Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science and Media

Identifying Comparative Questions on the Russian Web

Master's Thesis

Alexander Bondarenko

1. Referee: Prof. Dr. Benno Stein
2. Referee: Dr. rer. nat. habil. Andreas Jakoby

Submission date: April 12, 2018

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, April 12, 2018

.....
Alexander Bondarenko

Abstract

In this thesis, we study comparative questions on the Russian web—a specific type of inquiries users submit with the intent of comparing two or more objects. In particular, we develop approaches to distinguish comparative questions from other questions and evaluate the performance. As a result, we are able to estimate the amount of comparative questions among all questions submitted to the Russian search engine Yandex and posted on the question answering platform Otvety@Mail.ru in the year 2012. We have found about 1% of the questions asked in Yandex are comparative, which means every two seconds one comparative question is submitted to the search engine, and approximately 1.6% of the questions asked on Otvety@Mail.ru are comparative.

The contributions of this thesis comprise creating a collection of lexico-syntactic patterns to identify comparative questions in Russian, a pattern-based comparative question classification model, manually annotated datasets with comparative and non-comparative questions in Russian, an automated machine learning-based comparative question classifier, and qualitative analyses of comparative questions on the Russian web.

Contents

Acknowledgements

Without the help of Michael Völske from the Bauhaus-Universität Weimar and Matthias Hagen from the Martin-Luther-Universität Halle-Wittenberg, who guided me through the whole work, this thesis never would have seen the light of day.

I also want to thank my family for the distant support and those people in my life, who were near, especially Markus Lehmann, Milad Alshomary, Masoud Allahyari, and Payam Adineh.

Chapter 1

Introduction

> "We will not have understood an ability, such as the human mastery of a natural language, until we have found a theory, a formal system of rules, for describing this competence."
>
> — Hubert L. Dreyfus, *What Computers Can't Do.*

On a daily basis a single person faces a steadily expanding variety of choices, what device to buy, where to go on vacation, or what programming language to learn first. In many cases a decision is made based on the results of comparison of a preferred entity or object against one or more others. Individuals then often seek for advice from other persons or simply use a web search engine to compare objects. However, many of today's search engines often still simply output a list of ranked web pages regardless of the query intents. Unfortunately, conventional search engines such as Google or Yandex—the biggest media portal and the most popular search engine in Russian—do not yet recognize and treat requests for comparison or comparative queries any differently. Users eventually have to scan through a search engine result page and visit several web resources in order to find information on the compared objects in a form of advice and opinions containing reasons to choose one object over another or advantages and disadvantages. Thus, the common approach to the present search result representation can be inconvenient for the users who want to compare objects. However, a request for comparison could often be immediately satisfied with a particular representation on the search engine result page, though. Therefore, being able to recognize and differentiate a comparative search intent from others can be advantageous for search engines and other applications processing user requests.

In this thesis, we investigate the problem of identifying comparative questions and question-like queries on the Russian web. The study is based on a dataset containing questions submitted to the Russian search engine Yan-

dex and a dataset of questions posted on the community question answering platform Otvety@Mail.ru during the year 2012.

We differentiate between two main types of requests for comparison. The first class comprises fully formulated questions, i.e., questions humans formulate in natural language style to elicit information as in Example 1. Such questions are typical in forums and community question answering platforms like Yahoo! Answers or its Russian counterpart Otvety@Mail.ru ("Otvety" in Russian means "answers"). The questions are often verbose and include excessive information describing a particular problem users have.

Example 1:

Собираюсь купить телефон. Два варианта. Самсунг Гелакси С8 или Айфон 10, что лучше посоветуете? Где какие плюсы и минусы?

Going to buy a phone. Two options. Samsung Galaxy S8 or iPhone 10, which is better? Where and what are pluses and minuses?

The second class is question-like queries or simply question queries, which are formulated in a form of normal questions typical for natural language often starting with a question word and not as verbose as the first class (see Example 2). This type of questions is typical for search engines.

Example 2:

Чем альт отличается от скрипки?

What distinguishes a viola from a violin?

The importance of question queries has been explored in a number of studies by researchers including Pang and Kumar [2011], Völske et al. [2015], White et al. [2015] and Guy [2016]. They have shown that question queries submitted to a search engine, have been constantly growing and exceeded 3% of all the traffic for typed queries in 2012 and almost reached 12% in voice-assisted search in 2015. At the same time, a number of community-based question answering web platforms has grown; for example, Answers.com was founded in 1999, Yahoo! Answers in 2005, then Stack Overflow in 2008, Quora in 2010, etc. Both,

an expansion of voice search due to recent advances in speech recognition and a rise of community-based question answering influence the way people ask questions on the web and especially submit requests to search engines tending towards natural language question formulation. To conduct our study, we have available about 1.5 billion question-like queries from the Yandex logs and 11.2 million entries in the Otvety@Mail.ru from the year 2012.

Comparisons on the web usually serve two main purposes. On the one hand, questions with comparison request for examination in contrast or similarity of two or more objects. On the other hand, comparison is a popular approach used in natural language to give an opinion about objects expressing user's experience. Blogs, reviews, forums, social networks etc. provide a plethora of possibilities for human experts to express their opinions on entities and objects, often through comparison. The computational linguistics research community studies comparisons and comparative structures as part of opinion mining and sentiment analysis.

In order to establish a research path for this thesis, we have defined the following research questions:

> RQ1: What are the strong textual signals that distinguish comparative questions from other questions?
>
> RQ2: Can we build an effective classifier to automatically recognize comparative questions on the Russian web?
>
> RQ3: How often and in what categories do users ask comparative questions on the Russian web?

Our first research question contributes to understanding distinct lexical and syntactic structures of comparisons in questions. The second research question addresses building an effective web-scale comparative questions classifier. The last question studies the overall importance of the comparative questions on the Russian web.

In Chapter 2, we review the existing research in computational linguistics that shows that comparative questions and question queries are left out of sight; however, comparisons in statements as part of opinion mining and comparative structures in linguistics are well studied. We also give an overview of the state-of-the-art approaches to identifying comparatives in natural language as a linguistic phenomenon. We analyze what a current state of solving this problem is and what approaches exist and are applied by the research community. It will be also analyzed what questions in processing comparative structures still remain open and how our ideas can fill existing gaps.

In Chapter 3, we introduce an approach to identify comparative structures in Russian. In particular, a model for recognizing comparisons in questions is

proposed as a set of rules for searching strong textual signals. These rules are based on lexical and syntactic patterns, which are likely to occur in comparative questions. We use these patterns to build a pattern-based classifier to distinguishing comparative and non-comparative questions. The accuracy of the proposed approach reaches 72% for both classes on the Yandex corpus and 67% on Otvety@Mail.ru. However, the pattern-based approach gives a high recall producing almost no false negative classifications and a rather prevailing ratio of false positives. To the best of our knowledge, no such model for Russian has been introduced so far. However, approaches for mining comparisons in English have been studied in several works by researchers including Jindal and Liu [2006a,b] and Xu et al. [2011], as well as comparisons in Chinese by Shi et al. [2016a] and Wang et al. [2016], and in Korean by Yang and Ko [2009] and Gu and Yoo [2010]. The majority of the research, however, deals with comparisons in opinion mining and rather in affirmative sentences than in questions. Differently, we focus on the whole spectrum of questions on the Russian web capturing user intents looking for comparisons in the answers.

In Chapter 4, we describe how we created a manually labeled corpus of comparative and non-comparative question queries sampled from the questions submitted to the Russian search engine Yandex and question-answers pairs from the Otvety@Mail.ru. So far, we have labeled 10,000 questions, 35% of which are comparative and 65% are non-comparative. To create the dataset we first randomly sampled 1000 questions from each corpus. About 3% of the 2000 samples were comparative questions: approximately 1–2% in the Yandex and 4–5% in the Otvety@Mail.ru data. To increase the number of positive examples and better balance the dataset, in a second step, we used a pattern-based mining to fetch comparative questions from the corpora. Labeling of the questions as comparative or not was done by native Russian speakers of different background, educational level, and age.

Section 4.2 in Chapter 4 introduces a supervised machine learning approach to classify questions into comparative and non-comparative using for training our created labeled corpus. Our classification model achieves 73% accuracy and a precision of 98% significantly reducing the ratio of false positive predictions compared to the simple pattern-based classification. The chapter further addresses the more complex task of identifying comparative questions seeking for explanations in the answer, i.e., arguments, reasons, advices and human opinions and experience.

This work contributes two differentiation approaches and an annotated dataset to the field of comparative question identification and question categorization research. It also can be beneficial for information system applications dealing with information retrieval and question answering since the qualitative analyses provide insights into user intents for comparison encoded in questions.

Chapter 2

Background and Related Work

In this chapter, we first provide a theoretical background on comparisons and comparative structures in natural language with a focus on a cross-language domain, which is necessary for understanding specific aspects typical for comparatives in but not limited to the Russian language. Second, we review the state-of-the-art approaches to question classification and their importance for applications in information systems and analyze techniques in the automated identification of comparative sentences. We do not limit our observations specifically to the identification of comparatives in text documents but also briefly review adjacent research areas, which can benefit from the methods for comparative questions recognition.

2.1 Comparatives in Natural Language

In natural language, comparison is a specific linguistic structure, which is a main means of measurement according to Babarsad [2017]. Also, comparisons and comparative structures serve as a tool to correlate the degrees of properties of two or more objects through a shared feature (Zevakhina and Dzhakupova, 2015). Example 3a shows an affirmative comparative sentence, where one object *Галакси С* (Galaxy S) is compared against another one *Айфон* (iPhone) over the shared property *надежный* (reliable), where the degree of the reliability *d1* of the first object correlates with the degree of the reliability *d2* of the second object as $d1 > d2$ according to Zevakhina and Dzhakupova [2015].

Example 3a:

Галакси С надежнее, чем Айфон.

Galaxy S is more reliable than iPhone.

Differently, Example 3b demonstrates an interrogative sentence, i.e., question, requesting for comparison between two objects over the shared property *надежный* (reliable) in a natural language manner. It starts with a question word and has a particular word ordering.

Example 3b:

Что надежнее, Галакси С или Айфон?

What is more reliable, Galaxy S or iPhone?

Identification of comparative structures lies, first and foremost, in the domain of detecting specific textual signals and patterns indicating comparative clauses in sentences. Moltmann [1992] has analyzed comparatives in English and introduced a list of comparative operators paired with comparative clause introducers, which are obviously an indicator of comparative structures in sentences. An example of such an operator is the *-er* ending of English adjectives and adverbs in a comparative form. According to the work by Moltmann [1992] among many others, indicators of comparison in text are the following *operator-introducer* pairs: *-er–than* as in Example 4a; *as–as* as in Example 4b; *same–as* as in Example 4c and *different–than* as in Example 4d.

Example 4a:
Galaxy S is more reliable than iPhone.

Example 4b:
Galaxy S is as reliable as iPhone.

Example 4c:
Galaxy S has the same reliability as iPhone.

Example 4d:
Galaxy S has different reliability than iPhone.

The presence of comparative operators in questions thus forms a possible indicator of comparison, where among many other signals, adjectives and adverbs in a comparative form are often—however, not always—a strong signal in both English and Russian. Berezovsakaya and Hohaus [2015] have introduced the following two comparative cross-linguistic universal operators:

1. Comparative forms of adverbs and adjectives in comparison sentences with two or more objects.
2. *Greater–than* relation between objects in a sentence.

However, the Russian language has its own peculiarities due to six grammatical cases, which control relation between lexical units inside a sentence. For example, according to the research by Berezovskaya [2013], a genitive-marked comparative as in Example 5b is typical in Russian but not possible in English (compare Examples 5a and 5b).

Example 5a:

Галакси С надежнее, **чем** Айфон (Nominative).

Galaxy S (is) more reliable than iPhone.

Example 5b:

Галакси С надежнее Айфон**а** (Genitive).

Galaxy S (is) more reliable iPhone.

Both examples demonstrate a usual and equally frequent way to compare objects in Russian and constitute an important case when existing methodology in comparative structure identification developed primarily for English is transfered to non-English languages.

2.2 Question and Query Classification

Today's research in information systems distinguishes between question and web query classification approaches. The first is restricted to applications in automated question answering and follows a taxonomy different from that in query classification, which is primarily assumed for search engines. However,

as we discuss in Section 2.4, the number of question-like queries submitted to search engines has recently grown causing a mixing of the two taxonomies. Being able to identify the type of a question is often crucial for answering it or responding to the user's request. For example, questions demanding comparisons often require a distinct way to respond different for questions asking for one simple fact as, e.g., *Who is the current US president?*

Question Taxonomy

For the first time, a "comparison" class was introduced to the question taxonomy by Lauer and Peacock [1990] who observed that the previous taxonomies proposed by Graesser et al. [1988] and by Lehnert [1978] lacked a category for comparison questions. Overall, Lauer and Peacock introduce 12 types of comparison questions as listed below.

