically cut both the relevant units of knowledge and the candidate's answers.

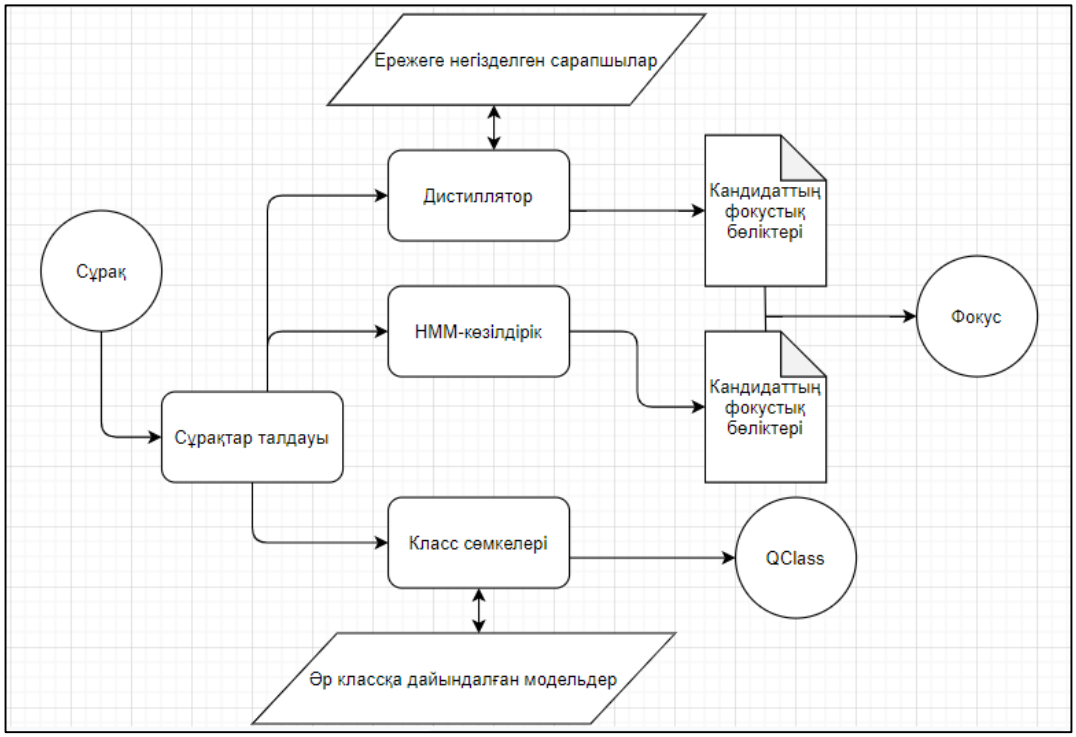

Figure 1 - Question Analysis Module

4 Methodology

To get the focus, we have special rules for dependency trees for all types of specific questions in the geographic domain, a fast-controlled focus extractor, and an HMM classifier that uses a distiller and variation, HMM-glasses. The Wetterby algorithm [17] to some extent makes it much more liberal than distillers. In addition to one general feature that affects the relationship between the words under consideration, their views on the main problem (ie, to get the focus) are based on completely different principles at different levels of solution. This feature is very important for our methodology, as it provides the necessary insight into the effective management of languages with a rich derivative structure, such as Kazakh. At the same time, a delicate balance is needed for the combination of these models. To this end, we take into account the individual reliability of both the distiller and the HMM-glasses in the set of exercises with personal data. In addition, to classify a question from a specific domain (geography in our case) into predefined classes, we have a rule-based classifier that produces a rough class based on hand-constructed sentence-based rules.

