Reformulations of Finnish questions for question answering

Lili Aunimo and Reeta Kuuskoski
Department of Computer Science
University of Helsinki
P.O. Box 68
FI-00014 University of Helsinki
aunimo|rkuuskos@cs.helsinki.fi

Abstract

This paper presents a series of experiments that were performed with a dataset of Finnish question reformulations. The goal of the experiments was to determine whether some reformulations are easier for a question answering system to deal with than others, and if so, how easy it is to transform a question into that form. A question answering system typically consists of several independent modules that are arranged into a pipeline architecture. In order to determine if some reformulations are easier for a question answering system to deal with, the performance of the question classifier component was analyzed. The experiments show that different question reformulations do affect the performance of the classifier significantly. However, the automatic transformation of a question into another form seems difficult.

1 Introduction

Question reformulations (or variants) are questions that have the same semantic content, i.e. that can be answered with the same answer, but whose form is different. Eight different reformulations of the same question from the TREC [1]-9 QA Track (Voorhees, 2000) question dataset [2] are listed in the following:

1. Name a film in which Jude Law acted.
2. Jude Law was in what movie?
3. Jude Law acted in which film?
4. What is a film starring Jude Law?
5. What film was Jude Law in?
6. What film or films has Jude Law appeared in?

As can be seen from the above reformulation examples, the differences between the questions can be syntactic or lexical. For example, question 2 has the word *movie* and question 3 the word *film*, and question 2 and question 5 have different word orders. Determining what is a question reformulation is not straightforward, because the set of all possible answers to a question is not always the same for different reformulations even though they do have at least one common answer (Voorhees, 2000). For example, among the above questions, question 6 accepts for answer in addition to a single movie name also a list of movie names. Thus, we define question reformulations as being a set of questions that have at least one similar answer.

There may exist similarity across questions with different semantic content that can be used in automatically generating or analyzing question reformulations. For example, if we can analyze the example questions above, we could also analyze the reformulations for the question *Name a film in which Woody Allen acted.* and for questions dealing with any other actor. We could also analyze the reformulations for the question *Name a play in which Jude Law acted.* and for questions dealing with any other things in which people act, such as a scene or TV-series. We call this a *similarity class*. The similarity class of the above example could be denoted by the verb based template *PERSON act in ACTED_THING*. The verb based templates of the similarity classes can be seen as semantic frames (see e.g. (Baker et al., 1998)).

[1] Text REtrieval Conference, http://trec.nist.gov

[2] The question reformulations data is available at: http://trec.nist.gov/data/qa/T9_QAdata/variants.key

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 12–21
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

Question answering (QA) systems are information access systems that receive as input a natural language question and that produce as output the answer. QA systems can be classified according to the type of data from which the answers are extracted. Text based QA systems extract the answer from plain text documents, FAQ (Frequently Asked Questions) based systems extract the answer from a dataset of question-answer pairs, and structure based systems extract the answer from a relational database or from a semistructured data repository such as text containing XML or HTML markup. Text based QA systems are the ones that have attracted most attention in the research community over the last years. A text based system typically consists of a pipeline architecture containing a question processing module, a document processing module and an answer extraction and formulation module (Harabagiu and Moldovan, 2003). One of the most important tasks of the question processing module is to determine the expected answer type of the question.

Text based QA systems have been systematically evaluated in evaluation campaigns such as the TREC, CLEF [3], NTCIR [4] and EQUER (Ayache, 2005). The evaluation datasets created in the above campaigns have had a significant impact in directing the research on QA systems. Some of the evaluation campaigns have had datasets consisting of question reformulations. In TREC-9, there were 54 questions, which all had from two to eight reformulations (Harabagiu et al., 2001). The total number of reformulations was 243. In the domain independent task of EQUER, 100 out of the total of 500 questions were reformulations, and in the domain specific task, 50 out of 200 were reformulations (Ayache, 2005).

In FAQ based QA systems, the main answering technique consists in measuring the similarity between a new question and the old ones and returning the answer that has been given to the old question that is most similar with the new one (Burke et al., 1997; Aunimo et al., 2003). In this kind of systems, the techniques for recognizing question reformulations are especially important.

