Changing the Level of Directness in Dialogue using Dialogue Vector Models and Recurrent Neural Networks

Louisa Pragst
Ulm University
Albert-Einstein-Allee 43
Ulm, Germany
louisa.pragst@uni-ulm.de

Stefan Ultes
Cambridge University
Trumpington Street
Cambridge, UK
su259@cam.ac.uk

Abstract

In cooperative dialogues, identifying the intent of ones conversation partner and acting accordingly is of great importance. While this endeavour is facilitated by phrasing intentions as directly as possible, we can observe in human-human communication that a number of factors such as cultural norms and politeness may result in expressing one's intent indirectly. Therefore, in human-computer communication we have to anticipate the possibility of users being indirect and be prepared to interpret their actual meaning. Furthermore, a dialogue system should be able to conform to human expectations by adjusting the degree of directness it uses to improve the user experience. To reach those goals, we propose an approach to differentiate between direct and indirect utterances and find utterances of the opposite characteristic that express the same intent. In this endeavour, we employ dialogue vector models and recurrent neural networks.

1 Introduction

An important part of any conversation is understanding the meaning your conversation partner is trying to convey. If we do not obscure our intent and phrase it as directly as possible, our conversation partner will have an easier time to recognise our goal and cooperate in achieving it. Thereby, we can enable a successful conversation. Nevertheless, there are countless instances in which humans choose to express their meaning indirectly, as evidenced by the work of Searle (1975) and Feghali (1997), among others. Answering the question 'How is the weather?' with 'Let's rather stay inside.' gives no concrete information about the weather conditions, but is commonly understood. There are several reasons why humans could choose to express their intent indirectly, such as cultural preferences, politeness, embarrassment, or simply using common figures of speech such as 'Can you tell me the time?'. Considering the frequency of indirectness in human-human communication, we need to anticipate the use of indirectness in human-computer communication and enable dialogue systems to handle it.

In this work, we introduce an approach to exchanging utterances with others that express the same intent in the dialogue but exhibit a differing level of directness. More concretely, our approach would replace the second utterance of the exchange 'What pizza do you want?' - 'I want a vegetarian pizza.' with an utterance like 'I don't like meat'. To this end, we employ models that can estimate the level of directness of an utterance on the one hand and the degree to which utterances express the same intent on the other.

Our approach can be applied to solve two challenges of indirectness for dialogue systems: On the side of the language analysis, the true intent of the user needs to be recognised so that the dialogue system can react in an appropriate, cooperative manner. If the language analysis is able to not only recognise the user's intended meaning, but also when the user is being indirect, this information can further be utilised by the dialogue manager, e.g. by scheduling a confirmation if the user is believed to have used indirectness. Our approach estimates the level of directness of an utterance as a first step. If the utterance is classified as indirect, this information can be provided to the dialogue manager. Furthermore, our approach exchanges the indirect utterance for a direct counterpart that more accurately reflects the users intent, thereby facilitating the task of the lan-

Proceedings of the SIGDIAL 2018 Conference, pages 11–19,
Melbourne, Australia, 12-14 July 2018.

guage analysis. The second area of dialogue system that can benefit from taking into account indirectness is the language generation. Studies could show that under specific circumstances indirectness is preferred not only from human conversation partners, but also in human-computer interaction (e.g. (Miehle et al., 2016; Pragst et al., 2017)). Therefore, dialogue systems that can adjust the level of directness in their output to the user and their circumstances should be able to provide an improved user experience. If a certain level of directness is determined to be desirable with regards to the current circumstances, our algorithm can determine whether the utterance chosen as system output possesses the targeted level of directness and exchange it for a more suitable alternative if it does not.

In the following, we will discuss related work, before presenting our general approach and its concrete implementation. This approach is evaluated in Section 4. Here, we introduce the dialogue corpus we created to obtain a reliable ground truth and discuss the results of our evaluation. Finally, we draw a conclusion in Section 5.

