Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 540–548
Marseille, 11–16 May 2020

Estimating User Communication Styles for Spoken Dialogue Systems

Juliana Miehle[1], Isabel Feustel[1], Julia Hornauer[1], Wolfgang Minker[1], Stefan Ultes[2]
[1] Institute of Communications Engineering, Ulm University, Ulm, Germany
{juliana.miehle, isabel.feustel, julia.hornauer, wolfgang.minker}@uni-ulm.de
[2] Mercedes-Benz Research & Development, Sindlefingen, Germany
stefan.ultes@daimler.com

Abstract

We present a neural network approach to estimate the communication style of spoken interaction, namely the stylistic variations *elaborateness* and *directness*, and investigate which type of input features to the estimator are necessary to achive good performance. First, we describe our annotated corpus of recordings in the health care domain and analyse the corpus statistics in terms of agreement, correlation and reliability of the ratings. We use this corpus to estimate the *elaborateness* and the *directness* of each utterance. We test different feature sets consisting of dialogue act features, grammatical features and linguistic features as input for our classifier and perform classification in two and three classes. Our classifiers use only features that can be automatically derived during an ongoing interaction in any spoken dialogue system without any prior annotation. Our results show that the *elaborateness* can be classified by only using the dialogue act and the amount of words contained in the corresponding utterance. The *directness* is a more difficult classification task and additional linguistic features in form of word embeddings improve the classification results. Afterwards, we run a comparison with a support vector machine and a recurrent neural network classifier.

Keywords: Dialogue management, User adaptation, Supervised learning.

1. Introduction

For humans, speech is the most natural form of interaction and it has been shown that people adapt their interaction styles to one another across many levels of utterance production when communicating, e.g. by matching each other's behaviour or synchronising the timing of behaviour (Burgoon et al., 2007; Niederhoffer and Pennebaker, 2002; Brennan, 1996; Pickering and Garrod, 2004; Nenkova et al., 2008). However, this adaptive behaviour has rarely been addressed and implemented in a live spoken dialogue system. With the aim of designing such a spoken dialogue system which adapts to the user's communication idiosyncrasies, we present a classification approach to automatically estimate the user's communication style during an ongoing dialogue. The estimated communication style can then be used in the dialogue management to adapt the system behaviour to the user, as depicted in Figure 1. To the best of our knowledge, this is the first work on automatic estimation of the user's communication style in a dialogue system. We present the task of estimating the user's communication style and investigate the influence of grammatical and linguistic features on its estimation. Our classifiers use only features that can be automatically generated during an interaction with a spoken dialogue system (i.e. without any manual annotation).

Even though intelligent assistants like Amazon Alexa, Apple Siri, Google Assistant or Microsoft Cortana are becoming increasingly popular, they do not consider different communication styles to adapt their behaviour. Instead, current research in the field of spoken dialogue systems focuses on general user adaptivity like satisfaction or general user groups (Honold et al., 2014; Ultes et al., 2015; Casanueva et al., 2015; Pragst et al., 2015; Miehle et al., 2019). However, various studies suggest that adapting the communication styles of spoken dialogue systems to the individual users in a similar way to what humans do will lead to more natural interactions (Cassell and Bickmore, 2003; Forbes-Riley et al., 2008; Stenchikova and Stent, 2007; Reitter et al., 2006; Mairesse and Walker, 2010).

To adapt the behaviour of the system to individual users, we consider the communication styles *elaborateness* and *directness* in this work as Pragst et al. (2019) have shown that they influence the user's perception of a dialogue and are therefore valuable candidates for adaptive dialogue management. The *elaborateness* thereby refers to the amount of additional information provided to the user and the *directness* describes how concretely the information that is to be conveyed is addressed by the speaker. This means that a *direct* and *concise* answer to the question "Can you tell me what the weather's gonna be like today?" is, for example:

> "It will rain."

It answers the question concretely and gives only the requested information in the shortest possible way. The *direct* and *elaborate* version of the same answer provides some additional information:

> "Most of the time it is cloudy and in the afternoon it will rain."

The *indirect* and *concise* version of this utterance also contains few information, yet addresses the fact that it is raining in a less concrete way:

> "Today is a good day for cosy activities at home."

In this case, the interlocutor can infer that the weather won't be nice as it is better to stay at home. The *elaborate* and *indirect* version provides some more details:

> "Today is a good day for cosy activities at home. In the afternoon you could get wet outside."

