KTH Computer Science
and Communication

A Machine Learning Approach to Dialogue Act Classification in Human-Robot Conversations

Evaluation of dialogue act classification with the robot Furhat and an analysis of the market for social robots used for education

NINA OLOFSSON & NIVIN FAKIH

Bachelor's Thesis at CSC
Supervisor: Joakim Gustafson & Bo Karlsson
Examiner: Alexander Baltatzis

Abstract

The interest in social robots has grown dramatically in the last decade. Several studies have investigated the potential markets for such robots and how to enhance their human-like abilities. Both of these subjects have been investigated in this thesis using the company Furhat Robotics, and their robot Furhat, as a case study.

This paper explores how machine learning could be used to classify *dialogue acts* in human-robot conversations, which could help Furhat interact in a more human-like way. Dialogue acts are acts of natural speech, such as questions or statements. Several variables and their impact on the classification of dialogue acts were tested. The results showed that a combination of some of these variables could classify 73 % of all the dialogue acts correctly.

Furthermore, this paper analyzes the market for social robots which are used for education, where human-like abilities are preferable. A literature study and an interview were conducted. The market was then analyzed using a SWOT-matrix and Porter's Five Forces. Although the study showed that the mentioned market could be a suitable target for Furhat Robotics, there are several threats and obstacles that should be taken into account before entering the market.

Sammanfattning

Maskininlärning för klassificering av talhandlingar i människa-robot-konversationer

Intresset för sociala robotar har ökat drastiskt under det senaste årtiondet. Ett flertal studier har undersökt hur man kan förbättra robotars mänskliga färdigheter. Vidare har studier undersökt potentiella marknader för sådana robotar. Båda dessa aspekter har studerats i denna rapport med företaget Furhat Robotics, och deras robot Furhat, som en fallstudie.

Mer specifikt undersöker denna rapport hur maskininlärning kan användas för att klassificera talhandlingar i människa-robot-konversationer, vilket skulle kunna hjälpa Furhat att interagera på ett mer mänskligt sätt. Talhandlingar är indelningar av naturligt språk i olika handlingar, såsom frågor och påståenden. Flertalet variabler och deras inverkan på klassificeringen av talhandlingar testades i studien. Resultatet visade att en kombination av några av dessa variabler kunde klassificera 73 % av alla talhandlingar korrekt.

Vidare analyserar denna rapport marknaden för sociala robotar inom utbildning, där mänskliga färdigheter är att föredra. En litteraturstudie och en intervju gjordes. Marknaden analyserades sedan med hjälp av en SWOT-matris och Porters femkraftsmodell. Fastän studien visade att den ovannämnda marknaden skulle kunna vara lämplig för Furhat Robotics finns ett flertal hot och hinder som företaget måste ta hänsyn till innan de tar sig in på marknaden.

Acknowledgements

This thesis was written in co-operation with the company Furhat Robotics. We therefore want to thank Gabriel Skantze, a co-founder of the company, for providing us with support and expertise throughout the project. We also wish to thank Preben Wik, CEO of Furhat Robotics, for participating in an interview. Finally, we want to express our gratitude to our supervisors Bo Karlsson and Joakim Gustafson.

Contents

Chapter 1

Introduction

There is a growing interest in robots with human-like behavior. This field of study, known as social robotics, is a fairly recent branch of the more general robotics field. The use of interactive robots could revolutionize several industries with numerous areas of application. Nevertheless, there are still many obstacles to overcome when developing social robots. One challenge is to make a robot not only understand the content of what is being said in a conversation, but also to follow social behavior and rules attached to its role. To achieve this, social robots must understand the intention of utterances and be able to classify them into different categories, such as questions, statements or agreements. This is also referred to as *dialogue act classification*. The robot's answer or feedback can then be based on which category the dialogue act belongs to. This is essential for human-like conversation that is expected to be natural, intuitive and effective.

Classifying the dialogue act of an utterance is a complex task. Humans often take several variables into account when participating in a conversation. For instance, such variables could be facial expressions, head nods or gestures. This can also be applied to dialogue systems. Such a system, which is able to collect information from more than one interactive channel, is called a *multimodal dialogue system*.

An example of a robot with a multimodal system is Furhat, a robotic head that has been developed at the department of Speech, Music and Hearing at KTH. Furhat manages multiparty dialogues by combining facial animation with physical embodiment. A company named Furhat Robotics was founded in 2014 with the purpose to create and sell such robots. [1] Although the robots are fully functional, they must be able to interact more appropriately with humans. One step closer to achieving this would be to develop a dialogue act classifier for Furhat, which will be the focus of this thesis.

Classifying dialogue acts with a robot like Furhat could theoretically be done by programming specific rules. However, this would be complicated since Furhat will have to understand a wide (or actually infinite) range of possible utterances. A more suitable approach is therefore to use *machine learning* to classify dialogue acts. Machine learning is a field of study focused on the development of algorithms that can learn from data. The purpose of such algorithms is to spot patterns or make predictions based on input provided to the system. This is transformed into a model which can later be used to make predictions for new and unseen data. The abovementioned properties of machine learning make it a suitable tool to use when developing a dialogue act classifier for a robot like Furhat.

A dialogue act classifier could make a robot like Furhat better understand natural language and thereby appear more human-like, which is an important aspect when introducing social robots to the market. However, the market for social robots is new and also strongly driven by innovation. It is therefore important for startup companies, like Furhat Robotics, to understand the market and analyze their own position in order to gain competitive advantages. One market which Furhat Robotics is considering to enter is the market for social robots to be used for educational purposes. Virtual text-based tutors already exist today. Several studies have investigated how such tutors could get a better understanding of student language input [2] [3], in order to make them more attractive on the market. This thesis has a similar approach, and will investigate both the possibility of developing Furhat's human-like appearance – with a dialogue act classifier – and the market for human-like robots used in education.

1.1 Report Structure

This thesis is divided into two parts – each part with their respective research questions, theoretical frameworks, methods, results and conclusions. Part I handles technical research questions and quantitative analysis whereas part II has an economic perspective and a qualitative analysis.

1.2 Purpose

The main purpose of part I was to investigate if a machine learning algorithm can be used to implement a dialogue act classifier suitable for human-robot conversations, similar to the ones that the robot Furhat participates in. This can be used in further studies to develop Furhat's ability to communicate in a more natural way. This study also aimed to examine which variables have a crucial role when developing the dialogue act classifier.

Furthermore, the purpose of part II was to examine if the identified market for social robots used for education is a suitable target for Furhat Robotics.

1.3 Problem Definition

The scope of the first part of this report can be summarized in the following questions:

- Is it possible, given the conditions, to develop a dialogue act classifier for the robot Furhat using machine learning? If so, with which accuracy can this be done?
- Which variables have a vital role when developing the classifier, and why?

The second part of this report will investigate the following questions:

- Can a company like Furhat Robotics create competitive advantages in the market for social robots used for education?
- Which are the possible domains within this market?

1.4 Method

The following steps summarize the methods used to answer the abovementioned research questions. More detailed descriptions can be found in the respective parts of the report.

1. Literature study of previous work within dialogue act classification and machine learning.
2. Transcription of data from human-robot conversations with Furhat.
3. Creation of suitable data structures to represent every utterance from the conversations.
4. Dialogue act classification of every utterance with a well-known machine learning algorithm (J48).
5. Investigation of the impact of different attributes on the accuracy of the dialogue act classification.
6. Literature study of different methods and models within the research field of operations strategy and marketing.
7. Literature study of the market for social robots used for educational purposes and of ongoing projects related to the field.
8. Interview with Preben Wik about Furhat Robotics and the potential markets for robots.
9. Evaluation of the market for social robots which are used for education from the perspective of a small startup company (Furhat Robotics).

Part I

Dialogue Act Classification with Machine Learning

Chapter 2

Theoretical Framework

This chapter provides the theoretical framework for this thesis. First, the previous work in this area and the robot Furhat is presented in more detail. Thereafter, the theory behind machine learning classification is explained. This is followed by a description of the J48 algorithm which was used in this study to implement a dialogue act classifier. Lastly, the evaluation methods and metrics used to evaluate the classifier are described.

2.1 Previous work

There are several research papers on the subject of dialogue act classification due to its importance for dialogue research. Much research has been made possible by multiple available corpora which have been tagged with dialogue acts (examples are HCRC Map Task, CallHome and Switchboard). To solve the task of classifying dialogue acts, researches have used numerous methods.

A big part of the previous work in this area is based on conversations in written form, such as online chat conversations or forums. Samei et al. (2014) [4] investigated the role of context in chat-based intelligent tutoring systems using the machine learning methods *Naïve Bayes* and *J48 Decision Tree*. Rus et al. (2012) [5] developed a method for automatically classifying dialogue acts from chat conversations extracted from educational games. The authors used clustering methods to find natural groupings in the utterances.

In addition to the studies based on conversations in written form, there are some recent reports which address the problems with human-robot and human-human interaction.

Chen and Di Eugenio (2013) [6] used human-human conversations to classify dialogue acts with several machine learning algorithms while Wilske and Kruijff (2006) [7] investigated how service robots could handle indirect dialogue acts using other methods than machine learning. The goal of this study was to explore dialogue act classification in conversations with the robot Furhat. What distinguishes these conversations is that it is a three-way conversation, that a robot is participating and the different features available, such as the many prosody features used in this study which are explained later in section 3.5.5.

2.2 Furhat

The robot head Furhat uses an animation system which is back-projected on a translucent mask from a small projector located in the neck. This creates the illusion of a living being and makes the robot appear more human-like. Furhat can also move his head in a natural way, perform gestures and address people with eye contact. The box contains an Intel NUC computer with all the necessary software [8].

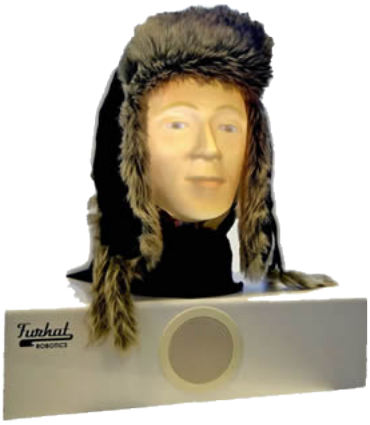


Figure 2.1. The robot head Furhat.

