An Approach to Off-talk Detection based on Text Classification within an Automatic Spoken Dialogue System

Oleg Akhtiamov, Roman Sergienko and Wolfgang Minker
Institute of Communications Engineering, Ulm University, Albert Einstein Allee 43, Ulm, Germany

Keywords: Human-human-machine Interaction, Text Processing, Term Weighting, Feature Transformation.

Abstract: This paper describes the problem of the off-talk detection within an automatic spoken dialogue system. The considered corpus contains realistic conversations between two users and an SDS. A two- (on-talk and off-talk) and a three-class (on-talk, problem-related off-talk, and irrelevant off-talk) problem statement are investigated using a speaker-independent approach to cross-validation. A novel off-talk detection approach based on text classification is proposed. Seven different term weighting methods and two classification algorithms are considered. As a dimensionality reduction method, a feature transformation based on term belonging to classes is applied. The comparative analysis of the proposed approach and a baseline one is performed; as a result, the best combinations of the text pre-processing methods and classification algorithms are defined for both problem statements. The novel approach demonstrates significantly better classification effectiveness in comparison with the baseline for the same task.

1 INTRODUCTION

As a rule, the interaction between a human and an automatic spoken dialogue system (SDS) is considered as a human-machine type of conversation. However, this is an ideal case; real SDSs deal with a mixed type of interaction, which may also include human-human conversations, particularly in the situations of collective problem solving, when one user talks to the system directly and to other people at the same time retelling and discussing the information obtained from the system. Such a system is supposed to be selective to react properly to each utterance: to respond immediately, to analyse without a direct response (for instance, for context definition or user image creation), or to ignore. Considering this, we specify three types of users' talk: the first one is on-talk; it comprises explicit problem-oriented requests to the system. The second type is problem-related off-talk, which includes phrases implicitly addressed to the system, for example, retelling or discussing the obtained information between users. Though such utterances are not intended to be a direct system input, they could be useful for the system adaptation to users' behaviour. The third type is irrelevant off-talk; this class is useless for the particular task and should be ignored in order not to confuse the system.

The obtained classification problem can be transformed into a text categorization task directly after speech recognition. In the vector space model (Sebastiani, 2002), text categorization is considered as a machine learning problem. The complexity of the text categorization with the vector space model is compounded by the need to extract numerical data from text information before applying machine-learning methods. Therefore, text pre-processing is required and can be performed using term weighting.

At first sight, we obtain an ordinary text classification task. However, off-talk detection faces some special challenges, since off-talk is a complex phenomenon, which is difficult to handle using only lexical information. There are some works on off-talk detection considering lexical (lexical *n*-gram approach) (Shriberg et al., 2012), acoustic-prosodic (prosodic *n*-gram approach), and visual (visual focus of attention) features (Batliner et al., 2006). The Bag-of-Words approach (lexical unigram) presented there does not use any advanced term weighting techniques, which could significantly improve classification effectiveness. In this research, we use only lexical information (Bag-of-Words model with advanced term weighting) in order to conclude how representative it can be for off-talk detection.

There exist different unsupervised and

Akhtiamov, O., Sergienko, R. and Minker, W.
An Approach to Off-talk Detection based on Text Classification within an Automatic Spoken Dialogue System.
DOI: 10.5220/0005977802880293
In *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016) - Volume 2*, pages 288-293
ISBN: 978-989-758-198-4

supervised term weighting methods. The most well-known unsupervised term weighting method is TF-IDF (Salton and Buckley, 1988). The following supervised term weighting methods are also considered in the paper: Gain Ratio (Debole and Sebastiani, 2004), Confident Weights (Soucy and Mineau, 2005), Second Moment of a Term (Xu and Li, 2007), Relevance Frequency (Lan et al., 2009), Term Relevance Ratio (Ko, 2012), and Novel Term Weighting (Gasanova et al., 2014).

The considered text pre-processing techniques usually lead to high dimensionality for a text classification task. Therefore, we apply the Feature Transformation Method Based on Term Belonging to Classes for dimensionality reduction (Sergienko et al., 2016).

As machine learning algorithms, we choose two approaches, which have demonstrated good results for the task of natural language call routing (Sergienko et al., 2016): the method of *k* Nearest Neighbours (KNN) (Zhou et al., 2009) and the Support Vector Machine-based algorithm Fast Large Margin (SVM-FLM) (Fan et al., 2008). This task and the current one are quite close to each other in the view of text classification.

This paper is organized as follows. In Section 2, we describe two problem statements and a corpus. Section 3 contains the description of term weighting and the feature transformation method. In Section 4, a baseline off-talk detection approach is compared with the proposed one for the same corpus. Finally, we provide concluding remarks and directions for future investigations in Section 5.

2 CORPUS DESCRIPTION

The corpus we have used for our research was created with a real SDS within the publicly funded German Smart Web project and contains human-human-machine interactions (audio and video) in the context of a visit to the Football World Cup in 2006 (Batliner et al., 2006). The recordings took place in situations, which were as realistic as possible. No instructions regarding off-talk were given. The user was carrying a mobile phone and was interrupted by another person. This way, a large amount of off-talk could be evoked. The user was asking for transport information, a competition program, which sights were worth visiting, etc. 2218 segmented turns of 99 different German speakers were recorded. There are 2970 user utterances in total; one utterance includes all successive sentences belonging to one class within one turn.