Question reformulations for English have been extensively studied in the field of QA because not only users tend to express the same information need as a different natural language question (Aunimo et al., 2003), but also because there are a variety of similarity classes among questions whose identification would lead to better question analysis results. However, only very little work has been done for processing question reformulations for QA in other languages than English. To the best of our knowledge, this paper presents the first experiments on question reformulations for Finnish questions.

Question reformulations in QA systems have been approached in two different ways. The first approach is to measure the similarity between questions, and the second approach is to generate reformulations for questions (Hermjakob et al., 2002). In order to measure the similarity between questions, many different similarity metrics have been developed (see e.g. (Harabagiu et al., 2001; Burke et al., 1997; Aunimo et al., 2003)). One of the approaches even transforms the questions into a completely different form, the semantic case frame representation, before measuring similarity between them (Tomuro, 2003).

Our approach can be seen as as a hybrid approach combining elements from the above mentioned question reformulations generation (Hermjakob et al., 2002) and semantic case frame generation (Tomuro, 2003) approaches. In our approach, questions are not transformed into an abstract semantic representation (as in the case frame generation approach), but into one real reformulation (like in the question reformulations generation approach) that is called the canonical reformulation. However, our approach differs from the questions reformulations generation approach in that we only produce at most one reformulation for a given question, which is called the canonical reformulation. In addition, the canonical question reformulations are real questions, while some reformulations of the reformulations generation approach are closer to answer reformulations than question reformulations.

The rest of the paper is organized as follows: Section 2 describes the question reformulations dataset that is used in the experiments. In Section 3, the method for determining the canoni-

[3] Cross-Language Evaluation Forum, http://www.clef-campaign.org

[4] NII-NACSIS Test Collection for IR Systems, http://research.nii.ac.jp/ntcir/workshop

cal (or centroid or best) reformulation for each question is presented. Section 4 presents the question classification method that is used to asses the effect of different question reformulations to QA. Section 5 describes the different transformations that are needed to transform a question variant into the canonical form. Finally, Section 6 presents the results of the experiments and an analysis.

2 The question variant data

The question variants dataset[5] consists of 200 Finnish questions from the Multieight-04 Corpus (Magnini et al., 2005) and of three variants for each question. Each question variant in the dataset was translated from English by a single translator who worked independently and without having seen the other variants. The variants may be similar or different with each other. As can be seen from Figure 1, there are 59 questions where all variants are different from each other, 46 with one or two pairs of similar variants, 55 with three similar variants and 40 questions where all four variants are similar. The dotted line in the column illustrating the 46 questions that have two similar variants shows the number of questions containing two pairs of similar variants (5) and the number of questions containing only one pair of similar variants (41).

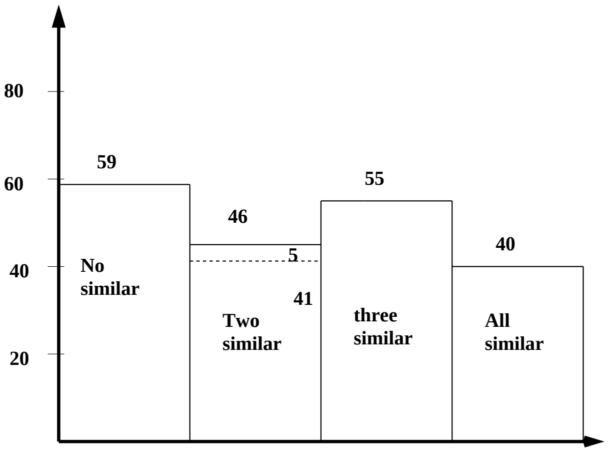

Figure 1: Number of questions with no similar variants, two similar variants, three similar variants and four similar variants.

3 Distance between question variants

The distance between two question variants, v_i and v_j, is measured using their edit distance (Wagner and Fischer, 1974) d, which is the minimum cost sequence of edit operations needed to change one variant into the other. The edit operations are insertion, deletion and substitution, and each of them is assigned a cost of 1. In order to be able to compare the edit distances with each other, they are normalized between 0 and 1. In the case where the cost of substitution is 1, the normalization is achieved by dividing the distance by the length of the longer string, more formally:

$$d_{normalized}(v_i, v_j) = \frac{d(v_i, v_j)}{max(|v_i|, |v_j|)},$$

where $d_{normalized}$ denotes the normalized distance d, and v_i and v_j are the question variants. As the basic unit of edit distance operations, both words and single characters are used, as will be explained in detail later in this section.

