EFFECT OF DIALOG ACTS ON WORD USE IN POLYLOGUE

Roland Roller[†⋆], *Shinji Watanabe*[†] *and Tomoharu Iwata*[†]

[†]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[⋆]German Research Centre for Artificial Intelligence (DFKI), Berlin, Germany

ABSTRACT

In this work we examine the effect of dialog acts on word use, in context of the influence of interlocutors in a polylogue on each other. The basic idea of this work is the extension of the cache model and the influence model by dialog act information. The cache model covers the re-usage of words and the influence model calculates the influence of interlocutors in a polylogue on each other. Both approaches could be used to improve the word prediction accuracy in a word generative model. We start to examine the usage of dialog acts to improve our word generative model in terms of perplexity. For the usage of dialog acts, a knowledge about the future dialog act is required. Therefore, we examine how dialog act miss-prediction influences the resulting performance. Further on, we introduce a new approach to generate artificial dialog acts which guarantees the knowledge about the following dialog act. Our final experiments present the improvements in terms of perplexity using our new approach in AMI, NIST and NTT meeting corpora.

Index Terms— Speech Entrainment, Influence Model, Dialog Act Model

1. INTRODUCTION

In conversations, people tend to be influenced by their interlocutor. They could be affected by their interactions, their viewing direction or also by their language. In this context for example, interlocutors tend to agree on common terms (speech entrainment) [1]. Another effect, which is also connected to that phenomena is, the re-usage of words in a conversation [2]. These phenomena could be utilized to improve the word prediction accuracy (WPA) of a word generative model (WGA), which aims to gain better results for language model applications (e.g. information retrieval or speech recognition). So far, the re-usage of words within a conversation could be covered quite well with the cache model [3]. Iwata and Watanabe [4] went further and showed that the influence of different speakers affect the re-usage of words in a polylogue scenario. Their approach calculates the influences of each speaker on each other and use this information to optimise the word prediction accuracy. In this work, we extend these approaches by using dialog act (DA) information. We assume that each DA contains characteristic word occurrences, depending on the definition of the dialog act. Therefore, we use this effect to improve the cache and influence estimation.

There are some studies about the improvement of the word prediction accuracy by using DA information, for example [5], [6] and [7]. However, one major problem of using DA information in that context of e.g. speech recognition, is the dialog act prediction. Woszczyna and Waibel [5], Nagata and Morimoto [6] and Reithinger *et al.* [8] face the problem on basis of Markov models by using previous DA information (n-gram) to predict the next DA. Other approaches to predict the following dialog act, presented by Alexandersson and Reithinger [9] and Geertzen [10] for example, are using grammar induction. In this paper, we present a method to group words by statistical, time dependent characteristics. That means, that we expect different word distributions at the beginning of a user turn, compared to the following words in that same turn. On bases of that, we divide the each user turn into two artificial dialog acts. A former dialog act, containing the first words and a latter dialog act containing the following words. On basis of this approach, we know about the future dialog act and do not need a dialog act estimation anymore. Finally, we extend the influence and cache model by these artificial dialog act models and examine the approaches on three different corpora in different languages. All DA extended models outperformed their baseline model.

In the following section we present the baseline approaches which are required for our further work, followed by the definitions of our dialog act models and the extensions of the cache and influence model in section 3. In section 4, we run experiments using our adapted influence model and analyse the effect of dialog act miss-estimation. Section 5 introduces the new artificial dialog act approach, followed by our final experiments in section 6. Section 7 presents the conclusion and gives an outlook for future work.

2. BASELINE

In this section we briefly introduce all baseline information we need in the further sections. All defined approaches are based on the following definitions. Let be $\boldsymbol{w} = \{w_1, \cdots, w_t, \cdots\}$ a word sequence of a polylogue, where w_t represents the tth word, and let be $\boldsymbol{s} = \{s_1, \cdots, s_t, \cdots\}$ the speaker sequence, where s_t indicates the speaker of the tth word, with $w_t \in \{1, \cdots, W\}$ and $s_t \in \{1, \cdots, M\}$. W represents the vocabulary size and M the number of participants. The uniform distribution is defined as $P_U(w)$ and the general word distribution is $P_G(w)$. We use a smoothing parameter β, to avoid zero probability, τ represents the cache size and $\delta(x, y)$ is Kronecker's delta, which means $\delta(x, y) = 1$, if $x = y$ and 0 otherwise. Although the formulations in this paper is based on a uni-gram word generative model, we can apply it to an n-gram model straightforwardly.