2 Related Work

Allen and Perrault (1980) propose a plan-based approach to understanding the intention of the speaker, explicitly mentioning indirect speech acts as application. Similarly, Briggs and Scheutz (2013) address both the understanding and the generation of indirect speech acts. Their approach combines idiomatic and plan-based approaches. In plan-based approaches, a planning model that contains potential goals as well as actions with pre-and post conditions needs to be defined manually in order to anticipate the user's plan and thereby identify the intent of an utterance. Our approach aims to eliminate the explicit preparation of the planning model, and instead relies on patterns learned from a large amount of examples.

In our work, we utilise a Dialogue Vector Model (DVM) (Pragst et al., 2018) to assess whether two utterances express the same intent in a dialogue. A number of different approaches to the representation of sentences in vector space have been proposed, e.g. utilising recurrent neural networks (Sutskever et al., 2014; Palangi et al., 2016; Tsunoo et al., 2017), convolutional neural networks (Shen et al., 2014; Kalchbrenner et al., 2014; Hu et al., 2014) and autoencoders (Socher et al., 2011). However, those approaches rely on the words in the sentence only to generate a vector representation. As a consequence, sentences that have the same meaning, but do not share the same words (which is often the case for utterances with different levels of directness) are not mapped in the vicinity of each other. In contrast, DVMs map functionally similar sentences close to each other and are therefore better suited for our needs.

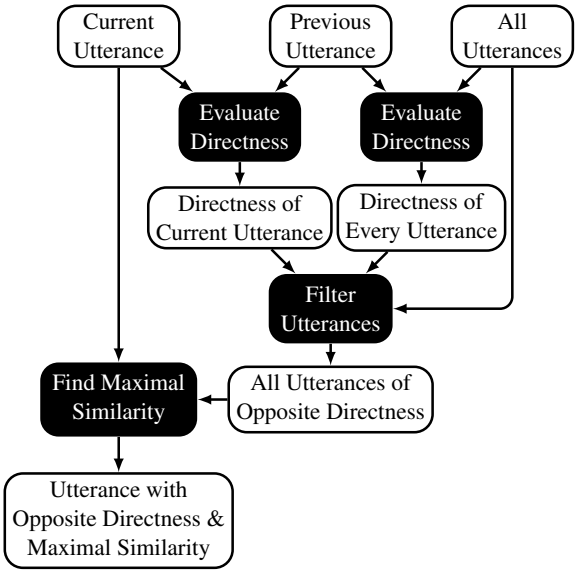

Figure 1: Flow chart of the steps taken to exchange an utterance with another one that is functionally similar and of the opposite directness.

Skip thought vectors (Kiros et al., 2015) are sentence embeddings that are generated in a similar manner as word vector representations, and therefore similar to dialogue vector models. Rather than using the words in the sentence itself as basis to create a vector representation, those vectors are generated taking into account surrounding sentences. However, this representation is trained on novels rather than dialogue, as opposed to DVMs, which focus specifically on dialogue and its peculiarities.

3 Changing the Level of Directness

Our work is concerned with the exchange of utterances for functionally similar ones with differing levels of directness. We define functional similarity as the degree to which two utterances can be used interchangeably in a dialogue as they express the same meaning. Substituting a direct/indirect utterance with its respective counterpart can be achieved by performing the following steps:

Algorithm 1: Pseudocode for exchanging one utterance for another that is functionally similar and of the opposite directness.