- *Comparison with verification.* Questions seeking for a verification of comparative relationship.
- *Comparison and disjunctive.* Questions asking to choose between two objects A or B.
- *Comparison with concept completion.* Questions comparing the results of two implied concept-completion questions.
- *Comparison and feature specification.* Questions asking for comparison of the features of an object.
- *Comparison with quantification.* Questions comparing two quantities.
- *Comparison and causal antecedent.* Questions asking for comparison of the causes that affect two different entities.
- *Comparison and causal consequence.* Questions asking for comparison of the effects of two causes.
- *Comparison and goal orientation.* Questions requesting to compare goals.
- *Comparison and enablement.* Questions seeking for comparison of capabilities of two entities.
- *Comparison and instrumental/procedural.* Questions requesting for differences between two plans or procedures.
- *Comparison and expectational.* Questions comparing what has occurred with some more likely result.

- *Comparison and judgmental.* Questions asking for a comparison of two judgments.

Despite the focus of the work by Lauer and Peacock [1990] on comparison questions in a narrow field of auditing and finances, it provides a detailed and precise insight to a large variety of ways how such questions can be asked.

Burger et al. [2001] introduces another question taxonomy based on the original Graesser classification scheme and enhanced by a "comparison" class, which distinguishes 18 question categories in total. Examples of the comparison questions given in the work are:

1. "How is X similar to Y?", e.g., "In what way is Florida similar to China?"
2. "How is X different from Y?", e.g., "How is an F-test different from a t-test?'

Question Classification Approaches

Question classification is usually done based on rules or using machine learning. The first approach aims to determine rules in order to classify questions according to the predefined taxonomy. The research work by Prager et al. [1999] introduces a set of 180 hand-crafted templates to match the questions in order to classify them into 20 categories. For example, "where"-questions are assigned a "place" category; the questions matching a keyword "how old" belong to "age" etc. Similarly, Singhal et al. [1999] for the question answering task at TREC-8 classified questions based on a simple coarse taxonomy, where, for example, questions starting with "who" and "whom" were assigned a type "person", "where"-questions a type "location", etc. Thoroughly designed rules can ensure an accurate question classification. However, Li and Roth [2006] argue that crafting rules by hand is only suitable for a coarse classification with a little number of categories but expensive and time-consuming for a fine categorization with many categories. In contrast, their study proposes to exploit a machine learning-based two-staged approach to classifying questions into 6 coarse as a first step and 50 fine classes afterwards in a hierarchical taxonomy. Both classifications are multi-class and built on a Winnow algorithm within the learning architecture SNoW created by Carlson et al. [1999].

Query Classification Approaches

Differently, query classification comprises two fundamental approaches: query intent classification and topical categorization. The first approach is based on a taxonomy proposed by Broder [2002], which includes the following query types.

- *Navigational*, which is an immediate user intent to reach a particular web page.
- *Informational*, which reflects an intent to acquire some information assumed to be present on one or more web pages.
- *Transactional*, which addresses an intent to perform some web-mediated activity.

Topical categorization of web queries has been well studied by different researchers including Kang and Kim [2003], Shen et al. [2006], Beitzel et al. [2007], and Jansen and Booth [2010]. Topical query categories usually vary depending on the research focus. For example, a study by Chang et al. [2014] proposes a rule-based query type classification into eight categories based on keywords and their synonyms as shown in Table 2.1. The category "versus" assumes requests for comparison of two or more objects. The simplistic rule defines a presence of the keyword "vs" or "difference" to be sufficient for assigning the class "versus".

Table 2.1: Query types according to Chang et al. [2014].

Class	Description	Pattern
WHY	Reason or motivation.	"Why"-questions, keywords: "reason", "motivate".
HOW	How to act or solve a problem.	"How"-questions.
VERSUS	Comparison of two or more objects.	Keywords: "vs", "difference".
PROS AND CONS	Advantages and disadvantages.	Keywords: "benefit", "difficulty".
IMPACT	Impact of an event.	Keywords: "impact", "influence", "effect".
PEOPLE	Queries about person or celebrity.	Contains a name of a person from a predefined list of celebrities.
LOCATION	Requests for geographical position, street address, phone number	Contains a location from a predefined location list.
OTHER	Any other queries.	Any other words.

Another approach has been developed by Shen et al. [2005] for the query classification task at the KDD Cup 2005. In a first step, the algorithm retrieves corresponding documents from the web considering query keywords and groups retrieved documents into topical categories. The retrieved web pages are used afterwards to extract additional features in order to expand the queries in each category. These data are used to train a support vector machine to classify the web queries into a set of target categories. A similar approach based on extracting topics from the web pages retrieved using a given query has been proposed by Broder et al. [2007]. The authors report that the method is able to classify queries into a large number of classes, thousands in particular. Very often, a query classification combines several different approaches. For example, the method proposed by Beitzel et al. [2007] comprises manual labeling, rule-based classification (selectional preferences) and supervised machine learning exploiting the Perceptron with Margins algorithm developed by Krauth and Mézard [1987]. Recently, neural networks have been actively exploited for query classification in several research studies; for instance, Liu et al. [2015] built a multi-task deep neural network and Shi et al. [2016b] used a deep long-short-term-memory convolutional neural network.

2.3 Automated Identification of Comparatives

Comparative sentence mining has drawn attention of the computational linguistics community mostly as part of sentiment analysis in user opinions on the web. This is possibly due to a high interest of companies in understanding customer reviews containing comparisons of different brands and items, which can be used for commercial promotion. There are about 10% of the opinionated sentences containing comparisons on the web as reported by Kessler and Kuhn [2013] who conducted the study onblog posts about cameras and cars. Identifying comparative structures automatically is an important but difficult task in processing large amounts of text data available on the web. In today's research, there are two main approaches used to identify comparisons in text:

- Machine learning algorithms.
- Rule-based mining approaches.

The most frequently used supervised learning in comparative sentences classification deploys support vector machines (SVM) and Naive Bayes. The classifier is trained on annotated data using features extracted from text, which usually are part-of-speech (POS) tags, keywords, and n-gram features. This approach needs manual assigning of classes by humans for training examples. Support vector machines are effective classifiers in a high-dimensional feature

space even when a number of dimensions exceeds a number of samples. For example, the SVM-based classifier trained on consumers' product reviews in Chinese is used by Wang et al. [2015] for binary classification of the sentences into comparative and non-comparative. The authors report that the ratio of the comparative sentences in the corpus is less than 25%, so that they use a keyword strategy to additionally mine comparative examples in order to balance the training dataset. We use a similar approach to balancing our dataset as we expect only about 3% of comparative questions in the corpora. The classifier proposed by Wang et al. [2015] is trained on keywords, sequence patterns and manual rules as features and achieves reported overall performance of 92% precision and 80% recall. Another research work by Xu et al. [2011] exploits a multi-class support vector machine for comparative opinion mining in the customers' reviews in English. Specifically, the SVM model is involved in the identification of the particular type of comparative relations in sentences, more precisely, distinguishing between "non-equal gradable", "same", and "superlative" comparisons. The researchers use as features for the classifier compared entities' types and their relative position, entity words, sentiment words, which express relation between entities, and grammatical sentence roles of the entities. The reported multi-class SVM performance measures are 61% accuracy, 62% precision, and 93% recall. Park and Blake [2012] pursue the goal to automatically identify comparative claims in scientific articles. Their study deals with full-text toxicology articles. The proposed method exploits several algorithms for classification including support vector machine reporting 93% classification accuracy. Thorough feature construction is the main focus of their research work. All in all, 35 features are used, including occurrences of particular lexical units, i.e., keywords from the predefined list, semantic words and syntax, and word dependencies in the sentence obtained using Stanford parser.

A Naive Bayesian classifier is another favored and frequent method used in text classification and in particular in mining comparative sentences. According to Manning et al. [2008], Naive Bayes is a simple yet efficient method for text classification, which needs a small amount of training data in order to estimate the necessary parameters. Zhang [2004] claims that regardless of how strong the dependencies among attributes are, the method can still be optimal. For example, in the study published by Yang and Ko [2009], a Naive Bayesian classifier is used to mine comparative sentences from text documents in Korean. The proposed algorithm firstly mines candidates using keyword matching and then applies a machine learning classifier to exclude non-comparative examples. This is a good example of the two-stage approach in comparative sentence classification. The classifier is trained over continuous word sequences within a radius of three, part-of-speech tags, and keywords.

Another research by Tkachenko and Lauw [2014] successfully deploys a Naive Bayesian classifier to identify comparative sentences and compared entities in customers' reviews on digital cameras. There are several other works, which also exploit both support vector machines and Naive Bayes. For example, in the study made by Park and Blake [2012], a Naive Bayesian classifier performs similar or slightly better than SVM in comparative sentence extraction from scientific texts. Jindal and Liu [2006a] in the study on identification of comparative sentences in text documents report that Naive Bayes outperforms a support vector machine.

Rule mining is a popular data mining technique to discover frequent, interesting and useful relations in data. In comparative sentence mining, two approaches in rule mining are mainly used: association rule mining and class sequential rules. Association rule mining finds co-occurrences of the items in the collection of the items, e.g., words in texts, and according to Slimani and Lazzez [2014] is a relation defined as $X \Rightarrow Y$, where X and Y are disjoint sets of items. This is a generic approach used in many areas of research, including marketing and product promotion, medical diagnosis and bioinformatics for DNA sequence analysis. Rule mining is used in the study by Ganapathibhotla and Liu [2008], where the researchers measure an association between a comparative word and an entity feature so that associations are utilized to identify a preferred entity in a comparative sentence. Other research works by Jindal and Liu [2006a,b] propose class sequential rules to find frequent patterns in comparative sentences. A class sequential rule is a sequence of items paired with a related class and can be represented as follows:

> $X \rightarrow y$, *where* X - *sequence of comparative patterns (keywords and POS tags),* $y \in Y$ = *{comparative, non-comparative}*

Their approach effectively combines class sequential rules together with manually-crafted rules to mine comparative sentences in forums, product reviews, and news articles. Another study by Liu et al. [2013] on a comparative sentences identification considers class sequential rules, which are using sequences of comparative words, particular adverbs, and syntactic patterns.

One of the most relevant and interesting studies has been conducted by Li et al. [2010] and proposes a method to identifying comparative questions and extracting compared entities from them. The work defines a comparative question as a question that tends to compare two or more entities and these entities are obligatory to be explicitly listed in the question. Thus, considering an example given in the paper, Examples 6a, 6b are not comparative questions and Example 6c is, where *iPod Touch* and *Zune HD* are compared entities, or

comparators.

Example 6a:

Which one is better?

Example 6b:

Is Lumix GH-1 the best camera?

Example 6c:

What's the difference between iPod Touch and Zune HD?

In order to identify comparative questions, Li et al. [2010] use sequential patterns $S_i(s_1 s_2 ... s_i ... s_n)$, where elements s_i of the sequence S_i are represented by words, part-of-speech tags, compared objects symbols and auxiliary symbols denoting beginning and end of the question. Classification into comparative and non-comparative questions is done following a simple rule — once a question matches one of the sequential patterns it is classified as comparative and non-comparative otherwise. In cases when a single question matches several sequences, the longest pattern is assigned as being the most specific and relevant to a given question. The evaluation of the comparative question mining approach was performed on 5200 questions collected from the Yahoo! Answers, from which approximately 3% were classified as comparative. The study reports the classifier performance as 81.7% recall, 83.3% precision, and 82.5% F1 measure.

2.4 Questions on the Web

In our work, we focus on question-like queries, in other words, queries formulated in the form of natural language questions. A recent analysis of queries on the web has shown a growing number of queries formulated in the form of a normal question starting with a question word as in Example 7a against a telegram style shortened forms as in Example 7b.

Example 7a:

What's the difference between iPhone 8 and iPhone 10?

Example 7b:

Difference iPhone 8 and iPhone 10.

This has happened due to a rapidly growing popularization of voice search in mobile devices as a result of significant improvements in speech recognition technologies. Guy [2016] states that upon analysis of 0.5 million random samples of the voice queries submitted to the Yahoo mobile search application in 2015, he discovered about 12% of queries were formulated as questions, i.e., starting with *wh*-word ("what", "why", "when" etc.) and *yes/no*-questions starting with "do", "does", "is", "can" etc. This is to contrast with the typed question-like queries, which also have shown an increasing proportion. According to the study by Pang and Kumar [2011], which analyzes queries submitted to the Yahoo search engine in the year 2010, the ratio of question queries in the whole query traffic was approximately 2%. Later, White et al. [2015] studied web queries collected through the years 2011-2013 and discovered 3.2% queries being typed in a form of a question. Question-like queries in an amount of 3–4% were found by Völske et al. [2015], who analyzed query logs of the Russian leading search engine Yandex collected in 2012. The rising number and popularity of community question answering platforms like Yahoo! Answers, Stack Exchange, Quora etc. on the web are another reason for the phenomenon. Both voice search and community question answering have influenced and triggered changes in the ways users request for information on the web tending to a more natural way of asking questions.