This example is taken from Miehle et al. (2018a) addressing the issues of how varying communication styles of a

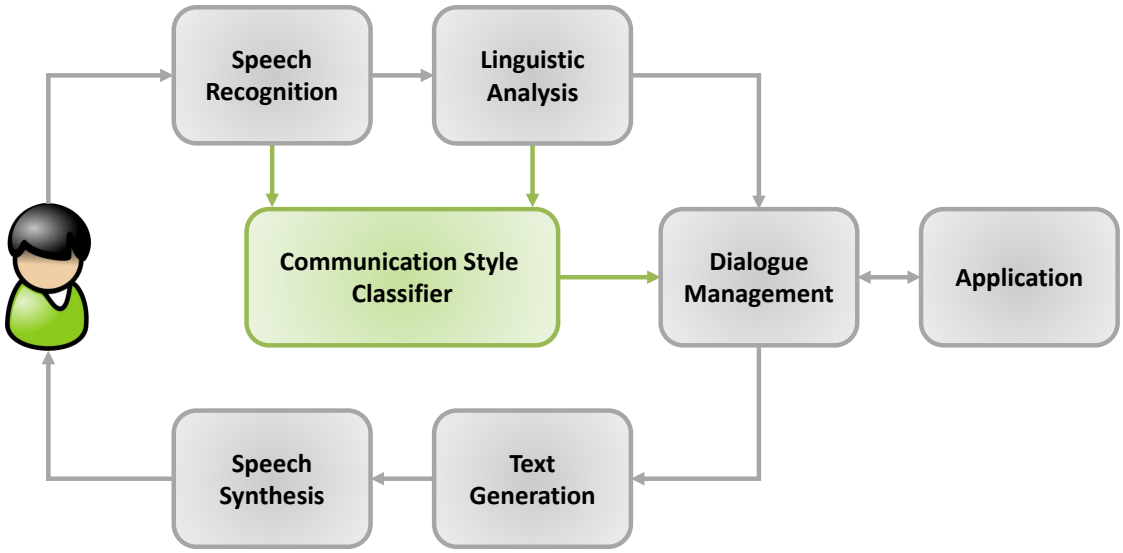

Figure 1: The estimated communication style, which is classified based on features from the speech recognition and the linguistic analysis, can be used in the dialogue management to adapt the system behaviour to the user.

spoken user interface are perceived by users and whether there exist global preferences in the communication styles *elaborateness* and *directness*. The authors could show that the system's communication style influences the user's satisfaction and the user's perception of the dialogue and that there is no general preference in the system's communication style. The authors conclude that spoken dialogue systems need to adapt their communication style to each user individually during every dialogue in order to achieve a high level of user satisfaction.

A study presented by Miehle et al. (2016) investigated cultural differences between the Germans and the Japanese. The results revealed that communication idiosyncrasies in human-human interaction may also be observed during human-computer interaction in a spoken dialogue system context. Moreover, Miehle et al. (2018b) presented another study examining five European cultures whose communication styles are much more alike than the German and Japanese communication idiosyncrasies. The study explores not only the influence of the user's culture but also of the gender, the frequency of use of speech based assistants as well as the system's role. The results show that the system's role significantly influences the user's preference in the system's communication style whereas the frequency of use of speech based assistants has no influence. Moreover, the findings show differences among the cultures and, depending on the culture, there are gender differences with respect to the user's preference in the system's communication style.

These studies show that adaptive behaviour regarding the user's communication style is an important aspect of spoken user interfaces. This is why we address the task of estimating user communication styles.

The structure of the paper is as follows: In Section 2., related work in the field of the estimation of communication styles will be discussed. Afterwards, the corpus that has been used in this work will be described in Section 3. and our classification approach as well as the results will be presented in Section 4., before concluding in Section 5.

2. Related Work

Previous work has already explored approaches for the classification of *elaborateness* and *directness* in the context of related applications.

Di Buccio et al. (2014) propose a methodology to automatically detect and process verbose queries submitted to search engines. It is shown that the information retrieval effectiveness can be significantly improved by considering the query verbosity. Moreover, Gharouit and Nfaoui (2017) suggest to use BabelNet as knowledge base in the detection of verbose queries and then present a comparative study between different algorithms to classify queries into two classes, verbose or succinct. However, both papers deal with the classification of queries submitted to search engines. To the best of our knowledge, there exists no previous work in the field of *elaborateness* classification for spoken language.

Goel et al. (2018) explore different supervised machine learning approaches to automatically detect indirectness in tutoring conversations. The authors collected a corpus of tutoring dialogues from 12 American-English speaking pairs of teenagers whereby the conversations include social interaction as well as tutoring periods. They annotated four types of indirectness for the tutoring periods, namely apologising (e.g. "Sorry, its negative 2."), hedging language (e.g. "You just add 5 to both sides."), the use of vague category extenders (e.g. "You have to multiply and stuff.") and subjectivising (e.g. "I think you divide by 3 here."). Each utterance was then classified as direct or indirect based on its inclusion in any of these categories. Afterwards, they used different classification approaches to detect indirectness based on textual and visual features, reaching an F1 sore of 62 %. However, we claim that there are more aspects than the four types of indirectness annotated