Furhat can interact with several persons at the same time and records multiple features from a conversation. Some of these features were used as attributes to help classify dialogue acts in this study. The attributes that were used are later explained in section 3.5.

2.3 Introduction to Machine Learning Classification

In this study, machine learning classification was used to implement a dialogue act classifier. Machine learning is one of many areas in the domain of artificial intelligence. It provides computers with a human-like ability to learn by remembering, adapting and generalizing. With the help of different algorithms, a learning system can spot patterns or make predictions based on provided input. The input is a set of data instances with a finite set of attributes and corresponding values. The output in the learning algorithms – being the parameter to predict – can vary in nature. For instance, one can attempt to predict quantitative measurements such as different numerical variables. Another approach is to predict categorical variables and thereby attempt to classify every data instance in the provided input. This is known as *machine learning classification* [9].

2.3.1 Supervised and Unsupervised Classification

Depending on the nature of the feedback available to the learning system, machine learning can be classified into different categories. There is one approach called *unsupervised learning* which often is referred to as "learning without a teacher". The task is to find hidden patterns and structures in the provided input data. The opposite is *supervised learning* [9]. Both methods can be used for classification of data, but when using unsupervised algorithms one can not decide which classes to use in advance. However, when using a machine learning classification algorithm in a supervised manner, one limits the amount of possible input and output values to a finite set of classes $Y = \{class_1, \ldots, class_N\}$ where $N \in \mathbb{N}_{>0}$ [9]. Therefore, the supervised approach is suitable in dialogue act classification when one has a limited amount of chosen dialogue acts for the natural language.

2.4 Machine Learning Algorithm and Evaluation

The following sections will explain the machine learning algorithm which was used for dialogue act classification in this study as well as the chosen evaluation method and evaluation metrics.

2.4.1 J48 Decision Tree

In this study, the J48 decision tree was used for classifying dialogue acts. The J48 algorithm is a Java implementation of the C4.5 algorithm which was developed by J. R. Quinlan [10]. C4.5 is a recursive partitioning algorithm which results in a decision tree that is used to classify new and unseen data. At each node of the tree, the algorithm computes the attribute value which best splits the dataset into parts where the subsets are rich in some class. In the C4.5 algorithm, Quinlan uses the concept of *entropy* to measure the goodness of a split. The splitting criterion is called the *gain ratio* [10] and is computed in a few steps which will be explained below.

Let there be a set of D cases and C different classes in total. Let $p(D, j)$ denote the proportion of cases in D belonging to the class j. The entropy of D can then be computed as:

$$E(D) = -\sum_{i=1}^{C} p(D, j) \times log_2(p(D, j)) \tag{2.1}$$

The corresponding gain in information on an attribute A with k different possible outcomes is defined as:

$$G(D, A) = E(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times E(D_i) \tag{2.2}$$

The number of outcomes affect the information entropy that the algorithm can gain. The gain is maximal when there is one case in every set D_i [10]. $G(D, A)$ thereby favors the attributes which have a large number of values. To compensate for this, Quinlan suggests to compute the *split information* $S(D, A)$, which is the entropy due to the split of D based on the value of the attribute A. The split information can be computed as:

$$S(D, A) = -\sum_{i=1}^{k} \frac{|D_i|}{|D|} \times log_2 \left(\frac{|D_i|}{|D|} \right) \tag{2.3}$$

Finally, the gain ratio – to be used as a splitting criterion in the construction of the decision tree – can be computed by combining equation 2.2 with 2.3 and computing a ratio:

$$GainRatio(D, A) = \frac{G(D, A)}{S(D, A)} \tag{2.4}$$

The gain ratio assesses the desirability of a test on an attribute A and is computed for every possible scenario. When splitting the decision tree, the algorithm chooses the split with the highest gain ratio. There are some special cases, for instance when none of the attributes provide any gain in entropy. The algorithm has additional methods for handling these special cases. In the end, the abovementioned steps – together with some additional methods and pruning of the tree – result in a decision tree which is consistent with the training data [10]. It can thereby be used for classification of similar data.

The C4.5 algorithm, in this study used in the J48 Java implementation, requires some pre-processing of data to work in an optimal way. In this study, conversations with Furhat were used as training data. This is further explained in chapter 3.

2.4.2 K-fold Cross-validation

To estimate the accuracy of a classifier, such as the one produced in this study, one can use *k-fold cross-validation.* The dataset D is split into k mutually exclusive folds: $D_1, D_2, \ldots, D_k$. The model is then trained and tested k times. Each time it is trained on the set difference $D \backslash D_t = \{x : x \in D \text{ and } x \notin D_t\}$ where $t = \{1, 2, \ldots, k\}$. It is then tested on D_t. The result of the cross-validation, or more precisely the accuracy of the model, is computed by dividing the number of correct classifications with the total number of instances in D [11].

Cross-validation is a way to reduce the variance of the estimation of accuracy. Kohavi [11] showed that *stratified cross-validation* can reduce this even further and that it also results in a low bias. Stratified cross-validation ensures that the division into k folds approximately has the same representation of class values in every fold. In this study, stratified 10-fold cross-validation was used to train and test the dialogue act classifier. When choosing how many folds one should use, there is a trade-off between accuracy of the model and variance of the accuracy. Kohavi showed that the number 10 seems to strike a good balance for many problems, which is why $k = 10$ has been used in this study.

2.4.3 Evaluation Metrics

A *confusion matrix* contains information about the actual and predicted class, making it easy to evaluate the performance of a classification system. The following table shows the structure of a confusion matrix. The system will predict if the instances belong to a specific class or not (represented by "other class" in the table). All off-diagonal elements (false positive and false negative) represent misclassified data and a good classifier will therefore have a dominant diagonal.

		Predicted	
		Class	Other class
	Class	True positive (tp)	False negative (fn)
Actual	Other class	False positive (fp)	True negative (tn)

Table 2.1. The confusion matrix which shows the result of a classification system which has classified a number of instances as either belonging to a specific class or not.

Precision and *recall* are measurements used to evaluate and interpret the results and their relevance. Precision is defined by:

$$Precision = \frac{tp}{tp + fp} \times 100 \qquad (2.5)$$

A perfect precision (100 %) for a specific class means that the system was able to classify every instance of the specific class correctly. This implies that all instances that were predicted to belong to the specific class actually belonged to that class.

Recall, on the other hand, is defined by:

$$Recall = \frac{tp}{tp + fn} \times 100 \qquad (2.6)$$

A score of 100 % for a class indicates that every instance that belongs to the specific class was categorized as that class.

There is an inverse relationship between precision and recall, which makes it hard to improve the performance of one measure without dropping the performance of the other one (referred to as the “recall-precision tradeoff”) [12].

A measure that combines both precision and recall is called *F-measure*. The combined metric gives an overall view of the system's performance and is defined by:

$$Fmeasure = 2 \times \frac{recall \times precision}{recall + precision} \tag{2.7}$$

Accuracy measures the overall performance of the system. A perfect score indicates that all the labeled data was predicted and categorized correctly. Accuracy is defined by:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \times 100 \tag{2.8}$$

Chapter 3

Implementation

This chapter explains the implementation of the dialogue act classifier. To begin with, a brief description of the delimitaions and limitations which have set the boundaries for part I is presented. After that, the collection and preprocessing of data from conversations with Furhat is described. This is followed by an explanation of how the machine learning algorithm was implemented. Finally, the approach used to evaluate the classifier is described.

3.1 Delimitations and Limitations

The main delimitations and limitations for part I are the following:

- Only a small sample of data from nine conversations with Furhat was used as input to the machine learning algorithm. This was a necessary delimitation because of time constraints, since the transcription of data (described in section 3.3) was time consuming. However, we believed this to be enough data to be able to draw conclusions regarding the different attributes and their impact on the classification.

- Only one algorithm, namely *J48 decision tree*, was used to implement a dialogue act classifier. Thereby no comparison was made between different algorithms – only between different sets of attributes.

3.2 Data Collection

The robot Furhat was exhibited during the Research Demonstration Week at Tekniska Museet in Stockholm, Sweden. For the purpose of this study we were supplied with data in the form of recorded video and audio. The data was collected from conversations between Furhat and visitors, where each had three participants (including Furhat). In the exhibition, the participants, together with Furhat, tried to rank the cards which were displayed on a digital screen based on a given criteria. As an example, the goal of one game was to rank buildings by height. Altogether, 9 conversations were used as material for this study, adding up to 687 separate utterances in total (excluding Furhat's utterances). All conversations were held in Swedish and therefore all data used as material in this study was in Swedish.

3.3 Transcription of Data

It was necessary to preprocess the data in order to prepare it for later analysis. The audio tracks were divided into different segments, where each segment represented an utterance. These were then systematically transcribed. Utterances with no obvious meaning but nonetheless of linguistic character (such as "mhm" and "ah") were kept. However, laughter, coughs, grunts and other non-linguistic utterances were removed. Although they can be an important part of the spoken language, they might not affect the classification to a great extent. Therefore, the decision to remove these utterances was made to limit the scope of this study.

3.4 Machine Learning Approach

A machine learning algorithm known as *J48 decision tree* was used to implement the dialogue act classifier. This algorithm was chosen since it has been used in many previous studies. These studies were, however, based on conversations in text. It was therefore interesting to investigate if it would classify as accurately on human-robot conversations. Another reason for choosing the algorithm was that decision trees are easy to understand and examine. This property was helpful when analyzing the different attributes and their impact on the classifier. A detailed description of how the algorithm works is in section 2.4.1. To gain access to the J48 algorithm, WEKA (an acronym for Waikato Environment for Knowledge Analysis) was used. This section explains what WEKA is and how it handles input.

3.4.1 WEKA

WEKA is a collection of well-known machine learning algorithms which can be applied to datasets to perform data mining tasks. The software is open source and issued under the GNU General Public License [13].

3.4.2 ARFF Format

The algorithms in WEKA can use the Attribute-Relation File Format (ARFF) as input file format to perform training and testing of classification. ARFF files are text files with ASCII encoding which declare a set of available attributes and a set of data instances – or attribute vectors – sharing these attributes [14].

