Frequently Asked Questions Retrieval for Croatian Based on Semantic Textual Similarity

Vanja Mladen Karan* **Lovro Žmak†** **Jan Šnajder***

*University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia

†Studio Artlan, Andrije Štangera 18, 51410 Opatija, Croatia
{mladen.karan,jan.snajder}@fer.hr lovro.zmak@studioartlan.hr

Abstract

Frequently asked questions (FAQ) are an efficient way of communicating domain-specific information to the users. Unlike general purpose retrieval engines, FAQ retrieval engines have to address the lexical gap between the query and the usually short answer. In this paper we describe the design and evaluation of a FAQ retrieval engine for Croatian. We frame the task as a binary classification problem, and train a model to classify each FAQ as either relevant or not relevant for a given query. We use a variety of semantic textual similarity features, including term overlap and vector space features. We train and evaluate on a FAQ test collection built specifically for this purpose. Our best-performing model reaches 0.47 of mean reciprocal rank, i.e., on average ranks the relevant answer among the top two returned answers.

1 Introduction

The amount of information available online is growing at an exponential rate. It is becoming increasingly difficult to navigate the vast amounts of data and isolate relevant pieces of information. Thus, providing efficient information access for clients can be essential for many businesses. Frequently asked questions (FAQ) databases are a popular way to present domain-specific information in the form of expert answers to users questions. Each FAQ consists of a question and an answer, possibly complemented with additional metadata (e.g., keywords). A FAQ retrieval engine provides an interface to a FAQ database. Given a user query in natural language as input, it retrieves a ranked list of FAQs relevant to the query.

FAQ retrieval can be considered half way between traditional document retrieval and question answering (QA). Unlike in full-blown QA, in FAQ retrieval the questions and the answers are already extracted. On the other hand, unlike in document retrieval, FAQ queries are typically questions and the answers are typically much shorter than documents. While FAQ retrieval can be approached using simple keyword matching, the performance of such systems will be severely limited due to the *lexical gap* – a lack of overlap between the words that appear in a query and words from a FAQ pair. As noted by Sneiders (1999), there are two causes for this. Firstly, the FAQ database creators in general do not know the user questions in advance. Instead, they must guess what the likely questions would be. Thus, it is very common that users' information needs are not fully covered by the provided questions. Secondly, both FAQs and user queries are generally very short texts, which diminishes the chances of a keyword match.

In this paper we describe the design and the evaluation of a FAQ retrieval engine for Croatian. To address the lexical gap problem, we take a supervised learning approach and train a model that predicts the relevance of a FAQ given a query. Motivated by the recent work on semantic textual similarity (Agirre et al., 2012), we use as model features a series of similarity measures based on word overlap and semantic vector space similarity. We train and evaluate the model on a FAQ dataset from a telecommunication domain. On this dataset, our best performing model achieves 0.47 of mean reciprocal rank, i.e., on average ranks the relevant FAQ among the top two results.

In summary, the contribution of this paper is twofold. Firstly, we propose and evaluate a FAQ retrieval model based on supervised machine learning. To the best of our knowledge, no previ-

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 24–33,
Sofia, Bulgaria, 8-9 August 2013.

ous work exists that addresses IR for Croatian in a supervised setting. Secondly, we build a freely available FAQ test collection with relevance judgments. To the best of our knowledge, this is the first IR test collection for Croatian.

The rest of the paper is organized as follows. In the next section we give an overview of related work. In Section 3 we describe the FAQ test collection, while in Section 4 we describe the retrieval model. Experimental evaluation is given in Section 5. Section 6 concludes the paper and outlines future work.

2 Related Work

Most prior work on FAQ retrieval has focused on the problem of lexical gap, and various approaches have been proposed for bridging it. Early work, such as Sneiders (1999), propose to manually enrich the FAQ databases with additional meta data such as the required, optional, and forbidden keywords and keyphrases. This effectively reduces FAQ retrieval to simple keyword matching, however in this case it is the manually assigned metadata that bridges the lexical gap and provides the *look and feel* of semantic search.