2.5 Question Answering

Question answering, in general, can be considered as part of information retrieval. Assume an application deploying a question answering system. The application receives a question formulated in natural language as an input, maps it into a word model, retrieves relevant information from available corpora and outputs correlated retrieved data as an answer to the question. Answering comparative questions needs to be treated specifically. The question should be recognized as comparative. Compared objects along with shared features should be extracted from the question. Relevant data needs to be retrieved from a huge amount of text. Retrieved information should be properly

matched with corresponding compared objects and presented to the questioner. Answering comparative questions is an inseparable part of question answering inventory. Automated question answering can be traced back to one of the earliest systems LUNAR, which was presented at the National Computer Conference in 1973 as described in the work by Woods [1973]. It was a domain-specific system, developed as a project supported by NASA. LUNAR was able to very precisely answer questions about lunar rocks and soil. The databases were provided by NASA and were a result of the analysis of the pictures collected by the Apollo 11 mission. LUNAR used a formal query language for retrieving information from the databases. The user typed a request or question in natural English language, which was converted into a specific query language.

An important move towards an open domain question answering was made when Lehnert [1977] proposed a mechanism of a conceptual analysis to map a lexical representation of the question into a sequence of concepts. This idea was implemented in the QUALM question answering system making a significant advancement in the theory of question answering, as it exploited a substantial idea of the semantic meaning of questions. Many research works have contributed to developing question answering including the following. Alfonseca et al. [2001] proposed a question answering algorithm used at TREC based on an information retrieval mechanism with indexing and ranking candidate documents. The question is used as a query to the information retrieval engine. As the first step in question analysis, the algorithm identifies to which of the 15 categories the question belongs, which is necessary for the answer extraction from the candidate documents. Another work by Surdeanu et al. [2011] introduces an approach to answering non-factoid questions, in which text representation is extended to a complex set of combined features, including bag-of-words, n-grams, syntactic dependencies and semantic roles. The ranking model after the candidate answer retrieval uses a batch of features from four feature groups, including a length-normalized BM25, question-to-answer translation model, words densities and frequencies, and a Corrected Conditional Probability. Also, in the study by Clark et al. [2016], the proposed method exploits a combination of information retrieval approaches along with statistical and probabilistic logical reasoning. This combination allows the algorithm to balance the weaknesses of each individual technique.

Chapter 3

Pattern-based Comparative Question Mining

> "In natural language, the ambiguities arise not only from the variety of structural groupings the words could be given, but also from the variety of meanings that can be assigned to each individual word."
>
> — Hubert L. Dreyfus, *What Computers Can't Do.*

Identification of comparisons in natural language is a difficult task both for syntactic and semantic reasons, states Friedman [1989]. Nevertheless, we show that it is possible to identify strong textual signals in sentences and questions in particular, which indicate the presence of comparisons. We develop a collection of lexical and syntactic patterns to mine comparative questions in a large corpus of question queries submitted on the Russian web.

3.1 Datasets

To conduct our study, we have at hand around 1.5 billion records from the Yandex query logs and approximately 11.2 million Otvety@Mail.ru entries from the year 2012. Both datasets include question queries in Russian; however, some of the questions contain words and phrases written in other languages, primarily English, like named entities including commercial brands or game and movie titles or inquiries for correct word spelling in English etc.

Yandex is a leading search engine and media portal on the Russian speaking web with a search engines market share exceeding 52% in Russia in January 2018 according to LiveInternet[1]. Otvety@Mail.ru is a community question

[1]https://www.liveinternet.ru/stat/ru/searches

answering platform and a counterpart of Yahoo! Answers for a Russian speaking audience. It allows users to both ask questions and leave answers. Both datasets were previously preprocessed as described in Völske et al. [2015] including:

- Filtering out question queries based on 58 combinations of question word and uni- or bigram.
- Removing spam and bots.
- Removing repeated, single-word and non-unique questions.

A single Yandex record consists of the query itself, a timestamp and a user ID. We have removed the two latter items of the records and left only question query strings. An Otvety@Mail.ru sample has a more complex structure with incorporated metadata as shown in Figure 3.1. Each question posted on the platform can receive several answers, from which either one or none can be chosen by users as *best*. In order to improve the quality and meaningfulness of the question examples in our study, we extract question-like queries from the *qtext* field only if there is a *best* answer associated with the question. The best answer can be only chosen if the question has at least two answers. All in all, we have extracted approximately 7.7 million question queries having a *best* answer. Each question is assigned by users with one of the 28 predefined categories including but not limited to *Auto*, *Cuisine*, *Computers*, *Beauty*, *Programming*, *Goods*, etc.

3.2 Pattern-based Classification

Even though natural language is diverse and the number of words and variations of word combinations is infinitely large, comparative units are built following a set of syntactical rules (Beck et al., 2009; Berezovskaya, 2013). We assume that the presence of a specific comparison marker is sufficient to retrieve comparative questions. We propose a simple *if–then* rule to perform classification. The algorithm checks whether a given question in the corpus contains one of the comparative patterns. If this condition is satisfied then the question is classified as comparative and non-comparative otherwise. We construct a collection of the comparative patterns based on the research in both linguistics and natural language processing.

3.2.1 Collection of Comparative Patterns

In order to minimize costs of the comparative questions pattern-based mining, during the process of developing comparative patterns, we tend to reduce the

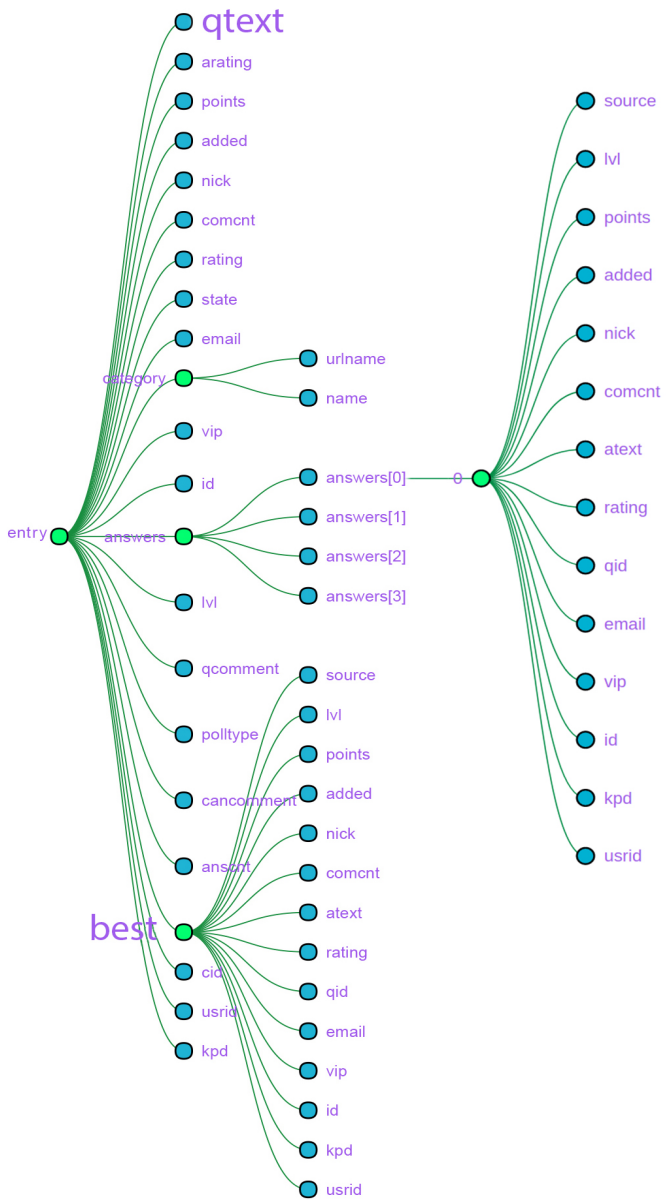

Figure 3.1: Tree structure of the Otvety@Mail.ru record.

number of the patterns and simultaneously maximize their inclusion. We claim that one of the typical requests for comparison is

Что лучше Единица 1 или Единица 2?

What (is) better Object 1 or Object 2?

Here, we are not restricted to the particular and specific objects or entities users can compare; consequently, we omit them and get the first comparative question pattern [**что лучше или**] [**what better or**]. We assume that questions containing all words in the pattern are comparative questions as in Example 8a.

Example 8a:

Что лучше, Турция **или** Кипр?

What is **better**, Turkey **or** Cyprus?

We also argue and show that the presence of the linguistic units themselves is more important than the order, in which they occur in the sentence. For example, Kallestinova [2007] asserts, "It is well known that Russian has a relatively free word order" and gives an example of only three words, which produce six syntactically and semantically correct valid sentences. Subject and object and dependency relations in sentences in Russian are governed by six cases. Summing it up, Examples 8a–d are exchangeable and likely to be present in our large corpora; however, the question as in Example 8a is expected to occur in the Russian language more frequently than the others.

Example 8b:

Лучше что, Турция **или** Кипр?

Better what, Turkey **or** Cyprus?

Example 8c:

Турция **или** Кипр, **что лучше**?

Turkey **or** Cyprus, **what better**?

Example 8d:

Что, Турция? **Или лучше** Кипр?

What, Turkey? **Or better** Cyprus?

In the next step, we generalize our first comparative search pattern to match as many records in the corpus as possible. In Example 8a, two entities are compared over a shared attribute *good*. However, natural language offers a large number of possible object's adjective attributes to carry out a comparison, for example, *bad–worse, reliable–more reliable, high–higher, awesome–more awesome* etc. We conclude that *better* in the pattern can be substituted by a part-of-speech tag *comparative adjective* or *comparative adverb*. Adjectives serve as a property of nouns and adverbs as an attribute of verbs as in Example 9.

Example 9:

Что лучше [COMP ADV] посетить Турцию **или** Кипр?

What is **better** [COMP ADV] to visit Turkey **or** Cyprus?

Furthermore, we consider that as subjects in comparison not only can act things but also persons or actions. Moreover, requests for comparison in the form of questions can contain a broad variety of aspects expressed in different question words, e.g., a request for comparing destinations as in Example 10a or persons as in Examples 10b, 10c.

Example 10a:

Where is it **cheaper** to go, Turkey **or** Cyprus?

Example 10b:

Who is a **better** choice Name1 **or** Name2?

Example 10c:

Which of the writers is **better** Name1 **or** Name2?

Considering that the available corpora consist of only questions, we can simply omit a question word in the comparative pattern. Thus, we claim that above-given Examples 8–10 would match the pattern **[COMP ADJ|ADV or]** independently of the word ordering and be classified as comparative.

Mining comparisons based on searching for comparative markers has been studied in the work by Jindal and Liu [2006a,b], who have developed an approach to automatically classify opinion sentences into comparative and non-comparative. The authors have created a list of 83 keywords and patterns in English used in class sequential rules to match affirmative comparative sentences. These comparative keywords and phrases include but are not limited to *POS tag COMP and SUPERL* ⟨*adjective and adverb*⟩*; as* ⟨*word*⟩ *as; similar*; verbs *choose, prefer, recommend, differ* etc. We adopt the list and fit it into the Russian language to experimentally identify those keywords, which can be used for mining comparative questions in Russian.

Another study has been done by Fiszman et al. [2007], which analyzes comparative constructions in biomedical texts. Although the study was conducted on a specific domain and limited to a particular type of texts, discovered comparative patterns are universal. To list several of them, we have borrowed, for instance, *more/less–than, compare Term1 with/to Term2* etc. Searching for patterns we have not only ourselves come up with examples of comparisons used in natural language but also sought for use cases in linguistic works made by researchers including Beck et al. [2009], Berezovskaya [2013], and Berezovsakaya and Hohaus [2015] as well as in the psychological studies by Tversky [1977].

A follow-up in building a collection of patterns is to investigate how the word "compare" is used for asking comparative questions. We test the pattern [**по сравнению** | **в сравнении**] [**in comparison** | **compared to(with)**]. We have mined questions matching the pattern from the Yandex dataset and found a large number of examples requesting for comparing time differences as in Example 11a, prices in different countries as in Example 11b and simple facts as in Example 11c.

Example 11a:

Сколько времени в Бельгии **по сравнению** с Абаканом?

What time is it in Belgium **in comparison** with Abakan?

Example 11b:

Дешевле ли цены на Украине **по сравнению** с Россией?

Are the prices lower in Ukraine **in comparison** with Russia?

Example 11c:

Сколько витамина с содержится в плоде киви **по сравнению** с лимоном?

How much of vitamin C does kiwi contain **in comparison** with lemon?

Further, in our study, we detach the questions as in Examples 11a–c and name them *non-reasoning comparative questions*. We distinguish such questions from ones seeking reasoning support. In the *reasoning comparative questions*, we include those, which cannot be answered with simple factual information but rather demand a human-touched feedback based on experience, opinion, feelings, and reasons. In Figure 3.2, we present a hierarchy of comparative questions based on the user intent encoded in the question for receiving reasoning support in the answer.