3.5 Attributes Used to Classify Dialogue Acts

In machine learning classification, every piece of data is represented as a set of attributes with corresponding values. The algorithm uses the attributes to build a model, so that one of the attributes can be predicted given all the others. In this study the purpose was to predict the dialogue act of an utterance. Apart from the words in the utterance, there is a variety of possible attributes that one could use depending on which information is available. Table 3.1 shows the attributes that were used in this study.

Attributes
Dialogue act
Transcript unigrams
Part-of-speech (POS) n-grams
Word count
Prosody attributes (13 in total)
Previous dialogue act

Table 3.1. The different attributes for every utterance used in the machine learning algorithm.

The following sections will explain the different attributes in table 3.1 in more detail.

3.5.1 Dialogue Act

One of the attributes – and the attribute which was predicted – was the dialogue act of an utterance. Which categories are most suitable to use depends on the situation. Previous research and studies have used a wide range of different categories, though many of these studies have been made on data which is very different from the conversations with Furhat. For example, as discussed in section 2.1 several studies have classified dialogue acts on written text, such as chat messages. The decision of which categories to use in this study was based on the information which would be important from Furhat's perspective. This resulted in 7 different categories which are explained below in table 3.2.

Greeting	Greeting phrases in the beginning of a conversation to say hello to each other.
Statement	A wide category containing utterances where the speaker spontaneously wishes to express something. Answers to questions and opinions about previous utterances belong to the opinion category.
Question	Utterances where it is obvious that the person speaking is expecting an answer. In spoken language, questions are not always as grammatically correct as questions in written text. For example, prosody can be taken into account when trying to understand the difference between questions and statements in spoken language.
Feedback	Often short words used in a conversation, such as "mm" or "mhm", to signal that one has understood what has been said or is expecting another person to continue speaking.
Interjection	Words or short sentences used to express emotions or sentiments such as "shit" or "oops". This category also contains filled pauses. Examples of such disfluencies are "eh", "ehm" and "hmm".
Opinion	Answers to questions and opinions about previously uttered statements. A person can either agree or disagree to utter an opinion about something a previous speaker has said. There are also some cases where the speaker is unsure and such utterances also belong to this category. Spontaneously uttered opinions are not a part of this category but belong to the statement category, since they do not refer to anything previously said.
Aborted	Incomplete utterances.

Table 3.2. The different categories of dialogue acts which were used in this study.

Table 3.3 shows a part of a transcript from a recorded conversation between the robot Furhat and two visitors at Tekniska Museet. At this point, the participants are discussing which building is the tallest between the ones presented on the digital screen in Furhat's game. Note that no punctuation marks such as dots, question marks or commas were used to represent the data. Only exactly what was said was transcribed.

Speaker	**Transcript**	**Dialogue Act**
Furhat	vilken byggnad tycker du vi ska börja med	Question
Speaker 1	öh vilken är högst	Question
Speaker 2	jag tror den	Statement
Speaker 1	ja den där ja ja	Opinion
Speaker 2	och sen kommer nog sen kommer nog den tror jag	Statement
Furhat	mm	Opinion
Speaker 1	ja	Opinion
Speaker 2	och sen kommer den tror jag	Statement
Speaker 1	ja det kan det nog	Aborted
Speaker 1	nja jag tror den är högre	Opinion
Furhat	mhm	Feedback
Speaker 2	så	Feedback
Speaker 1	så kanske	Statement

Table 3.3. A part of a transcript from a conversation between Furhat and two visitors at Tekniska Museet.

This is an example of the recorded data and the dialogue acts that were chosen for each utterance. It might not be obvious for the reader which dialogue act categories the utterances belong to just by reading the transcript. However, in human-human interaction other variables are taken into consideration, such as prosody, body language and context. This is one of the challenges in human-robot interaction. Another significant challenge when understanding spoken language is that it is not always grammatically correct. The transcript is an example of this where one utterance is incomplete (and tagged as “Aborted”) and some contain repetitive parts.

3.5.2 Transcript Unigrams

The words spoken in each utterance had to be represented by attributes. Instead of having only one attribute with a string value, a suitable method was to use *n-grams*. An n-gram, in this context, is a sequence of *n* words, extracted from all the given utterances in the training data. In this study, unigrams, or single words, were chosen to represent the utterances. Higher orders of n-grams could be used to better represent the structure of the utterances. However, this was covered in the part-of-speech n-gram attributes described in section 3.5.3. Therefore, only unigrams were chosen to represent the content of the utterances. Each word unigram in the entire training set was associated with an attribute. The following is an example of unigrams used in the machine learning algorithm, declared as attributes in ARFF format (this format is described in section 3.4.2). The label "numeric" declares that the attribute for a certain utterance can be given a value of 0 or 1, depending on if it contains the unigram or not.

```
@attribute   'ganska'   numeric
@attribute   'du'       numeric
@attribute   'tar'      numeric
@attribute   'vi'       numeric
@attribute   'byta'     numeric
@attribute   'gaffel'   numeric
```

3.5.3 Part-of-Speech N-grams

Part-of-speech (POS) tagging was used to further represent the grammatical structure and content of the utterances. Every sentence (or utterance) was transformed into a POS sentence with the help of Stagger – an open source POS tagger for Swedish [15]. These POS sentences were later transformed into unigrams, bigrams or trigrams (sequences of one, two or three POS tags) and tested separately to compare the results. Below is an example of POS bigrams that were used in the training of the dialogue act classifier. The START and END tags symbolize where the utterances start and end. VB represents verbs, NN represents nouns, JJ adjectives and KN conjunctions.

```
@attribute   START VB   numeric
@attribute   JJ KN      numeric
@attribute   NN END     numeric
@attribute   VB END     numeric
```

3.5.4 Word Count

The length of an utterance could be a useful aspect to take into account when classifying dialogue acts. Therefore, this was used as an attribute. Every word in every separate utterance was counted and declared as a positive integer to represent the length of the utterance.

3.5.5 Prosody Attributes

Prosody is the rhythm, intonation and stress of speech. This was represented as several attributes in the speech act classifier.

The *pitch* (or intonation) is measured by Furhat's system in Hertz, representing how fast the speaker's vocal folds are vibrating. This only happens when the speaker is uttering vowels and voiced consonants (such as "M" and "L", but not for example "S" and "T"). The Hertz value is then converted into semitones (a logarithmic scale) in order to make it into a normal distribution. This is later normalized with Z-scores for every speaker. Thereby, a value of 0 indicates the normal intonation (the mean) of the speaker in question. A value of -1 means one standard deviation below the mean and 1 means one above the mean.

The *energy* of every utterance is measured by the system to indicate if the speaker is speaking with a strong of weak voice. This is – in the same way as with the pitch – normalized by z-scores for every speaker.

Below is an explanation of the different categories that were measured for each utterance. The variable X indicates a certain period in an utterance.

1. *pitch.X.mean:* The average Z-score in the period X. Measures if the utterance ended in a low or high pitch.
2. *pitch.X.stdev:* The standard deviation of the Z-score in the period X. Measures how much the pitch changes in the end of the utterance.
3. *pitch.X.max:* The maximum Z-score of the period X.
4. *pitch.X.slope:* Positive or negative change of the Z-score during the period X which measures in which way the pitch changes.
5. *energy.X.max:* The maximum energy during the period X to indicate if the speaker used a strong or weak voice.

In this study, three values of X were used: X = the last 200 ms, X = the last 500 ms and X = the whole utterance. This resulted in 13 different prosody attributes in total, shown in table 3.4.

Prosody Attributes
pitch.200.mean
pitch.200.stdev
pitch.200.max
pitch.200.slope
energy.200.max
pitch.500.mean
pitch.500.stdev
pitch.500.max
pitch.500.slope
energy.500.max
pitch.all.stdev
pitch.all.max
energy.all.max

Table 3.4. The 13 different prosody attributes for every utterance used in the machine learning algorithm.

3.5.6 Previous Dialogue Act

Context is important when it comes to understanding and following a conversation. One can imagine that the dialogue act of a previous utterance has an impact on the dialogue act of the following one. For instance, if a question is asked the following utterance is likely to be of the class "opinion" (described in table 3.2). However, this could not be fully implemented in the dialogue act classifier since Furhat is not able to beforehand know the dialogue acts of the other speakers' utterances. He can only know the dialogue acts of his own ones. Therefore, an attribute was created for every utterance containing the information about the previous dialogue act from Furhat. The three following utterances were chosen to contain this contextual information, and the following ones were labeled with the attribute value "noprev", meaning no previous dialogue act was available.

3.6 Attribute Vectors in ARFF

To represent the information contained within each utterance, attribute vectors were created in the file format ARFF. Programs were written in XQuery and Python to extract and create the necessary attributes as well as parse, modify and put together the entire data set in ARFF. The constructed programs will not be explained in more detail since this would not contribute to the understanding of the classifier.

Below is an example of an ARFF file with two examples of attribute vectors (under the label "@data") which were used as training material in this study for the machine learning algorithm. Note that some attributes have been left out.

```
@relation training_data

@attribute da {greeting, statement, opinion, ...}
@attribute wordcount numeric
@attribute transcript string
@attribute pos string
@attribute pitch200mean numeric
...
@attribute energyallmax numeric
@attribute previous {greeting, statement, ..., noprev}

@data
greeting, 2, 'hej Furhat', 'START PM PM END', −1.20, 0.19,
0.00, −0.26, 0.00, −1.01, 0.27, 0.00, −0.39, 0.00, 0.36,
0.00, 0.00, statement

statement, 3, 'han är läskig', 'START PN VB JJ END', 0.32,
0.20, 0.69, 0.05, 0.00, 0.28, 0.21, 0.69, 0.06, 0.00, 0.21,
0.69, 0.00, statement
```

The string values for every utterance and every POS sentence were later represented as numeric n-gram attributes (with values 0 or 1), as described in sections 3.5.3 and 3.5.2.

3.7 Evaluation of the Dialogue Act Classifier

After pre-processing the data and creating an input file for WEKA, different attribute sets were chosen and passed through the J48 algorithm to create dialogue act classifiers.

3.7.1 Comparison Between Different Attribute Sets

Different sets of attributes were chosen and evaluated in order to compare the results. For every attribute set, 10-fold cross-validation was used to train and test the model for classification of dialogue acts. A J48 decision tree was thereby built for every different attribute set.