For anything but a small-sized FAQ database, manual creation of metadata is tedious and cost intensive, and in addition requires expert knowledge. An alternative is to rely on general linguistic resources. FAQ finder (Burke et al., 1997) uses syntax analysis to identify phrases, and then performs matching using shallow lexical semantic knowledge from WordNet (Miller, 1995). Yet another way to bridge the lexical gap is smoothing via clustering, proposed by Kim and Seo (2006). First, query logs are expanded with word definitions from a machine readable dictionary. Subsequently, query logs are clustered, and query similarity is computed against the clusters, instead of against the individual FAQs. As an alternative to clustering, query expansion is often used to perform lexical smoothing (Voorhees, 1994; Navigli and Velardi, 2003).

In some domains a FAQ engine additionally must deal with typing errors and noisy user-generated content. An example is the FAQ retrieval for SMS messages, described by Kothari et al. (2009) and Contractor et al. (2010).

Although low lexical overlap is identified as the primary problem in FAQ retrieval, sometimes it is the high lexical overlap that also presents a problem. This is particularly true for large FAQ databases in which a non-relevant document can "accidentally" have a high lexical overlap with a query. Moreo et al. (2012) address the problem of false positives using case based reasoning. Rather than considering only the words, they use phrases ("differentiator expressions") that discriminate well between FAQs.

The approaches described so far are essentially unsupervised. A number of supervised FAQ retrieval methods have been described in the literature. To bridge the lexical gap, Xue et al. (2008) use machine translation models to "translate" the user query into a FAQ. Their system is trained on very large FAQ knowledge bases, such as Yahoo answers. Soricut and Brill (2004) describe another large-scale FAQ retrieval system, which uses language and transformation models. A good general overview of supervised approaches to ranking tasks is the work by Liu (2009).

Our system falls into the category of supervised methods. In contrast to the above-described approaches, we use a supervised model with word overlap and semantic similarity features. Taking into account that FAQs are short texts, we use features that have been recently proposed for determining the semantic similarity between pairs of sentences (Šarić et al., 2012). Because we train our model to output a relevance score for each document, our approach is essentially a *pointwise* learning-to-rank approach (Qin et al., 2008).

3 Croatian FAQ test collection

The standard procedure for IR evaluation requires a test collection consisting of documents, queries, and relevance judgments. We additionally require an annotated dataset to train the model. As there currently exists no standard IR test collection for Croatian, we decided to build a FAQ test collection from scratch. We use this collection for both model training and retrieval evaluation.

To obtain a FAQ test collection, we crawled the web FAQ of Vip,[1] a Croatian mobile phone operator. For each FAQ, we retrieved both the question and the answer. In the Vip FAQ database questions are categorized into several broad categories (e.g., by type of service). For each FAQ, we also extract the category name assigned to it. We obtained a total of 1344 FAQs. After removing the

[1] `http://www.vipnet.hr/pitanja-i-odgovori/` (accessed Sep 2009)

Query	FAQ question	FAQ answer
Kako se spaja na internet? (*How to connect to the internet?*)	Što mi je potrebno da bih spojio računalo i koristio se internetom? (*What do I need to connect my computer and use the internet*)	Morate spojiti računalo sa Homebox uređajem LAN kabelom. . . (*You must connect your computer to the Homebox device using a LAN cable . . .*)
Putujem izvan Hrvatske i želim koristiti svoj Vip mobilni uređaj. Koliko će me to koštati? (*I am traveling abroad and want to use my Vip mobile device. How much will this cost?*)	Koja je mreža najpovoljnija za razgovore, a koja za slanje SMS i MMS poruka u roamingu? (*Which network is the best for conversations, and which one for SMS and MMS messages in roaming?*)	Cijene za odlazne pozive u inozemstvu su najpovoljnije u mrežama Vodafone partnera. . . (*Outgoing calls cost less on networks of Vodafone partners . . .*)
Kako pogledati e-mail preko mobitela? (*How to check e-mail using a mobile phone?*)	Koja je cijena korištenja BlackBerry Office usluge? (*What is the price of using the BlackBerry Office service?)*	. . . business e-mail usluga uračunata je u cijenu. . . (. . . *business e-mail is included in the price . . .*)

Table 1: Examples of relevant answers to queries from the dataset

duplicates, 1222 unique FAQ pairs remain.