Data: $origU$, the utterance to be exchanged
$prvU$, the utterance occurring previous to $origU$
$allU$, the set of all available utterances
`DVM`, a function that maps an utterance to its corresponding dialogue vector
`evalInd`, a function that returns the estimated level of directness, ranging from one to three
Result: $excU$, the substitute for $origU$

$origDirectness \longleftarrow \texttt{evalInd}(prvU, origU)$;
if $origDirectness \leq 1$ **then**
  $oppU \longleftarrow \{u \in allU : \texttt{evalInd}(prvU, u) > 1\}$;
else
  $oppU \longleftarrow \{u \in allU : \texttt{evalInd}(prvU, u) \leq 1\}$;
$excU \longleftarrow$
$\text{argmin}_{u \in oppU}\, \texttt{euclDist}(\texttt{DVM}(origU), \texttt{DVM}(u))$;

1. Determine the level of directness of the utterance.
2. Gather the remaining known utterances that are of the opposite directness level.
3. From those, choose the utterance that is functionally most similar to the original utterance.

Figure 1 shows this procedure on an abstract level, while a more detailed pseudo-code is depicted in Algorithm 1. Two challenges need to be addressed in order to perform this approach: The first one is to correctly determine the level of directness of an utterance, the second one is to identify utterances that perform a similar semantic functionality in a dialogue. To solve those challenges, we utilise established approaches, namely recurrent neural networks (RNN) and dialogue vector models (DVM). In the following, we take a closer look at how we apply those approaches to solve the presented challenges.

To determine which utterances can be exchanged without altering the intended meaning, a suitable similarity measure is needed. In our work, we utilise DVMs (Pragst et al., 2018) to that end. DVMs are representations of sentences as vectors that captures their semantic meaning in the dialogue context. They are inspired by word vector models (Mikolov et al., 2013a) and generated in a similar manner: The mapping of utterances to their vector representations is trained akin to autoencoding. However, rather than training against the input utterance itself, utterances are trained against their adjacent utterances in the input corpus, either using the utterance to predict its context or using the context to predict the utterance. The resulting vector representation groups sentences that are used in a similar context and therefore likely to fulfil the same conversational function in close vicinity to each other, as could be shown by Pragst et al. (2018). Therefore, DVMs are well suited to determine whether utterances perform a similar function in a dialogue. Our algorithm calculates the euclidean distance between the dialogue vector representations of two utterances and chooses the utterance with the minimal distance as the most functionally similar.

For the estimation of the level of directness an utterance possesses, we choose a supervised learning approach with a RNN. RNNs are a popular supervised machine learning approach to find complex relationships in large amounts of sequential data. As indirectness relies on the context of the conversation, the use of RNNs seems promising for the estimation the level of directness an utterances possess. The architecture of our RNN is depicted in Figure 2. It is a time delay network that uses the previous input in addition to the current one. To obtain a numerical representation of an utterance that can be used as input to the network, we utilise word vector models (Mikolov et al., 2013a) and DVMs (Pragst et al., 2018). The input for an utterances then consists of its dialogue vector representation and the sum of the word vector representations of its words. Furthermore, the word and dialogue vectors of the previous utterance are provided as recurrent data to reflect the dialogue context. The target value is given by corpus annotations of the level of directness of the utterance. As we are trying to solve a classification problem, the network is designed to provide the probability that the utterance belongs to each of the classes as its result. After training, the network constitutes the core part of the function that estimates the level directness of an utterance.

4 Evaluation

This section presents the evaluation of the proposed approach. We first introduce a dialogue corpus that is suitable to train the required models and provides a reliable ground truth to compare the results of our approach to. Afterwards, the setup of the evaluation is described and its results presented and discussed.

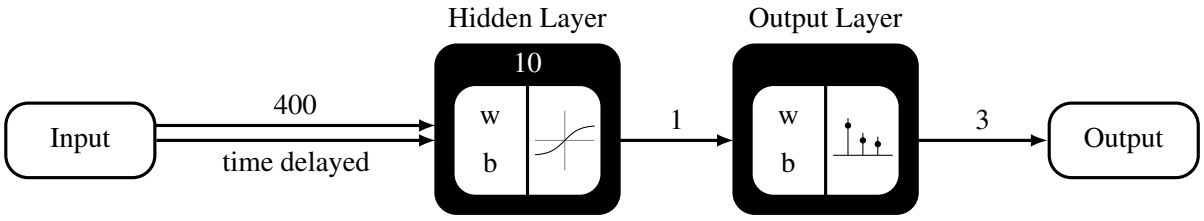

Figure 2: The architecture of the RNN used for the estimation of directness. It is a time-delay network with a one step delay from the input layer to the hidden layer, which contains ten nodes. The output layer gives the probability that the input belongs to a class for each of the three classes.