The original corpus has the following labels: the 1st class – on-talk (a normal request), the 2nd one – reading off-talk (a user reads the system response from the display aloud), the 3rd class - paraphrasing off-talk (a user retells others the information from the display), the 4th one – spontaneous off-talk (other off-talk, for instance, thinking aloud or interruptions).

We observe two different problem statements: the first one contains three classes (classes 2 and 3 are merged into one - problem-related off-talk), the second statement is a simplified version of the first one and contains two classes (classes 2-4 are merged into one – off-talk). We perform a speaker-independent validation, since it allows us to obtain more representative results, especially when the real system has a wide user audience.

We have split the corpus into 15 random subsets for the cross-validation. After that, we establish a training and a test set for each subset; different test sets have no intersection. For each training set, we have designed a vocabulary of unique words, which appear in the set. The size of the vocabulary varies from 1,355 to 1,421 words for different subsets.

3 TEXT PRE-PROCESSING

After the generation of training and test samples, we performed term weighting. As a rule, term weighting is a multiplication of two parts: the part based on term frequency in a document (TF) and the part based on term frequency in the whole database. The TF-part is fixed for all considered term weighting methods and calculated in the following way:

$$TF_{ij} = \log(tf_{ij} + 1);\ tf_{ij} = \frac{n_{ij}}{N_j},$$

where n_{ij} is the number of times the i^{th} word occurs in the j^{th} document, N_j is the document size (number of words in the document).

The second part of term weighting is calculated once for each word from the vocabulary and does not depend on an utterance for classification. We consider seven different methods for the calculation of the second part of term weighting.

3.1 Inverse Document Frequency (IDF)

IDF is a well-known unsupervised term weighting method, which was proposed in (Salton and Buckley, 1988). There are some modifications of IDF, and we use the most popular one:

$$idf_i = \log \frac{|D|}{n_i},$$

where $|D|$ is the number of documents in the training set, and n_i is the number of documents that have the i^{th} word.

3.2 Gain Ratio (GR)

Gain Ratio (GR) is mainly used in term selection (Yang and Pedersen, 1997). However, it was shown in (Debole and Sebastiani, 2004) that it could also be used for weighting terms, since its value reflects the importance of a term. The definition of GR is as follows:

$$GR(t_i, c_j) = \frac{\sum_{c \in \{c_j, \overline{c_j}\}} \sum_{t \in \{t_j, \overline{t_j}\}} P(t,c) \cdot \log \frac{P(t,c)}{P(t) \cdot P(c)}}{-\sum_{c \in \{c_j, \overline{c_j}\}} P(c) \cdot \log P(c)},$$

where *P(t, c)* is the probability estimation that a document contains the term t and belongs to the category c; *P(t)* is the probability estimation that a document contains the term t, and *P(c)* is the probability estimation that a document belongs to the category c.

Then, the weight of the term t_i is calculated as the max value between all categories:

$$GR(t_i) = \max_{c_j \in C} GR(t_i, c_j),$$

where C is a set of all classes.

3.3 Confident Weights (CW)

The method uses the special value *Maxstr* as an analogy of IDF.

First of all, the method estimates the probability P that a document contains the term t with the confidence interval for every category c_j, to get $P(t \mid c_j)$ and $P(t \mid \overline{c_j})$ with a confidence interval. Let M denote the lower bound of $P(t \mid c_j)$ and N denote the upper bound of $P(t \mid \overline{c_j})$. The strength of the term t_i considering c_j is defined as follows:

$$str(t,c) = \begin{cases} \log_2\left(\frac{2M}{M+N}\right), if\left(M > N\right) \\ 0, otherwise \end{cases}.$$

The maximum strength (*Maxstr*) of the term t is calculated in the following way:

$$Maxstr(t) = \max_{c \in C} \left(str(t,c)\right)^2.$$

where C is a set of all classes.

3.4 Second Moment of a Term (TM2)

Let $P(c_j \mid t)$ be the probability estimation that a document belongs to the category c_j with the condition that the document contains the term t and belongs to the category c; $P(c_j)$ is the probability estimation that a document belongs to the category c without any conditions. The idea is as follows: the more $P(c_j \mid t)$ is different from $P(c_j)$, the more important the term t_i is. Therefore, we can calculate the term weight in the following way:

$$TM(t_i) = \sum_{j=1}^{|C|} \left(P(c_j \mid t) - P(c_j)\right)^2.$$

3.5 Relevance Frequency (RF)

The RF value is calculated as follows:

$$rf(t_i, c_j) = \log_2\left(2 + \frac{a_j}{\max\{1, \overline{a_j}\}}\right),$$

$$rf(t_i) = \max_{c_j \in C} rf(t_i, c_j),$$

where a_j is the number of documents of the category c_j which contain the term t_i, and $\overline{a_j}$ is the number of documents of all the remaining categories which also contain this term.