Next, we asked ten annotators to create at least twelve queries each. They were instructed to invent queries that they think would be asked by real users of Vip services. To ensure that the queries are as original as possible, the annotators were not shown the original FAQ database. Following Lytinen and Tomuro (2002), after creating the queries, the annotators were instructed to rephrase them. We asked the annotators to make between three and ten paraphrases of each query. The paraphrase strategies suggested were the following: (1) turn a query into a multi-sentence query, (2) change the structure (syntax) of the query, (3) substitute some words with synonyms, while leaving the structure intact, (4) turn the query into a declarative sentence, and (5) any combination of the above. The importance of not changing the underlying meaning of a query was particularly stressed.

The next step was to obtain the binary relevance judgments for each query. Annotating relevance for the complete FAQ database is not feasible, as the total number of query-FAQ pairs is too large. On the other hand, not considering some of the FAQs would make it impossible to estimate recall. A feasible alternative is the standard pooling method predominantly used in IR evaluation campaigns (Voorhees, 2002). In the pooling method, the top-k ranked results of each evaluated system are combined into a single list, which is then annotated for relevance judgments. For a sufficiently large k, the recall estimate will be close to real recall, as the documents that are not in the pool are likely to be non-relevant. We simulate this setting using several standard retrieval models: keyword search, phrase search, tf-idf, and language modeling. The number of combined results per query is between 50 and 150. To reduce the annotators' bias towards top-ranked examples, the retrieved results were presented in random order. For each query, the annotators gave binary judgments ("relevant" or "not relevant") to each FAQ from the pooled list; FAQs not in the pool are assumed to be not relevant. Although the appropriateness of binary relevance has been questioned (e.g., by Kekäläinen (2005)), it is still commonly used for FAQ and QA collections (Wu et al., 2006; Voorhees and Tice, 2000). Table 1 shows examples of queries and relevant FAQs.

The above procedure yields a set of pairs (Q_r, F_{rel}), where Q_r is a set of query paraphrases and F_{rel} is the set of relevant FAQs for any query paraphrase from Q_r. The total number of such pairs is 117. From this set we generate a set of pairs (q, F_{rel}), where $q \in Q_r$ is a single query. The total number of such pairs is 419, of which 327 have at least one answer ($F_{rel} \neq \emptyset$), while 92 are not answered ($F_{rel} = \emptyset$). In this work we focus on optimizing the performance on answered queries and leave the detection and handling of unanswered queries for future work. The average number of relevant FAQs for a query is 1.26, while on average each FAQ is relevant for 1.44 queries. Test collection statistics is shown in Table 2. We make the test collection freely available for research purposes.[2]

For further processing, we lemmatized the query and FAQ texts using the morphological lexicon from Šnajder et al. (2008). We removed the stopwords using a list of 179 Croatian stopwords.

[2]Available under CC BY-SA-NC license from `http://takelab.fer.hr/faqir`

	Word counts			Form	
	Min	Max	Avg	Quest.	Decl.
Queries	1	25	8	372	47
FAQ questions	4	63	7	287	4
FAQ answers	1	218	30	–	–

Table 2: FAQ test collection statistics

We retained the stopwords that constitute a part of a service name (e.g., the pronoun *"me"* (*"me"*) in *"Nazovi me"* (*"Call me"*)).

4 Retrieval model

The task of the retrieval model is to rank the FAQs by relevance to a given query. In an ideal case, the relevant FAQs will be ranked above the non-relevant ones. The retrieval model we propose is a confidence-rated classifier trained on binary relevance judgments, which uses as features the semantic textual similarity between the query and the FAQ. For a given a query-FAQ pair, the classifier outputs whether the FAQ is relevant (positive) or irrelevant (negative) for the query. More precisely, the classifier outputs a confidence score, which can be interpreted as the degree of relevance. Given a single query as input, we run the classifier on all query-FAQ pairs to obtain the confidence scores for all FAQs from the database. We then use these confidence scores to produce the final FAQ ranking.