4.1 Dialogue Corpus

Our approach requires a dialogue corpus for several task: as a source for alternative utterances, as training data for the directness classifier, as training data for the DVM and as ground truth for the evaluation. To fulfil those tasks, the employed corpus has to meet two requirements: it needs to contain a sufficient amount of examples for functionally similar direct and indirect utterances, and the utterances need to be annotated with their dialogue act and level of directness.

We considered several existing dialogue corpora, none of which suited our needs. Furthermore, we dismissed the option to collect and annotate a dialogue corpus ourselves, considering the difficulty to make sure that speakers would use different levels of directness for the same purpose without inhibiting the naturalness of the dialogues. Instead, we decided to generate a suitable dialogue corpus automatically.

The advantages an automatically generated corpus offers for our work are the certainty that it contains a number of examples for functionally similar direct and indirect variants, as well as a dependable ground truth for the evaluation. However, automatically generated corpora come with certain limitations. After introducing our dialogue corpus in the following, we will discuss the potential advantages and limitations of automatically generated corpora.

4.1.1 Description of the Dialogue Corpus

Our corpus contains dialogues with two different tasks: ordering pizza and arranging joint cooking. Example dialogues can be found in Figure 3. The dialogues incorporate typical elements of human conversation: different courses of the dialogue, over-answering, misunderstandings as well as requests for confirmation and corrections, among others. The example dialogues also show instances of different wordings for the same purpose, such as several indirect variants of 'Yes.', such as 'Great.', 'I'm looking forward to it.' and 'That sounds delicious.' that can be found across the dialogues, and the direct 'I would like to order pizza.' in Dialogue 3 that is exchanged for the indirect 'Can I order pizza from you?' in Dialogue 4. Additionally, the same utterance can have a different level of directness depending on the context: in Dialogue 1, the utterance 'I haven't planned anything.' as response to 'Do you have time today?' is indirect, whereas it is direct as response to 'Do you have plans today?' in Dialogue 2. Overall, the corpus contains more than 400000 different dialogue flows and about four wordings per dialogue action.

As first step of the corpus generation, we defined a dialogue domain in a similar manner to the ones often employed by dialogue managers (e.g. OwlSpeak (Ultes and Minker, 2014)). It contains all system and user actions foreseen for the dialogues, and defines rules about feasible successions of those. Furthermore, each system and user action is assigned a number of different utterances that can be used to express their intent. Each utterance incorporates a level of directness ranging from one to three, with one being direct (e.g. 'I want vegetarian pizza.') and three indirect (e.g. 'I don't like meat.'). A rating of two is assigned if the utterance is indirect, but still very close to the direct one, or a common figure of speech (e.g 'Can I get vegetarian pizza?'). The directness level depends not only on the utterance itself, but also on the dialogue context. Therefore, the utterance 'I have time today.' receives a rating of three if the previous utterance was 'Do you have plans today?', and a rating of one if the previous utterance was 'Do you have time today?'.

In the next step, all dialogue flows are generated by recursively picking a dialogue action, gen-

Dialogue 1

SPEAKER 1: Hello.
SPEAKER 2: Hello.
SPEAKER 1: Do you have time today?
SPEAKER 2: I haven't planned anything.
SPEAKER 1: How hungry are you?
SPEAKER 2: Just a little.
SPEAKER 1: Would you share some food with me?
SPEAKER 2: Yes.
SPEAKER 1: Do you have any food preferences?
SPEAKER 2: I like pineapple.
SPEAKER 1: You probably would like pineapple salad.
SPEAKER 2: Great.
SPEAKER 1: We could cook that together.
SPEAKER 2: I'm looking forward to it.
SPEAKER 1: Byebye.
SPEAKER 2: Byebye.