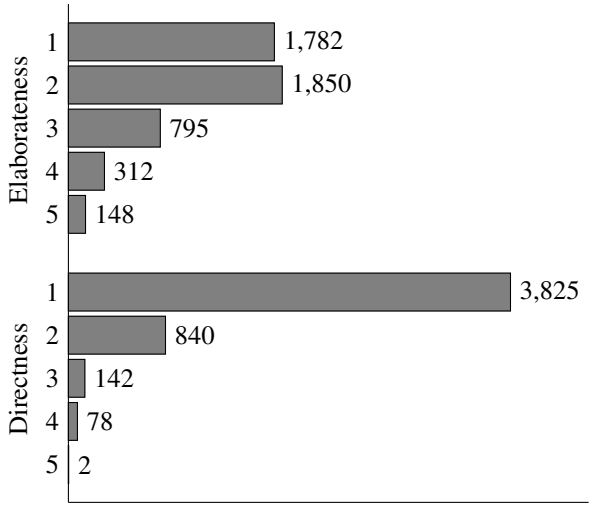

Figure 2: The class distribution of the annotated *elaborateness* (top) and *directness* (bottom) scores (median of the three ratings).

Elaborateness (5 classes)				
	R1/R2	R1/R3	R2/R3	Av.
κ	0.551	0.425	0.420	**0.465**
ρ	0.832	0.752	0.730	**0.771**
ICC				**0.914**
Elaborateness (3 classes)				
	R1/R2	R1/R3	R2/R3	Av.
κ	0.651	0.513	0.510	**0.558**
ρ	0.816	0.741	0.704	**0.754**
ICC				**0.901**
Directness (5 classes)				
	R1/R2	R1/R3	R2/R3	Av.
κ	0.305	0.311	0.305	**0.307**
ρ	0.391	0.405	0.393	**0.396**
ICC				**0.643**
Directness (3 classes)				
	R1/R2	R1/R3	R2/R3	Av.
κ	0.325	0.327	0.320	**0.324**
ρ	0.391	0.405	0.392	**0.396**
ICC				**0.659**
Directness (2 classes)				
	R1/R2	R1/R3	R2/R3	Av.
κ	0.377	0.395	0.387	**0.386**
ρ	0.380	0.396	0.387	**0.388**
ICC				**0.655**

Table 1: Agreement (κ), correlation (ρ) and reliability (ICC) in *elaborateness* and *directness* of the three ratings (R1, R2, R3). All results are significant at the 0.001 level.

in this corpus and that indirectness cannot be broken down to rather simple key word spotting (e.g. "sorry", "just", "and stuff", "I think"). Neuliep (2011) describes the indirect style as "one where the speaker's intentions are hidden or only hinted at during interaction". For our corpus annotation, we used this definition and annotated the directness/indirectness in a global way and not based on fixed structures or key words.

Other work in this field only focuses on a specific phenomena of indirect speech, like hedge detection (Prokofieva and Hirschberg, 2014; Ulinski et al., 2018), politeness detection (Danescu-Niculescu-Mizil et al., 2013; Aubakirova and Bansal, 2016) and uncertainty detection (Liscombe et al., 2005; Forbes-Riley and Litman, 2011; Adel and Schütze, 2017).

3. Corpus Description

Our data set is based on recordings on health care topics containing spontaneous interactions in dialogue format between two participants: one is taking the role of the system while the other one is taking the role of the user. Each dialogue turn contains one or more dialogue acts. These dialogue acts are chosen out of a set of 43 distinct dialogue acts which have been predefined. A list of all dialogue acts can be found in Table 9 in the appendix. Along with the dialogue acts, the respective utterances are also added to the data set. Overall, the corpus covers 135 German dialogues containing 4,887 annotated dialogue acts. The average number of dialogue acts per dialogue is 36.20.

We have annotated each dialogue act with the two communication styles *directness* and *elaborateness*. Both are assigned scores between 1 and 5 with 1 = extremely direct/concise and 5 = extremely indirect/elaborate. Each dialogue act has been annotated by three different raters. The class distribution of the annotated *elaborateness* and *directness* scores (median of the three ratings) is shown in Figure 2. It can be seen that it is quite hard to distinguish between different levels of *elaborateness* and *directness*. As the classes 3, 4, and 5 contain utterances which are elaborate/indirect to a greater or lesser extent, we have combined them to one new class, thus reducing the corpus to three classes. For *directness*, the annotation has shown that it even makes sense to see it as a binary decision between direct/indirect utterances. As the classes 2-5 contain different degrees of indirectness (from slightly indirect to extremely indirect), we additionally combined these classes to one indirect class for binary classification.

In order to analyse the quality of the annotated scores, we have used the following measures: **Cohen's Kappa** κ measures the relative agreement between two sets of ratings and is defined as

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{1}$$

where p_0 is the observed agreement, and p_e is the chance agreement (Cohen, 1960). Hence, $\kappa = 1$ for perfect agreement and $\kappa = -1$ for perfect disagreement.

Spearman's rank correlation coefficient Rho ρ is a non-parametric measure for the rank correlation between two variables and describes how well one variable can be ex-

	Elaborateness	*Directness*
	Mean/Median	Mean/Median
κ	0.951	0.802
ρ	0.991	0.884

Table 2: Agreement (κ) and correlation (ρ) between the mean and the median of the three ratings for the *elaborateness* and the *directness*. All results are significant at the 0.001 level.

pressed by the other (Spearman, 1904). It is defined as

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \quad (2)$$

where x_i and y_i are corresponding ranked ratings, while $\bar{x}$ and $\bar{y}$ are the mean ranks. Thus, $\rho = 1$ if observations have identical ranks and $\rho = -1$ if observations have fully opposed ranks.