3.7.2 Baseline

All sets of attributes were compared against a baseline algorithm available in WEKA known as ZeroR. This algorithm has a naïve approach which only predicts the most frequent class for every unseen data instance. In this study, the most frequent class was "statement", consisting of 287 instances out of 687 in total.

Chapter 4

Results

This chapter presents the results of the machine learning classification of dialogue acts. The proportion of the different classes is shown in a diagram and the accuracy of the produced models is presented with different sets of attributes. Additionally, evaluation metrics are shown for the POS n-grams, to understand why one n-gram was better than the other. Lastly, detailed measurements for the prosody attributes are shown.

4.1 Classes of Dialogue Acts

Figure 4.1 shows the proportion of the different classes in the input data.

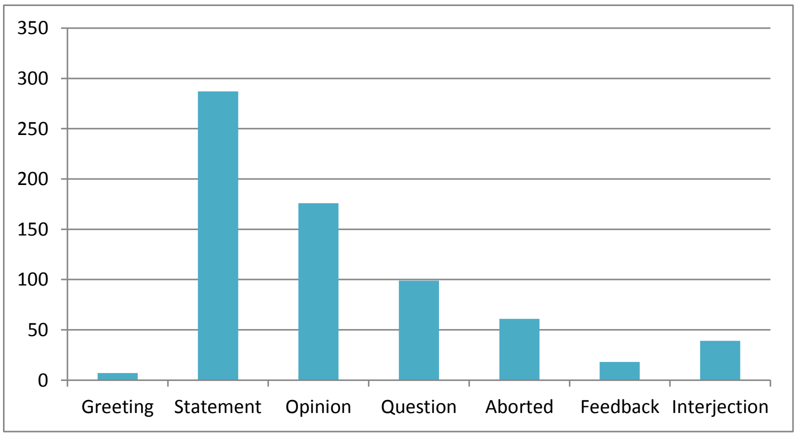


Figure 4.1. The proportion of every class in the input data.

4.2 Comparison Between Attribute Sets

Figure 4.2 shows the accuracy of the dialogue act classifier with different sets of attributes. As described in section 2.4.3, the accuracy denotes the amount of correctly classified instances out of the total amount. A baseline (ZeroR) has also been added to the chart to show which attributes resulted in a performance gain compared to the baseline.

Number 1-7 in figure 4.2 shows the accuracy of every individual attribute category. Number 8 and 9 are combinations of attributes, where 8 is a combination of number 1-7. Number 9 shows the best combination, where the best alternative for representing POS tags was chosen together with the other attributes which performed better than the baseline.

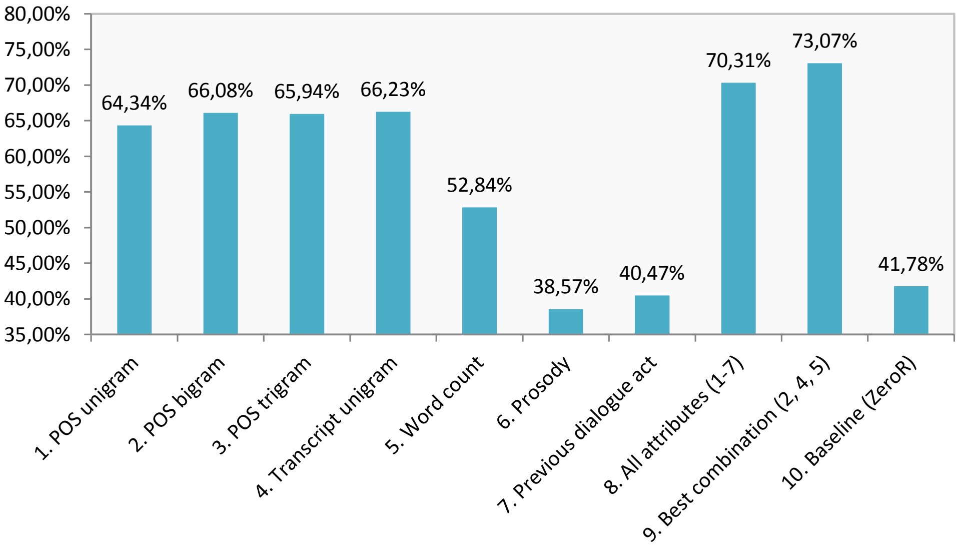


Figure 4.2. The accuracy of predicting the dialogue act attribute with different attribute sets. Evaluated with 10-fold cross-validation with a J48 decision tree.

4.3 Different POS N-grams

Three different n-grams of POS tags (unigram, bigram and trigrams) were used in this study to represent grammatical information in the utterances. The tests showed that noticeable changes in precision, recall and F-measure only could be seen in the "question" class. The difference can be seen in table 4.1.

Classifying questions

Attributes	Precision	Recall	F-Measure
POS unigrams	0.767	0.333	0.465
POS bigrams	0.741	0.636	0.685
POS trigrams	0.763	0.616	0.682

Table 4.1. Precision, recall and F-measure when using different n-grams of POS tags as attributes. The results are taken from the classification of the "question" class.

To understand why bigrams and trigrams would be better for classifying questions, some decision rules were extracted from the J48 decision tree for bigrams. This indicates that these bigrams are commonly used in questions. Further discussion and analysis on this subject is provided in section 5.2.

Leaf rule	Instances at leaf	Incorrectly classified
"HP VB" > 0	26	1
"START HD" > 0	9	1
"JJ KN" > 0	3	0
"PP DT" > 0	3	1
"NN END" <= 0	49	16
"PN VB" <= 0	2	0

Table 4.2. Decision rules extracted from the decision tree for classifying dialogue acts with POS bigrams. The tags are commonly used annotations for part-of-speech tags.

4.4 The Best Attribute Set

The attribute set which gave the best performance in terms of an accuracy of (73,07 %) was a combination of:

- Part-of-speech (POS) bigrams
- Transcript unigrams
- Word count

This combination is shown in figure 4.2 as combination number 9. Table 4.3 shows the confusion matrix of this result.

		Predicted						
		G	S	O	Q	A	F	I
Actual	G	6	1	0	0	0	0	0
	S	0	248	14	11	11	0	3
	O	0	29	142	3	0	2	0
	Q	0	22	2	75	0	0	0
	A	0	42	6	8	5	0	0
	F	0	4	6	0	0	8	0
	I	0	12	6	0	0	3	18

Table 4.3. The confusion matrix of the best combination of attributes (number 9 in table 4.2), where G, S, O, A, F and I denote Greeting, Statement, Opinion, Question, Aborted, Feedback and Interjection.

Table 4.4 shows the precision, recall and F-measure for the abovementioned combination of attributes.

Class	Precision	Recall	F-Measure
Greeting	1.000	0.857	0.923
Statement	0.693	0.864	0.769
Opinion	0.807	0.807	0.807
Question	0.773	0.758	0.765
Aborted	0.313	0.082	0.130
Feedback	0.615	0.444	0.516
Interjection	0.857	0.462	0.600

Table 4.4. The different measures for every class in the dialogue act classifier with the best combination of attributes (number 9 in table 4.2).

4.5 Prosody

As shown in figure 4.2, the accuracy when using prosody attributes to classify dialogue acts was lower than the baseline (38,57 % compared to 41,78 %). The prosody attribute category is a combination of 13 different attributes. These are explained in section 3.5.5. The following figure shows the accuracy for each individual attribute, making it easy to distinguish the ones with a lower accuracy than the baseline.

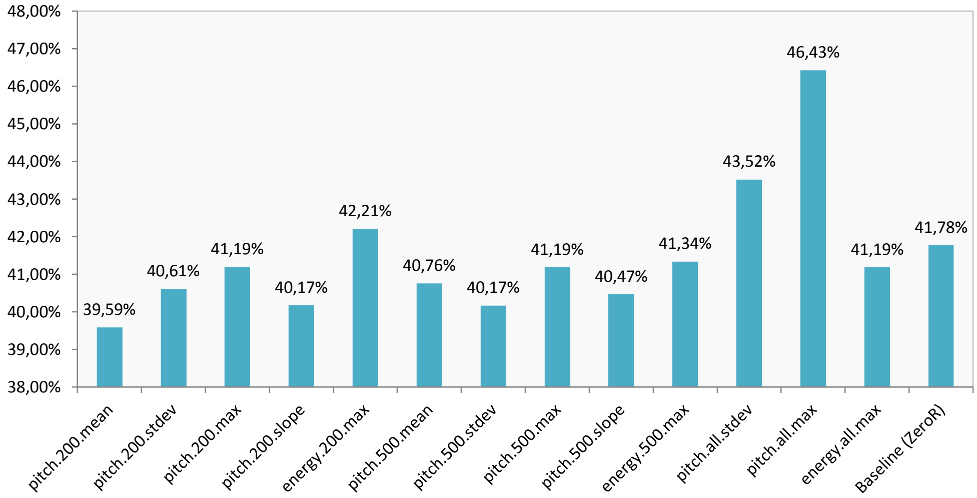


Figure 4.3. The accuracy of all individual attributes that together form the prosody. Evaluated with 10-fold cross-validation with a J48 decision tree.

Chapter 5

Discussion

This chapter discusses and analyzes the results in the previous chapter and attempts to seek answers to the research questions presented in section 1.3. First, the different attribute sets are discussed with a focus on the two attribute categories (prosody and previous dialogue act) that did not result in a performance gain. Furthermore, the distinction between the different POS n-grams is analyzed with the help of decision rules from the decision tree. Finally, the best combination of attributes in terms of accuracy is discussed.

5.1 Different Attribute Sets

The different attributes were tested and the accuracy of each attribute was compared to the baseline in figure 4.2. All attributes except for the 13 combined prosody attributes and the previous dialogue act attribute gained higher accuracy than the baseline. The two attributes which did not help the classifier are discussed below.

5.1.1 Prosody

Previous studies [16] indicate that prosody is important for dialogue act recognition and that it can help to distinguish dialogue acts which have identical word sequences but is uttered in different ways – meaning that they have different prosodic values. However, the results in this study indicated that the combination of the 13 prosody attributes resulted in a lower accuracy than the baseline.