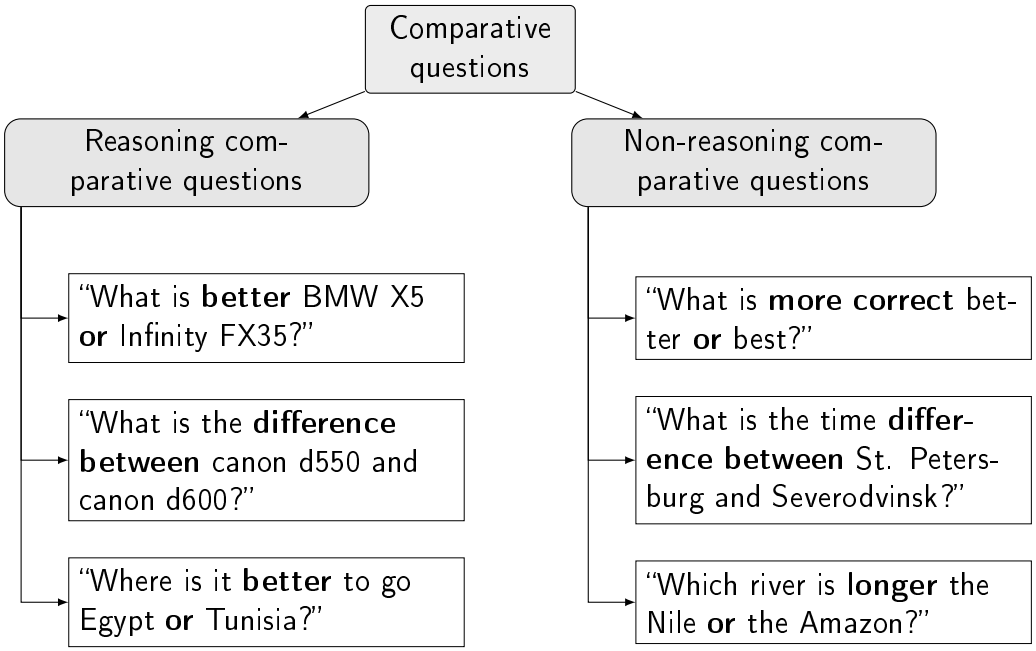

Figure 3.2: Subdivision of comparative questions based on the user need for a reasoning support.

Among others, we consider as non-reasoning any inquiries about correct spelling, time or price differences in various locations, simple facts (for example, comparison of factual physical parameters of the objects) or any other questions, which can be answered with factual data without providing an explanation, reasons or arguments. According to the restriction mentioned above, the total ratio of comparative questions for the *compare*-pattern has surprisingly amounted to as little as 55% of all mined questions matching the given pattern. This pattern and all others are presented in Table A.1 in Appendix A. To evaluate the ratio of comparative questions mined using textual patterns, we have randomly sampled 100 examples from the corpus and manually labeled them with classes.

Many works on comparison identification by researchers including Jindal and Liu [2006a,b], Ganapathibhotla and Liu [2008], Xu et al. [2011], Kessler and Kuhn [2014], Saritha and Pateriya [2014] and Wang et al. [2017] have added superlative forms of adjectives and adverbs in a list of comparative keyword indicators. This type of comparison is seen as a relation *better/greater, worse/less than all others* expressed by contrasting one object against all possible objects of the same or similar type or possessing same or similar properties.

There are three ways to build superlatives in Russian: a simple alteration of the word suffix and ending as in Example 12a and adding adjectives "самый" and "наиболее" in compound superlatives as in Examples 12b, 12c.

Example 12a:

умный (smart) - умн**ейший** (smartest)

Example 12b:

умный (smart) - **самый** умный (most smart)

Example 12c:

умный (smart) - **наиболее** умный (most smart)

Thus, we have created three patterns based on the *superlative* part-of-speech tag and presence of "самый" and "наиболее" and mined question examples from the Yandex corpus. The ratio of comparative questions with a superlative POS tag achieves 61% and in the more specific *most*-form superla-

tives 83%. These statistics are summarized in Table A.1 in Appendix A.

The peculiarity and main distinctness of the *superlative*-form questions in contrast to *comparative*-form ones are that the first type does not provide the entities to be compared (in most cases), however, the second does. In other words, the *superlative* comparison can be seen as follows.

⟨Object1⟩ compared to ⟨abstract set of all the others⟩

Whereas the *comparative*-form (in most cases) can be described as the following model, where the compared entities or objects are explicitly listed.

⟨Object1⟩ compared to ⟨Object2 ... ObjectN⟩

According to the mentioned findings, we distinguish between direct comparative questions and indirect ones. The corresponding hierarchy of the comparative questions is presented in Figure 3.3.

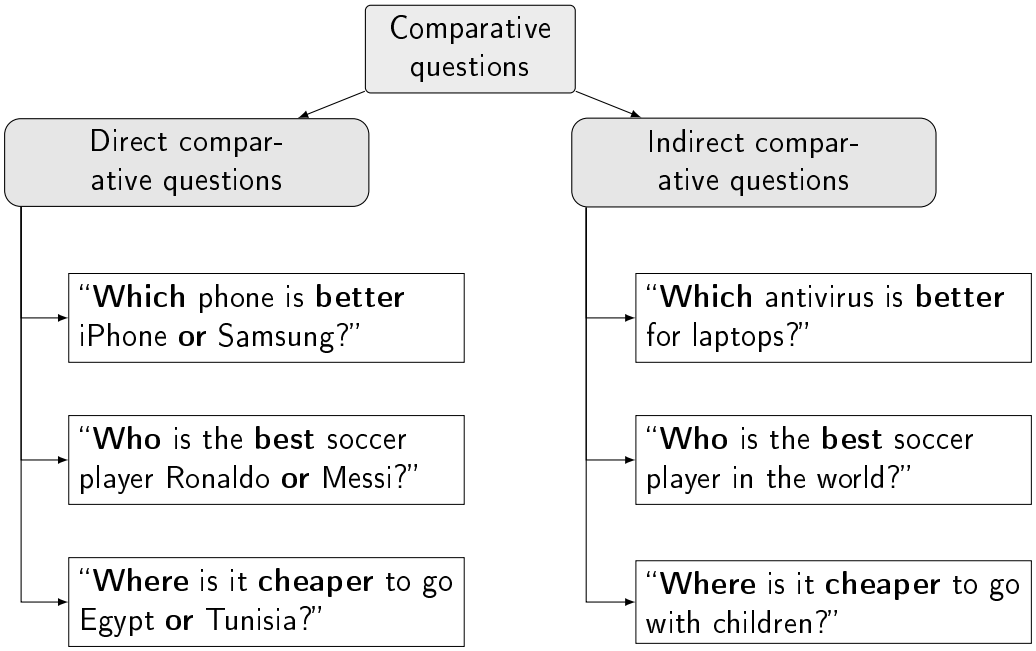

Figure 3.3: Subdivision of comparative questions based on inclusion of the compared objects.

While creating a collection of the comparative patterns, we found that

53% of the questions in the Yandex dataset containing *or* conjunction were comparative. Thus, *or* can be a good comparative marker. However, it is a much weaker indicator of the comparative questions in the Otvety@Mail.ru corpus, which gives less than half of that in Yandex, i.e., 21%. We calculated a ratio of comparative questions for the given pattern in the Otvety@Mail.ru dataset by random sampling 100 examples and manual annotation. All results for the Otvety@Mail.ru corpus are presented in Table A.2 in Appendix A.

The sparsity of natural language makes encoding all possible word combinations in textual patterns rather difficult. That is why we undertake an attempt to reduce pattern dependency on the words themselves. We focus on auxiliary lexical items participating in building language constructions such as conjunctions and quantifiers like comparative forms of adverbs and adjectives and structural parameters of the sentences. This attempt bring us to investigating the average length of the comparative questions. The length, as we consider, can also be an important parameter in identifying comparative questions, especially in cases, when users simply omit *or* when listing objects to be compared in the search engine queries as in *Which is better to use Java Python?* In such questions, a single comparative adjective can be rather a weak comparison indicator. We have calculated that there are approximately 28 million questions from the Yandex corpus, which contain a comparative adverb or adjective and about 0.5 million such questions are in the Otvety@Mail.ru dataset. However, only around 20% of them are in fact comparative. Whereas, the majority of the retrieved questions are in a form of asking for advice or according to our classification indirect comparative as in Examples 13a–c.

Example 13a:

Какой телефон лучше купить в 2012?

Which phone is it better to buy in 2012?

Example 13b:

Где дешевле отдыхать в августе в Европе?

Where is it cheaper to go on vacation in Europe in August?

Example 13c:

Подскажите, пожалуйста, какую программу лучше использовать при качании мышц?

Your opinion please, which program is it better to use for building muscles?

On the contrary, the request for a comparison between objects can be also submitted without using comparative adjectives and adverbs, which usually serve as quantifiers in questions and describe a particular feature or property, over which comparison or contrast is performed. For example, *what to buy* and *what to choose*-questions as well as *similar between* and *difference between*-questions do not usually contain comparative adjectives and adverbs.

Based on the experiments with different patterns, we model a generic representation of the comparative questions typical in our corpora of the filtered questions asked by users on the Russian web. The generalized comparative question frame is presented below.

⟨**Question word**⟩ ⟨**generalization term**[1]⟩ ⟨**comp. keyword**[2]⟩ ⟨**Object1**⟩ ⟨**conjunction**⟩ ⟨**Object2**⟩ ⟨**conjunction**⟩ ⟨**Object3**⟩**?**

Under *generalization term*[1] we understand a hypernym, i.e., a superordinate word or more general term, of the compared entities or objects. As Li et al. [2010] states, for example, "Ford" and "BMW" can be compared as automobile manufacturers (or brands). In the question *Which mobile phone is better iPhone or Samsung?*, the generalization term is *mobile phone* for the compared objects iPhone and Samsung. *Comparative keyword*[2] means adverbs and adjectives in a comparative (sometimes superlative) form and such words as *difference, advantage, similarity, choose, compare* etc. All comparative patterns are collected together and displayed in Table A.1 and Table A.2 in Appendix A.

The comparative question model presented above is a generalized encoding of comparative questions, in which the elements can occur in different combinations, and represents possible forms of the requests for comparison. Based on the model, we have created two patterns exploiting the length of the questions as the main feature. In the Yandex corpus, we have counted approximately 5 million questions matching the pattern [**5-6 words or**] with 71% of them being comparative and almost 8 million examples matching the pat-

tern [**5-6 words COMP ADJ|ADV** exclude **or**], which gives an 81% ratio of comparative questions. It is important to note that we have not considered the patterns based on the question length for the Otvety@Mail.ru corpus. This is due to the verbosity of the questions asked on the community question answering platform. This verbosity often includes an introductory part like *What would you recommend* followed by a question, *I am interested in your opinion* and question, or *I am going to buy a new mobile phone* and question. This objectively makes the average length of the questions bigger and rather difficult to predict.

3.2.2 Classification Results

The works on identification of comparative sentences made by Jindal and Liu [2006a,b] report that keyword-based comparison mining has a high recall and a low precision; thus, it is used for candidate mining as the first step in a comparative sentences classification. In contrast, our pattern-based approach uses words and part-of-speech tags combinations to match questions rather than single keywords. We have designed the comparative patterns to cover as many question examples as possible on the one hand and obtain a high inclusion of comparative examples matching a given pattern on the other hand. We use our collection of patterns to classify questions in two corpora, Yandex and Otvety@Mail.ru, into comparative and non-comparative based on simple rules as shown in Algorithm 3.1. Each question is checked on matching one of the comparative patterns in the collection. If a given question matches the pattern it is assigned with a *comparative* class and *non-comparative* otherwise. We omit a pattern containing single *or* conjunction because of its small accuracy as shown in Table A.1 in Appendix A. For classification, the comparative patterns are encoded using *regular expressions* in Python. The questions are preprocessed by punctuation removal and lowercasing.

Algorithm 3.1: Annotate questions in the dataset with classes

```
inputs : A single, un-labeled question
output: A label 'comparative' or 'non-comparative'
for pattern ← pattern_ collection do
    if match(pattern, question) then
        return 'comparative'
    end
end
return 'non-comparative'
```

We used a set of comparative patterns for a binary classification task and obtained 44,170,477 comparative questions in the Yandex corpus, which amounted to 2.9% of all question queries. To evaluate classifier performance, we randomly sample 1000 classified examples and manually check the correctness of classification. We utilize standard performance measures such as accuracy, precision, recall, and F1. For the classifier performance evaluation we use the following terms to describe the results of the classifier predictions:

		Actual class	
Predicted class		Comparative	Non-comparative
	Comparative	True positive (TP)	False positive (FP)
	Non-comparative	False negative(FN)	True negative(TN)

Accuracy refers to how close overall predicted classes to actual or how accurately the classifier can distinguish between comparative and non-comparative questions in a binary classification task. Accuracy is a good measure only if the amounts of false positive and false negative predictions are almost same. It is calculated as

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

Precision refers to how close to each other measurements for each class or, in other words, how good the classifier performs in predicting comparative and non-comparative questions. Low precision usually indicates a high number of the false positive predictions. It is defined as

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

Recall describes how many of the truly comparative questions our classifier labeled. Low recall indicates many false negative predictions. It is calculated as

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

F1 score refers to the harmonic mean between the precision and the recall and is useful to use when classes are distributed unevenly. It is calculated as

$$F1\ score = 2 \times \frac{recall \times precision}{recall + precision}$$

Table 3.1 demonstrates our pattern-based classifier performance on the whole Yandex corpus based on 1000 randomly sampled and manually checked classification results.

Table 3.1: Pattern-based classifier performance evaluation on the Yandex corpus.