The training set consists of pairs (q, f) from the test collection, where $q \in Q_r$ is a query from the set of paraphrase queries and $f \in F_{rel}$ is a FAQ from the set of relevant FAQs for this query (cf. Section 3). Each (q, f) pair represents a positive training instance. To create a negative training instance, we randomly select a (q, f) pair from the set of positive instances and substitute the relevant FAQ f with a randomly chosen non-relevant FAQ f'. As generating all possible negative instances would give a very imbalanced dataset, we chose to generate only $2N$ negative instances, where N is the number of positive instances. Because $|F_{rel}|$ varies depending on query q, number of instances N per query also varies; on average, N is 329.

To train the classifier, we compute a feature vector for each (q, f) instance. The features measure the semantic textual similarity between q and f. More precisely, the features measure (1) the similarity between query q and the question from f and (2) the similarity between query q and the answer from f. Considering both FAQ question and answer has proven to be beneficial (Tomuro and Lytinen, 2004). Additionally, ngram overlap features are computed between the query and FAQ category name.

As the classification model, we use the Support Vector Machine (SVM) with radial basis kernel. We use the LIBSVM implementation from Chang and Lin (2011).

4.1 Term overlap features

We expect that FAQ relevance to be positively correlated with lexical overlap between FAQ text and the user query. We use several lexical overlap features. Similar features have been proposed by Michel et al. (2011) for paraphrase classification and by Šarić et al. (2012) for semantic textual similarity.

Ngram overlap (NGO). Let T_1 and T_2 be the sets of consecutive ngrams (e.g., bigrams) in the first and the second text, respectively. NGO is defined as

$$ngo(T_1, T_2) = 2 \times \left(\frac{|T_1|}{|T_1 \cap T_2|} + \frac{|T_2|}{|T_1 \cap T_2|} \right)^{-1} \quad (1)$$

NGO measures the degree to which the first text covers the second and vice versa. The two scores are combined via a harmonic mean. We compute NGO for unigrams and bigrams.

IC weighted word overlap (ICNGO). NGO gives equal importance to all words. In practice, we expect some words to be more informative than others. The informativeness of a word can be measured by its information content (Resnik, 1995), defined as

$$ic(w) = \ln \frac{\sum_{w' \in C} freq(w')}{freq(w)} \quad (2)$$

where C is the set of words from the corpus and $freq(w)$ is the frequency of word w in the corpus. We use the HrWaC corpus from Ljubešić and Erjavec (2011) to obtain the word counts.

Let S_1 and S_2 be the sets of words occurring in the first and second text, respectively. The IC-weighted word coverage of the second text by the first text is given by

$$wwc(S_1, S_2) = \frac{\sum_{w \in S_1 \cap S_2} ic(w)}{\sum_{w' \in S_2} ic(w')} \quad (3)$$

We compute the ICNGO feature as the harmonic mean of $wwc(S_1, S_2)$ and $wwc(S_2, S_1)$.

4.2 Vector space features

Tf-idf similarity (TFIDF). The tf-idf (term frequency/inverse document frequency) similarity of two texts is computed as the cosine similarity of their tf-idf weighted bag-of-words vectors. The tf-idf weights are computed on the FAQ test collection. Here we treat each FAQ (without distinction between question, answer, and category parts) as a single document.

LSA semantic similarity (LSA). Latent semantic analysis (LSA), first introduced by Deerwester et al. (1990), has been shown to be very effective for computing word and document similarity. To build the LSA model, we proceed along the lines of Karan et al. (2012). We build the model from Croatian web corpus HrWaC from Ljubešić and Erjavec (2011). For lemmatization, we use the morphological lexicon from Šnajder et al. (2008). Prior to the SVD, we weight the matrix elements with their tf-idf values. Preliminary experiments showed that system performance remained satisfactory when reducing the vector space to only 25 dimensions, but further reduction caused deterioration. We use 25 dimensions in all experiments.

LSA represents the meaning of a w by a vector $v(w)$. Motivated by work on distributional semantic compositionality (Mitchell and Lapata, 2008), we compute the semantic representation of text T as the semantic composition (defined as vector addition) of the individual words constituting T:

$$v(T) = \sum_{w \in T} v(w) \tag{4}$$

We compute the similarity between texts T_1 and T_2 as the cosine between $v(T_1)$ and $v(T_2)$.