Figure 4.3 shows the accuracy for the 13 individual attributes. Previous studies [16] indicate that some dialogue acts can be distinguished by their final F0 raise. Since F0 measures the pitch, this could explain the high accuracy value for pitch.X.stdev and pitch.all.max. These attributes, and energy.200.max, performed higher than the baseline, which indicate that they can be used separately to classify different dialogue acts.

The reason why a combination of these attributes, together with the other prosody-related ones, resulted in a much lower accuracy can be discussed. Although some of them have a lower accuracy than the baseline, all the attributes performed better individually than when combined (38,57 %). This indicates that they are redundant and that the large amount of information confuses the system instead of improving it.

5.1.2 Previous dialogue act

The previous dialogue act attribute also gave a lower accuracy than the baseline. In this study, this attribute assigned the robot's dialogue acts to the following three utterances, regardless of who the utterances belonged to (speaker1 or speaker2). This was an attempt at representing some form of context, though the assigned value might not always provide meaningful information which makes it a relatively unreliable attribute. However, this attribute could have been of great use if it had been built on better data. The problem is that it was not possible to predict the dialogue acts of speaker1 and speaker2 and create contextual attributes from this at the same time as running the J48 algorithm. This would have required creating an iterative function which could create new attribute values continuously, making it possible to assign the right (or at least predicted) previous dialogue act to all the utterances. A small improvement such as only assigning the previous dialogue act one time to speaker1 and speaker2, instead of assigning them to the following three utterances independently, might have resulted in a better outcome. This was, however, not investigated in this study.

5.2 Different POS N-grams

Three different POS n-grams were used as attributes – unigrams, bigrams and trigrams. Figure 4.2 shows the difference between them in terms of accuracy, where bigrams proved to be the best choice. There was not a significant difference between the three when looking at precision, recall and F-measure, except for the "question" class. This difference is shown in table 4.1. It appears that bigrams and trigrams help the classifier make a distinction between questions and other classes. Table 4.2 contains information which could help answer why. Two of the different leaf node rules contain *relative* POS tags. These are HP (relative pronouns) and HD (relative determiners). Examples of these in Swedish are "vad", "vilket" and "vilken", which appear frequently in questions. The rules ["HP VB" > 0] and ["START HD" > 0] denote such tags to either be followed by a verb or to start a sentence, and these are common ways to structure questions in Swedish. One can then understand why bigrams and trigrams would perform better in this case, since they can contain information of tags in sequence which are common in questions. Another rule which had an impact on the classification was ["NN END" <= 0], meaning that an utterance was classified as a question if it did not end with NN (a noun). This might not be as intuitive, but might be a result of the fact that many interjections and feedbacks were tagged as NN by Stagger. The classifier could thereby sort these away, which might be a reason why bigrams would be a better alternative.

5.3 The Best Combination

The best combination of attributes in terms of accuracy of the dialogue act classifier was described in section 4.4. These attributes were chosen from figure 4.2 since they performed better than the baseline. Together they resulted in an accuracy of 73,07 %, which is far better than the baseline of 41,78 %. Apparently, grammatical and lexical information together with the length of an utterance provide enough information to classify dialogue acts to a quite high accuracy in conversations similar to Furhat's.

Table 4.4 shows that the classifier performed well in almost all categories. However, the class "aborted" was difficult to predict. The F-measure for this class was as low as 0.130. Spoken language contains many incomplete utterances and this is obviously a challenge. One might think that the POS bigrams could help here, since they can contain information about the end of an utterance, but this was not the case. Further investigation is therefore needed to find better attributes to use which could solve the problem with low F-measure for incomplete utterances.

The confusion matrix (see table 4.3) shows if some classes were confused with others. In this case, it shows that greetings and opinions were not as much confused with other categories as for example aborted utterances, which were often confused with statements. Furthermore, a quite large proportion of actual questions and interjections were classified as statements.

5.4 Conclusions

This study showed that machine learning can be used to classify dialogue acts with a quite high accuracy in human-robot conversations. Although the results were based on data collected from conversations with the robot Furhat, they can be generalized to similar human-robot conversations. Different attributes, and their impact on the classification, were tested. While POS unigram, bigram, and trigrams, transcript unigrams and word count performed better than the baseline, the prosody attributes (all 13 used together) and the previous dialogue act had a lower accuracy.

The best combination of attributes out of the ones which were used proved to be a combination of POS bigrams, transcript unigrams and word count, which resulted in an accuracy of 73,07 %. This means that grammatical and lexical attributes together with the length of an utterance provided enough information to classify dialogue acts with a high accuracy. The results from this study can be used as a basis for further studies on how a dialogue act classifier could be implemented in Furhat or similar robots. Although we did not use a large amount of data in this study, we do not believe that more data would improve the classifier to a great extent. Further studies could instead investigate the impact of other attributes which were not a part of this thesis.

Part II

The Market for Social Robots which are Used for Education

Chapter 6

Methods

This chapter presents the different methods used to answer the research questions for part II which were presented in section 1.3. First, the structure of the literature study is explained. This is followed by a description of the interview which was conducted and finally the limitations and delimitations of this study are presented.

6.1 Literature Study

A literature study was performed in two phases. Phase one focused on collecting information about the market for social robots used for educational purposes. It also focused on studying articles about robots in society and ongoing projects about social robots in schools. One helpful roadmap to use when analyzing the market for robots in general is the *Robotics 2020 Multi-Annual Roadmap* presented by SPARC – the Partnership for Robotics in Europe [17]. This roadmap, which presents different markets and categories of robots, was used to support the discussion and conclusions in this study.

Phase two of the literature study focused on gathering information about different models to be used for analyzing the market from the company's perspective. The information was collected from research papers and books on the subject. The result of this part of the literature study – more precisely the chosen models for analysis – is presented in chapter 8.

6.2 Interview

A semi-structured interview was conducted with Preben Wik, the CEO of Furhat Robotics. A couple of discussion questions were prepared beforehand. The following list shows the different subjects which were discussed during the interview:

- Potential markets
- Existing competitors and potential new ones

- Strengths and weaknesses of Furhat Robotics
- The need for robots in society (and particularly human-like robots)
- Patents and rights

6.3 Delimitations and Limitations

Instead of investigating many similar robot companies, Furhat Robotics was used as a case study. Because of the qualitative nature of the analysis, the conclusions cannot be generalized to cover all companies similar to Furhat Robotics. Additionally, as a result of the time constraints, the analysis in this part was based on one interview only which was conducted with a co-founder of Furhat Robotics. The conclusions drawn in part II are therefore strongly affected by this perspective, even if additional research in the form of a literature study has been made. The analysis should therefore serve as inspiration for further studies of the market and not be used as a basis for decision-making.

Chapter 7

Background

This chapter presents an overview of previous studies and work related to social robots which are used for education. First, a summary of related studies is given. This is followed by a description of the educational robot market and examples of robots available on the market. Finally, the project EMOTE is presented.

7.1 Previous Studies of Social Robots for education

The use of social robots has increased dramatically during the last decade. Progress regarding social behavior and growing capabilities has resulted in humanoid and complex robots that are finding their way to medical centers, museums and living rooms, to name a few applications. The roadmap presented by SPARC [17] calls this category of robots *Consumer Robots*, containing several sub-domains. One sub-domain is the education field.

Previous studies within the field of robot-aid learning indicate that robots can be used to help students learn subjects as mathematics, science and languages. Furthermore, researchers have provided substantial evidence that robots help students to develop their collaboration skills and problem-solving abilities [18].

Robots which can serve for educational purposes can be divided into two categories: educational robots (also referred to as hands-on robots) and educational service robots, which are social and anthropomorphized robots [19]. Educational service robots are used as a subset of educational technology such as audiotapes, tablets and books. Studies show that young children perform better on examination and gain more interest when learning languages with the help of robots compared with previously used technology [20]. The robots' ability to add social interaction to learning context and in that way help students develop their collaboration skills, also give them an advantage compared to purely software-based learning.

Further researches investigate the different roles that a robot can take during the learning activity. Three main categories have been defined [21] [22]: *tool*, *peer* and *tutor*. Studies indicate that age and subject are two important factors that have to be taken into account when choosing what role the robot should have. For example, older children preferred a tutoring style when learning language while younger children were content with robots behaving as peers. The degree of necessary social behavior of the robot was also linked to the role of the robot. However, studies and experiments show that social robots in general generate much more interest compared to a less social agent – regardless of its role [21]. Studies also show that social robots led to a higher post-test score when teaching a foreign language to primary school students.

7.2 Educational Robot Market

Several social robots have been introduced to the education market. The different robots have their own specialization and purpose. For example, some robots can be used for teaching foreign languages while others can be used to teach science. The robots have various features which help them appear more human-like. Examples of such features are voice and facial recognition. However, the most well-known social robots, which are presented in this study, have a robot-like appearance. These can be seen in figure 7.1.

The following section gives two examples of social robots which have been introduced to the market and can be used for educational purposes.

7.2.1 Examples of Robots Used for Educational Purposes

The robot *Robovie* is an example of a humanoid robot that can be used to teach English. It has been preloaded with facial photos and voiceprints from 119 teacher and students. Robovie is also equipped with knowledge from a fifth-grade science book, which makes it suitable for younger students [23].

Another well established robot is *Nao.* A special edition called Nao Academics Edition was developed for educational purposes. The robot is supposed to help students with science and is used as an interactive tool for students in primary and secondary school. It can also be used as a platform to stimulate creativity and innovation in higher education [24]. Nao is an open platform, which allows and encourages developers to build and add features to the robot. This also makes it suitable to use as a platform for developing new generation of humanoid robots. In addition to the powerful brain, Nao can also recognize shapes, people and voices, which allows it to communicate in 19 different languages [25]. The robot is used worldwide with over 5000 sold units in over 50 countries [26].