2.1. Cache Model

The cache model (CC) is an efficient approach to cover the re-usage of words within a specific period τ. The original cache model deals with monologue word generative models. We use this approach to deal with polylogue WGAs. The word use of speaker n at position t could be described as follows:

$$P_{CC}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) = \lambda_1 P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}) + \lambda_2 P_{\mathrm{G}}(w) + \lambda_3 P_{\mathrm{U}}(w) \quad (1)$$

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

with

$$P_{\mathrm{C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}) = \frac{\sum_{t'=t-\tau}^{t-1} \delta(w, w_{t'}) + \beta}{\sum_{w'} \sum_{t'=t-\tau}^{t-1} \delta(w', w_{t'}) + \beta W}. \qquad (2)$$

Equation 2 represents the cache model itself and τ indicates the cache size. The linear interpolation coefficients λ_1, λ_2, and λ_3 are obtained by using maximum a posteriori (MAP) estimation, with $0 \leq \lambda_i \leq 1$ and $\sum_i \lambda_i = 1$.

2.2. Influence Model

The influence model is an extension of the cache model. It uses an individual cache for every speaker and identifies, after a certain amount of training steps, the influence of the interlocutors on each other within a polylogue. The resulting influence parameter are used to weight the different cache models to calculate the resulting WGA and was defined in [4]. The word use of speaker n at position t can be modelled as follows:

$$\begin{aligned} P_{\mathrm{I}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) \quad &= \quad \sum_{m=1}^{M} \lambda_{\mathrm{nm}} P_{\mathrm{C'}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) \\ &\quad + \lambda_{\mathrm{nM+1}} P_{\mathrm{G}}(w) + \lambda_{\mathrm{nM+2}} P_{\mathrm{U}}(w) \qquad (3) \end{aligned}$$

where

$$\begin{aligned} &P_{\mathrm{C'}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m) \\ &= \frac{\sum_{t'=t-\tau}^{t-1} \delta(w, w_{t'})\delta(m, s_{t'}) + \beta}{\sum_{w'} \sum_{t'=t-\tau}^{t-1} \delta(w', w_{t'})\delta(m, s_{t'}) + \beta W} \qquad (4) \end{aligned}$$

is the cache of speaker m. λ_{nm} represents the influence of speaker m on speaker n. λ_{nm}, λ_{nM+1}, and λ_{nM+2} are also obtained by using maximum a posteriori estimation with $0 \leq \lambda_{nm} \leq 1$ and $\sum_m \lambda_{nm} = 1$.

3. DIALOG ACT EXTENDED MODELS

The basic idea of the influence model is, that people tend to use more or less often words of their interlocutor, depending on the strength of influence. The dialog act adapted influence model works even on a more precise level. We suppose, that the influence on re-using words, also depends on specific dialog acts. For example a person is re-using words of an interlocutor in a polylogue. On the other side, both persons using different words as *backchannel* or *confirmation*. In that case, the influence on a dialog act covering general information, has a stronger influence, than their *backchannel* or *confirmation*.

In the following, we present the extension of the cache and the influence model by employing a dialog act approach. We will use three different DA models to extend the given approaches. In addition to the general word distribution, we first extract a general word distribution for every dialog act based on the training corpus (we call it: DA-WGA). We define $\boldsymbol{d} = \{d_1, \cdots, d_t, \cdots\}$ as a dialog act sequence, where d_t indicates the dialog act of the tth word, with $d_t \in \{1, \cdots, D\}$ where D indicates the number of different dialog acts and L is the size of the training corpus. The word use of speaker n at position t with dialog act d is defined as follows:

$$P_{\mathrm{DA_W}}(w|d) = \frac{\sum_{i=1}^{L} \delta(w, w_i)\delta(d, d_i) + \beta}{\sum_{w'} \sum_{i=1}^{L} \delta(w', w_i)\delta(d, d_i) + \beta W} \qquad (5)$$

Since the influence model is trained on the previous dialog turns within a polylogue, we also define for every dialog act a cache model (DA cache):

$$\begin{aligned} &P_{\mathrm{DA_C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, d) \\ &= \frac{\sum_{t'=t-\tau}^{t-1} \delta(w, w_{t'})\delta(d, d_{t'}) + \beta}{\sum_{w'} \sum_{t'=t-\tau}^{t-1} \delta(w', w_{t'})\delta(d, d_{t'}) + \beta W} \qquad (6) \end{aligned}$$