The **Intraclass Correlation Coefficient** ICC measures the reliability between the ratings. We use the *One-Way Random Average Measures* $ICC(1, k)$ as defined in Shrout and Fleiss (1979) with $k = 3$ raters who are randomly selected from a larger population of raters. $ICC = 1$ indicates maximum reliability, $ICC = -1$ maximum unreliability.

The **results** can be seen in Table 1. The original ratings (five classes) achieve an overall inter-rater agreement of $\kappa = 0.47$ for *elaborateness* and $\kappa = 0.31$ for *directness*, a correlation of $\rho = 0.77$ for *elaborateness* and $\rho = 0.40$ for *directness* and a inter-rater reliability of $ICC = 0.91$ for *elaborateness* and $ICC = 0.64$ for *directness*. If we reduce the classes to three or two (in case of *directness*), we obtain a higher agreement while the correlation and the inter-rater reliability do not change significantly. Overall, we have a good inter-rater reliability for both communication styles given the difficulty of the annotation task.

In order to use the communication style annotations as target for our classification task, we need a **final score** to be calculated from the three ratings. Typical candidates are the mean and the median. To decide which one to use, we have calculated the mean and the median of the three ratings and analysed which of the two metrics better reflects the individual ratings. The comparison of the mean and the median in terms of Cohen's Kappa κ and Spearman's rank correlation coefficient Rho ρ is shown in Table 2. A strong correlation ρ can be observed. The values of κ indicate a high level of agreement, but still there seem to be some differences, especially for the *directness*. Therefore, we have compared the agreement and correlation between the mean and the median of the three ratings with the individual ratings (see Table 3). The values of κ indicate that the median better reflects the individual ratings, what might be explained by the fact that we have applied an ordinal scale. Therefore, the median of the ratings is used as final score for the classification in the following section.

4. Estimation of Communication Styles

The main contribution of this work is to address the problem of automatic estimation of the communication styles

Elaborateness				
	M/R1	M/R2	M/R3	Av.
κ	0.751	0.700	0.589	**0.680**
ρ	0.917	0.916	0.843	**0.892**
	Md/R1	Md/R2	Md/R3	Av.
κ	0.795	0.732	0.611	**0.713**
ρ	0.923	0.913	0.831	**0.889**
Directness				
	M/R1	M/R2	M/R3	Av.
κ	0.517	0.534	0.511	**0.521**
ρ	0.668	0.736	0.675	**0.693**
	Md/R1	Md/R2	Md/R3	Av.
κ	0.630	0.604	0.614	**0.616**
ρ	0.702	0.676	0.701	**0.693**

Table 3: Agreement (κ) and correlation (ρ) between the mean (M) and the median (Md) of the three ratings for the *elaborateness* and the *directness* and the individual ratings (R1, R2, R3). All results are significant at the 0.001 level.

elaborateness and *directness* and to investigate how grammatical and linguistic features influence the performance. After defining and evaluating a baseline approach, different setups for adding linguistic information is described and their performance analysed. Afterwards, we run a comparison between different classifiers.

4.1. The Dialogue Act Baseline

As a baseline approach, an artificial neural network (ANN) classifier with one hidden layer has been trained using only *dialogue act features (DA)* that can directly be derived from the data[1]. These features contain the dialogue act and the amount of words in the utterance of the corresponding dialogue act. Note that the dialogue act is the output of the linguistic analysis while the text representation of the utterance is the output of the speech recogniser (see Figure 1). Hence, both features in this feature set can be automatically derived during an ongoing interaction in every spoken dialogue system and no annotation is necessary.

The neural net is trained and evaluated with a 10-fold cross-validation setting on the corpus described in Section 3. Grid search was used to find the best set of hyper parameters (i.e. the amount of nodes, the amount of epochs, the optimiser, the output function and the loss function). To take account for the imbalanced data during the grid search optimisation, the unweighted average recall (UAR) was used, which is the arithmetic average of all class-wise recalls. The results are shown in Table 4.

Classification of the 3-class *elaborateness* reaches an UAR of 84 % only using dialogue act features, which is quite promising. Classification of the 3-class *directness* results in an UAR of 56 %, and the binary *directness* reaches an UAR of 75 %. These results clearly show the difficulty

[1] During our experiments, we have also tested additional annotated features, but this led to worse results.

		Elaborateness (3 classes)	*Directness* (3 classes)	*Directness* (2 classes)
DA	UAR	0.840	0.555	0.753
	ACC	0.838	0.832	0.848
	F1	0.838	0.582	0.761
	κ	0.749	0.467	0.527
	ρ	0.862	0.523	0.541

Table 4: The classification results using the ANN classifier and the dialogue act features (DA) in terms of the Unweighted Average Recall (UAR), the Accuracy (ACC), the F1-Score, Cohen's Kappa κ and Spearman's rank correlation coefficient Rho ρ.