4.1 Focus extraction

Distiller. We have noticed that in the geography domain we have chosen, there are certain rules for asking common (predicate-based) questions to many questions. We identified each such model (type of question) and manually determined the rules (experts) for focusing on the dependency analysis of each question. This set of rules is called a Distiller. We currently have an expert on seven rules, as well as a general expert who handles rare cases using one general rule. The main reason for adding an expert is the lack of data. However, we would like to make this optional, as the presence of a specific general expert and a number of experts may result in the accuracy of the fine instead of a small or increased recall depending on the size of the data set, which is not always the case. option required in practice. The data set of all experts and their aggregates are given in Table 1. The rules contain navigation instructions in the dependency

section of the question. For example, the rule for the "what" expert is, and the rule for the "given" expert, as well as the general rule is as follows (Figure 2).

not: (what is…)

Get a sentence on the question (SENTENCE)

- Get traceback from the subject and collect only the owner (POSSESSOR) and the classifier (CLASSIFIER)

given: (… is given…)

- Remove the subject from the sentence in question (SENTENCE)
- Save and control the first degree DATIVE.ADJUNCT of the sentence (SENTENCE) and assemble only the first degree modifier (MODIFIER)

general:

- Remove the subject from the sentence in question (SENTENCE)
- Observe from the subject and get the first rank of the owner (POSSESSOR) and / or classifier (CLASSIFIER) together with their owner (POSSESSOR) and / or classifier (CLASSIFIER)

Each rule-based expert has a level of confidence based on the results of his or her work to distinguish the correct focus from the questions related to the examination. This score is later used to indicate the reliability of the expert's opinion when combined with HMM-glasses. In addition to the focal parts of the question, both the distiller and HMM-glasses report the reliability in the form of a trio:

$$\langle fpt, fpd, fpc \rangle_n$$

here n {{1 ..| Q/}, FPT (focus Part text) represents the text of the focus part, fpd (focus part dependency tag) represents the focus part, and fpc (focus part confidence score) represents the confidence of the focus part. | Q / Q represents the number of words in the question.

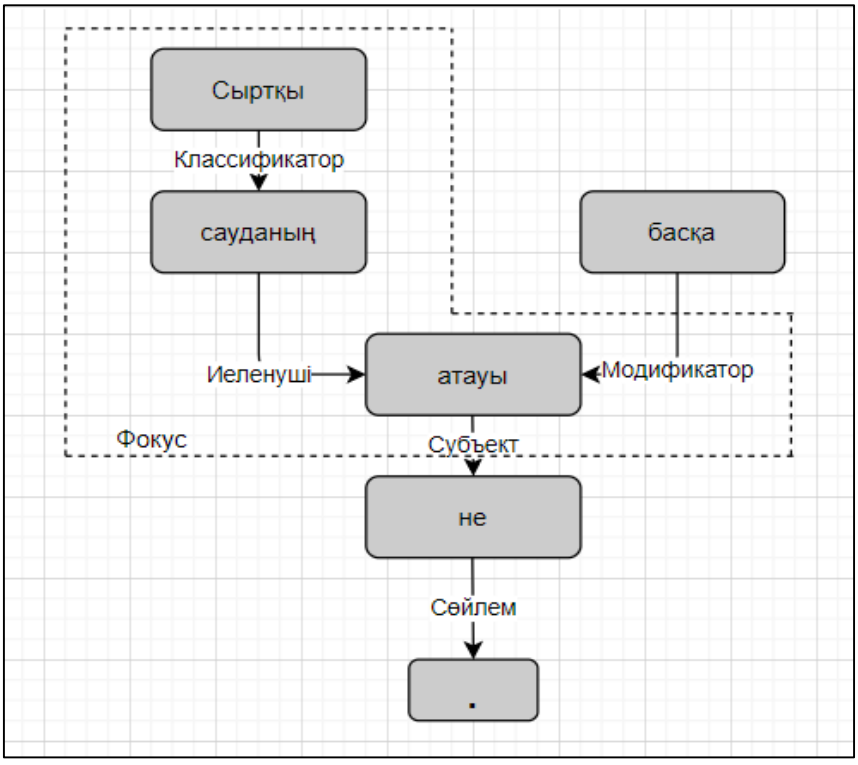

Figure 2 - The expert" What "indicates that the focus of the question is" the name of foreign trade

Table 1. Experts and the frequency of questions in their training data

Expert Type	Frequency (%)
general	25.6
which	19.5

what is	15.0
is called	9.6
how many	9.6
is given	7.2
which	7.2
how many	6.3

Both models produce similar triplets for each part of the focus. However, there is a significant difference between the rules and statistical models in the way in which confidential data is presented for each part of the focus. As explained in detail in Section 4.1, the HMM-glasses work on individual parts of the question, while the distiller sub-trees the dependence of the question on the tree.