For the complete extension of the influence model, we further define for every interlocutor a user specific set of dialog act cache models (DA influence):

$$\begin{aligned} &P_{\mathrm{DA_I}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m, d) \\ &= \frac{\sum_{t'=t-\tau}^{t-1} \delta(w, w_{t'})\delta(m, s_{t'})\delta(d, d_{t'}) + \beta}{\sum_{w'} \sum_{t'=t-\tau}^{t-1} \delta(w', w_{t'})\delta(m, s_{t'})\delta(d, d_{t'}) + \beta W} \qquad (7) \end{aligned}$$

Based on these dialog act models, we extend the cache and the influence model in the following way. The word use of speaker n, expecting dialog act d, at position t, could be described by these extensions:

$$P_{\mathrm{CC_1}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) = P_{\mathrm{CC}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) + \lambda_4 P_{\mathrm{DA_W}}(w|d) \qquad (8)$$

$$\begin{aligned} &P_{\mathrm{CC_2}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) \\ &= P_{\mathrm{CC_1}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) + \lambda_5 P_{\mathrm{DA_C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m, d) \qquad (9) \end{aligned}$$

Equation 8 and 9 are extensions of the cache model. The first equation extends the standard cache by a DA-WGA and the second one by DA-WGA and a DA cache model. The following three models extend the influence model.

$$P_{\mathrm{I_1}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) = P_{\mathrm{I}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n) + \lambda_{nM+3} P_{\mathrm{DA_W}}(w|d) \qquad (10)$$

$$\begin{aligned} &P_{\mathrm{I_2}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) \\ &= P_{\mathrm{I_1}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) + \lambda_{nM+4} P_{\mathrm{DA_C}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, d) \qquad (11) \end{aligned}$$

Equation 10 and 11 are similar to the adaptions of the cache model. The first one includes the DA-WGA and the second one integrates a DA-WGA and a DA cache model. Equation 12 represents the final dialog act influence model. It contains all previous defined dialog act models. Beside the DA-WGA and the DA cache, it additionally integrates the influence on the dialog act:

$$\begin{aligned} P_{\mathrm{I_3}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) &= P_{\mathrm{I_2}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, n, d) \\ &+ \sum_{m=M+5}^{2M+5} \lambda_{nm} P_{\mathrm{DA_I}}(w|\boldsymbol{w}_{t-\tau}^{t-1}, m, d) \end{aligned} \qquad (12)$$

The last part of equation 12 represents the influence of dialog act d of speaker $m - (M + 5)$ on speaker n. Thus, we extend the original influence model to involve DA information. In the next section we analyse the effect of dialog act information on the word prediction accuracy in terms of perplexity and the problem of dialog act miss-estimation.

4. DIALOG ACT ANALYSIS

To examine the efficiency of our new approach, we ran model (8) - (12) on the transcription of the AMI [11] corpus, a four person meeting corpus in English. It contains around 100 hours transcribed audio material and covers 16 different dialog acts, grouped into six groups. For these and the following experiments, we divided the corpus into a training (93 dialogs) and a test (46 dialogs) set. Further on, we examined the negative effect of dialog act miss-estimation. The baseline was the DA prediction accuracy of Reithinger *et al.* [8] using n-grams with 75.8% for the best three and Geertzen [10] using grammar induction with 78.9% for the best DA. This experiment aimed to simulate different kind of the estimation accuracies, to examine the negative effect on its perplexity. The simulation of the estimation was realized, by choosing the correct or a incorrect dialog act, based on the accuracy parameter. That means, for example, that we chose a wrong DA-WGA by a probability of 30%, if we use an accuracy model of 70%. Since we randomly selected one of the other dialog acts, we ran each test 10 times and calculated the mean perplexity value at the end. 100% accuracy means, that we always predict the dialog act in the right way and stands for the *aspired value*. Baseline for the results in table 1 is the perplexity of the cache ($\boldsymbol{CC = 334.51}$) and influence model ($\boldsymbol{I = 327.35}$). Within all following experiments we are using the cache parameter $\tau = 300$, as suggested in [4] and a smoothing parameter $\beta = 10^{-8}$.

model / eq.	60%	70%	80%	90%	100%
CC_1 / (8)	351.06	333.31	318.30	306.29	292.66
CC_2 / (9)	355.58	335.89	318.30	303.14	288.24
I_1 / (10)	336.05	325.60	312.42	299.65	287.17
I_2 / (11)	340.98	325.81	311.07	297.06	283.50
I_3 / (12)	345.09	**325.59**	**310.67**	**296.00**	**282.10**

Table 1. Perplexities using different DA estimation accuracies

If we consider the *aspired value* (100%), we achieve good results. All dialog act adapted models outperform their baseline model by far. The adapted influence model extension I_3 achieves the best results. Unfortunately, the performance of current techniques are still far away from a 100% DA estimation accuracy. Considering the accuracy between 70-80%, we still achieve an improvement, compared to the baseline models. On the other side, these results are much worse, than the results using 100% accuracy. The problematic of increasing perplexity, predicting the wrong dialog act, brings us to a new approach, presented in the next section.