Dialogue 2

SPEAKER 1: Hello.
SPEAKER 2: Hello.
SPEAKER 1: Do you have plans today?
SPEAKER 2: I haven't planned anything.
SPEAKER 1: What did you eat today?
SPEAKER 2: Just a little.
SPEAKER 1: Would you share some food with me?
SPEAKER 2: I don't need much.
SPEAKER 1: Which food do you like?
SPEAKER 2: I don't like meat.
SPEAKER 1: You probably would like pineapple salad.
SPEAKER 2: That sounds delicious.
SPEAKER 1: We could cook that together.
SPEAKER 2: Great.
SPEAKER 1: Byebye.
SPEAKER 2: Byebye.

Dialogue 3

SPEAKER 1: Hello.
SPEAKER 2: I am listening.
SPEAKER 1: I would like to order pizza.
SPEAKER 2: We offer different sizes.
SPEAKER 1: A small one sounds good.
SPEAKER 2: I have noted a small pizza.
SPEAKER 1: Great.
SPEAKER 2: What would you like on top?
SPEAKER 1: I like pineapple.
SPEAKER 2: You're getting a Hawaiian pizza.
SPEAKER 1: I don't like meat.
SPEAKER 2: Do you want a salad?
SPEAKER 1: You can't live just on pizza.
SPEAKER 2: So you want a small vegetarian pizza with a salad?
SPEAKER 1: That sounds delicious. Byebye.
SPEAKER 2: Byebye.
SPEAKER 1: Byebye.

Dialogue 4

SPEAKER 1: Hello.
SPEAKER 2: Hello. Is there anything I can help you with?
SPEAKER 1: Can I order pizza from you?
SPEAKER 2: We offer Hawaiian, peperoni and vegetarian.
SPEAKER 1: I choose peperoni pizza. I love salad. I'm thinking about a large one.
SPEAKER 2: I have noted a large pepperoni pizza with a salad.
SPEAKER 1: This is going to be good.
SPEAKER 2: Byebye.
SPEAKER 1: Byebye.

Figure 3: Example dialogues from the automatically generated corpus. The dialogues encompass different tasks, over-answering, misunderstandings, confirmations and corrections. Furthermore, they contain several examples of exchangeable utterances with differing directness levels, as well as examples of the same utterances changing its level of directness due to the dialogue context.

erating a list of its possible successors as stated by the rules in the dialogue domain and repeating the procedure for each of the successors. If a dialogue action does not have successor, the sequence of dialogue actions that have been chosen to get to that point are saved as a complete dialogue. The wording is chosen randomly from the utterances associated with the respective dialogue action.

4.1.2 Discussion of Automatically Generated Corpora

The use of automatically generated corpora is not widely adopted in the research community of human-computer interaction. Due to their artificial nature, they have obvious limitations: they possess less flexibility than natural conversations, regarding both the dialogue flow and the different wordings. As a result, both dialogue flow and wording are much more predictable for automatically generated corpora and it is highly likely that machine learning approaches and similar procedures will perform better on generated dialogues than they would on natural ones. Nevertheless, we believe that generated dialogues have their benefits: they should not be used to gauge the actual performance of approaches in an applied spoken dialogue system, but rather to appraise their potential.

The comparison of natural and automatically generated dialogue corpora bears parallels to the discussion regarding laboratory experiments and field experiments, and their respective advantages and limitations (as discussed by Berkowitz and Donnerstein (1982), Harrison and List (2004) and Falk and Heckman (2009), among others). While natural dialogues more accurately represent conversations in the real world, automatically generated dialogues offer more control. In particular, that means specific questions can be tested in a structured and systematic manner, the generation ensuring that relevant data is incorporated in the corpus and irrelevant data that might interfere with the experiments is excluded, as well as the presence of a dependable ground truth. Therefore, we can reliably assess whether an approach is viable to solve a given task.