		Elaborateness (3 classes)	*Directness* (3 classes)	*Directness* (2 classes)
DA+G	UAR	0.841	0.558	0.753
	ACC	0.840	0.834	0.848
	F1	0.839	0.588	0.761
	κ	0.753	0.470	0.526
	ρ	0.864	0.521	0.540

Table 5: The classification results using the ANN classifier and the dialogue act features as well as the grammatical features (DA+G) in terms of the Unweighted Average Recall (UAR), the Accuracy (ACC), the F1-Score, Cohen's Kappa κ and Spearman's rank correlation coefficient Rho ρ.

of the task, which has already been shown by the corpus creation. There, it has been quite hard for the annotators to distinguish between different levels of *directness* so that the class distribution of the *directness* is sub-optimal for the classification task. However, comparing the results to a majority-class classifier clearly shows that there is still a lot of information encoded in the DA feature set achieving higher UAR. The majority-class classifier always predicts the most frequent class in the training set and achieves an UAR of 33 % for three classes and an UAR of 50 % for two classes.

Furthermore, comparing the achieved κ and ρ with the results obtained for the human annotators (Table 1) shows that all, the results of the classifier for the *elaborateness* (three classes) and the results of the classifiers for the *directness* (three and two classes), outperform the agreement and correlation between the three human annotators. Hence, all trained classifiers for both the *elaborateness* and the *directness* constitute a reasonable baseline.

4.2. The Contribution of Grammatical and Linguistic Features

To address the question of whether grammatical features improve the estimation of the communication style, a second feature set is used containing the dialogue act features (as have been used for the baseline) as well as grammatical features. The *grammatical features (G)* are represented by Part-of-speech (POS) tags which have been assigned to the utterances by the RDRPOSTagger (Nguyen et al., 2014). As the utterance is the output of the speech recognition and this tagger can be used online during an ongoing interaction, there is also no annotation necessary for this feature set. The results are shown in Table 5. It can be seen that there is no improvement in comparison to the baseline.

In addition to grammatical features, *linguistic features* may majorly contribute to the overall classification performance. In order to encode the linguistic features, a Bag-of-Words (BoW) approach has been used in combination with unigrams (U), unigrams and bigrams (UB) and word embeddings (WE). Using BoW, two distinct vocabularies have been created:

- The BoW-U vocabulary contains every word occurring in the database of spoken dialogues.
- The BoW-UB vocabulary contains the BoW-U vocabulary (single words) as well as every two-word-sequence in the database.

These vocabularies and the combination with word embeddings led to three different linguistic feature sets:

- U: This feature set contains a BoW-U vector for each utterance, thus encoding the number of times each word (of the overall vocabulary) appears in the corresponding utterance.
- UB: This feature set contains a BoW-UB vector for each utterance, thus encoding the number of times each word and each two-word-sequence (of the overall vocabulary) appear in the corresponding utterance.
- WE: For this feature set, the BoW-U vocabulary has been combined with the German pre-trained fastText word vectors by Grave et al. (2018)[2]. Matrix X of dimension $u \times w$ contains the BoW-U vectors (dimension $1 \times w$ with w the amount of words in vocabulary BoW-U) for each utterance, where u is the total number of utterances. Matrix W of dimension $w \times p$ contains the fastText word vectors (dimension $1 \times p$ with p the length of each word vector) for each word. By multiplying these matrices a new matrix $Z = X \cdot W$ of dimension $u \times p$ is obtained, containing a vector representation for each utterance. These utterance vectors of dimension $1 \times p$ can then be used as feature vectors for the classification task.

In addition to using these linguistic feature sets individually, we have used them in combination with the dialogue act features (DA) and the grammatical features (G). All results are shown in Table 6.

For *elaborateness*, the best results are achieved with the baseline feature set. Grammatical and linguistic features do not seem to have any effect on the classification performance. This leads to the conclusion that for the *elaborateness*, analysing the utterance length dependent on the dialogue act seems to contain enough information to achieve good classification performance.

[2]During our experiments, we have also tested self-trained word vectors, but this led to worse results.