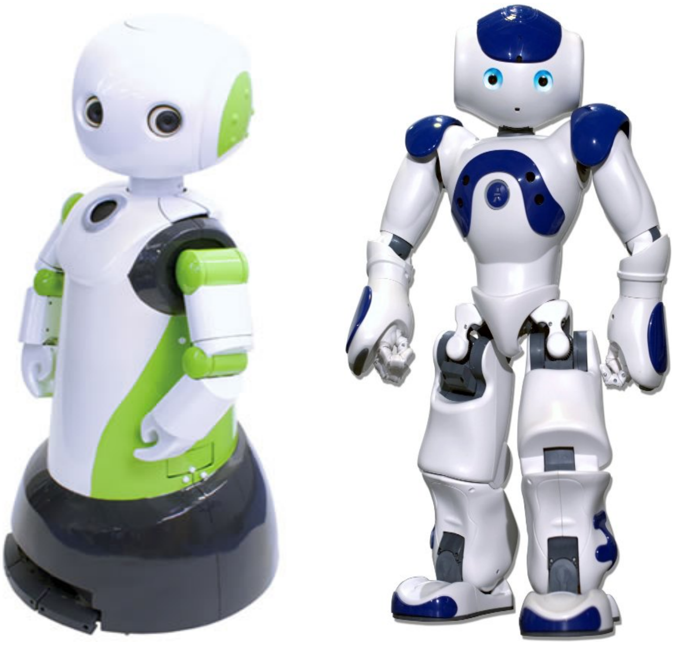


Figure 7.1. The humanoid robots Robovie and Nao.

7.3 EMOTE

EMOTE is a EU-funded project that started in December 2012. It was funded with 2,9 million Euros and aims to design, develop and evaluate if robots' ability to interpret and act on emotions can enhance the learning of school children [27]. In addition to creating an empathic robot the project will examine which emotions are relevant to the given situation and how these are expressed. Moreover, the project team will determine the abilities the robot needs to possess to be able to act as an emphatic robot tutor.

The project is a collaboration between six European institutions and companies. Sweden, Germany and England are some of the represented countries among the collaborators [27]. Research and studies are conducted in different schools with children in the age of 11-13. The robot which is used in the project is NAO T14 – a version of the Nao robot described in section 7.2.1 [28].

Chapter 8

Theoretical Framework

This chapter provides the theoretical framework necessary to understand the results shown in chapter 9. To begin with, a description of market segmentation is presented. This is followed by an explanation of the theory behind Porter's five forces and the SWOT-method.

8.1 Market Segmentation

A marketing strategy which enables a company to define and divide a large market into smaller segments with similar needs, interests and priorities is called *market segmentation* [29]. Market segmentation is used to gain a better understanding of the company's target audience and to make the marketing more effective. There are different market segmentation strategies which can be used when dividing the market into segments. The following is a summary of the four basic strategies:

1. **Geographic segmentation:** Divides customers into segments based on geographical areas such as nations, states, regions and cities.
2. **Behavioral segmentation:** Division based on costumers' attitude, response and use of a product.
3. **Demographic segmentation:** Divides customers into segments based on values such as age, gender, family, income, education, religion, race etc.
4. **Psychographic segmentation:** Used as a supplement to geographic and demographic segmentation and divides people according to their attitudes, values, lifestyles, interest and opinions.

In addition to identifying segments, the company must examine the different segments and decide which segment or segments to target. Factors which need to be taken into consideration when deciding the target group can for example be the target size and how competitive the segment is.

8.2 Porter's Five Forces

In order to stake out a position on the market, which is profitable and not too vulnerable to attack, a company must take a couple of aspects into consideration. Michael E. Porter summarized his thoughts on this subject in a model known as Porter's Five Forces. According to Porter, having awareness of these five forces can help a company understand the structure of its industry [30].

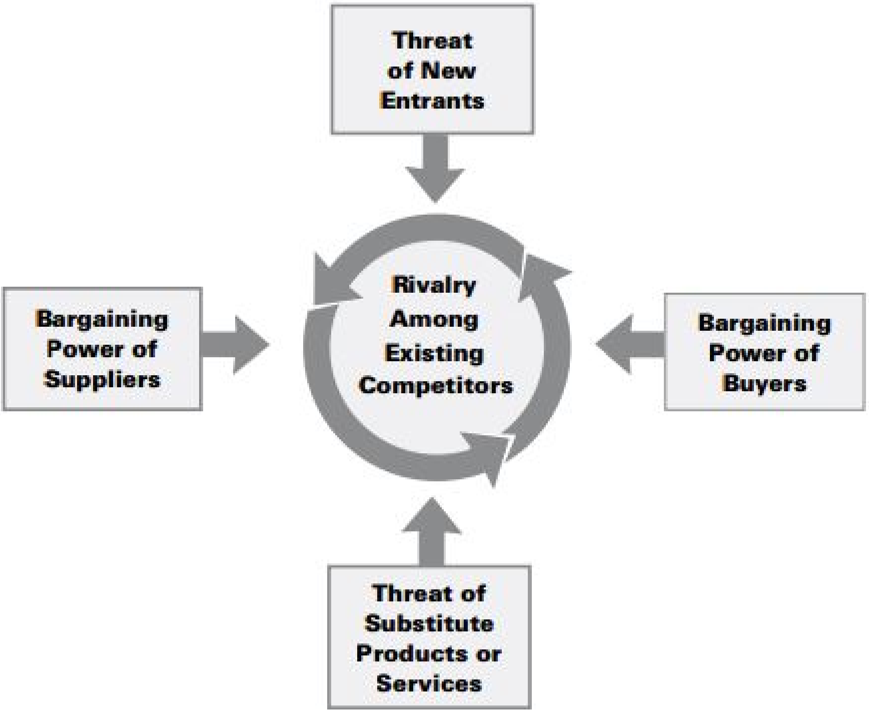

Threat of New Entrants
Bargaining Power of Suppliers
Rivalry Among Existing Competitors
Bargaining Power of Buyers
Threat of Substitute Products or Services

Figure 8.1. The five competitive forces that shape strategy, presented by Michael E. Porter in 1979.

The following is a summary of the five forces:

1. **Threat of New Entrants**. New entrants to an industry intensifies the existing competition. Entry barriers have an effect on how much of a threat new entrants could pose. Porter presented a number of different factors, such as capital requirements, government policy and access to distribution channels.

2. **Threat of Substitute Products or Services**. Substitutes are always present. However, they can be easy to overlook since they appear to be so different from the product in question. The threat of a substitute is high if it has an attractive price-performance trade-off to the product and the buyer's cost of switching to the substitute is low.

3. **Bargaining Power of Buyers**. Buyers on the market can have a powerful role. They can for instance force down prices, demand higher quality or better service – thereby lowering industry profitability. This is especially a threat if buyers are price sensitive in this area.

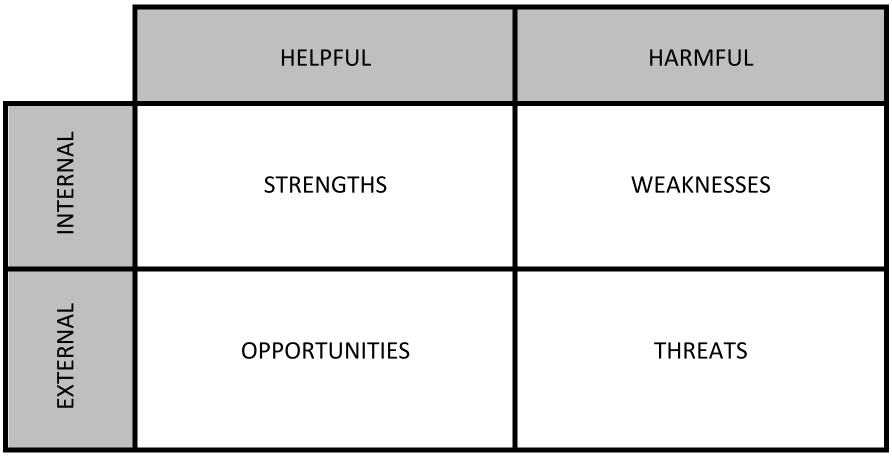
4. **Bargaining Power of Suppliers**. In the same way as buyers, suppliers can force down industry profitability. Suppliers can become powerful if the supplier group is more concentrated than the group it sells to.

5. **Rivalry Among Existing Competitors**. Rivalry within the industry comes in many different forms, such as new product innovations, price discounting, powerful advertising and service improvements.

8.3 SWOT

A SWOT-analysis is a method used to evaluate a company, project or product, based on four different perspectives: *strengths, weaknesses, opportunities* and *threats* [29]. Strengths are characteristics which give the company competitive advantages while weaknesses are characteristics leading to a disadvantage relative to other companies. Opportunities can be described as elements that can be exploit and lead to competitive advantages and threats as elements that can jeopardize the company's position.

The aim of the analysis is to give an overview of the company's current situation and to identify both internal and external factors to develop a better understanding of the company and its potential markets. SWOT can also be used when choosing a strategy and to gain competitive advantages.



	HELPFUL	HARMFUL
INTERNAL	STRENGTHS	WEAKNESSES
EXTERNAL	OPPORTUNITIES	THREATS

Figure 8.2. The SWOT-matrix containing the four elements; strengths, weaknesses, opportunities and threats. The matrix is used to get an overview of the company's current situation.

Chapter 9

Results

This chapter presents the results of part II, consisting of the conducted interview, Porter's Five Forces and a SWOT-analysis as well as a market overview.

9.1 Interview with Preben Wik

Preben Wik is the CEO of Furhat Robotics. During the interview, Wik stated that their first costumers in the startup phase have been research centers. However, they are now considering to enter the education market. He noted that it is not difficult do develop and adapt the existing software. Furhat could therefore be used to teach a wide range of subjects. In addition to the education market, there are more potential markets for the robot head such as nursery homes, shops and airports. During the discussion of potential markets, Wik also expressed that there has been a revolution in social robotics but that people may not be so positive to the change. Wik stated that previous work has focused on the motor functions instead on focusing on developing the humanoid features. This is an advantage for Furhat Robotics which focuses on improving those skills. He stated that their unique interface and Furhat's human-like features, as eye contact and lip-syncing, differ it from the rivals and that companies such as Google, Apple and Microsoft have showed interest in the advanced robot head.

The interview also revealed that the company can not specialize towards a specific market which requires a lot of improvements (such as the medical market) because of the lack of financial resources, which is needed to hire competent people that could develop their product.