Additionally, by being able to provide the complete data set for a smaller scale use case as defined by the dialogue domain, we can get an idea about the potential performance of an approach given a large amount of data that approaches the state of total coverage. While this amount of data is usually unobtainable for most researchers, large companies have the resources to collect a suitably big corpus and are likely already working towards it. Therefore, it is beneficial to examine the full potential of a given approach. However, in our considerations regarding the availability of large amounts of data we need to take into account that even large companies typically do not have access to a large amount of *annotated* data.

In summary, we believe that automatically generated dialogues, while not providing us with an accurate performance measure of an approach in the real world, can help us to assess its general viability to solve a specific task and to estimate its performance given enough data.

4.2 Setup of the Evaluation

For the evaluation of our approach we determine its accuracy in finding an utterance that shares the dialogue action with the original utterance and is of the opposite level of directness. The ground truth for both criteria is given by the previously presented dialogue corpus. In addition, we also evaluate the performance of the trained classifier and investigate how it influences the overall performance. As the ability of DVM to group utterances that share a dialogue action has already been shown in (Pragst et al., 2018), it will not be part of this evaluation.

To investigate the effects of the amount of available data, we use several DVMs that are trained on only a fraction of the complete corpus. Corpus sizes of 0.1, 0.2, 0.4, 0.6, 0.8 and of course the full corpus are considered. The dialogues that are part of the reduced corpora are chosen at random.

Another aspect we study is the impact of the amount of available annotated training data for the classifier on its performance. As usual, we use ten-fold cross-validation in our evaluation. However, instead of only using 90% of the utterances for training and 10% for testing, we also evaluate our approach using 10% of the utterances for training and 90% for testing. With this, we want to investigate how our approach performs given only a limited amount of annotated data.

Finally, we compare the performance of the classifier when using only dialogue vectors as input and when using both dialogue vectors and the sum of word vectors. As DVMs map functionally similar utterances in close vicinity to each other, direct and indirect utterances should be hard to

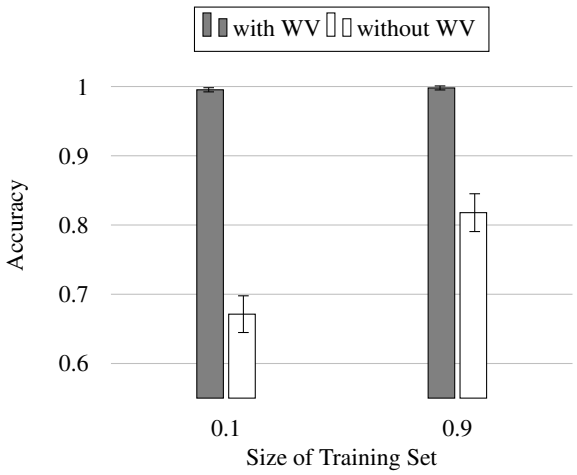

Figure 4: The mean accuracy and SD achieved by different classifiers.

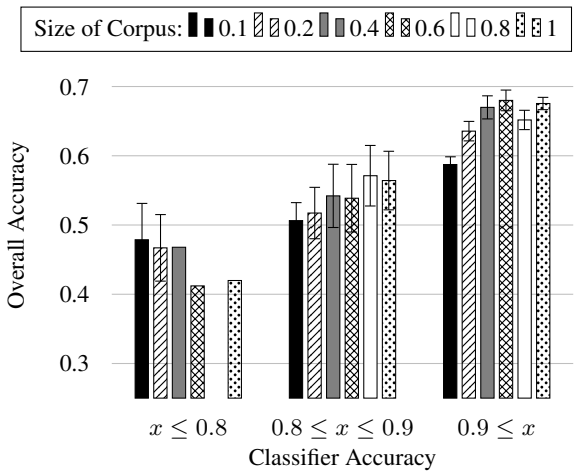

Figure 5: The mean accuracy and SD achieved by with different DVMs and Classifiers.

distinguish with just the information from those models. On the other hand, the sum of word vectors might be missing important context information for the identification of the directness level. We believe that the combination of both the sum of word vectors and dialogue vectors will improve the performance of the classifier.