		Elaborateness (3 classes)	*Directness* (3 classes)	*Directness* (2 classes)
U	UAR	0.747	0.485	0.729
	ACC	0.752	0.822	0.842
	F1	0.742	0.478	0.744
	κ	0.618	0.430	0.492
	ρ	0.779	0.490	0.503
U+DA	UAR	0.809	0.484	0.743
	ACC	0.811	0.823	0.846
	F1	0.807	0.477	0.755
	κ	0.708	0.433	0.512
	ρ	0.831	0.507	0.522
U+DA+G	UAR	0.817	0.484	0.746
	ACC	0.818	0.822	0.846
	F1	0.814	0.476	0.757
	κ	0.719	0.431	0.516
	ρ	0.841	0.505	0.524
UB	UAR	0.745	0.520	0.748
	ACC	0.742	0.751	0.822
	F1	0.734	0.497	0.740
	κ	0.607	0.354	0.481
	ρ	0.776	0.411	0.485
UB+DA	UAR	0.786	0.533	0.748
	ACC	0.785	0.761	0.826
	F1	0.781	0.511	0.742
	κ	0.669	0.387	0.485
	ρ	0.811	0.452	0.490
UB+DA+G	UAR	0.799	0.542	0.756
	ACC	0.796	0.757	0.827
	F1	0.793	0.513	0.747
	κ	0.687	0.391	0.495
	ρ	0.827	0.458	0.500
WE	UAR	0.757	0.493	0.727
	ACC	0.755	0.783	0.828
	F1	0.749	0.495	0.729
	κ	0.626	0.364	0.464
	ρ	0.786	0.414	0.479
WE+DA	UAR	0.825	**0.589**	**0.762**
	ACC	0.821	0.803	0.842
	F1	0.819	0.589	0.759
	κ	0.726	0.443	0.522
	ρ	0.855	0.498	0.535
WE+DA+G	UAR	0.827	**0.594**	**0.765**
	ACC	0.823	0.794	0.843
	F1	0.821	0.588	0.762
	κ	0.729	0.432	0.528
	ρ	0.857	0.480	0.544

Table 6: The classification results using the ANN classifier and the linguistic features (separately and in combination with the dialogue act features and the grammatical features) in terms of the Unweighted Average Recall (UAR), the Accuracy (ACC), the F1-Score, Cohen's Kappa κ and Spearman's rank correlation coefficient Rho ρ. Linguistic features are beneficial for the estimation of the *directness*, but not for the estimation of the *elaborateness*.

For *directness*, the overall performance could be improved by using linguistic information encoded as word embeddings. This in combination with grammatical and dialogue act features (WE+DA+G) led to UARs of 59 % and 76 % for the estimation of *directness* using three classes and two classes, respectively. Using the BoW approach in combination with unigrams and bigrams could not improve the classification performance.

To sum up, linguistic features are beneficial for the estimation of the *directness*, but not for the estimation of the *elaborateness*. For the latter, the dialogue act features (i.e. the dialogue act and the amount of words in the utterance) seem to be sufficient.

4.3. Comparison with a SVM Classifier

In a further step, we have compared our ANN classifier with a support vector machine (SVM) classifier. For this evaluation, we have used the dialogue act features (DA), the grammatical features (G) as well as linguistic features encoded as word embeddings (WE) as these feature sets provided the best results with the ANN classifier. We have trained and evaluated our SVM with a 10-fold cross-validation setting (same as with the ANN classifier) and used grid search to find the best set of parameters (i.e. C and γ). The results are shown in Table 7. For the classification of the *elaborateness*, the SVM classifier performs comparable to the ANN classifier and reaches an UAR of 84 % when using the dialogue act features (DA) or the dialogue act features in combination with the grammatical features (DA+G). The *directness* classification using the SVM classifier yields worse results than the ANN classifier, for both classification in three and two classes.

4.4. Comparison with a RNN Classifier

Lastly, we have compared our ANN classifier with a recurrent neural network (RNN) classifier consisting of two long short-term memory (LSTM) layers followed by two hidden perceptron layers and one output layer. The LSTM layers extract and store temporal information that might be beneficial for the communication style estimation. As in Section 4.3., we have used the dialogue act features (DA), the grammatical features (G) as well as linguistic features encoded as word embeddings (WE). We have trained and evaluated our RNN classifier with a 10-fold cross-validation setting (same as with the ANN and SVM classifier) and used grid search to find the best set of parameters (i.e. the amount of nodes of the LSTM layers, the dropout, the amount of nodes of the hidden perceptron layers, the amount of epochs, the optimiser, the output function and the loss function). The results are shown in Table 8. For the classification of the 3-class *elaborateness* and the 3-class *directness*, the RNN classifier yields worse results than the ANN classifier. However, for the estimation of the binary *directness*, the RNN classifier outperforms the ANN classifier, reaching an UAR of 78 % when using linguistic features encoded as word embeddings in combination with grammatical and dialogue act features (WE+DA+G). This shows that temporal information is beneficial for the estimation of the *directness*, but not for the estimation of the *elaborateness*.