Performance measures, %	
Measurement	Value
Accuracy	72
Precision	44
Recall	100
F1 score	61

In the Otvety@Mail.ru, the pattern-based classifier labeled 852,652 questions as comparative, i.e., 11.1% in the corpus. The performance measures based on randomly sampled 1000 classified examples are slightly lower than those obtained on the Yandex dataset and shown in Table 3.2. Overall, the questions in Otvety@Mail.ru are more verbose and have weaker inner-sentence structure, which makes a process of matching with patterns more complex and less accurate.

Our pattern-based classifier demonstrates a very good coverage of all possible comparative questions producing a very low rate of false negative predictions, i.e., comparative questions, which are classified as non-comparative, and a high rate of false positive, which are non-comparative questions classified as comparative. This is reflected in a high recall and a rather low precision of the classification.

Table 3.2: Pattern-based classifier performance evaluation the Otvety@Mail.ru corpus.

Performance measures, %	
Measurement	Value
Accuracy	67
Precision	34
Recall	99
F1 score	51

We discuss in Section 3.3 a classification performance on a positive comparative class separately, i.e., how well the classifier is able to identify comparative questions.

3.3 Discussion

This chapter introduces a list of lexical and syntactic patterns used for identifying comparative questions in Russian. The collection of the patterns was built based on available corpora with questions submitted by users on the Russian web in the year 2012 representing a full variety of questions users can ask. Identification of comparative questions using hand-crafted rules can be quite accurate yet inefficient time-wise. Pattern-based classification gives a very high recall that guarantees all possible variations of comparative questions are covered; however, non-comparative questions are also very often classified as comparative. A combined confusion matrix of the pattern-based classifier results calculated on overall 2000 random samples of positively and negatively classified examples on both datasets Yandex and Otvety@Mail.ru is shown in Figure 3.4. The classifier gives a very low rate of false negative and relatively high rate of false positive predictions.

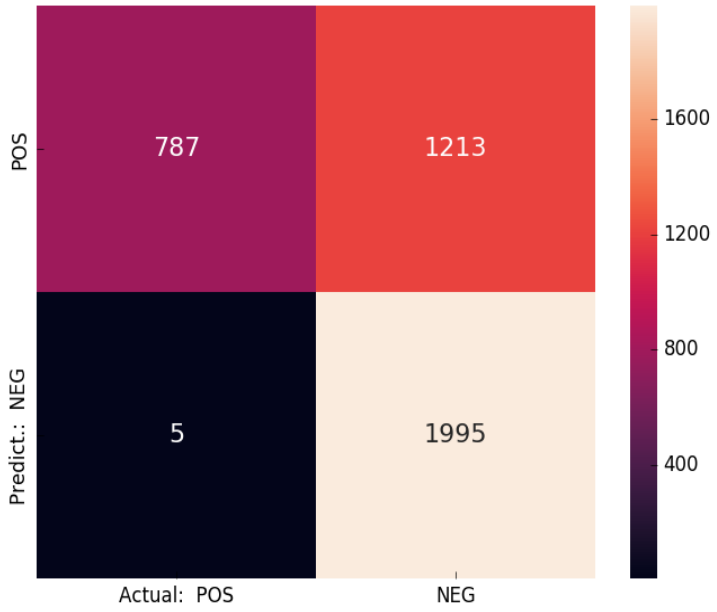

Figure 3.4: Confusion matrix of the combined Yandex and Otvety datasets pattern-based classification.

Nevertheless, a pattern-based identification is a good way to mine candidates for further application of machine learning algorithms and can be assumed as the first step in a two-level comparative question mining. Several examples of misclassifications are presented in Table 3.3. The misclassified examples show that the pattern-based comparative question classifier very well captures structural peculiarities of the comparatives as a linguistic phenomenon, yet cannot distinguish between comparative questions seeking for reasoning support and asking for facts or between direct and indirect comparative questions.

Table 3.3: Misclassified examples using the pattern-based classifier.

Misclassified examples	
Yandex	
False positives	False negatives
Какой материал для кухни лучше? **Which material is better for the kitchen?**	– –
Как правильно писать дверьми или дверями? **What is correct spelling doors or dors?**	– –
Otvety@Mail.ru	
False positives	False negatives
Очень важно! Поссорились с парнем очень серьёзно! Помиримся? Или всё, не будет больше отношений?! **Very important! Broke up with a boyfriend! Will it come over it? Or it's the end and there won't be relationship anymore?**	Что, прокладка или головка? ВАЗ 2114 **Layer pad or butt end? VAZ 2114.**
Посоветуйте, как лучше удалить установленную игру? **Advice, how is it better to remove the installed game?**	Кто, Путин или Медведев? **Putin or Medvedev?**

Chapter 4

Machine Learning-based Comparative Question Mining

Due to the sparsity and variability of natural language, the pattern-based classifier proposed in Chapter 3 can be satisfactory only to some extent since it gives a high recall and low precision. We also observe that it is rather difficult to distinguish comparative questions seeking reasoning support using only a pattern matching. Thus, exploiting machine learning is a logical next step in mining comparative questions. The present chapter introduces a supervised machine learning algorithm trained on a manually labeled dataset.

4.1 Building a Labeled Corpus

In order to teach computers to understand natural language and assign classes to text documents, it does not suffice to simply provide machines with a large amount of raw text. Availability of data examples manually labeled into different classes by human experts is crucial not only for building a classifier based on a supervised machine learning algorithm but also important and necessary for validation and evaluation of the classifier performance.

4.1.1 Dataset Preparation

We have at hand 1.5 billion records from the Yandex logs and 11.2 million entries from Otvety@Mail.ru from the year 2012. Both datasets were preprocessed as described in the research work by Völske et al. [2015] including filtering out queries in the form of questions, removing spam and bots, excluding repeated and non-unique questions. From the Otvety@Mail.ru corpus, we have extracted 7.7 million questions, which have an answer chosen by the user as the best. In order to prepare the data for annotation with classes, we,

first, have randomly sampled 1000 questions from each of both datasets. We have manually labeled questions and calculated that the number of comparative ones in the Yandex corpus is around 1–2% and approximately 4% in the Otvety@Mail.ru. As we aim to build a labeled dataset to train an automated machine learning classifier, we need to increase the ratio of comparative questions in the dataset, ideally making the number of positive and negative examples equal. At the second step, we use a collection of comparative patterns to mine comparative examples from the raw data. Our pattern-based approach gives a high recall, meaning that almost all possible variations of comparative questions are covered. At the same time, a lower accuracy guarantees to fill the dataset with difficult examples for classification. Random sampling, in turn, ensures mining questions free from predefined comparative patterns. In order to decrease the bias in the dataset, we also have mined examples, which exclude comparative patterns, applying *if-not-in-patterns* condition as shown in Figure 4.1. The final dataset ready for manual annotation contains 10,000 questions, where roughly equal amounts were mined from the Yandex and Otvety@Mail.ru datasets.

4.1.2 Data Labeling

Even though our pattern-based comparative question mining approach performs with a satisfactory accuracy, there still remain weaknesses such as difficulties with filtering out comparative questions, which seek for reasoning or argumentation support in answers. To solve this task, we exploit supervised machine learning trained on a manually labeled dataset. The labeling was performed by volunteers, native Russian speakers of different age, occupation and educational background who resided both in Germany and Russia. To prepare data for class annotation, we randomly shuffle the questions and group them into chunks of 100 examples each. The questions are collected in a table with two columns. The first column includes the question examples and the second is kept empty for assigning labels "yes" for comparative questions, "no" for non-comparative, "not clear" or "nc" for cases when annotators cannot decide which category to assign. We consider as comparative only the questions seeking for reasoning support, which are reasoning comparative questions according to the classification in Figure 3.2 and simultaneously direct comparative according to the classification in Figure 3.3. The questions, which are comparative from a purely linguistic point of view, i.e., containing syntactic and lexical comparative structures but do not seek for argumentation or reasoning in the answer and do not assume several objects being compared in the answer, have been asked to be labeled as non-comparative. The annotators are instructed to

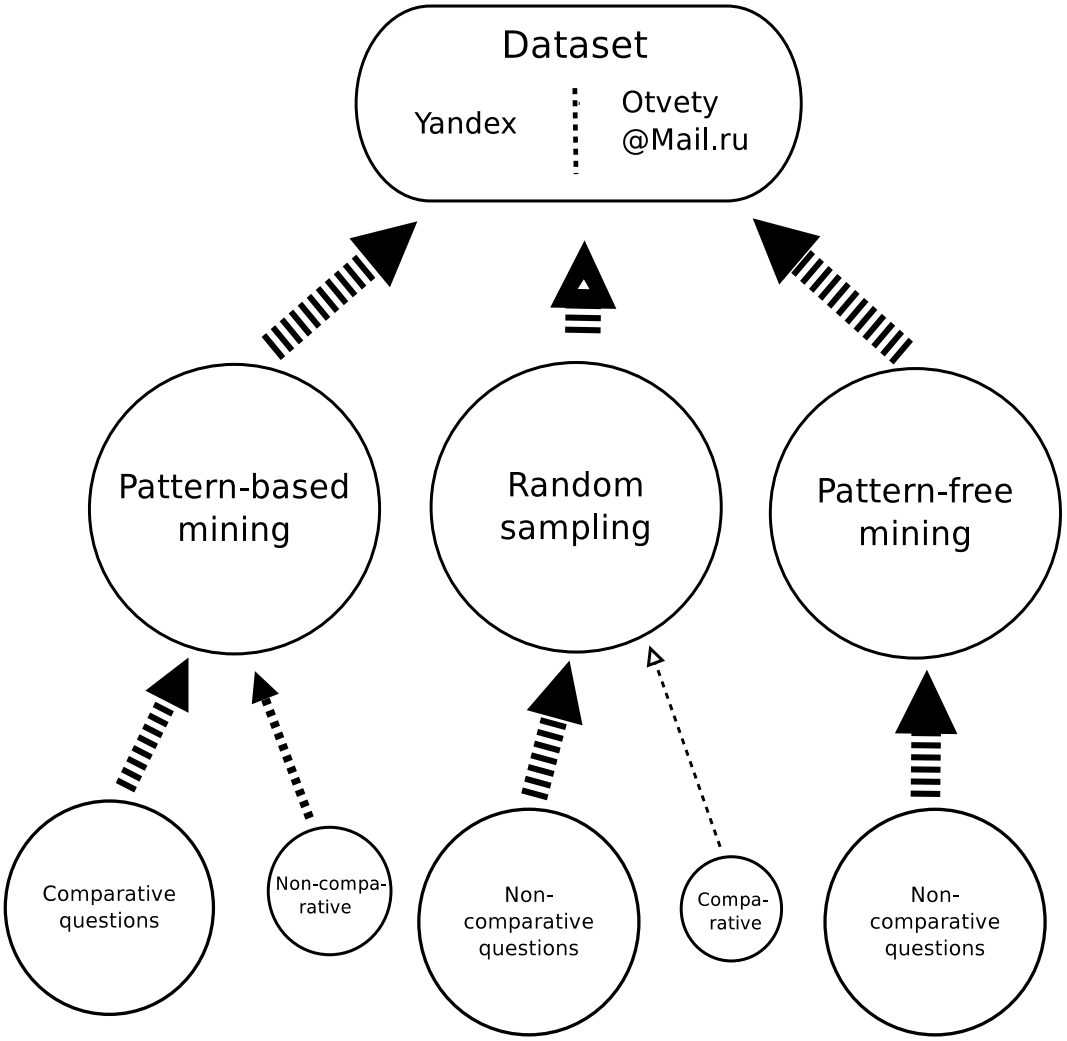

Figure 4.1: The process of mining questions for the labeled dataset.

label questions according to the following principle—if the question has an intent to compare two or more objects and the asker's expectation for reasoning, argumentative or opinionated support in the answer is clearly present in the question, this question has to be marked as comparative. Any other questions, which do not allow reasons, arguments, and opinion in the answer and rather can be answered with facts have to be labeled as non-comparative. We have provided the annotators with several complex examples as presented below.

Yes (comparative)	No (non-comparative)
Questions containing two or more objects or entities to be compared listed often (but not necessarily) with a conjunctive *or*.	Questions asking for correct spelling, for time differences or requesting for factual information.
Which phone is more reliable Samsung or iPhone?	**Which river is longer the Nile or the Amazon?**
Questions with superlative adjectives and adverbs with compared objects listed.	Questions with superlative adjectives and adverbs, which do not contain compared objects.
Who is the best soccer player Messi or Ronaldo?	**Who is the best soccer player in the world?**
Questions asking for differences or commonalities between the objects, which allow sharing opinion, experience, reasons and arguments in the answer.	Questions requesting information about differences or commonalities in the time zones, fixed definitions and concepts, currency rates and other subjects, which do not allow personal opinion, reasons or arguments.
What is the difference between lenses in canon d550 and d600?	**What is the difference between adverbs and adjectives?**
Questions seeking for advice to choose between the objects.	Questions asking for advice in general not assuming concrete options.
What is it better to go for vacation Turkey or Cyprus?	**Where is it better to go for vacation with children?**

When post-processing labeled examples, we inspected the entries marked as "not clear" and assigned either "comparative" or "non-comparative" class to the examples. The labeled datasets were merged together in a single class-annotated corpus. Overall, we obtained 35% of comparative and 65% of non-comparative labeled questions in the corpus.