IC weighted LSA similarity (ICLSA). In the LSA similarity feature all words occurring in a text are considered to be equally important when constructing the compositional vector, ignoring the fact that some words are more informative than others. To acknowledge this, we use information content weights defined by (2) and compute the IC weighted compositional vector of a text T as

$$c(T) = \sum_{w_i \in T} ic(w_i) v(w_i) \tag{5}$$

Aligned lemma overlap (ALO). This feature measures the similarity of two texts by semantically aligning their words in a greedy fashion. To compare texts $T1$ and $T2$, first all pairwise similarities between words from $T1$ and words from $T2$ are computed. Then, the most similar pair is selected and removed from the list. The procedure is repeated until all words are aligned. The aligned pairs are weighted by the larger information content of the two words:

$$sim(w_1, w_2) = \max(ic(w_1), ic(w_2)) \times ssim(w_1, w_2) \tag{6}$$

where $ssim(w_1, w_2)$ is the semantic similarity of words w_1 and w_2 computed as the cosine similarity of their LSA vectors, and ic is the information content given by (2). The overall similarity between two texts is defined as the sum of weighted pair similarities, normalized by the length of the longer text:

$$alo(T_1, T_2) = \frac{\sum_{(w_1,w_2) \in P} sim(w_1, w_2)}{\max(length(T_1), length(T_2))} \tag{7}$$

where P is the set of aligned lemma pairs. A similar measure is proposed by Lavie and Denkowski (2009) for machine translation evaluation, and has been found out to work well for semantic textual similarity (Šarić et al., 2012).

4.3 Question type classification (QC)

Related work on QA (Lytinen and Tomuro, 2002) shows that the accuracy of QA systems can be improved by question type classification. The intuition behind this is that different types of questions demand different types of answers. Consequently, information about the type of answer required should be beneficial as a feature.

To explore this line of improvement, we train a simple question classifier on a dataset from Lombarović et al. (2011). The dataset consists of 1300 questions in Croatian, classified into six classes: *numeric*, *entity*, *human*, *description*, *location*, and *abbreviation*. Following Lombarović et al. (2011), we use document frequency to select the most frequent 300 words and 600 bigrams to use as features. An SVM trained on this dataset achieves 80.16% accuracy in a five-fold cross-validation. This is slightly worse than the best result from Lombarović et al. (2011), however we use a smaller set of lexical features. We use the question type classifier to compute two features: the question type of the query and the question type of FAQ question.

Feature	RM1	RM2	RM3	RM4	RM5
NGO	+	+	+	+	+
ICNGO	+	+	+	+	+
TFIDF	–	+	+	+	+
LSA	–	–	+	+	+
ICLSA	–	–	+	+	+
ALO	–	–	+	+	+
QED	–	–	–	+	+
QC	–	–	–	–	+

Table 4: Features used by our models

4.4 Query expansion dictionary (QED)

Our error analysis revealed that some false negatives could easily be eliminated by expanding the query with similar/related words. To this end, we constructed a small, domain-specific query expansion dictionary. We aimed to (1) mitigate minor spelling variances, (2) make the high similarity of some some cross-POS or domain-specific words explicit, and (3) introduce a rudimentary "world knowledge" useful for the domain at hand. The final dictionary contains 53 entries; Table 3 shows some examples.

5 Evaluation

5.1 Experimental setup

Because our retrieval model is supervised, we evaluate it using five-fold cross-validation on the FAQ test collection. In each fold we train our system on the training data as described in Section 4, and evaluate the retrieval performance on the queries from the test set. While each (q, F_{rel}) occurs in the test set exactly once, the same FAQ may occur in both the train and test set. Note that this does not pose a problem because the query part of the pair will differ (due to paraphrasing).

To gain a better understanding of which features contribute the most to retrieval performance, we created several models. The models use increasingly complex feature sets; an overview is given in Table 4. We leave exhaustive feature analysis and selection for future work.

As a baseline to compare against, we use a standard tf-idf weighted retrieval model. This model ranks the FAQs by the cosine similarity of tf-idf weighted vectors representing the query and the FAQ. When computing the vector of the FAQ pair, the question, answer, and category name are concatenated into a single text unit.