The centroid of each question variant set is the variant which is closest to all other variants. More formally, it is

$$i^* = arg \min_i \sum_{j=0}^{n} d(v_i, v_j),$$

where n is the total number of variants for a question. The centroid variant id i^* is thus the variant id whose sum of distances from all the other variants is the smallest.

For the experiments described in section 6, also the set of the most different variants was created. The most different variant, which is called the *worst* variant, is calculated in the same way as the centroid except that instead of taking the variant with the smallest sum of distances to all other variants, the variant with the biggest sum is taken. More formally, the worst variant id $i^{'}$ is calculated as:

$$i^{'} = arg \max_i \sum_{j=0}^{n} d(v_i, v_j).$$

Table 1 gives an example of four different question variants and of the sums of their distances from the other variants. Both word and character based distance sums are given. As can be seen from the table, variant 1 is the centroid if word based distance is used and variant

[5] The question variants dataset is available at http://www.cs.helsinki.fi/research/doremi/interests/QAResources.shtml

Id	Question variants	Sum of Distances	
		Word	Char
1	Kuinka monta asukasta on Etelä-Afrikassa?	2.2	1.56
2	Mikä on Etelä-Afrikan väkiluku?	3	1.87
3	Paljonko Etelä-Afrikassa on asukkaita?	2.8	1.89
4	Kuinka monta asukasta Etelä-Afrikassa on?	2.4	1.44

Table 1: Question variants and sums of their normalized distances from the other variants.

4 is the centroid if the character based distance is used. Variant 2 is the worst variant according to the word based distance and variant 3 according to the character based distance.

In some cases, the above described method does not yield a unique centroid or worst variant. In such cases, the edit distance for the variants having the same score is calculated based on words or characters - depending on which one was used in the first phase. (If it was words in the first phase, characters is used in the second phase, and vice versa.) If no unique centroid or worst variant is still not obtained, it is determined manually.

The principles guiding the manual selection are based on the set of automatically selected centroids. For instance, if the case morphemes of an abbreviation denoting an organization are separated with a colon in the automatically generated set, the manual selections apply the same rule. For example, NATO:lle would be selected instead of NATOlle. If the decision cannot be taken based on the automatically generated set of centroids, the linguistic instinct of the human expert is followed.

4 Question classification

Question classification means the classification of natural language questions according to the expected answer type of the question. In the experiments described in Subsection 6.2, the set of nine question classes listed in Table 2 are used. The classes are taken directly from the Multieight-04 Corpus (Magnini et al., 2005). The table also shows the frequencies of the classes in the dataset that is used in the experiments presented in this paper.

Class	#
TIME	30
LOCATION	26
MANNER	17
PERSON	27
ORGANIZATION	25
MEASURE	20
DEF_ORGANIZATION	11
OTHER	35
DEF_PERSON	9
$\sum$	200

Table 2: Question class frequencies of the 200 Finnish questions in the Multieight-04 Corpus (Magnini et al., 2005).

The question classifier used in the experiments is the C4.5 decision tree classifier (Quinlan, 1993). The nodes of the tree are tests for attribute values that are extracted from the questions. A decision tree classifier is a natural choice when dealing with *nominal data*, i.e. data where the instance descriptions are discrete and don't have any natural notion of similarity or ordering. For example, the attribute named *first* may take the values *Kuka* and *Mikä*. These values have no order relation. Nominal data is often represented as a list of attribute-value pairs. Another benefit of using a decision tree classifier is its interpretability. It is straightforward to render the information it contains as logical expressions.

In order to be able to induce a question classifier from the question data and to be able to apply the classifier to unseen questions, the questions have to be transformed into lists of attribute-value pairs. This transformation is not at all straightforward and the choice of attributes and values has a significant effect on classifier accuracy (Aunimo, 2005). The impact of attribute or feature selection is significant also when performing question classification with other classifiers than C4.5 (see e.g. (Suzuki et al., 2003; Li et al., 2004)). The attribute set and the transformation of questions into lists of attribute-value pairs is described in the following.