Therefore, the distiller's solution is not sufficient to take into account the individual probabilities for each part in focus. Thus, the distiller collects the particles as a focus, and a total confidence score (total confidence score) reported by the responsible expert, which compares the f pc scores of all the particles, making all the particles equal in terms of the distiller in focus.

HMM-glasses. HMM-glasses model the focus capture as HMM (Hidden Markov Model) and perform a sequential classification of words in the question using the Witherby algorithm. There are only two latent states, namely FOC (i.e. the controlled part is part of the focus) and NON (i.e. the controlled part is not part of the focus), which treats each part of the problem as a follow-up and is the controlled part part of the question focus? First, we serialize the dependency tree of the question and send the serialized tree. Serialization (or coding) of a tree is its systematic expression, which is mainly used in the fields of applied mathematics, databases and networks [18,14]. Of course, the method of wood serialization has a significant impact on the characteristics of the algorithm results. We have studied and empirically tested this effect using two common approaches to serialization (see Section 5). Commonly accepted methods of tree serialization try to effectively serialize the tree within the information-theoretical resource boundaries (in terms of time and space). On the other hand, we are only interested in the coherence of the tree-like structure. In other words, the dependency relationship must be agreed between all methods of serialization. Therefore, we considered the simplest methods, direct mode and reverse mode.

Direct and reverse modes. When creating a chain from the dependency tree in the direct mode, the left children (according to the reverse visualization of the dependency tree) prevail over the children who are taken in turn. So first the left branch on the left is taken, then the branch on the right right is taken, and so on. Finally, the parent joins. The reverse is true - positive children outnumber those children. Any difference in serialization changes the whole learning process, thereby reflecting the specifics of a particular series. In this way, the serialization depending on the method of teaching provides a significant diversity of characteristics. Figure 2 below shows the serialization of the question. Recall that we consider only the morphemes of words (ie, excluded from all corrections).

direct serialization (- >)

what is another name for foreign trade

FOC FOC NON FOC NON

reverse serialization (< -)

or the name of another foreign trade

NON FOC NON FOC FOC

In essence, the direct mode of serialization corresponds to the reading of the question from left to right (or from beginning to end), and the reverse mode corresponds to its reading to the end. Different approaches to serialization allow the creation of ensembles of different types of models that process different parts of the question, as they explore different features of the data during the study. Thus, a complex model can be obtained by combining several HMM-glasses with different approaches to serialization. We model the focus task as HMM, first by calculating the previous

probability of our latent state (i.e., FOC and NON), and second, by studying the probability from this set of serialized questions as follows $a_{jk} = P(t^j|t^k)$ $b_{ij} = P(w_i|t^j)$ then a_{jk} previous condition t^k given that, but b_{ij} current control w_i the word indicates the probability that, current status is t^j . Decoding is performed using the Whiterby algorithm, in which the state corresponds to the nodes in the Whiterby strings that represent the most probable judgments for each part that is the focal part of the question. In addition, the control probability b_{ij} is used as a three-fold reliable estimate (ie fpc). Recall that all results are presented in the form of triplets (see Section 4.1).

Dependency tags compared to Word text. In all parts of problem analysis, when possible, it is easier to determine the possible relationship of large distances in problem-solving, for example, by using dependent relationships between words, compared to simple syntactic approaches for languages with a rich derivation structure. Therefore, the very first design of HMM-glasses is designed to study and evaluate the sequence of dependence tags of the problem, which corresponds to the study of the shape of the tree, rather than the sequence of words. However, this approach leads to model confusion, as some tags are more common than others, for example, the problem has only one suggestion tag, and it has many tag-converters. More importantly, focus is often a small part of the question. For example, the idea that a modifier is part of the focus is greatly distorted by the fact that the number of cases without a modifier is higher than the opposite. In addition, working with normalized frequencies requires a large amount of training data to be a statistically significant learning experience. Thus, HMM-glasses are currently studying the probability of a part of the text (ie words) in question. This leaves the template with no information about the manual dependency relationship. However, this is offset by the distiller, as experts only determine the rules of dependence for extraction.