Model	P	R	F1
RM1	14.1	68.5	23.1
RM2	25.8	75.1	37.8
RM3	24.4	75.4	36.3
RM4	**25.7**	**77.7**	**38.2**
RM5	25.3	76.8	37.2

Table 5: Classification results

5.2 Results

Relevance classification performance. Recall that we use a binary classifier as a retrieval model. The performance of this classifier directly determines the performance of the retrieval system as a whole. It is therefore interesting to evaluate classifier performance separately. To generate the test set, in each of the five folds we sample from the test set the query-FAQ instances using the procedure described in Section 4 (N positive and $2N$ negative instance).

Precision, recall, and F1-score for each model are shown in Table 5. Model RM4 outperforms the other considered models. Model RM5, which additionally uses question type classification, performs worse than RM4, suggesting that the accuracy of question type classification is not sufficiently high. Our analysis of the test collection revealed that this can be attributed to a domain mismatch: the questions (mobile phone operator FAQ) are considerably different than those on which the question classifier was trained (factoid general questions). Moreover, some of the queries and questions in our FAQ test collection are not questions at all (cf. Table 2); e.g., *"Popravak mobitela."* (*"Mobile phone repair."*). Consequently, it is not surprising that question classification features do not improve the performance.

Retrieval performance. Retrieval results of the five considered models are given in Table 6. We report the standard IR evaluation measures: mean reciprocal rank (MRR), average precision (AP), and R-precision (RP). The best performance was obtained with RM4 model, which uses all features except the question type. The best MRR result of 0.479 (with standard deviation over five folds of ± 0.04) indicates that, on average, model RM4 ranks the relevant answer among top two results.

Performance of other models expectedly increase with the complexity of features used. However, RM5 is again an exception, performing worse than RM4 despite using additional question

Query word	Expansion words	Remark
face	facebook	A lexical mismatch that would often occur
ograničiti (*to limit*)	ograničenje (*limit*)	Cross POS similarity important in the domain explicit
cijena (*price*)	trošak (*cost*), koštati (*to cost*)	Synonyms very often used in the domain
inozemstvo (*abroad*)	roaming (*roaming*)	Introduces world knowledge
ADSL	internet	Related words often used in the domain

Table 3: Examples from query expansions dictionary

Model	MRR	MAP	RP
Baseline	0.341	21.77	15.28
RM1	0.326	20.21	17.6
RM2	0.423	28.78	24.37
RM3	0.432	29.09	24.90
RM4	**0.479**	**33.42**	**28.74**
RM5	0.475	32.37	27.30

Table 6: Retrieval results

type features, for the reasons elaborated above.

Expectedly, classification performance and retrieval performance are positively correlated (cf. Tables 5 and 6). A noteworthy case is RM4, which improves the F1-score by only 5% over RM3, yet improves IR measures by more than 10%. This suggest that, in addition to improving the classifier decisions, the QED boosts the confidence scores of already correct decisions.

A caveat to the above analysis is the fact that the query expansion dictionary was constructed base on the cross-validation result. While only a small amount of errors were corrected with the dictionary, this still makes models RM4 and RM5 slightly biased to the given dataset. An objective estimate of maximum performance on unseen data is probably somewhere between RM3 and RM4.

5.3 Error analysis

By manual inspection of false positive and false negative errors, we have identified several characteristic cases that account for the majority of highly ranked irrelevant documents.

Lexical interference. While a query does have a significant lexical similarity with relevant FAQ pairs, it also has (often accidental) lexical similarity with irrelevant FAQs. Because the classifier appears to prefer lexical overlap, such irrelevant FAQs interfere with results by taking over some of the top ranked positions from relevant pairs.

Lexical gap. Some queries ask a very similar question to an existing FAQ from the database, but paraphrase it in such a way that almost no lexical overlap remains. Even though the effect of this is partly mitigated by our semantic vector space features, in extreme cases the relevant FAQs will be ranked rather low.

Semantic gap. Taken to the extreme, a paraphrase can change a query to the extent that it not only introduces a lexical gap, but also a semantic gap, whose bridging would require logical inference and world knowledge. An example of such query is *"Postoji li mogućnost korištenja Vip kartice u Australiji?"* (*"Is it possible to use Vip sim card in Australia?"*). The associated FAQ question is *"Kako mogu saznati postoji li GPRS/EDGE ili UMTS/HSDPA roaming u zemlji u koju putujem?"* (*"How can I find out if there is GPRS/EDGE or UMTS/SPA roaming in the country to which I am going?"*).