5. ARTIFICIAL DIALOG ACTS

In the previous sections we presented promising results using DA information for the improvement of the word prediction accuracy within a polylogue. Unfortunately, a reliable prediction, which is not guaranteed yet, is required to use their advantage. This brings us to an alternative idea, about grouping user replies in a way, to reach a better result. The first approach aimed on grouping specific dialog acts in a meaningful way together, e.g. *accepts* and *rejects*. But our preliminary results did not seem to be promising. Examination the different dialog acts and replies in our investigation indicates that especially at the beginning of a user turn, interlocutors tend to use quite similar words e.g. *backchannels* or *confirmation* ('*yes*', '*yeah*'). Table 2 present an abridgment of a dialog between of speaker A, B and C within the AMI corpus. The origin annotated dialog acts are marked with square brackets.

sp.	AMI dialog sequence
A	[I know and it becomes ridiculous yes I know]
C	[and it becomes][yeah]
B	[Or a speech recognition]
C	[yeah speech recognition]
A	[yeah]
B	[which is extremely expensive]
C	[but][yeah]
B	[I think that's the only way that you kind of avoid]
A	[yes mm-hmm]
B	[that kind of issue]
A	[Do we really have to initially um you know looking ...]

Table 2. AMI DA annotated dialog example

First, our approach tries to take advantage of the expectation that specific words seem to occur more often at the beginning of a user turn. In addition, our approach avoid the problematic of imperfect dialog act estimation. These bring us to the idea to group all words within a user turn into one of two remaining 'dialog acts'. The first amount of words b will be grouped into the first dialog act d_{start} and all other words into the second dialog act d_{rest}. By doing this, we can describe our DA-WGA in the following way:

$$P_{\mathrm{DA_W}}(w|d_{start}) = \frac{\sum_{i=1}^{T}\sum_{j=1}^{b}\delta(w, w_j) + \beta}{\sum_{i=1}^{T}\sum_{w'}\sum_{j=1}^{b}\delta(w', w_j) + \beta W} \qquad (13)$$

$$P_{\mathrm{DA_W}}(w|d_{rest}) = \frac{\sum_{i=1}^{T}\sum_{j=b+1}^{T_i}\delta(w, w_j) + \beta}{\sum_{i=1}^{T}\sum_{w'}\sum_{j=b+1}^{T_i}\delta(w', w_j) + \beta W} \qquad (14)$$

with T as the amount of user turns in our training corpus and T_i as the amount of words in turn i. The parameter b could be maximised by using a statistical approach. Based on average length of the short dialog acts at the beginning of a user turn, we set the border parameter $b = 3$. In other words, we define all first three words of a user turn to dialog act d_{start} and all following words to d_{rest}. Table 2 shows an example dialog with the original and the artificial generated annotation. The grey coloured text represent the dialog act d_{start} after creation of artificial dialog acts. By reconstruction of the corpus to artificial dialog acts, we always know about the following dialog act.

DA	yeah	I	okay	so	you
start (%)	10.71	4.54	4.08	3.13	2.41
rest (%)	0.62	1.77	0.34	1.20	2.11

Table 3. AMI: top five words DA start

DA	to	a	and	it	you
rest (%)	2.71	2.42	2.18	2.11	2.11
start (%)	0.60	1.16	2.06	1.79	2.41

Table 4. AMI: top five words DA rest

In table 3 and 4 we list the top five words for each artificial dialog act, compared the the occurrence in the other one. Obviously, both dialog act models contain a different word occurrence. An interesting fact is, that the word '*yeah*' has a very high confidence score compared to all other words. Further on, the word 'you' seems to have a similar occurrence probability in both models.