4.2 Supervised Classification

Except for a manual text classification, where a decision whether a given example belongs to the one or other class is made by humans, and rule-based classification, where the rule establishes a combination of words to assign a class, there exist machine learning classification algorithms, where rules are automatically learned from training data. The first two approaches are inefficient and expensive for large amounts of text data. Differently, supervised machine learning can be seen as an imitation of humans in performing a text classification task at much higher speed and on much larger amounts of data.

4.2.1 Computational Model

Raw text cannot be directly used as an input for supervised learning. In order for the machine learning algorithm to learn effectively and efficiently from the labeled data, input text must be carefully and accurately annotated. Annotation of the text assumes converting raw text into machine-readable form by means of adding metadata. For example, Pustejovsky and Stubbs [2012] recommend a multi-stage annotation of the data over several cyclic iterations in order to achieve the highest classification results; this includes revising the annotation after testing and evaluating the classifier performance.

To perform annotation, we are interested in a morphological parser for Russian to do mainly part-of-speech tagging and lemmatization. Based on comparative studies of the most popular Russian parsers by Dereza et al. [2016] and Kotelnikov et al. [2017], we have chosen *MyStem* built by Segalovich [2003] and *pymorphy2* created by Korobov [2015] as very accurate. For example, according to the study by Kotelnikov et al. [2017], *pymorphy2* demonstrates a high accuracy for lemmatization and a high F1 measure for POS tagging. The package is written in Python and is an open source software. We use the parser for running experiments on the relatively small labeled dataset of 10,000 questions. *MyStem* provides a dictionary-based analysis and able to solve a disambiguation problem as well as infer a morphology for unknown words. According to evaluation in the study by Kotelnikov et al. [2017], *MyStem* demonstrates the best results for both lemmatization and POS tagging in comparison with other popular morphological parsers for Russian. We use *MyStem* to run experiments on the whole available corpora of more than 1.5 billion records on the cluster with 150 distributed nodes due to a large amount of data. The parser allows a console input – output and is more convenient for distributed computations on large amounts of data. Unfortunately, part-of-speech taggers for Russian tend to fail in recognizing words written with non-Cyrillic letters and assign a *none* tag to such words. We have found in total about 160 million

questions containing words written with Latin characters in the Yandex corpus, i.e., 11% of the total amount of questions, and approximately 0.5 million records in Otvety@Mail.ru, i.e., 7%. Mostly, they are brands and trademarks or original names of computer games, movies, books, personalities, celebrities etc. To overcome this, we have performed a character-by-character transliteration and tested the chosen parsers, which have tagged transliterated words as nouns. Several results of the part-of-speech tagging of the transliterated words from Latin into Cyrillic characters are presented below.

Original word	Transliteration	POS tag
iphone	ипхон	Noun
samsung	самсунг	Noun
nokia	нокиа	Noun

The Russian language has six grammatical cases and three genders, which make the endings of the verbs, adjectives, and nouns inflate. For example, the noun "mother" takes the following forms according to the cases.

Word	Case
мам**а**	Nominative
мам**ы**	Genitive
мам**е**	Dative
мам**у**	Accusative
мам**ой**	Prepositional
мам**е**	Instrumental

Considering single and plural forms of the nouns, each word takes in total 12 forms. Adjectives change their endings according to 3 genders, 2 numbers and 6 cases amounting to 36 possible variations. To avoid unnecessary excessive size of the vocabulary used in classification algorithm, we lemmatize each noun and adjective; the verbs are brought to the infinitive form. On the one hand, we aim to decrease the dependency of the model from the words themselves and increase the importance of the comparative form of the adverbs and adjectives on the other hand. To achieve this, we include in the annotation

algorithm a substitution of the comparative adjectives and adverbs by their part-of-speech-tag as presented in Algorithm 4.1.

Algorithm 4.1: Annotate questions in the dataset

```
Pos (Question)
    inputs : Questions in corpus
    output: Annotated questions
    Initialize sent ← ∅
    foreach word in question do
        if len(word) > 1 then
            if POS.tag(word) = COMP then
                sent ← (sent + 'COMP')
            end
            else
                sent ← (sent + lemma(word) + POS.tag(word)
            end
        end
    end
    return sent
```

Each annotated question contains a sequence of words followed by their POS tags; if the word is a comparative adjective or adverb then only a POS tag "COMP" is placed. The whole annotation pipeline includes the following steps.

- Punctuation removal and lowercasing of the words.
- Removal of the one-character words.
- Transliteration of the Latin characters into Cyrillic.
- Word lemmatization.
- Adding part-of-speech tags to the words and substitution of the comparative adjectives and adverbs by a POS tag.

An example of a few questions annotated with part-of-speech tags and used further to feed the classifier is presented below.

Original question	Annotated question
Что лучше iphone 8 или Samsung nokia?	что CONJ COMP ипхон NOUN или CONJ самсунг NOUN нокиа NOUN
Какая разница между canon d550 и canon d500?	какой ADJF разница NOUN между PREP цанон NOUN д550 NONE цанон NOUN д500 NONE

Feature Engineering

Choosing features for the text annotation to feed a classifier is as important as choosing a proper machine learning algorithm. We have compared performances of the linear support vector machine based on different sets of the features and show results in Table 4.1.

Table 4.1: Performance of the linear SVM trained with different feature sets.

	Performance measures			
Measure	Words, POS tags, subst. comp, unigrams	POS tags only	Words, POS tags, without subst. comp, unigrams	Words, POS tags, subst. comp, uni-, bi-trigrams
Accuracy	0.9259	0.7928	0.9223	**0.9261**
Precision	0.9176	0.7556	0.9144	**0.9187**
Recall	**0.9194**	0.7436	0.9145	0.9192
F1 score	0.9180	0.7870	0.9142	**0.9183**

The classifier performs surprisingly well already only on the part-of-speech tags without using actual words. We have also noticed that substitution of the comparative adverbs and adjectives as shown in Algorithm 4.1 gives only a slight improvement, yet can be useful in the classification of large amounts of unseen data, where with a high probability new words, which are not in the model's vocabulary, can occur. Even though a combination of uni-, bi- and trigrams together gives a tiny improvement, the drawback of these features is a huge vocabulary leading to the increase of the computational costs. Considering a trade-off between classifiers efficiency and effectiveness, we choose for the future experiments the feature model as presented in Algorithm 4.1, corresponding to the first column in Table 4.1.

4.2.2 Machine Learning Algorithm

Machine learning algorithms are assumed to substitute humans in many common tasks. For text classification, choosing a proper algorithm from many available is crucial. As previously discussed, we build a classifier based on a supervised learning approach. According to the review in Chapter 2, support vector machines and Naive Bayes are the most frequently used methods in text classification.

We address previous research done on evaluation of different approaches. In the article "Machine Learning in Automated Text Categorization", Sebastiani [2002] concludes that boosting-based classifiers, support vector machines and regression methods perform best in text classification. In another research work "Methods for Identifying Comparative Sentences", Saritha and Pateriya [2014] use a Naive Bayesian classifier, SVM, and Bayesian network. The study "Identifying Comparative Sentences in Text Documents" conducted by Jindal and Liu [2006a] reports exploiting both naive Bayesian classification and support vector machines and states that SVM performs unsatisfactorily. The book "Natural Language Processing with Python" written by Bird et al. [2009] recommends decision trees, Naive Bayes and Maximum Entropy classifiers for text classification. Based on common practices in the research community, we choose to train classifiers built on support vector machines with linear and RBF kernels, decision trees, Bernoulli Naive Bayes suitable for binary classification, logistic regression and gradient boosting. In order to evaluate classifier performance, we use a random split of the labeled dataset with 90% for training and 10% for testing. The question examples from the labeled dataset are represented as feature vectors using bag-of-words unigram features with a vocabulary built over the annotated data as described in Section 4.2.1. We evaluate classifiers based on a five-fold cross-validation using standard measures of accuracy, precision, recall, and F1. The performance evaluation results are collected in a tabular form and presented in Table 4.2.

Table 4.2: Comparison of the classifiers' performance on the labeled dataset.

	Performance measures					
Measure	SVM Linear	SVM RBF	Decision Tree	Bernoul. NB	Logistic Regres.	Grad. Boost
Accuracy	**0.9259**	0.9197	0.9051	0.9008	0.9241	0.9250
Precision	0.9176	0.9116	0.8953	0.9052	0.9165	**0.9207**
Recall	0.9194	**0.9117**	0.8982	0.8744	0.9165	0.9131
F1 score	**0.9180**	0.9115	0.8950	0.8867	0.9164	0.9167

All algorithms perform very well on the manually labeled 10,000 questions and their performances are very close to each other. However, in our experiments, support vector machines, logistic regression, and gradient boosting perform the best in the task of identifying comparative questions.

4.2.3 Classification Results

Our supervised machine learning-based comparative classifier trained on the annotated and manually labeled data is exploited to identify comparative questions in the whole available corpora with the questions submitted by users on the Russian web. We also investigate a distribution of comparative questions over different topical categories of the question queries. First, we use a question category classifier as described in the paper by Völske et al. [2015] and split the questions into 14 topic-like categories. Since support vector machine and linear regression algorithms have shown better results, we run both classifiers on the questions in the category "consumer electronics" in order to evaluate their performances and choose the most accurate method for classification of a large amount of text data. We randomly sample 100 positive and 100 negative predictions and calculate performance. Even though the support vector machine demonstrated a very high recall of 100%, its overall performance was rather poor with 33% F1 measure. Linear regression, in contrast, has shown much better performance as 85% F1 measure. Consequently, we choose linear regression to conduct classification on the whole Yandex and Otvety@Mail.ru corpora.

We have calculated a number of questions in each category and run the comparative classifier over the questions in the Yandex corpus to estimate the number of comparative examples. We arrange the results by comparative ratio—a proportion in percent of the comparative questions in each category—

in a descending order and collect them in Table 4.3. In total, the classifier identified 1.8% of comparative questions in the whole Yandex dataset (compare with 2.9% obtained by the pattern-based approach).

Table 4.3: Comparative questions in the Yandex corpus sorted by categories.

Yandex: 1,500,825,102 questions.			
Category	Questions in category	Comparative questions	Comparative ratio, %
education	98,450,656	3,134,609	3.2
society_culture	97,443,356	2,563,935	2.6
consumer_electronics	99,363,186	2,414,295	2.4
home_garden	170,496,347	4,056,252	2.3
health	129,091,266	2,654,259	2.1
family_relationships	72,357,503	1,467,107	2.0
cars_transportation	146,408,887	2,829,809	1.9
entertainment_music	94,118,056	1,603,967	1.7
sports	43,613,556	695,861	1.6
business_finance	137,092,094	2,104,212	1.5
beauty_style	98,075,280	1,312,472	1.3
adult	55,936,554	717,097	1.3
computers_internet	145,029,127	1,162,806	0.8
games_recreation	113,349,234	814,506	0.7
Total/Average		27,531,187	1.8

To evaluate classifier performance, we randomly sample 200 questions from each category separately, 100 from positively classified examples, i.e., comparative questions, and 100 from negative. We manually check all labels predicted by the classifier in order to estimate its performance. Due to a high cost of the manual processing, we evaluate a classifier performance on four categories. Table 4.4 shows the evaluation results.

Table 4.4: Classifier performance evaluation on the Yandex corpus for several categories.

Performance measures, %				
Category	Accuracy	Precision	Recall	F1 score
consumer_electronics	87	74	100	85
family_relationship	78	56	100	71
cars_transportation	68	36	97	53
computers_internet	66	33	100	50
Average	75	50	99	65

We then apply the comparative question classifier to the categories in the Otvety@Mail.ru. The classification results are presented in Table 4.5. One can see the difference in allocations of the comparative questions in the categories between the two datasets. The community-based question answering platform is more often used by users to ask comparative questions about electronic devices and cars. Users seek for advice or opinion from other humans when choosing goods to buy and, in contrast, educational questions or social and political questions are more often submitted to the search engine.

Table 4.5: Comparative questions in the Otvety@Mail.ru corpus sorted by categories.

Otvety@Mail.ru: 7,671,254 questions.			
Category	Questions in category	Comparative questions	Comparative ratio, %
consumer_electronics	280,688	17,133	6.1
cars_transportation	395,263	21,522	5.4
home_garden	565,664	25,766	4.6
sports	278,926	12,067	4.3
society_culture	943,082	39,796	4.2
education	363,226	14,744	4.1
adult	981,292	39,062	4.0

Continues on the next page

Table 4.5: Comparative questions in the Otvety@Mail.ru corpus sorted by categories (continued).

Otvety@Mail.ru: 7,671,254 questions.			
Category	Questions in category	Comparative questions	Comparative ratio, %
beauty_style	311,010	11,324	3.6
business_finance	409,096	13,580	3.3
health	419,388	13,039	3.1
family_relationships	1,250,761	37,987	3.0
entertainment_music	704,830	19,385	2.8
computers_internet	476,390	10,245	2.2
games_recreation	291,638	4625	1.6
Total/Average		280,275	3.7

Overall, the comparative classifier demonstrates a slightly worse performance on the Otvety@Mail.ru (see Table 4.6), which is similar to the pattern-based classification. This is due to the verbosity of the questions typical in the community question answering.