Word matching errors. In some cases words which should match do not. This is most often the case when one of the words is missing from the morphological lexicon, and thus not lemmatized. A case in point is the word *"Facebook"*, or its colloquial Croatian variants *"fejs"* and *"face"*, along with their inflected forms. Handling this is especially important because a significant number of FAQs from our dataset contain such words. An obvious solution would be to complement lemmatization with stemming.

5.4 Cutoff strategies

Our model outputs a list of all FAQs from the database, ranked by relevance to the input query. As low-ranked FAQs are mostly not relevant, presenting the whole ranked list puts an unnecessary burden on the user. We therefore explored some strategies for limiting the number of results.

First N (FN). This simply returns the N best ranked documents.

Measure threshold criterion (MTC). We define a threshold on FAQ relevance score, and re-

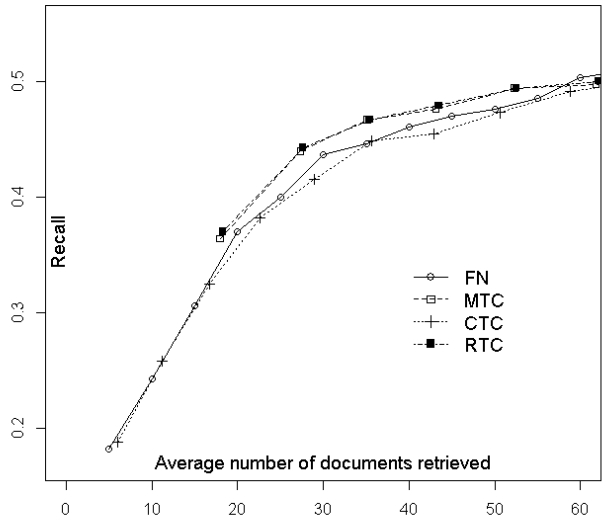

Figure 1: Recall vs. average number of documents retrieved (for various cutoff strategies)

turn only the FAQs for which the classifier confidence is above a specified threshold.

Cumulative threshold criterion (CTC). We define a threshold for cumulative relevance score. The top-ranked FAQs for which the sum of classifier confidences is below the threshold are returned.

Relative threshold criterion (RTC). Returns all FAQs whose relevance is within the given percentage of the top-ranked FAQ relevance.

A good cutoff strategy should on average return a smaller number of documents, while still retaining high recall. To reflect this requirement we measure the recall vs. average number of retrieved documents (Fig. 1). While there is no substantial difference between the four strategies, MTC and RTC perform similarly and slightly better than FN and CTC. As the number of documents increases, the differences between the different cutoff strategies diminish.

5.5 Performance and scalability

We have implemented the FAQ engine using in-house code in Java. The only external library used is the Java version of LIBSVM. Regarding system performance, the main bottleneck is in generating the features. Since all features depend on the user query, they cannot be precomputed. Computationally most intensive feature is ALO (cf. Section 4.2), which requires computing a large number of vector cosines.

The response time of our FAQ engine is acceptable – on our 1222 FAQs test collection, the results are retrieved within one second. However, to retrieve the results, the engine must generate features and apply a classifier to every FAQ from the database. This makes the response time linearly dependent on the number of FAQs. For larger databases, a preprocessing step to narrow down the scope of the search would be required. To this end, we could use a standard keyword-based retrieval engine, optimized for high recall. Unfortunately, improving efficiency by precomputing the features is impossible because it would require the query to be known in advance.

6 Conclusion and Perspectives

We have described a FAQ retrieval engine for Croatian. The engine uses a supervised retrieval model trained on a FAQ test collection with binary relevance judgments. To bridge the notorious lexical gap problem, we have employed a series of features based on semantic textual similarity between the query and the FAQ. We have built a FAQ test collection on which we have trained and evaluated the model. On this test collection, our model achieves a very good performance with an MRR score of 0.47.