The DVMs we utilise in our evaluation as similarity measure and as input to the RNN are trained on the presented dialogue corpus. The network additionally receives the sum of the word vectors of an utterance, based on the Google News Corpus model (Mikolov et al., 2013b), as input.

4.3 Results

Overall, our results show that the proposed approach has a high potential. The best mean accuracy reaches a value of 0.68 , and the classifier predicts the right class with 0.87 accuracy on average. In the following, we discuss the results and their implications in more detail, starting with the results of the classifier, before assessing the overall performance.

4.3.1 Classification of Directness

The baseline performance our classifier should surpass the prediction of the majority class. With the given data, such a classifier can achieve an accuracy of 0.5291. Our trained classifier achieves a significantly better accuracy of 0.8710 ($t(203) = 35.366$, $p < .001$) averaged over all test cases. Even the worst classifier, with an accuracy of 0.6354, performs more than 10% better than choosing the majority class.

As expected, significant differences exist for the size of the training set ($t(159.425) = -4.008$, $p < .001$), with a larger training set leading to better results. Furthermore, adding the linear combination of the word vectors as input improves the performance of the classifier significantly ($t(101.347) = 32.434$, $p < .001$). The mean performances can be seen in Figure 4. The corpus size the DVMs were trained on does not have a significant impact.

Those results suggest that the amount of labelled training data greatly affects the performance of a classifier using RNN. If the goal is a large scale application, the necessary amount of labelled data might be difficult to achieve. Future work should therefore consider the possibility of unsupervised training approaches or approaches with better scalability. In addition to a larger amount of training data, using the sum of word vectors as additional input greatly improves the performance. As a number of extensive word vector models exist for several languages (e.g. (Bojanowski et al., 2016)), this data is easily available irrespective of the scale of the targeted dialogue domain.

4.3.2 Exchange of Utterances

Our approach for choosing a valid replacement for an utterance was able to achieves a high accuracy of 0.70 at its best performance. However, this performance is significantly influenced by both the accuracy of the classifier for the level of directness ($F(2, 29.090) = 141.564$, $p < .001$) and the amount of data the DVM was trained on ($F(5, 52.864) = 4.304$, $p < .003$). Depending

on the quality of the employed components, the accuracy ranges from 0.41 to 0.70. A graphical representation can be found in Figure 5.

The results show the high potential of our approach, but also emphasize the importance of both a good classifier to estimate the level of directness and a good measure of the functional similarity of utterances. If either component under performs, the accuracy declines to undesirable levels. DVMs depend on a large amount of data being available. However, this data does not need to be annotated. Hence, suitable DVMs for our approach can be trained with the amount of data usually available to big companies. Training a good classifier presents a more severe challenge, as annotated data is needed. An unsupervised approach to the training of a classifier for the level of directness would therefore be highly beneficial for the viability of our approach.

4.4 Limitations of the Evaluation

The evaluation of our approach yields promising results and shows its high potential. However, we need to take into account that those results were achieved using an artificially generated corpus. Furthermore, we tested the performance of our approach in a theoretical setting, not its impact in an actual application. This section discusses the limitations of our evaluation.

Natural dialogue possess a greater variability than automatically generated dialogue, and therefore finding reliable patterns in them is a more difficult task. It is likely that the quality of both the classifier and the DVMs decreases if they are trained on a comparable amount of natural dialogue data compared to artificially generated data. We could show in the evaluation that the quality of the classifier and DVM has a major impact on the performance of our approach. This implies that more data is needed for natural dialogues than for automatically generated dialogues to achieve comparable results.