After having discussed the open platform which the team has developed, it is clear that they are not frightened of the possible threat of someone stealing their research. Wik expressed that people should have the opportunity to contribute to the platform and help each other develop robots for the future. However, they have considered to patent some of their work but are unsure of what they should and can protect. It is also expensive to patent and the company cannot afford to fund such costs.

9.2 Porter's Five Forces

1. Threat of New Entrants

Since the market is in an initial stage, the threat of new entrants is quite high. However, the entry barriers are also high. Porter presented different factors which affect this. Below are a few.

1. *Demand-side benefits of scale.* Since the market for educational social robots is very new, and products are relatively expensive, customers might be insecure and not so well-informed – thereby choosing well-known and established suppliers. The demand-side benefits of scale could therefore be quite big in this market.

2. *Capital requirements.* Developing a robot is costly. However, this phase is already passed for Furhat Robotics. Additional capital would be needed to enter the market for robots used for education. In addition to research and development costs, Furhat Robotics may have to take other factors, such as patent costs, into account.

3. *Incumbency advantages independent of size.* Although Furhat Robotics is not as established as other companies, they have quality advantages which are not available to potential rivals. This includes 15 years of related research and cumulative experience leading to efficiencies.

2. Threat of Substitute Products or Services

The threat of substitute products or services is big for educational robots. The following is a list of the substitutes which were found in this study:

1. Substitutes with an attractive price.
   - Educational computer games
   - Tablets
   - Non-digital educational games
2. Substitutes with (in most cases) higher quality and performance.
   - Additional teachers
   - Remedial teachers

3. Bargaining Power of Buyers

Buyers have a strong position on the market since the product is not completely necessary for them. There are also numerous substitutes.

4. Bargaining Power of Suppliers

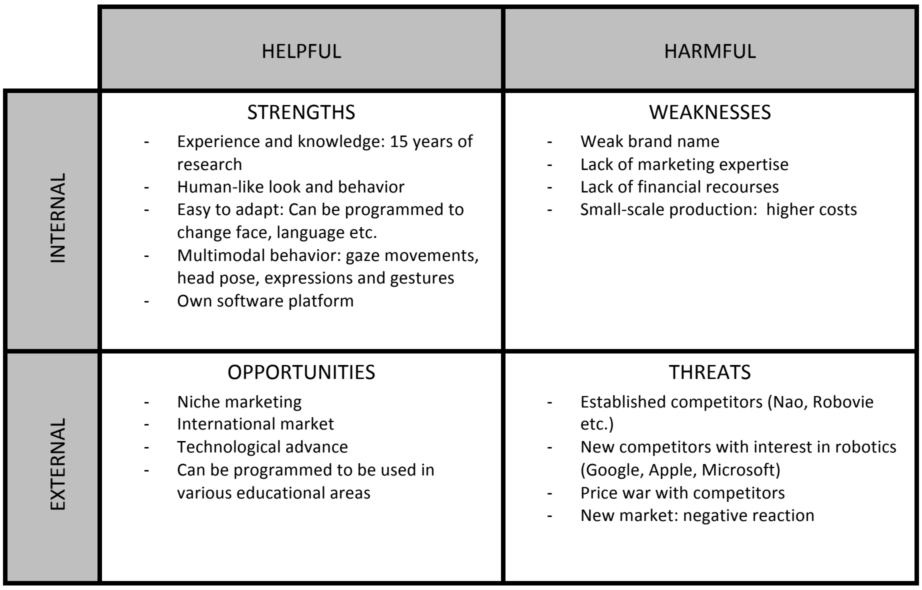
The supplier group is more concentrated than the buyer group, which normally leads to a strong bargaining power. However, because of the very nature of the product (mentioned above) the suppliers might not have a very strong position on the market at this point.

5. Rivalry Among Existing Competitors

Although there are competitors on the market, they are not numerous. Exit barriers are not necessarily high since Furhat can be programmed to do other tasks than to assist with educational tasks. The industry is not yet driven by price competition but by quality competition.

9.3 SWOT

A SWOT-analysis was made based on the results given from the literature study and the conducted interview. The matrix shows Furhat Robotics' strengths, weaknesses, opportunities and threats.



	HELPFUL	HARMFUL
INTERNAL	STRENGTHS - Experience and knowledge: 15 years of research - Human-like look and behavior - Easy to adapt: Can be programmed to change face, language etc. - Multimodal behavior: gaze movements, head pose, expressions and gestures - Own software platform	WEAKNESSES - Weak brand name - Lack of marketing expertise - Lack of financial recourses - Small-scale production: higher costs
EXTERNAL	OPPORTUNITIES - Niche marketing - International market - Technological advance - Can be programmed to be used in various educational areas	THREATS - Established competitors (Nao, Robovie etc.) - New competitors with interest in robotics (Google, Apple, Microsoft) - Price war with competitors - New market: negative reaction

Figure 9.1. A SWOT-analysis of Furhat Robotics in the educational market. The matrix is explained and discussed in section 10.1.

9.4 Domains in the Area of Robots for Education

In addition to the abovementioned roles (tool, peer and tutor), which a social robot can take in the classroom, four domains in the area of robots used for education were detected [21]. A brief description of the four domains is shown in table 9.2.

Technical education	Aims to give the students knowledge of robots and technology. The purpose of such education is to introduce computer science and programming to undergraduate students.
Non-technical subjects	Robots used as an intermediate tool to help students with subjects as mathematics and geometry.
Foreign language	The purpose of robots within this domain is to help students learn a second language. Studies [18] show that children are less hesitant to speak to robots in a foreign language than talking to a human instructor.
Assistive robotics	Robots within this field are used for the cognitive development of children and teenagers. Previous studies indicate that robots generate a high degree of motivation and engagement in individuals which otherwise are unlikely to interact with human therapists [31]. Assistive robots can therefore be used to help students with cognitive disabilities such as autism.

Table 9.2. The four domains in the area of robots for education

9.5 Schools in Sweden

The result from the research of the market for social robots used in education is presented in two diagrams, where the first one shows the number of Swedish schools and the second one shows the expenditures for the different sectors. These can be used to get an overview of the potential market.

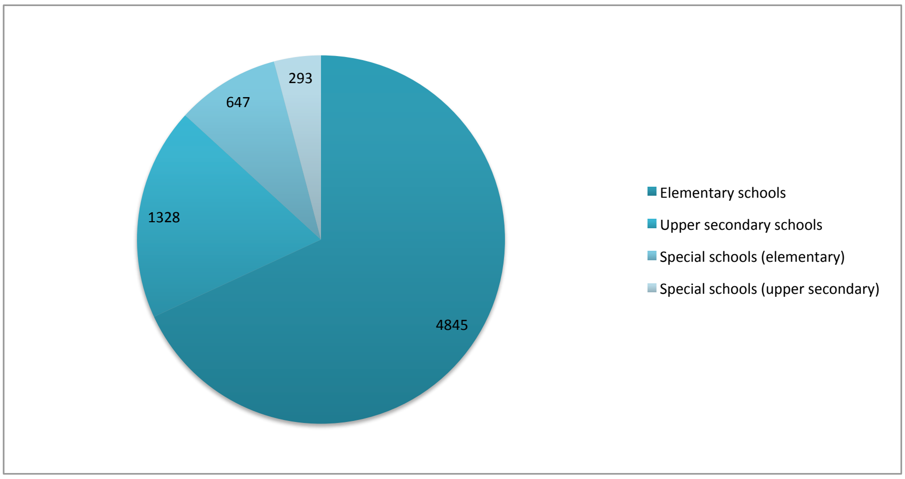


Figure 9.2. A diagram showing the number of schools in Sweden within the different sectors.

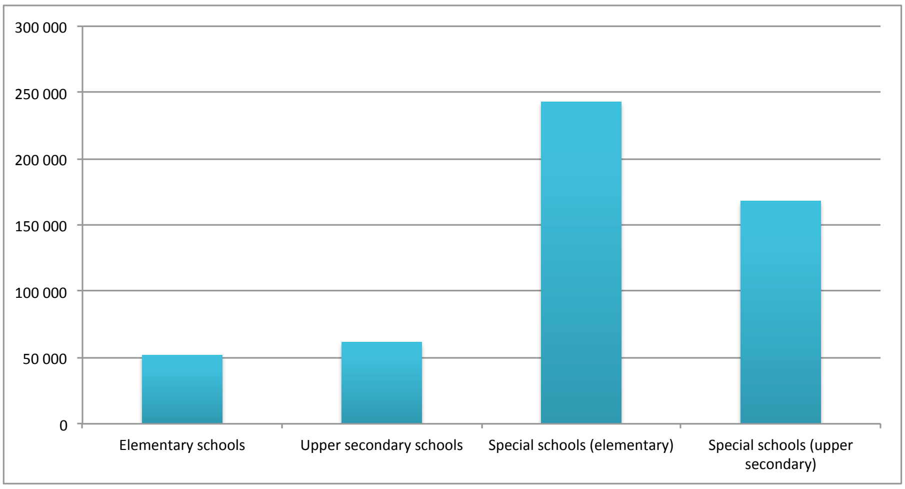


Figure 9.3. A diagram showing the expenditure for education, teaching equipment and school library per student in Swedish Kronor (SEK).

Chapter 10

Discussion

This chapter analyzes the results presented in chapter 9. First, an analysis of the SWOT-matrix and Porter's Five Forces will be presented. This is followed by a discussion of the use of Furhat in the four domains of learning areas. Finally, a conclusion based on the analysis is given.

10.1 Evaluation Using the SWOT-matrix

As seen in figure 9.1, Furhat Robotics has strengths that differentiates them from their rivals. Their greatest strengths are their knowledge, gained from 15 years of research, and their unique product. Furhat is more human-like compared to the well-known robots presented in this study. In addition to the human-like face, Furhat is able to seek and keep eye contact with the person it is addressing and synchronize the lip movements with speech.

The software platform which is used in Furhat allows high level of abstraction and efficient coding, which makes it easy to develop and adapt the robot. This can be used for changing language or face, which results in more potential educational areas to explore. This is a great opportunity for the company.

Despite the strengths and opportunities, the company has weaknesses and threats that might prevent them from successfully entering the market. The biggest disadvantages are the company's lack of financial resources and marketing experience. Because of the fact that Furhat Robotics is not established in the education market and not a strong brand, people might consider buying from another, more well-known company. This leads to less sold units which in turn leads to less produced items. Small-scale production means that the cost per unit is higher and the company must therefore take a higher price to compensate for the costs. The lack of generated money results in less money for development and marketing, which in turn leads to less sold units. This would lead to a downward spiral.