The attribute set contains six attributes: *first*, *second*, *third*, *fourth*, *fifth* and *last* word. As the names of the attributes suggest, they are derived from the first, second, third, fourth, fifth and last words of the question. Punctuation is treated in the same way as any word. If the question contains less than six words, the missing word attributes are given the value *NIL*. However, each question has to contain at least two words. Table 3 lists the attributes, their type, the five most common attribute values in the question variants dataset with their frequencies and the total number of possible values for each attribute in the dataset. The type of the first attribute is *plain word*, which means that the first word of the question is taken as such. The type of the second and last word is *lemmatized word*, which means that the lemmatized form of the word is used. For lemmatizing, the Functional Dependency parser from Connexor[6] was used. The type of the third, fourth and fifth words is *POS or lemmatized word*, which means that for open word classes (noun, verb, adjective, adverb) and numerals, the part of speech (POS) is used and for the rest of the word classes the lemmatized word form is used. All attribute values can be symbols for punctuation, such as quotation mark, comma, etc. An example: The question *Minä vuonna Thomas Mann sai Nobelin palkinnon?*[7] is transformed into the following list of attribute-value pairs: *first=Minä, second=vuosi, third=noun, fourth=noun, fifth=verb, last=palkinto*. The class of the question is *TIME*.

5 Question transformations

The question variants were examined and classified based on their differences from the centroid. 27 different transformation classes were found. They are listed in Table 4. Examples of the classes are given in Table 5. The classes are described in detail in the following text.

The transformation categories are divided into subclasses, and the POS of the altering term is inside brackets. `lex` classes are those transformations that are achieved by lexical changes. For instance, `lex(n)` means that a noun has been replaced by another noun, its lexical variant. This is the case in variant v3 of question 168 (or Q168/v3) in Table 5, where the noun *isku* of the centroid has been replaced with its lexical variant *hyökkäys*.

Morphological changes are categorized into the class `morph`. `morph(v)` typically denotes a change in the tense of a verb (for instance imperfect is replaced by perfect) or mode (active replaced by passive). The class `morph(n)` is recorded when a common noun is subject to a morphological alternation. This happens in Q168/v3, where the case of the noun *metroasema* changes.

The class `pos` refers to those alternations where the POS of a word changes, but the base word remains the same. An example of this is in in Q22/v4, where the adjective *skotlantilainen* has been replaced by the inessive form of the proper noun *Skotlanti*, yielding *Skotlannissa*. This transformation is classified as (`pos(n/a)`). When both the POS and the base word are different, the transformation is classified `lexpos`.

Some transformations are of conventional type, and belong to class `conv`. For instance, changes that reflect the source language of the translation are of class `conv(trans)`. Conventions that deal with the way in which abbreviations are written belong to class `conv(abbr)`. In Q24/v4 , the non-capital letter in the beginning of the name of the mosque produces `conv(case)`, and the existence of the hyphen `conv(ortograph)`.

The transformations listed in the column `other` in Table 4 do not belong to the above mentioned subclasses. When the specificity differs, the transformation is of type `spec`, the difference in the order of the terms belongs to class `order`. Examples of these transformations can be seen in Q22/v4, where the specifying noun *kieltä* is missing from the canonical form, and the noun *Skotlannissa* appears after the verb.

The changes in the case of proper noun modifiers are categorized into their own classes. The selection of a genitive attribute instead of the locative form of a proper noun signifying a location is of class (`genattr/locative`). An example of this is Q168/v3, where in the centroid, the locative form *Pariisissa* is used, whereas in the variant, the location is expressed using the genitive attribute.

When the syntactic structure differs due to the selection of the word that bears the contents of the question, the transformation is of class `struct`. This is the case in Q22/v3, for instance, where the verb is very general, *on*, and the essential meaning is contained in the attributes of the noun: *gaelia puhuvia ihmisiä*; in the corresponding centroid (Q22/c) the con-

[6] http://www.connexor.com

[7] What year was Thomas Mann awarded the Nobel Prize?