Table 4.6: Classifier performance evaluation on the Otvety@Mail.ru corpus for several categories.

Performance measures, %				
Category	Accuracy	Precision	Recall	F1 score
consumer_electronics	77	57	95	71
family_relationship	72	44	100	61
cars_transportation	69	39	98	56
computers_internet	68	36	97	53
Average	72	44	98	60

The averaged performance of the comparative classifier on both corpora Yandex and Otvety@Mail.ru is presented in Table 4.7.

Table 4.7: Overall averaged classifier performance evaluation on the web corpora.

Performance measures, %	
Measurement	Value
Accuracy	73
Precision	98
Recall	47
F1 score	63

4.3 Comparative Questions on the Russian Web

The classifier based on supervised machine learning identifies a total number of approximately 27.5 million as comparative questions in the Yandex corpus. As a reminder, out of all search engine logs from the year 2012, we consider only the queries submitted in the form of natural language questions; we consider as comparative only the questions seeking reasoning support and do not count factoid questions and questions with superlatives as well as indirectly comparative ones. In Section 4.4, we discuss in more detail the performance of the classifier on the comparative class only. According to the confusion matrix presented in Figure 4.2, only half of the questions classified as comparative are in fact reasoning comparative, which results in approximately 14 million examples or close to 1% in the Yandex corpus. This means that every two seconds, one comparative question is submitted to the Russian search engine, which most frequently occur in the query categories such as *education*, *society and culture*, and *consumer electronics* (see Table 4.3). Following the same methodology, we estimate a ratio of comparative questions asked on a community question answering platform as 1.6%, with the most frequent occurrences in the categories *consumer electronics*, *cars and transportation*, and *home and garden* (see Table 4.5). This means every minute, users ask a comparative reasoning question on the community question answering web platform.

We have analyzed the basic structure of the questions submitted to the search engine according to comparative and non-comparative classes and calculated occurrences of the question words. The five most frequent question words in the Yandex corpus are grouped in Table 4.8, which presents a relative ratio of questions containing a given question word.

Table 4.8: Question words distribution in questions in Yandex.

Comparative		Non-comparative	
Question word	Ratio, %	Question word	Ratio, %
Что **what**	33	Как **how**	50
Как **how**	23	Что **what**	11
Какой **which**	14	Какой **which**	7
Кто **who**	3	Сколько **how many/much**	5
Почему **why**	3	Где **where**	5

Half of the non-comparative questions start with *how*, thus we can assume *how*-questions including *how to*-questions are the most non-comparative, and, in contrast, the most comparative are *what*-questions.

4.4 Discussion

This chapter describes the construction of a comparative question classifier exploiting supervised machine learning trained on the manually annotated dataset with comparative and non-comparative classes. Creating a labeled dataset is another contribution of this work. So far, we have annotated 10,000 questions submitted on the Russian web. We introduced an effective computational model of the question representation to feed a machine learning classifier. The proposed classifier performed very well on the labeled dataset exceeding 91% precision, recall, and F1 measure. The pre-trained classifier was applied to classify questions in the whole Yandex and Otvety@Mail.ru corpora achieving 65% F1 measure in classifying questions in the Yandex corpora. We used supervised machine learning to solve a more complex task of identifying comparative questions seeking reasoning support, which could not be solved by simply matching questions with comparative patterns. The classifier covers all possible reasoning comparative questions making almost no false negative predictions (see Figure 4.2 and Figure 4.3). However, half of the predicted comparative questions in the Yandex corpora and 44% in Otvety@Mail.ru are in fact non-comparative.

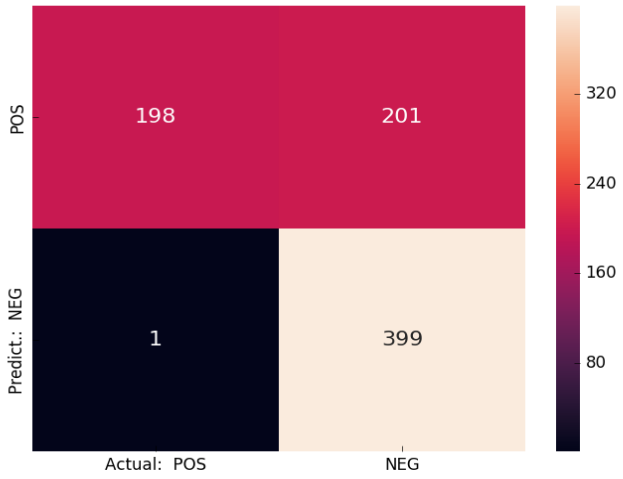

Figure 4.2: Confusion matrix for the machine learning classification of the Yandex dataset.

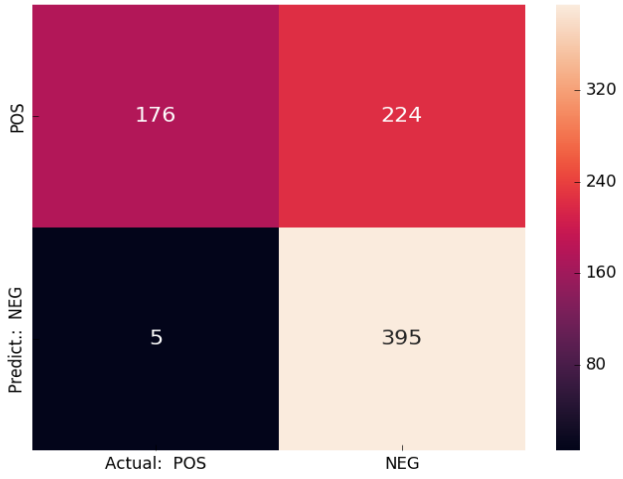

Figure 4.3: Confusion matrix for the machine learning classification of the Otvety@Mail.ru dataset.

We offer an approach to dealing with the rather high rate of false positive comparative predictions. The machine learning classification implementation

allows considering a confidence score of the classifier in making a decision when predicting classes. Any score values lower than 0.5 trigger a negative prediction and from 0.5 cause positive ones. We now establish a threshold of 0.6 as a decision boundary. Then, the classifier identifies 8.7 million questions as comparative in the Yandex data, from which only 25% are misclassified false positive matches denoting a 50% drop in false positive predictions (compare with data in Figure 4.2). We propose that a proper classifier confidence score threshold should be more thoroughly studied in future work in order to improve a comparative questions automated classifier. Here are several examples of false positive predictions:

1. Кто выиграет Россия или Португалия?
   Who will win Russia or Portugal?

2. В чем главное отличие переменных электрических и магнитных полей?
   What is the main difference between dynamic electrical and magnetic fields?

3. Чем старше женщина – тем больше преимуществ? Она мудрей, хитрее и нежней?
   The older woman the more advantages she has? She (is) wiser, slier and tenderer?

The first example is difficult for the classifier because its lexical and syntactic content is close to possibly comparative questions. The second is a comparative factoid questions. This is probably the most complex task to distinguish between reasoning and non-reasoning comparative questions. The third example simply contains five comparative adjectives, which contribute most in comparative prediction.

This chapter also studies a distribution of the comparative questions in different categories in the search engine logs and community question answering queries. The majority of comparative questions submitted to the Russian search engine belong to *education* and asked on the community questions answering platform—to *consumer electronics* categories. Three most frequent question words, which occur in comparative questions, are *what*, *how* and *which*.

Chapter 5

Conclusion

We have discussed a research problem of identification of comparative questions in the Russian language based on the queries submitted to Yandex, the biggest search engine for a Russian speaking audience, and community question answering website Otvety@Mail.ru. We consider comparatives in questions both as a linguistic phenomenon as well as user intent for comparing objects explicitly with reasoning support. In order to identify comparatives seen from a purely linguistic perspective, we search for strong textual signals in questions in the form of lexical and syntactic patterns in questions. When trying to capture a user information need encoded in a given question to receive objects comparison supported by opinions, arguments, reasons and human experience, we exploit supervised machine learning.

Research Focus and Main Findings

In order to establish a research path, we defined three research questions at the beginning of this thesis. The first triggered a search for textual patterns in order to identify comparative questions in Russian. We created a collection of lexical and syntactic patterns to perform a classification of a large amount of text data containing question queries submitted on the Russian web. Our pattern-based approach introduced in Chapter 3 performed satisfyingly well in the binary classification into comparative and non-comparative questions. However, the limitation of the approach is its inability to go beyond the perception of comparison as a purely linguistic phenomenon. Nevertheless, we can be sure that such textual patterns exist that can help in recognizing comparisons in texts. We also identified different types of comparative questions and proposed hierarchies, e.g., direct and indirect, reasoning and non-reasoning comparative.

We redirected the task of capturing user intent for reasoning support in

comparative questions to a supervised machine learning problem, where human experts were employed to provide a dataset annotated with classes, from which an automated machine classifier can learn a classification function. Chapter 4 addresses the second research question about building an effective comparative question classifier. The chapter first provides a methodology for creating a class-annotated dataset with comparative and non-comparative questions in the Russian language. Second, it proposes a supervised machine learning based classification model pre-trained on a manually labeled dataset. And finally, Chapter 4 introduces a method to classify a large dataset with more than 1.5 billion questions into comparative and non-comparative classes. The main challenge for the machine learning approach is to capture user intent for reasoning support in comparative questions. A binary classification of the large corpora performed very well in terms of standard classifier measures. However, a still high false positive prediction rate is the classifier's main weakness. We propose a methodology of setting a threshold for the classifier's confidence rate to significantly—as it has been confirmed experimentally—reduce false positive predictions and increase the classifier's reliability. Section 4.3 also answers the third research question and identifies that every two seconds users typed a comparative question in the Russian search engine Yandex in the year 2012, resulting in 1% of all question queries submitted. We believe that the overall ratio of question queries including comparative questions on the web has increased by today.

Future Work

Future works in several research areas named in Chapter 1, we believe, can benefit from the results of this thesis. Our proposed approaches can help better understand user intents when they submit question queries on the web. Moreover, a request for comparison when correctly and properly recognized can be immediately satisfied by a search engine with a proper output on the results page. Throughout our study, we saw that comparative questions possess distinct textual patterns that make it possible to identify such questions. Moreover, we found that supervised machine learning is able to successfully capture a user need in reasoning. To improve classifier performance, we must firstly enhance the feature representation of text by including more complex features reflecting syntactical and semantical relations between linguistic units in reasoning comparative questions. Secondly, a carefully thought-out inclusion of the classifier's degree of confidence will improve the quality of comparative questions identification. And lastly, an improvement can be achieved by enriching the training dataset with diverse examples.

We have made first observations of the specific type of user need in comparing objects supported by arguments, which they submit on the web. However, many possible directions in the research around this problem still remain.

Appendix A

Comparative Patterns for Pattern-based Mining

Table A.1: Comparative patterns and statistics in the Yandex corpus.

Yandex: 1,500,825,102 questions			
Pattern in Russian	In English	Number of hits	Comparative ratio, %
Отлич(аться)[1] & и|от|или|между	Differentiate|distinguish & and|from|or|between	5,556,748	100
Что|Кто лучше[2], exclude или[3]	What|Who better, exclude or	466,343	100
Различ(аться)[4] & и|от|или|между	Distinct & and|from|or|between	455,828	98

Continues on the next page

[1]Stem "отлич" is used. This pattern covers "отличие" ("difference", Noun) and also conjugated forms of "отличаться" ("differ", Verb).

[2]Indirect type of comparison, in which objects for comparison are not specified unambiguously.

[3]Questions with "или" (or) are excluded.

[4]Stem "различ" is used. This pattern covers "различие" ("distinction", Noun) and also conjugated forms of "различаться" ("distinguish", Verb).

Table A.1: Comparative patterns and statistics in the Yandex corpus (continued).

Yandex: 1,500,825,102 questions			
Pattern in Russian	In English	Number of hits	Comparative ratio, %
Преимуществ(о)[5] | недостат(ок) & перед|над|сравнени(е)	Advantage|flaw & over	34,308	98
COMP ADV|ADJ & или	COMP ADV|ADJ & or	5,513,516	97
Как(ой) & COMP ADV|ADJ, exclude или[2]	Which & COMP ADV|ADJ, exclude or	8,978,548	97
Что общего|сходство| схож(и)	What common| similar	336,321	95
Как правильно| пишется & или[6]	How correct| to write & or	849,555	94
Плюс(ы) & минус(ы)	Plus(es) & minus(es)	15,249	94
Разница & от|между|и|или[7]	Difference & from|between|and|or	600,252	93
Выбрать|купить|взять & или|между	Choose|buy|take & or|between	627,900	90
Сам(ый)|наиболее & ADV|ADJ, exclude лучший[2]	Most & ADV|ADJ, exclude best	2,841,788	83
5-6 words[8] & COMP ADV|ADJ, exclude или	5-6 words & COMP ADV|ADJ, exclude or	7,880,305	81

Continues on the next page

[5]Excluding endings of the nouns; assumed inclusion of all inflected noun forms.