We discussed a number of open problems. Error analysis suggests that our models prefer the lexical overlap features. Consequently, most errors are caused by deceivingly high or low word overlap. One way to address the former is to consider not only words themselves, but also syntactic structures. A simple way to do this is to use POS patterns to detect similar syntactic structures. A more sophisticated version could make use of dependency relations obtained by syntactic parsing.

We have demonstrated that even a small, domain-specific query expansion dictionary can provide a considerable performance boost. Another venue of research could consider the automatic methods for constructing a domain-specific query expansion dictionary. As noted by a reviewer, one possibility would be to mine query logs collected over a longer period of time, as employed in web search (Cui et al., 2002) and also FAQ retrieval (Kim and Seo, 2006).

From a practical perspective, future work shall focus on scaling up the system to large FAQ databases and multi-user environments.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia under the Grant 036-1300646-1986. We thank the reviewers for their constructive comments.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 385–393. Association for Computational Linguistics.

Robin D. Burke, Kristian J. Hammond, Vladimir Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg. 1997. Question answering from frequently asked question files: experiences with the FAQ Finder system. *AI magazine*, 18(2):57.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Danish Contractor, Govind Kothari, Tanveer A. Faruquie, L. Venkata Subramaniam, and Sumit Negi. 2010. Handling noisy queries in cross language FAQ retrieval. In *Proceedings of the EMNLP 2010*, pages 87–96. Association for Computational Linguistics.

Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2002. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web*, pages 325–332. ACM.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6).

Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Distributional semantics approach to detecting synonyms in Croatian language. In *Information Society 2012 - Eighth Language Technologies Conference*, pages 111–116.

Jaana Kekäläinen. 2005. Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems. *Information processing & management*, 41(5):1019–1033.

Harksoo Kim and Jungyun Seo. 2006. High-performance FAQ retrieval using an automatic clustering method of query logs. *Information processing & management*, 42(3):650–661.

Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarthy, and L. Venkata Subramaniam. 2009. SMS based interface for FAQ retrieval. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2, pages 852–860.

Alon Lavie and Michael J. Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Nikola Ljubešić and Tomaž Erjavec. 2011. HrWaC and SlWaC: compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer.

Tomislav Lombarović, Jan Šnajder, and Bojana Dalbelo Bašić. 2011. Question classification for a Croatian QA system. In *Text, Speech and Dialogue*, pages 403–410. Springer.

Steven Lytinen and Noriko Tomuro. 2002. The use of question types to match questions in FAQ Finder. In *AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 46–53.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *Proceedings of ACL-08: HLT*, pages 236–244.

Alejandro Moreo, Maria Navarro, Juan L. Castro, and Jose M. Zurita. 2012. A high-performance FAQ retrieval method using minimal differentiator expressions. *Knowledge-Based Systems*.

Roberto Navigli and Paola Velardi. 2003. An analysis of ontology-based query expansion strategies. In *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia*, pages 42–49.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. 2008. How to make letor more useful and reliable. In *Proceedings of the ACM Special Interest Group on Information Retrieval 2008 Workshop on Learning to Rank for Information Retrieval*, pages 52–58.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *eprint arXiv: cmp-lg/9511007*, volume 1, page 11007.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: systems for measuring semantic text similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 441–448. Association for Computational Linguistics.

Jan Šnajder, Bojana Dalbelo Bašić, and Marko Tadić. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management*, 44(5).

Eriks Sneiders. 1999. Automated FAQ answering: continued experience with shallow language understanding. In *Question Answering Systems. Papers from the 1999 AAAI Fall Symposium*, pages 97–107.

Radu Soricut and Eric Brill. 2004. Automatic question answering: beyond the factoid. In *Proceedings of HLT-NAACL*, volume 5764.

Noriko Tomuro and Steven Lytinen. 2004. Retrieval models and Q and A learning with FAQ files. *New Directions in Question Answering*, pages 183–194.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207. ACM.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *SIGIR'94*, pages 61–69. Springer.

Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 355–370. Springer.

Chung-Hsien Wu, Jui-Feng Yeh, and Yu-Sheng Lai. 2006. Semantic segment extraction and matching for internet FAQ retrieval. *Knowledge and Data Engineering, IEEE Transactions on*, 18(7):930–940.

Xiaobing Xue, Jiwoon Jeon, and W Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 475–482. ACM.