6. EXPERIMENTS

In the following we present our final results. In addition to the AMI corpus, we ran our DA apapted models also on the Japanese NTT [12] (6 meetings, 4 speaker) and English NIST RT-07 [13] (8 meetings, 4-6 speaker) corpora. Since these two corpora do not contain so many different sessions, we run an n-fold cross-validation evaluation and calculate the mean perplexity value. The results of our models (equation: (1), (3) and (8) - (12)) using the artificial dialog act approach are presented in table 5.

model / eq.	AMI	NTT	RT-07
CC / (1)	334.51	359.85	339.49
I / (3)	327.35	334.54	328.17
CC_1 / (8)	305.76	347.89	308.69
CC_2 / (9)	302.05	330.79	297.07
I_1 / (10)	300.91	332.54	303.37
I_2 / (11)	297.89	324.01	294.15
I_3 / (12)	**295.45**	**317.96**	**290.09**

Table 5. Dialog act extended experiments

As we can see, using the artificial dialog act extension increased the perplexity compared to the baseline models in all experiments, even if all corpora include different interlocutors, different amount of interlocutors and two different languages. The DA-extended influence approach I_3 achieved on all corpora the best results. Compared to the accuracy experiments in table 1, we achieved on the AMI corpus similar results to a dialog act estimation accuracy of about 90%.

7. CONCLUSION AND FUTURE WORK

In this work, we examined the effect of dialog act information on word usage in a polylogue and how it improves the word prediction accuracy. We defined a new and simple way to group words to artificial dialog acts and extended the cache and influence model by these DA models. The extended models were tested on different corpora, with a different amount of interlocutors and two different languages. The presented results show that our extensions outperformed their baseline approaches, which confirms the effectiveness of the proposed method.

So far, the approach works on text level. An integration into a real system (e.g., automatic speech recognition, statistical machine translation, and information retrieval) could be one of the next steps. Further on, we focussed in our work on uni-gram word models. One of the next important steps would be the extension to n-gram. In addition, since we set the artificial DA border parameter to $b = 3$, an automatic estimation would be desirable. The development of a clustering algorithm would be highly interesting, which generates a certain amount of (clustered) dialog act based on specific word frequencies and reliable estimation of the following DA.

8. REFERENCES

[1] Ani Nenkova, Agustn Gravano, and Julia Hirschberg, "High frequency word entrainment in spoken dialogue," in *In Proceedings of ACL-08: HLT. Association for Computational Linguistics*, 2008.

[2] Gang Ji and Jeff Bilmes, "Multi-speaker language modeling," in *Proceedings of HLT-NAACL 2004: Short Papers*, Stroudsburg, PA, USA, 2004, HLT-NAACL-Short '04, pp. 133–136, Association for Computational Linguistics.

[3] R. Kuhn and R. de Mori, "A cache based natural language model for speech recognition," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1999, vol. 14, pp. 570–583.

[4] Tomoharu Iwata and Shinji Watanabe, "Learning influences from word use in polylogue," in *In Proceedings of Interspeech*, 2011, pp. 3089–3092.

[5] M. Woszczyna and A. Waibel, "Inferring linguistic structure in spoken language," in *In Proceedings of the International Conference on Spoken Language Processing*, 1994, vol. 2, pp. 847–850.

[6] Masaaki Nagata and Tsuyoshi Morimoto, "First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance," in *Speech Communication*, 1994, vol. 15, pp. 193–203.

[7] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van, and Ess dykema Marie Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," 2000, vol. 26, pp. 339–373.

[8] Norbert Reithinger, Ralf Engel, and Martin Klesen, "Predicting dialogue acts for a speech-to-speech translation system," in *In Proceedings of the International Conference on Spoken Language Processing*, 1996, pp. 654–657.

[9] Jan Alexandersson and Norbert Reithinger, "Learning dialogue structures from a corpus," in *In Proceedings of EuroSpeech*, 1997, pp. 2231–2235.

[10] Jeroen Geertzen, "Dialogue act prediction using stochastic context-free grammar induction," in *Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference*, Stroudsburg, PA, USA, 2009, CLAGI '09, pp. 7–15, Association for Computational Linguistics.

[11] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," in *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*, 2005.

[12] Takaaki Hori, Shoko Araki, Takuya Yoshioka, Masakiyo Fujimoto, Shinji Watanabe, Takanobu Oba, Atsunori Ogawa, Kazuhiro Otsuka, Dan Mikami, Keisuke Kinoshita, Tomohiro Nakatani, Atsushi Nakamura, and Junji Yamato, "Real-time meeting recognition and understanding using distant microphones and omni-directional camera," in *IEEE Workshop on Spoken Language Technology (SLT)*, 2010.

[13] Jonathan G. Fiscus, Jerome Ajot, and John S. Garofolo, "Multimodal technologies for perception of humans," chapter The Rich Transcription 2007 Meeting Recognition Evaluation, pp. 373–389. 2008.