Combination of distiller and HMM-glasses. Recall that the distiller produces focal parts with a single total expert confidence score. In addition, we have HMM-glasses products:

$$\begin{matrix} \text{HMM} & \text{Distiller} \\ \left\{\begin{matrix} \langle fpn_1, fpt_1, fpc_1\rangle & \langle fpn_1, fpt_1, fpc\rangle \\ \langle fpn_2, fpt_2, fpc_2\rangle & \langle fpn_2, fpt_2, fpc\rangle \\ \vdots & \vdots \\ \langle fpn_p, fpt_p, fpc_p\rangle & \langle fpn_q, fpt_q, fpc\rangle \end{matrix}\right. \end{matrix}$$

The combination of possible focal parts with different models is done by parts. In other words, the models try to convince each other that each part is within the last parts of the focus. To do this, we use fpc points, measure them by individual f points of the model according to the training data and get the maximum. Note that only if one detail is defined as a potential focus particle is one of the M1 models (ie, another M2 model assumes that this part is not part of the focus), then we calculate the confidence score M1 and compare it with the M2 f score as described above. If the confidence score M1 is greater than M2, the word is classified as part of the focus, otherwise it is removed from the focus.

4.2 Class extraction

To classify the questions, we identified two types of manual classes, namely [12,13] adapted, coarse and fine classes with different semantic permissions. The thin class of the question establishes a strong connection with a particular subject area, and its coarse class is essentially incorporated into a generalization model, which creates a classification used in areas other than geography. We currently have seven rough classes (Table 2) and a total of 57 fine classes. In this study, we focused only on rough grades. We plan to group thin classes using statistical methods, which requires a full number of questions in each thin class.

Table 2. Rough classes for the domain of geography

Question class	Frequency (%)
Description	25,2

Digital	24,2
The essence	19,6
Temporary	12,4
Location	11,9
Abbreviation	3,8
Man	2,4

To classify this question into one of the rough classes, we created a set of general phrases for each class that is unique to this class. For example, for the NUMERIC class, we have two sentences: "how many" and "how many". The classifier searches for the given patterns in the given question and classifies them accordingly. We additionally implement a statistical classifier that uses a tf-idf-based “word bag” strategy as a basic model for comparison in a rule-based manner. In the basic model, the weight of the word w for class c is calculated as follows:

$$tf\_idf_{w,c} = tf_{w,c} \times idf_w$$

where $f_{w,c}$ indicates the number when the word is in class c, and idf_w is calculated as shown below:

$$idf_w = \log \frac{class\ number}{w\ number\ of\ classes\ available}$$

Then, for the given Q question, we assign it to a class that multiplies the sum of tf-idf points: $argmax_c \sum_{w \in Q} tf\_idf_{w,c}$.

5 Evaluation and results

One of the main problems we faced was the lack of a suitable baseline (from previous studies, etc.) to demonstrate the real severity of the problem and the actual effectiveness of our solutions. Therefore, we implemented a basic model to obtain a focus that matches the words adjacent to the key word of the question for a certain proximity as part of the focus. The proximity model is slightly worse, but the results are similar to the tf.idf model. We selected only the original data with the best results (ie tf.idf) for the actual comparison. It's worth noting that the initial models should be made in a simple way, as no preliminary research has been conducted in the analysis of statistical issues in the Kazakh language. Therefore, the rationale for setting the lower limits of the problem is simple. In addition, a basic statistical model based on tf-idf is introduced to classify the questions, which uses a strategy of word set. All results are presented in the form of comparisons with the basic models given in tables 3 and 4.