[6]Non-reasoning type of comparison, asking for facts, correct spelling, time differences and so on.

[7]Records with "время" or "часовой" (e.g. "Какая разница во времени между Москвой и Киевом?" - "What is time difference between Moscow and Kiev?") are excluded.

[8]We assume the length the question length is 5 to 6 words.

Table A.1: Comparative patterns and statistics in the Yandex corpus (continued).

Yandex: 1,500,825,102 questions			
Pattern in Russian	In English	Number of hits	Comparative ratio, %
Почему & чем	Why & than	271,968	81
Когда|куда & ADV|ADJ comp., exclude или[2]	When|Whereto & ADV|ADJ comp., exclude or	431,223	73
5-6 words & или	5-6 words & or	4,946,512	71
SUPERL ADV|ADJ, exclude как[2]	SUPERL ADV|ADJ, exclude how	2,295,896	61
В сравнении| по сравнению	In comparison	74,625	55
Или	Or	12,205,139	53

Table A.2: Comparative patterns and statistics in the Otvety@Mail.ru corpus.

Otvety@Mail.ru: 7,671,254 questions			
Pattern in Russian	In English	Number of hits	Comparative ratio, %
Плюс(ы) & минус(ы)	Plus(es) & minus(es)	2821	99
Что|Кто лучше[9], exclude или	What|Who better, exclude or	8064	97
Преимуществ(о)[10] | недостат(ок) & перед|над|сравнени(е)	Advantage|flaw & over	761	97
Различ(аться)[11] & и|от|или|между	Distinct & and|from|or|between	5098	95
Отлич(аться)[12] & и|от|или|между	Differentiate|distinguish & and|from|or|between	51,946	91
Как правильно| пишется & или[13]	How correct| to write & or	6349	87
Что общего|сходство| схож(и)	What common| similar	4860	86
COMP ADV|ADJ & или	COMP ADV|ADJ & or	112,891	85
Разница & от|между|и|или[14]	Difference & from|between|and|or	10,741	85

Continues on the next page

[9]Indirect type of comparison, in which objects for comparison are not specified unambiguously.

[10]Excluding endings of the nouns; assumed inclusion of all inflected noun forms.

[11]Stem "различ" is used. This pattern covers "различие" ("distinction", Noun) and also conjugated forms of "различаться" ("distinguish", Verb).

[12]Stem "отлич" is used. This pattern covers "отличие" ("difference", noun) and also conjugated forms of "отличаться" ("differ", verb).

[13]Non-reasoning type of comparison, asking for facts, correct spelling, time differences and so on.

[14]Questions with "время" or "часовой" (e.g. "Какая разница во времени между Москвой и Киевом?" - "What is time difference between Moscow and Kiev?") are excluded.

Table A.2: Comparative patterns and statistics in the Otvety@Mail.ru corpus (continued).

Otvety@Mail.ru: 7,671,254 questions			
Pattern in Russian	In English	Number of hits	Comparative ratio, %
Как(ой) & COMP ADV|ADJ, exclude или[9]	Which & COMP ADV|ADJ, exclude or	63,887	72
В сравнении| по сравнению	In comparison	1972	69
Выбрать|купить|взять & или|между	Choose|buy|take & or|between	20,834	68
Почему & чем	Why & than	16,985	68
Сам(ый)|наиболее & ADV|ADJ, exclude лучший[9]	Most & ADV|ADJ, exclude best	66,895	51
Когда|куда & COMP ADV|ADJ, exclude или[9]	When|Whereto & COMP ADV|ADJ, exclude or	4227	51
SUPERL ADV|ADJ, exclude как[9]	SUPERL ADV|ADJ, exclude how	54,246	29

Bibliography

Enrique Alfonseca, Marco De Boni, José-Luis Jara-Valencia, and Suresh Manandhar. A prototype question answering system using syntactic and semantic information for answer retrieval. In *Proceedings of TREC 2001*.

Omid Bakhshandeh Babarsad. *Language Learning Through Comparison*. PhD thesis, University of Rochester, 2017.

Sigrid Beck, Sveta Krasikova, Daniel Fleischer, Remus Gergel, Stefan Hofstetter, Christiane Savelsberg, John Vanderelst, and Elisabeth Villalta. Crosslinguistic variation in comparison constructions. *Linguistic Variation Yearbook*, 9(1):1–66, 2009.

Steven M. Beitzel, Eric C. Jensen, David D. Lewis, Abdur Chowdhury, and Ophir Frieder. Automatic classification of web queries using very large unlabeled query logs. *ACM Trans. Inf. Syst.*, 25(2):9, 2007.

Polina Berezovsakaya and Vera Hohaus. The crosslinguistic inventory of phrasal comparative operators: Evidence from Russian. In *Proceedings of FASL 2015*.

Polina Berezovskaya. Acquisition of Russian comparison constructions: Semantics meets first language acquisition. In *Proceedings of ConSOLE 2013*, pages 45–65.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009.

Andrei Z. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

Andrei Z. Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. Robust classification of rare queries using web knowledge. In *Proceedings of SIGIR 2007*, pages 231–238.

John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, Ralph Weischedel. Issues, tasks and program structures to roadmap research in question & answering (Q&A). In *Roadmapping Documents of DUC 2001*, pages 1–35.

Andrew Carlson, Chad Cumby, Jeff Rosen, and Dan Roth. The SNoW learning architecture. Technical report, UIUCDCS, 1999.

Karol Chia-Tien Chang, Yu-Hsuan Wu, Yi-Lin Tsai, and Richard Tzong-Han Tsai. Improving iUnit retrieval with query classification and multi-aspect iUnit scoring: The IISR system at NTCIR-11 mobileclick task. In *Proceedings of NTCIR 2014*.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of AAAI 2016*, pages 2580–2586.

O. V. Dereza, D. A. Kayutenko, and A. S. Fenogenova. Automatic morphological analysis for Russian: A comparative study. In *Proceedings of Dialogue 2016*.

Marcelo Fiszman, Dina Demner-Fushman, Francois M. Lang, Philip Goetz, and Thomas C. Rindflesch. Interpreting comparative constructions in biomedical text. In *Biological, translational, and clinical language processing*, pages 137–144, 2007.

Carol Friedman. A general computational treatment of the comparative. In *Proceedings of ACL 1989*, pages 161–168.

Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of COLING 2008*, pages 241–248.

Arthur C. Graesser, Kathy Lang, and Dianne Horgan. A taxonomy for question generation. *Questioning Exchange*, 2(1):3–15, 1988.

Yeong Hyeon Gu and Seong Joon Yoo. Searching a best product based on mining comparison sentences. In *Proceedings of SCIS & ISIS 2010*, pages 929–933.

Ido Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of SIGIR 2016*, pages 35–44.

Bernard J. Jansen and Danielle L. Booth. Classifying web queries by topic and user intent. In *Proceedings of CHI 2010*, pages 4285–4290.

Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. In *Proceedings of SIGIR 2006*, pages 244–251 a.

Nitin Jindal and Bing Liu. Mining comparative sentences and relations. In *Proceedings of AAAI-06 and IAAI-06*, pages 1331–1336 b.

Elena Dmitrievna Kallestinova. *Aspects of word order in Russian*. PhD thesis, The University of Iowa, 2007.

In-Ho Kang and Gil-Chang Kim. Query type classification for web document retrieval. In *Proceedings of SIGIR 2003*, pages 64–71.

Wiltrud Kessler and Jonas Kuhn. Detection of product comparisons – How far does an out-of-the-box semantic role labeling system take you? In *Proceedings of EMNLP 2013*, pages 1892–1897.

Wiltrud Kessler and Jonas Kuhn. Detecting comparative sentiment expressions - A case study in annotation design decisions. In *Proceedings of Konvens 2014*, pages 165–170.

Mikhail Korobov. Morphological analyzer and generator for Russian and Ukrainian languages. In *Proceedings of AIST 2015*, pages 320–332.

Evgeny Kotelnikov, Elena Razova, and Irina Fishcheva. A close look at Russian morphological parsers: Which one is the best? In *Proceedings of AINL 2017*, pages 131–142.

Werner Krauth and Marc Mézard. Learning algorithms with optimal stability in neural networks. *Journal of Physics A: Mathematical and General*, 20(11):L745, 1987.

Thomas W. Lauer and Eileen Peacock. An analysis of comparison questions in the context of auditing. *Discourse Processes*, 13(3):349–361, 1990.

Wendy G. Lehnert. The process of question answering. Technical report, Yale University, Dept. of Computer Science, 1977.

Wendy G. Lehnert. *The process of question answering: A computer simulation of cognition.* Lawrence Erlbaum Associates, 1978.

Shasha Li, Chin-Yew Lin, Young-In Song, and Zhoujun Li. Comparable entity mining from comparative questions. In *Proceedings of ACL 2010*, pages 650–658.

Xin Li and Dan Roth. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249, 2006.

Chengxiang Liu, Ruifeng Xu, Jie Liu, Peng Qu, He Wang, and Chengtian Zou. Comparative opinion sentences identification and elements extraction. In *Proceedings of ICMLC 2013*, pages 1886–1891.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. Representation learning using multi-task deep neural networks for semantic classification and Information Retrieval. In *Proceedings of NAACL 2015*, pages 912–921.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

Friederike Moltmann. *Coordination and Comparatives.* PhD thesis, Massachusetts Institute of Technology, 1992.

Bo Pang and Ravi Kumar. Search in the lost sense of "query": Question formulation in web search queries and its temporal changes. In *Proceedings of ACL 2011*, pages 135–140.

Dae Hoon Park and Catherine Blake. Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the 2012 Workshop on Detecting Structure in Scholarly Discourse*, pages 1–9.

John M. Prager, Dragomir R. Radev, Eric W. Brown, Anni Coden, and Valerie Samn. The use of predictive annotation for question answering in TREC8. In *Proceedings of TREC 1999*.

James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning - A Guide to Corpus-Building for Applications*. O'Reilly, 2012.

S.K. Saritha and R.K. Pateriya. Methods for identifying comparative sentences. *International Journal of Computer Applications*, 108(19):23–26, 2014.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

Ilya Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *Proceedings of MLMTA 2003*, pages 273–280.

Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Q^2C@UST: Our winning solution to query classification in KDD Cup. *SIGKDD Explorations*, 7(2):100–110, 2005.

Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. In *Proceedings of SIGIR 2006*, pages 131–138.

Lulin Shi, Shi Li, Ping Jiang, and Hongsen Liu. Improving comparative sentence extraction of Chinese product reviews by sentiment analysis. *Journal of Engineering Science & Technology Review*, 9(6), 2016a.

Yangyang Shi, Kaisheng Yao, Le Tian, and Daxin Jiang. Deep LSTM based feature mapping for query classification. In *Proceedings of NAACL 2016*, pages 1501–1511 b.

Amit Singhal, Steven P. Abney, Michiel Bacchiani, Michael Collins, Donald Hindle, and Fernando C. N. Pereira. At&T at TREC-8. In *Proceedings of TREC 1999*.

Thabet Slimani and Amor Lazzez. Efficient analysis of pattern and association rule mining approaches. *CoRR*, abs/1402.2892, 2014.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383, 2011.

Maksim Tkachenko and Hady Wirawan Lauw. Generative modeling of entity comparisons in text. In *Proceedings of CIKM 2014*, pages 859–868.

Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327, 1977.

Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. What users ask a search engine: Analyzing one billion Russian question queries. In *Proceedings of CIKM 2015*, pages 1571–1580.

Hongwei Wang, Song Gao, Pei Yin, and James Nga-Kwok Liu. Competitiveness analysis through comparative relation mining: Evidence from restaurants' online reviews. *Industrial Management and Data Systems*, 117(4):672–687, 2017.

Wei Wang, TieJun Zhao, GuoDong Xin, and YongDong Xu. Exploiting machine learning for comparative sentences extraction. *International Journal of Hybrid Information Technology*, 8(3):347–354, 2015.

Wei Wang, TieJun Zhao, and GuoDong Xin. Learning extraction of chinese comparative sentences for evaluative text. *International Journal of Grid and Distributed Computing*, 9(3):53–62, 2016.

Ryen W. White, Matthew Richardson, and Wen-tau Yih. Questions vs. queries in informational search tasks. In *Proceedings of WWW 2015*, pages 135–136.

W. A. Woods. Progress in natural language understanding: An application to lunar geology. In *Proceedings of AFIPS 1973*, pages 441–450.

Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. Mining comparative opinions from customer reviews for competitive intelligence. *Decision Support Systems*, 50(4):743–754, 2011.

Seon Yang and Youngjoong Ko. Extracting comparative sentences from Korean text documents using comparative lexical patterns and machine learning techniques. In *Proceedings of IJCNLP 2009*, pages 153–156.

Natalia Zevakhina and Svetlana Dzhakupova. Russian metalinguistic comparatives: A functional perspective. HSE Working papers WP BRP 39/LNG/2015, National Research University Higher School of Economics, 2015.

Harry Zhang. The optimality of Naive Bayes. In *Proceedings of FLAIRS 2004*, pages 562–567.