		Elaborateness (3 classes)	*Directness* (3 classes)	*Directness* (2 classes)
DA	UAR	**0.842**	0.553	0.735
	ACC	0.839	0.829	0.849
	F1	0.839	0.588	0.753
	κ	0.752	0.438	0.510
	ρ	0.863	0.487	0.525
DA+G	UAR	**0.843**	0.534	0.749
	ACC	0.841	0.831	0.818
	F1	0.840	0.569	0.739
	κ	0.754	0.422	0.480
	ρ	0.864	0.493	0.485
WE	UAR	0.748	0.512	0.743
	ACC	0.748	0.801	0.818
	F1	0.741	0.511	0.735
	κ	0.614	0.415	0.471
	ρ	0.777	0.458	0.475
WE+DA	UAR	0.820	0.544	0.762
	ACC	0.815	0.807	0.835
	F1	0.813	0.554	0.757
	κ	0.717	0.417	0.515
	ρ	0.847	0.473	0.520
WE+DA+G	UAR	0.815	0.551	0.762
	ACC	0.810	0.811	0.835
	F1	0.808	0.562	0.757
	κ	0.709	0.426	0.515
	ρ	0.842	0.481	0.520

Table 7: The classification results using the SVM classifier in terms of the Unweighted Average Recall (UAR), the Accuracy (ACC), the F1-Score, Cohen's Kappa κ and Spearman's rank correlation coefficient Rho ρ. For the classification of the *elaborateness*, the SVM classifier performs comparable to the ANN classifier.

		Elaborateness (3 classes)	*Directness* (3 classes)	*Directness* (2 classes)
DA	UAR	0.805	0.507	0.765
	ACC	0.794	0.778	0.821
	F1	0.797	0.466	0.747
	κ	0.686	0.409	0.496
	ρ	0.822	0.465	0.502
DA+G	UAR	0.811	0.511	0.771
	ACC	0.800	0.789	0.817
	F1	0.799	0.472	0.746
	κ	0.696	0.432	0.495
	ρ	0.837	0.500	0.503
WE	UAR	0.731	0.502	0.736
	ACC	0.743	0.794	0.806
	F1	0.734	0.473	0.723
	κ	0.604	0.416	0.447
	ρ	0.781	0.454	0.451
WE+DA	UAR	0.808	0.513	0.770
	ACC	0.807	0.808	0.838
	F1	0.806	0.483	0.762
	κ	0.704	0.456	0.527
	ρ	0.837	0.523	0.534
WE+DA+G	UAR	0.814	0.521	**0.782**
	ACC	0.814	0.806	0.843
	F1	0.814	0.485	0.772
	κ	0.714	0.462	0.545
	ρ	0.842	0.521	0.550

Table 8: The classification results using the RNN classifier in terms of the Unweighted Average Recall (UAR), the Accuracy (ACC), the F1-Score, Cohen's Kappa κ and Spearman's rank correlation coefficient Rho ρ. Temporal information is beneficial for the estimation of the *directness*, but not for the estimation of the *elaborateness*.

5. Conclusion and Future Directions

In this work, we have presented a classification approach which, for the first time, addresses the estimation of the user's communication style in a spoken dialogue. We have considered the communication styles *elaborateness* and *directness*. First, we have described the annotated corpus based on recordings in the health-care domain which contain spontaneous interactions in dialogue format between two human participants. Each dialogue act has been annotated with the two communication styles *elaborateness* and *directness*. By analysing the corpus statistics in terms of the agreement, correlation and reliability of the ratings, we have achieved acceptable labels for the communication style annotations. Afterwards, we have used the corpus to estimate the *elaborateness* and the *directness* of each utterance. We have tested different feature sets as input for an ANN classifier and performed classification in two and three classes. The results show that the *elaborateness* can be classified quite well by only using the dialogue act and the amount of words contained in the corresponding utterance. The *directness* seems to be a more difficult classification task and additional linguistic features in form of word embeddings give improvement in the classification results. Our classifiers for both the *elaborateness* and the *directness* use only features that can be automatically recognised during an ongoing interaction in any spoken dialogue system, without any prior annotation. Thereafter, we have run a comparison with a SVM and a RNN classifier. The results show that the SVM classifier performs comparable to the ANN classifier for the *elaborateness* estimation, but worse for the *directness* estimation. The RNN classifier outperforms the ANN classifier in the binary *directness* classification showing that temporal information is beneficial for the estimation of the *directness*, but not for the estimation of the *elaborateness*.

In future work, we will consider semantic features in form of sentence embeddings for our classification task. Moreover, we will evaluate our methods in a multilingual test set as we also have Polish, Spanish and Turkish data available.

6. Data

The feature sets used within this work will be made publicly available and can be downloaded from `http://dx.doi.org/10.18725/OPARU-26061`. Unfortunately,

we are not able to publish the original corpus due to privacy reasons.

7. Acknowledgements

This work is part of a project that has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 645012. We thank our colleagues from the University of Tübingen, the German Red Cross in Tübingen and semFYC in Barcelona for organizing and carrying out the corpus recordings. Additionally, this work has received funding within the BMBF project "RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with Learning Autonomous Robots" and the technology transfer project "Do it yourself, but not alone: Companion Technology for DIY support" of the Transregional Collaborative Research Centre SFF/TRR 62 "Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

8. Bibliographical References

Adel, H. and Schütze, H. (2017). Exploring different dimensions of attention for uncertainty detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 22–34.

Aubakirova, M. and Bansal, M. (2016). Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2035–2041.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44.

Burgoon, J. K., Stern, L. A., and Dillman, L. (2007). *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

Casanueva, I., Hain, T., Christensen, H., Marxer, R., and Green, P. (2015). Knowledge transfer between speakers for personalised dialogue management. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 12.