Name	**Type**	**Example Values and their frequencies**	**#**
first	plain word	Mikä (152), Kuka (121), Missä (80), Kuinka (75), Milloin (69)	23
second	lemmatized word	olla (219), moni (49), vuosi (41), jokin (37), maa (13)	168
third	POS or lemmatized word	noun (415), NIL (207), verb (107), adjective (53), quotation mark (3)	14
fourth	POS or lemmatized word	NIL (411), noun (217), verb (78), adjective (49), quotation mark (15)	10
fifth	POS or lemmatized word	NIL (593), noun (118), verb (36), adjective (25), quotation mark (13)	8
last	lemmatized word	olla (57), nimi (33), kuolla (21), quotation mark (15), tehdä (14)	261

Table 3: The set of six attributes used in question classification. For each attribute, the table lists its name, type, five most common values with their frequencies and the total number of possible values.

lexical		**morphological**		**pos**		**convention**		**other**	
lex(n)	76	morph(n)	34	pos(n/a)	6	conv(ortograph)	23	order	89
lex(v)	74	morph(v)	25	pos(n/v)	1	conv(trans)	23	spec	45
lex(q)	28	morph(q)	10			conv(case)	18	struct	19
lex(a)	8	morph(a)	3			conv(abbr)	5	nounattr/genattr	9
lex(part)	4	morph(pron)	3			conv(accent)	2	genattr/locative	8
lex(postpos)	4							meaning	4
lex(pron)	1							add(adverb)	1
lexpos(n/a)			1						

Table 4: The question transformation classes and their frequencies in the question variants dataset.

tents are expressed in the verb and its attributes: *puhuu gaelia*. Similarly, in Q168/v1 the verb *iskettiin* contains the meaning, but in the centroid, it is in the noun acting as the object of the sentence, *isku*.

Class `meaning` is used when the semantics of the two questions differ, as can happen when the translators have interpreted the original English question differently. The interpolation of an additional word results a transformation of class `add`.

In general, the difference between two variants can consist of multiple concurrent transformations. The transformation classes `struct` and `meaning`, however, occur alone. When there is a fundamental difference in either the semantics or the syntax of the variants, no alternations on the surface level are recorded.

Basically, one question transformation can consist of multiple simple alternations. For example, the selection of a synonymous verb can cause multiple modifications to the morphology of the other words due to government, i.e. different verbs require different cases from their dependents. This influence has been ignored in the categorization of the transformation, and recorded solely by the transformation class `morph(v)`.

6 Results and analysis

6.1 Distances between question reformulations

The datasets of question centroids, or *best* questions, and of the *worst* questions were created using the metrics described in Section 3. Both word and character based metrics were applied to the data, and thus two different datasets for *best* and *worst* were created. Among the *best* variants calculated using first the word based distance and then, if needed, the character based distance, only 6 variants out of 200 were different from the ones obtained by only calculating the character based distance. The names of these two datasets are *Best WordChar* and *Best Char*, respectively. Among the *worst* variants, 28 out of 200 questions were different in the datasets *Worst WordChar* and *Worst Char*. When using the character based distance, only human judgments were used in order to select the best or worst variant among equally scored variants, because us-

Question		Transformations
o	**Q22: How many people speak Gaelic in Scotland?**	
c	*Kuinka moni skotlantilainen puhuu gaelia?*	
v1-v2	how many ScottishNOM speakS3 GaelicPART Kuinka moni skotlantilainen puhuu gaelia?	-
v3	how much PART(Gaelic speaking people) INESS existS3 Kuinka paljon gaelia puhuvia ihmisiä Skotlannissa on?	struct
v4	how many speakS3 GaelicGEN languagePART INESS Kuinka moni puhuu gaelin kieltä Skotlannissa?	pos(n/a) order spec
o	**Q24: Where is the Al Aqsa Mosque?**	
c	*Missä on Al Aqsa -moskeija?*	
v1-v3	where beS3 NOM mosque Missä on Al Aqsa -moskeija?	-

v4	where beS3 GEN mosque
	Missä on al-Aqsan moskeija?
		conv(case)
		conv(ortograph)
		nounattr/genattr

o	**Q168: When did the attack at the Saint-Michel underground station in Paris occur?**	
c	*Milloin tapahtui isku Saint-Michelin metroasemalle Pariisissa*	
v1	when GEN GEN Milloin Pariisin Saint-Michelin metro_stationALL attackIMP_PASS metroasemalle iskettiin?	struct
v2	when happenIMP_S3 attackNOM GEN Milloin tapahtui isku Saint-Michelin metro_stationALL INESS metroasemalle Pariisissa?	-
v3	when happenIMP_S3 GEN GEN Milloin tapahtui Pariisin Saint-Michelin metro_stationGEN attackNOM metroaseman hyökkäys?	lex(n) morph(n) genattr/locative 2×order
v4	when attackNOM GEN metroGEN stationALL Milloin isku Saint-Michelin maanalaisen asemalle INESS happenIMP_S3 Pariisissa tapahtui?	lex(n) order