In addition to the established competitors, new competitors may be a threat. Firms like Google and Apple have shown interest in Furhat, which means that they can be potential rivals — or that this can be an opportunity for Furhat Robotics, meaning that they may want to collaborate or buy the company. An additional threat is the market itself. Although studies indicate that robots have a positive impact when used in schools, the education market may not react as expected.

A way for Furhat Robotics to differ even more from their rivals is to target a niche market within the education market. Together with their ability to adapt their product to different needs, this might be an opportunity to reach the specific market before their rivals and offer a robot that is custom-made to the special niche. They can also reach a bigger market if they choose to enter the international market and sell to schools outside of Sweden.

10.2 Evaluation Using Porter's Five Forces

From the analysis made using Porter's Five Forces shown in section 9.1 it is clear that the threat of new entrants is quite high. Factors such as capital requirements and demand-side benefits of scale are disadvantages that Furhat Robotics should take into account before entering the education market. In addition to the threats in the form of existing and new competitors and the bargaining power of buyers, the analysis states threats of substitute products and services. If Furhat wants to enter the education market, it has to add more value than the substitutes. The company must consider the factors that differ their robot from the cheaper substitutes, such as tablets and non-digital educational tools, and use these factors as selling arguments. Their main strength here (shown in the SWOT-matrix) is the robot's social abilities, which the existing products cannot offer.

Previous studies indicate that robots generate higher test scores and help students in their learning process. This means that robots are fully capable to act as a peer or a tutor, together with a human teacher. Although the robots have to develop further before they can replace teachers, they can be a good substitute to additional or remedial teachers. According to Organisation for Economic Cooperation and Development (OECD) Swedish schools do not lack money. In fact, seen from an international perspective, the Swedish schools have substantial financial resources. However, the schools have problems finding competent teachers [32]. If Furhat has the competence needed, this could be a great sales argument.

10.3 Domains in the Area of Robots for Education

One of the conclusions, which can be drawn by the SWOT-analysis, is that Furhat can be re-programmed to adapt to different situations, meaning that it can, with some minor changes, be used in all four domains presented in table 9.2. Although Furhat may not be suitable for technical education in the meaning of hands-on assignments as building the robot, it can be used to teach programming by letting students program it. Wik pointed out that the open-source framework used to program Furhat is quite simple and easy to understand. The program itself is free, but to encourage students to keep programming and make them get a better understanding of what they are doing, they will need to buy Furhat and try the code on him. Although this is a possibility, it is not the original purpose of Furhat.

The domains of teaching non-technical subjects as well as foreign languages are two current fields within the education market. The rival Nao has already been introduced to the market and is getting publicity in Europe by being a part of EMOTE. Although this is negative for Furhat, EMOTE may also make it easier to avoid the possible threat of any negative market reaction by highlighting the advantages with robots in education. Non-technical subjects and foreign languages are therefore two possible domains which Furhat can exploit. However, to be able to beat the competition, Furhat Robotics must market their robot in the right way by clarifying what differs it form the rivals. Being a part of a project or collaborating with a well-known company such as Google or Apple may be a way for the company to market their product and create their own brand.

Assistive robotics is another field which is being investigated. As shown in figure 9.3 the Swedish special schools have more money per student in comparison to other schools, which means that they have more resources to spend on helpful tools like robots. This domain can therefore be a possible niche market within the educational field. Nao has been used in several related studies and has therefore already been introduces to this domain. Studies indicate that it is easier for children with autism to interact with robots because they find them less threatening. It can be problematic for children with special needs to connect with humans and Furhat's human-like features could therefore be a disadvantage in this domain. To be able to draw any conclusions regarding this hypothesis, further research must be done.

10.4 Conclusions

The market for social robots used for education could be a suitable target for a company like Furhat Robotics. However, there are already a few well-known robots on the market, like Nao, which have managed to create a strong brand name. These could impose a threat. A challenge for Furhat Robotics is therefore to point out to customers why a more human-like robot would be a better choice.

In total, four different possible domains within the education market have been found in this study. Because of the adaptable nature of Furhat Robotics' software, there is a possibility of targeting all four. However, some might be more suitable alternatives and further studies must investigate this.

If Furhat Robotics chooses to develop Furhat into an educational robot, the market could be much bigger than just schools. The same product could be used in for instance museums, summer camps and technology fairs. This possibility has, however, not been covered in this study.

In summary, no matter which market segment the company wishes to target, the risk would be quite high since the entire market for social robots barely exists yet. However, since several studies indicate that schools could benefit from using social robots, the education market might be a good target.

Bibliography

[1] Furhat Robotics. *About Us.* http://www.furhatrobotics.com/about-us/. Accessed: 2015-04-29.

[2] P. W. Jordan, M. Makatchev, and K. VanLehn. *Combining Competing Language Understanding Approaches in an Intelligent Tutoring System.* Tech. rep. Intelligent Systems Program and Computer Science Department, University of Pittsburgh, 2004.

[3] C. Rosé et al. "Overcoming the knowledgeengineering bottleneck for understanding student language input". In: *Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies.* 2003.

[4] B. Samei et al. *Context-Based Speech Act Classification in Intelligent Tutoring Systems.* Tech. rep. Institute for Intelligent Systems, University of Memphis, 2014.

[5] V. Rus et al. "Automated Discovery of Speech Act Categories in Educational Games". In: *Proceedings of the 5th International Conference on Educational Data Mining.* 2012.

[6] L. Chen and B. Di Eugenio. "Multimodality and Dialogue Act Classification in the RoboHelper Project". In: *Proceedings of the SIGDIAL 2013 Conference.* 2013.

[7] S. Wilske and G.J. Kruijff. "Service Robots Dealing with Indirect Speech Acts". In: *International Conference on Intelligent Robots and Systems.* 2006.

[8] Furhat Robotics. *Technology.* http://www.furhatrobotics.com/technology/. Accessed: 2015-04-29.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* Springer, 2009.

[10] J. R. Quinlan. *Improved Use of Continuous Attributes in C4.5.* Tech. rep. Basser Department of Computer Science, University of Sydney, 2006.

[11] R. Kohavi. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* Tech. rep. Computer Science Department, Stanford University, 1995.

[12] S. A. Alvarez. *An exact analytical relation among recall, precision, and classification accuracy in information retrieval.* Tech. rep. Department of Computer Science, Boston College, 2002.

[13] Waikato. *Weka 3: Data Mining Software in Java.* http://www.cs.waikato.ac.nz/ml/weka/. Accessed: 2015-04-27.

[14] Waikato. *Attribute-Relation File Format (ARFF).* http://www.cs.waikato.ac.nz/ml/weka/arff.html. Accessed: 2015-05-11.

[15] Robert Östling. *Stagger: an Open-Source Part of Speech Tagger for Swedish.* Tech. rep. Department of Linguistics, Stockholm University, 2013.

[16] A. Stolcke et al. “Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech”. In: *Computational Linguistics* (2000).

[17] SPARC. *Robotics 2020 Multi-Annual Roadmap.* Tech. rep. SPARC, 2015.

[18] C.-W. Chang et al. “Exploring the Possibility of Using Humanoid Robots as Instructional Tools for Teaching a Second Language in Primary School”. In: *Educational Technology & Society* (2010).

[19] J. Han. *Robot-Aided Learning and r-Learning Services.* Tech. rep. Department of Computer Education, Cheongju National University of Education, 2010.

[20] J. Han et al. “Comparative Study on the Educational Use of Home Robots for Children”. In: *Journal of Information Processing Systems* (2008).

[21] O. Mubin et al. “A Review of the Applicability of Robots in Education”. In: *Technology for Education and Learning* (2013).

[22] N. Shin and S. Kim. *Learning about, from, and with robots: Learning about, from, and with robots: students’ perspectives.* Tech. rep. Department of Education, Dongguk University, 2007.

[23] D. Shah. *Robovie talking robot joins science class at Higashihikari Elementary School in Japan.* http://fareastgizmos.com/robotic/robovie-talking-robot-joins-science-class-at-higashihikari-elementary-school-in-japan.php. Accessed: 2015-04-27.

[24] Aldebaran. *Teach with NAO.* https://www.aldebaran.com/en/robotics-solutions/educational-robots. Accessed: 2015-04-27.

[25] Aldebaran. *More about NAO.* https://www.aldebaran.com/en/more-about. Accessed: 2015-04-27.

[26] Aldebaran. *Unveiling of NAO Evolution: a stronger robot and a more comprehensive operating system.* https://www.aldebaran.com/en/press/press-releases/unveiling-of-nao-evolution-a-stronger-robot-and-a-more-comprehensive-operating. Accessed: 2015-04-27.

[27] Ituniv. *Gothenburg University Scientists to Develop Pedagogical Robots.* http://www.ituniv.se/english/current/news/Nyhet$_{d}$*etalj*/?*languageId* = 100001&*contentId* = 1158632&*disableRedirect* = *true*&*returnUrl* = *http*%3*A*%2*F*%2*Fwww.ituniv.se*%2*Faktuellt*%2*Fnyheter*%2*Ffulltext*%2*F*%2*Fforskare* − *vid* − *goteborgs* − *universitet* − *utvecklar* − *pedagogiska* − *robotar* − *i* − *nytt* − *eu* − *projekt.cid*1158632. Accessed: 2015-04-27.

[28] Robotnyheter. *Skolrobot ska anpassa sig efter elevernas känslor.* http://robotnyheter.se/2013/04/19/skolrobot-ska-anpassa-sig-efter-elevernas-kanslor/#more-20794. Accessed: 2015-04-27.

[29] P. Kotler and G. Armstrong. *Principled of Marketing.* Prentice Hall, 2011.

[30] M. E. Porter. “How Competitive Forces Shape Strategy”. In: *Harvard Business Review* (1979).

[31] J. Burns. “Robots in the classroom help autistic children learn”. In: *BBC News education reporter* (2012).

[32] R. Elmore et al. *Improving schools in swden: An OECD perspective.* Tech. rep. OECD, 2015.