Table 3. Results of evaluation of all models of focus

Model	Accuracy	Withdrawal (recall)	F-Assessment
Basic (tf.idf model)	0,769	0,197	0,290
Distiller (Generic Enabled)	0,714	0,751	0,732
Distiller (Generic Disabled)	0,816	0,623	0,706
HMM-Glasses (Backward Mode)	0,839	0,443	0,580
HMM-Glasses (Forward Mode)	0,847	0,495	0,625
HMM-Glasses (Forward and Backward Mode)	0,821	0,515	0,633
Combined	0,734	0,841	0,784

(Generic Enabled, Backward)			
Combined (Generic Enabled, Forward)	0,732	0,846	0,785
Combined (Generic Enabled, Forward & Backward)	0,721	0,851	0,781
Combined (Generic Disabled, Backward)	0,821	0,759	0,789
Combined (Generic Disabled, Forward)	0,818	0,765	0,791
Combined (Generic Disabled, Forward & Backward)	0,802	0,776	0,788

Table 4. Results of QClass classification. The upper part is a model based on tf-idf, the lower part is a rule-based model

Class	Accuracy	Recall	F-Assessment
Description	0,662	0,908	0,764
Temporary	0,767	0,618	0,670
Digital	0,801	0,758	0,776
The essence	0,100	0,025	0,040
Cut back	0,933	0,766	0,823
Location	0,759	0,212	0,312
Man	0,600	0,600	0,600
Tf.Idf total	0,660	0,555	0,569
Description	0,874	0,732	0,797
Temporary	1,000	1,000	1,000
Digital	0,995	0,911	0,951
The essence	0,603	0,817	0,694
Cut back	0,871	0,894	0,883
Location	0,944	0,880	0,911
Man	0,869	0,833	0,851
Rule-based total	0,879	0,867	0,869

Since the data to be evaluated on the basis of our models were prepared in this study course, we are building our evaluation strategy around the concept of hygiene, in which we provide two fundamental principles. First, at any point and for each model, the scores are derived from the result obtained for questions that the model has not previously crossed. Secondly, for a reasonable comparison of models, single points are calculated using the same questions in each price iteration for different models with different parameters. To evaluate the distillery, experts based on the rules are developed using only the first 107 issues that we had at the beginning. Thus, the remaining questions are safely considered as test data, since after receiving a large number of questions, no changes have been made. For all models, ratings are performed using 10-fold cross-validation, which is stratified for all questions. To get focus, the final results (i.e. accuracy, feedback, and f-Score) are obtained by macrosaving individual results. The distiller has the ability to turn the

Universal expert on and off, and HMM-Glasses has forward, backward, and forward-backward modes that calibrate the serialization of the dependency tree. All different combinations of these parameters are evaluated individually for each model, as well as individually in the combination, in each iteration of the accumulation process. To get the focus and classify the questions, the results are presented in Tables 3 and 4, respectively.

5.1 Results of getting the focus

As a result of individual evaluation of the distiller, comparison accuracy and lower return scores (compared to mixed models) were obtained. An important result of the evaluation of the distiller is the actions of a general expert. The results show that the general expert reduces the accuracy of the results obtained when enlarging the sample (i.e. the return) (i.e. the accuracy). However, both results do not compensate, because the results obtained show that the f-Score of the distiller with the general expert turned on is higher than that of the general expert turned off. Individual evaluation of the effect of serialization methods shows that it is slightly better than in the reverse mode, taking into account the f-scores in the forward and reverse modes. The reverse mode probably increases the negative impact of any model it is connected to, but the f-scores indicate that it is not useful to increase the recall, as it reduces the performance of the mixed models when it is connected. In general, although the individual accuracy of the models is quite reasonable, increasing the coverage (recall) for all combined models with both a distiller and HMM-glasses shows that the combination is more useful than individual scores, as it does not destroy the actual scores we observe in individual assessments. , f-scores. Therefore, it can be concluded that the models complement each other well.

5.2 Results of class rules

The results show that the use of knowledge in the subject area has led to significant success, which could not be approached by the basic statistical model. However, a set of manual rules is a big problem when changing a domain. Therefore, it is planned to further develop statistical reading, which automatically learns phrases related to these domains, as each instance requires many instances.