One of the main reasons to use an automatically generated dialogue corpus was to ensure the presence of pairs of direct and indirect utterances. This is important not only for the training of the classifier and DVM, but also to ensure that a suitable substitute is known. As our approach searches for a replacement in a set of established utterances, it can only be successful if the set does contain a suitable utterance. While the likelihood for the presence of a suitable substitute increases with the size of the dialogue corpus, it cannot be guaranteed that a replacement is present in natural dialogues. When transferring our approach to actual applications, this might present a challenge. To address this challenge, the generation of suitable utterances rather than their identification should be investigated.

While our evaluation shows what accuracy our approach can achieve given different circumstances, we did not yet investigate what accuracy it needs to achieve in actual applications to positively impact the user experience. Without this information, it is difficult to estimate which level of accuracy should be targeted and, as a consequence, the amount of training data needed.

5 Conclusion

In this work, we introduced an approach to exchange utterances that express the same meaning in the dialogue, but possess a differing level of directness. In this endeavour, we utilised supervised training with RNNs for the estimation of directness levels, and DVMs as basis for the similarity measure of the meaning of two utterances in a dialogue. A dialogue corpus that provides a sufficient amount of direct/indirect utterance pairs as well as annotations of the dialogue act and level of directness was generated automatically and utilised to show the high potential of our approach in an evaluation.

Although the results seem promising overall, we identified several challenges that need to be addressed in future work. The chosen classifier for the level of directness relies on a large amount of annotated data. Unsupervised learning approaches will be investigated to eliminate this need. Our evaluation did not incorporate the variability of natural dialogues. We will test our approach on natural dialogues to verify its applicability on more noisy data than an automatically generated corpus provides. Furthermore, the presence of direct/indirect pairs in natural dialogue corpora cannot be guaranteed. It might become necessary to explore the generation of suitable utterances if we find that natural dialogue data does not contain a sufficient amount of direct/indirect utterance pairs. Finally, the integration of our approach in an actual dialogue systems can confirm its beneficial effects on the user satisfaction.

References

James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial intelligence* 15(3):143–178.

Leonard Berkowitz and Edward Donnerstein. 1982. External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American psychologist* 37(3):245.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Gordon Michael Briggs and Matthias Scheutz. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *AAAI*.

Armin Falk and James J Heckman. 2009. Lab experiments are a major source of knowledge in the social sciences. *science* 326(5952):535–538.

Ellen Feghali. 1997. Arab cultural communication patterns. *International Journal of Intercultural Relations* 21(3):345–378.

Glenn W Harrison and John A List. 2004. Field experiments. *Journal of Economic literature* 42(4):1009–1055.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*. pages 2042–2050.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* .

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. pages 3294–3302.

Juliana Miehle, Koichiro Yoshino, Louisa Pragst, Stefan Ultes, Satoshi Nakamura, and Wolfgang Minker. 2016. Cultural communication idiosyncrasies in human-computer interaction. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, Los Angeles, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24(4):694–707.

Louisa Pragst, Wolfgang Minker, and Stefan Ultes. 2017. Exploring the applicability of elaborateness and indirectness in dialogue management. In *Proceedings of the 8th International Workshop On Spoken Dialogue Systems (IWSDS)*.

Louisa Pragst, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2018. On the vector representation of utterances in dialogue context. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France.

John R Searle. 1975. *Indirect speech acts*. na.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, pages 101–110.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 151–161.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Emiru Tsunoo, Peter Bell, and Steve Renals. 2017. Hierarchical recurrent neural network for story segmentation. *Proc. Interspeech 2017* pages 2919–2923.

Stefan Ultes and Wolfgang Minker. 2014. Managing adaptive spoken dialogue for intelligent environments. *Journal of Ambient Intelligence and Smart Environments* 6(5):523–539.