Table 5: The question variants for questions 22, 24 and 168 in the question variant dataset. The original question (o), the canonical form, or centroid, (c) and the variants (v1-v4) with their transformations. When there are no tranformations, the variant equals the centroid of the variant set.

ing the word based distance did not distinguish among the variants in any of the cases in the data. All in all, there were six questions in the dataset whose best and worst variants had to be determined manually. Five of these could be determined following the automatically made choices, and only one case was determined using the intuition of the human expert.

6.2 Classification of question reformulations

The extent to which different question reformulations affect the accuracy of question classification was investigated by evaluating the performance of a classifier using different question reformulation datasets. The different datasets are listed in Table 6 in the column named *Dataset Name*. The creation of the datasets *Best WordChar*, *Best Char*, *Worst WordChar* and *Worst Char* was explained in Section 3. The dataset *Mixed* is a dataset containing 50 randomly selected different questions from each variant set. The datasets *Variant 1, 2, 3 and 4* only contain questions created by one author. The dataset *All* consists of all question variants. The results table reports the accuracy of the classifier both on training data and on unseen data. The accuracy of the classifier on training data is given because it illustrates how well the features selected can classify the data at hand. The results obtained by testing the classifier with unseen data naturally give a more realistic picture of the performance of the classifier. These results were obtained by 10-fold cross-validation (see e.g. (Duda et al., 2001), page 483). The accuracy on unseen data for the dataset *All* has two figures, 86.4 and 70.8, in parentheses. The different figures are obtained by using a different sampling in the creation of the training and test sets. The higher accuracy is obtained when the different variants of the same questions appear in both the training and test datasets. The lower accuracy is obtained when all variants of a specific question have to appear in either the training set or the test set, but not in both.

Dataset Name	**Accuracy in %**	
	training data	unseen data
Datasets of 200 questions		
Best WordChar	80.8	75.0
Best Char	80.8	75.0
Worst WordChar	77.9	68.5
Worst Char	77.9	70.0
Mixed	79.1	70.6
Variant 1	75.5	71.5
Variant 2	80.8	73.0
Variant 3	80.2	73.5
Variant 4	78.8	72.0
Dataset of 800 questions		
All	90.9	86.4 (70.8)

Table 6: Accuracy of classifiers inducted from different datasets. Classification accuracy is reported both on training data and on unseen data.

Analysis of the classification results shows that different question reformulations do make a significant difference in the accuracy of a decision tree classifier that uses the features described in Section 4. Among the datasets of 200 questions, the highest classification accuracy, 75,0%, was achieved using the datasets *Best WordChar* and *Best Char* that consist of the centroid of each question. The lowest classification accuracy, 68,5%, was achieved using the dataset *Worst WordChar*. The classification accuracies of the datasets consisting of questions produced by a sole author (*Variants 1 , 2, 3 and 4*) are higher than that of the dataset that contains a mixture of variants from all four authors (*Mixed*). This shows that the authors have some author specific traits that have been captured by the classification features. However, the fact that the centroid datasets have a higher classification accuracy than any of the author specific datasets shows that all centroids are somehow more similar with each other than the questions created by the same author.

6.3 Question transformations and the dataset

As can be seen from Table 4, lexical variations and different ordering of words are the most frequent transformation classes in the dataset. In addition, the transformations related to conventions are very common. Some translator specific features can be seen from the data. For instance, the translator of variant set number four prefers placing the verb as the last item in the question, while the other translators generally use the question word - verb - noun pattern. For this reason, there are 14 questions where all the other variants are equal, but the translation done by the fourth translator differs by one transformation of class `order`. Also orthographic and translation conventions tend to be author specific.

The differences between variants vary significantly. Some of them consist only of one transformation, while others have multiple transformations. Even though there is the same amount of transformations, the complexity of different transformations is not equal.

The taxonomy of the transformation classes described in this paper is created through a data driven approach and as such, is based on a specific data set. It is suitable for the given data, but it is not certain that it can be generalized to other collections of data. A question paraphrasing taxonomy with only six classes has been constructed (Tomuro, 2003), and the classes cannot be straightforwardly mapped into our system.