This shortage is the reason for leaving a definition of good classes for future study. Table 4, along with the results of the tf-idf-based classification, shows the macro-accuracy, recall, and f-scores of the rule-based classifier's rough class identification.

6 Conclusion

In this study, we Kazakh language, such as agglyutïnatïvti closed domain system used to answer the questions for the proposed combination of new rules and statistical methods to analyze the question. Question analysis consists of focusing and classifying questions. To get the focus, we have several experts based on the rules for frequently asked questions in the Kazakh language. In addition, we described the method of sequential classification of the novel based on HMM, as well as combined the results of the rules and statistical models on the individual reliability of each model. To classify the questions, we used a rule-based classifier that uses phrases that are not appropriate for each class. We used basic models for both issues and reported on the comparison here. In addition to the proposed methodology, we offer a set of handwritten questions for reproduction and subsequent research. The work herein has been executed in the frame of the project AP09259556 «Development of methods and systems for integrated learning and natural language processing, based on artificial intelligence technologies».

References:

[1] Allam A.M.N., Haggag M.H. The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS). 2012. 2

[2] Benoit D., Demaine E.D., Munro J.I., Raman V. Representing trees of higher degree. In: Dehne, F., Gupta, A., Sack, J.-R., Tamassia, R. (eds.) WADS. LNCS, 1999.1663. 169–180. Springer, Heidelberg

[3] Bunescu R., Huang Y. Towards a general model of answer typing: Question focus

identification. In: International Conference on Intelligent Text Processing and Computational Linguistics (CICLING) (2010)

[4] Dominguez-Sal D., Surdeanu M. A machine learning approach for factoid question answering. Procesamiento de Lenguaje Natural (2006)

[5] Er N.P., Çiçekli A factoid question answering system using answer pattern matching. In: International Joint Conference on Natural Language Processing. 2013. 854–858

[6] Eryiğit G. The impact of automatic morphological analysis & disambiguation on dependency parsing of turkish. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey (2012)

[7] Eryiğit G., Nivre J., Oflazer K. Dependency parsing of turkish. Computational Linguistics. 2008. 34. 357–389

[8] Ferrucci D.A. Introduction to "this is watson". IBM Journal of Research and Development. 2012. 56. 1–15

[9] Gupta P., Gupta V. A survey of text question answering techniques. International Journal of Computer Applications. 2012. 53. 1–8

[10] Katz B. Annotating the world wide web using natural language. In: Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet, pp. 136–159 (1997)

[11] Lally, A., Prager, J.M., McCord, M.C., Boguraev, B.K., Patwardhan, S., Fan, J., Fodor, P., Chu-Caroll, J.: Question analysis: How watson reads a clue. IBM Journal of Research and Development 56, 2:1–14 (2012)

[12] Li X., Roth D. Learning question classifiers: the role of semantic information. Natural Language Engineering. 2006. 12. 229–249

[13] Metzler D., Croft B.W. Analysis of statistical question classification for fact-based questions. Information Retrieval. 2005. 8. 481–504

[14] Munro J.I., Raman V. Succinct representation of balanced parentheses and static trees. SIAM J. Comput. 2002. 31. 762–776

[15] Nivre J., Hall J., Nilsson J., Chanev A., Eryiğit G., Kübler S., Marinov S., Marsi E. Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering Journal. 2007.13. 99–135

[16] Şahin M., Sulubacak U., Eryiğit G. Redefinition of turkish morphology using flag diacritics. In: Proceedings of The Tenth Symposium on Natural Language Processing (SNLP 2013) (2013)

[17] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory. 13. (1967)

[18] Wen L., Amagasa T., Kitagawa H. An approach for XML similarity join using tree serialization. In: Haritsa, J.R., Kotagiri, R., Pudi, V. (eds.) DASFAA 2008. LNCS. 4947. 562–570. Springer, Heidelberg (2008)

[19] Zheng Z. Answerbus question answering system. In: Proceedings of the Second International Conference on Human Language Technology Research (HLT). 399–404. (2002)