Cassell, J. and Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Di Buccio, E., Melucci, M., and Moro, F. (2014). Detecting verbose queries and improving information retrieval. *Information Processing & Management*, 50(2):342–360.

Forbes-Riley, K. and Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9):1115 – 1136.

Forbes-Riley, K., Litman, D., and Rotaru, M. (2008). Responding to student uncertainty during computer tutoring: An experimental evaluation. *Lecture Notes in Computer Science*, 5091:60–69.

Gharouit, K. and Nfaoui, E. H. (2017). A comparison of classification algorithms for verbose queries detection using babelnet. In *2017 Intelligent Systems and Computer Vision (ISCV)*, pages 1–5. IEEE.

Goel, P., Matsuyama, Y., Madaio, M., and Cassell, J. (2018). "i think it might help if we multiply, and not add": Detecting indirectness in conversation. In *Proceedings of the International Workshop Series on Spoken Dialogue Systems Technology (IWSDS)*.

Grave, E., Bojanowski, P., Bojanowski, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Honold, F., Bercher, P., Richter, F., Nothdurft, F., Geier, T., Barth, R., Hoernle, T., Schüssel, F., Reuter, S., Rau, M., Bertrand, G., Seegebarth, B., Kurzok, P., Schattenberg, B., Minker, W., Weber, M., and Biundo, S. (2014). Companion-technology: Towards user- and situation-adaptive functionality of technical systems. In *10th International Conference on Intelligent Environments (IE 2014)*, pages 378–381. IEEE.

Liscombe, J., Hirschberg, J., and Venditti, J. J. (2005). Detecting certainness in spoken tutorial dialogues. In *Proceedings of the Ninth European Conference on Speech Communication and Technology*.

Mairesse, F. and Walker, M. A. (2010). Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

Miehle, J., Yoshino, K., Pragst, L., Ultes, S., Nakamura, S., and Minker, W. (2016). Cultural communication idiosyncrasies in human-computer interaction. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 74.

Miehle, J., Minker, W., and Ultes, S. (2018a). Exploring the impact of elaborateness and indirectness on user satisfaction in a spoken dialogue system. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 165–172. ACM, July.

Miehle, J., Minker, W., and Ultes, S. (2018b). What causes the differences in communication styles? a multicultural study on directness and elaborateness. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Miehle, J., Bagci, I., Minker, W., and Ultes, S. (2019). A social companion and conversational partner for the elderly. In *Advanced Social Interaction with Agents*, volume 510 of *Lecture Notes in Electrical Engineering*, pages 103–109. Springer International Publishing.

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language tech-*

nologies: Short papers, pages 169–172. Association for Computational Linguistics.

Neuliep, J. W. (2011). *Intercultural Communication: A Contextual Approach*. SAGE Publications.

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., and Pham, S. B. (2014). Rdrpostagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20.

Niederhoffer, K. G. and Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Pragst, L., Ultes, S., Kraus, M., and Minker, W. (2015). Adaptive dialogue management in the kristina project for multicultural health care applications. In *Proceedings of the 19thWorkshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, pages 202–203.

Pragst, L., Minker, W., and Ultes, S. (2019). Exploring the applicability of elaborateness and indirectness in dialogue management. In *Advanced Social Interaction with Agents*, volume 510 of *Lecture Notes in Electrical Engineering*, pages 189–198. Springer International Publishing.

Prokofieva, A. and Hirschberg, J. (2014). Hedging and speaker commitment. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*, pages 10–13.

Reitter, D., Keller, F., and Moore, J. D. (2006). Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124. Association for Computational Linguistics.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Stenchikova, S. and Stent, A. (2007). Measuring adaptation between dialogs. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*.

Ulinski, M., Benjamin, S., and Hirschberg, J. (2018). Using hedge detection to improve committed belief tagging. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 1–5.

Ultes, S., Kraus, M., Schmitt, A., and Minker, W. (2015). Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 374–383.

Appendix

Dialogue Acts
Accept
Acknowledge
Advise
AfternoonSayGoodbye
AnswerThank
AskMood
AskPlans
AskTask
AskWellBeing
CheerUp
Console
Declare
EveningGreet
EveningSayGoodbye
ExplicitlyConfirmRecognisedInput
ImplicitlyConfirmRecognisedInput
IndividualisticallyOrientedMotivate
MeetAgainSayGoodbye
MorningGreet
MorningSayGoodbye
Order
PersonalAnswerThank
PersonalApologise
PersonalGreet
PersonalSayGoodbye
PersonalThank
ReadNewspaper
Reject
RepeatPreviousUtterance
RephrasePreviousUtterance
Request
RequestAdditionalInformation
RequestMissingInformation
RequestNewspaper
RequestReasonForEmotion
RequestWeather
ShareJoy
ShowWeather
SimpleApologise
SimpleGreet
SimpleMotivate
SimpleSayGoodbye
SimpleThank

Table 9: List of predefined dialogue acts.