7 Conclusion

This paper presents the first experiments with Finnish question reformulations. In order to study the different reformulations of natural language questions that arise spontaneously, a question variants dataset was created. The effect of different reformulations of the same question on the performance of a question answering system was evaluated by creating ten different question datasets and by measuring the accuracy of the question classifier component of a question answering system using each of the datasets. Among the different question reformulation datasets created is the set of canonical reformulations. In order to create this dataset, a similarity metric between question reformulations was devised. The experiments show that the question classification accuracy is 75% when the canonical reformulations set is used, but only 70.6% when a mixed dataset reflecting a realistic set of incoming questions to a question answering system is used. Thus, the performance of the question processing component and most likely also the performance of the whole question answering system would improve if the incoming questions were first transformed into a canonical form.

The next step in the research that is presented in the paper consists in analyzing the different transformations that are needed in order to transform a question reformulation into a canonical form. Based on a careful study of the question reformulations dataset, a set of 27 transformation classes were defined. The analysis of these transformations shows that the automatic transformation of a question reformulation into a canonical form, or even the automatic recognition of questions that already are in canonical form would be very challenging. However, research has been done on transforming English question reformulations into a canonical form, and this research indicates that further research on Finnish question reformulations is needed in order to determine the feasibility of transforming Finnish questions into a canonical form. The work presented in this paper constitutes a starting point for this further research, and it already shows that various types of question reformulations spontaneously arise and that their effect on the performance of the question analysis component of a question answering system is significant.

References

Lili Aunimo, Oskari Heinonen, Reeta Kuuskoski, Juha Makkonen, Renaud Petit, and Otso Virtanen. 2003. Question answering system for incomplete and noisy data: Methods and measures for its evaluation. In *Proceedings of the 25th European Conference on Information Retrieval Research (ECIR 2003)*, pages 193 – 206, Pisa, Italy.

Lili Aunimo. 2005. A typology and feature set for questions. In *Proceedings of the Workshop on Knowledge and Reasoning for Answering Questions held in conjunction with the 19th International Joint Conference on Artificial Intelligence*, Edinburgh, United Kindom, August.

Christelle Ayache. 2005. Campagne evalda/equer, evaluation en question réponse. Technical report, ELDA - Evaluations and Language resources Distribution Agency.

Collin Baker, Charles Fillmore, and John Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, Montreal, Canada, August.

R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg. 1997. Question answering from frequently asked question files: Experiences with the faqfinder system. *AI Magazine*, 18(2):57–66.

Richard O. Duda, Peter E. Hart, and David G. Stork. 2001. *Pattern Classification*. Wiley-Interscience.

Sanda Harabagiu and Dan Moldovan. 2003. Question answering. In Ruslan Mitkov, editor, *Computational Linguistics*. Oxford University Press.

Sanda M. Harabagiu, Dan I. Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan C. Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. 2001. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics*, pages 274–281.

Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural language based reformulation resource and web exploitation for question answering. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of TREC-2002*, Gaithersburg, Maryland, November. Department of Commerce, National Institute of Standards and Technology.

Xin Li, Dan Roth, and Kevin Small. 2004. The role of semantic information in learning question classifiers. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, Hainan Island, China.

B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. 2005. Overview of the CLEF 2004 Multilingual Question Answering Track. In C. Peters, P. D. Clough, G. J. F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, volume 3491 of *Lecture Notes in Computer Science*. Springer Verlag.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA.

Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, and Eisaku Maeda. 2003. Question classification using HDAG kernel. In *Proceedings of the Workshop on Multilingual Summarization and Question Answering, held in conjunction with the 41th Annual Meeting of the Association for Computational Linguistics.*, Sapporo, Japan.

Noriko Tomuro. 2003. Interrogative reformulation patterns and acquisition of question paraphrases. In *Proceedings of the International Workshop on Paraphrasing Workshop , held in conjunction with the 41th Annual Meeting of the Association for Computational Linguistics.*, Sapporo, Japan.

Ellen M. Voorhees. 2000. Overwiew of the TREC-9 Question Answering Track. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of TREC-9*, Gaithersburg, Maryland, November. Department of Commerce, National Institute of Standards and Technology.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.