3.6 Term Relevance Ratio (TRR)

The TRR method uses *tf* weights and is calculated as follows:

$$TRR(t_i, c_j) = \log_2\left(2 + \frac{P(t_i \mid c_j)}{P(t_i \mid \overline{c_j})}\right),$$

$$P(t_i \mid c) = \frac{\sum_{k=1}^{|T_c|} tf_{ik}}{\sum_{l=1}^{|V|} \sum_{k=1}^{|T_c|} tf_{lk}},$$

$$TRR(t_i) = \max_{c_j \in C} TRR(t_i, c_j),$$

where c_j is the class of a document, $\overline{c_j}$ is all the other classes of c_j, V is the vocabulary of the training data, and T_c is the document set of the class c.

3.7 Novel Term Weighting (NTW)

This method was proposed in (Gasanova et al., 2014).

Let L be the number of classes; n_i is the number of documents which belong to the i^{th} class; N_{ij} is the number of occurrences of the j^{th} word in all articles from the i^{th} class. $T_{ij} = N_{ij} / n_i$ is the relative frequency of occurrences of the j^{th} word in the i^{th} class; $R_j = \max_i T_{ij}$; $S_j = \arg(\max_i T_{ij})$ is the class which we assign to the j^{th} word. The term relevance C_j is calculated in the following way:

$$C_j = \frac{1}{\sum_{i=1}^{L} T_{ij}} \left(R_j - \frac{1}{L-1} \sum_{\substack{i=1 \\ i \neq S_j}}^{L} T_{ij}\right).$$

3.8 Feature Transformation Method

We propose a feature transformation method based on term belonging to classes (Sergienko et al., 2016). The idea is to assign each term from the vocabulary to the most appropriate class. Such an assignment is performed during the calculation of GR, CW, RF, TRR and NTW. With TF-IDF and TM2, we can also assign one class for each term using the relative frequency of the word in classes:

$$S_j = \arg\max_{c \in C} \frac{n_{jc}}{N_c},$$

where S_j is the most appropriate class for the j^{th} term, c is the index of a class, C is a set of all classes, n_{jc} is the number of documents of the c^{th} class which contain the j^{th} term, N_c is the number of all documents of the c^{th} class.

After assigning each word to one class and term weighting, we can calculate the sums of term weights in a document for each class. We can put these sums as new features of the text classification problem. Therefore, such a method reduces the dimensionality significantly; the dimensionality of the classification problem equals the number of classes.

4 OFF-TALK DETECTION APPROACHES

4.1 Baseline Off-talk Detection Approach

The authors of the corpus performed their own research on off-talk detection (Batliner et al., 2006). They processed the audio data using an automatic speech recogniser with manual proofreading and extracted two groups of features: the first one is prosodic features, which evaluate 95 different speech characteristics such as speech rate, pause duration, accents and many others. A detailed overview of prosodic features is given in (Batliner et al., 2003). The second group is part-of-speech features. There are 5 different part-of-speech classes considered (nouns, adjectives, verbs, etc.); one of them is assigned to each word from the vocabulary (Batliner et al., 1999). After that, an n-gram approach is implemented for both groups of features with n=5.

As a classification method, the authors used Linear Discriminant classification (LDA) and estimated its performance as unweighted mean recall (R). They obtained the same two- and three-class problem statement and tested them using the speaker-independent validation (Batliner et al., 2006).

The highest classification performance is reached with all available features: R equals 0.681 and 0.600 for the two- and three-class task respectively.

4.2 Proposed Off-talk Detection Approach

We implement the proposed approach based on text classification using SVM-FLM (Fan et al., 2008) and KNN (Zhou et al., 2009). These methods show successful results for text classification tasks (Sergienko et al., 2016), are able to solve high-dimensionality problems, and possess moderate resource consumption. There are effective implementations of these algorithms built in the *RapidMiner* free software package (Shafait et al., 2010), which we use to solve our classification tasks.

As the main criterion of classification effectiveness, we have to use the same estimation based on unweighted mean recall (R) as the authors did in their research, since it allows us to compare the final results correctly. However, as the main criterion of classification effectiveness during the process of the parametric optimization for the parameter k in KNN, we use the macro F-score (Baeza-Yates and Ribeiro-Neto, 1999). It is a more representative performance estimation than unweighted mean recall; in some cases, the recall value can be equal to 1, but in fact, the classifier effectiveness is not equal to 100%, since the precision value remains low. The macro F-score does not possess this disadvantage; it is calculated as the geometric mean of mean precision and mean recall:

$$P_i = \frac{|Dr_i \cap Df_i|}{|Df_i|}, \; R_i = \frac{|Dr_i \cap Df_i|}{|Dr_i|},$$

$$F(a) = \frac{1}{a\frac{1}{P} + (1-a)\frac{1}{R}}, \qquad a \in [0,1],$$

where i is the number of a class, Dr_i is the set of objects in a test set, which belong to this class, Df_i is the set of objects in the test set classified by the system to this class. We calculate the macro F_1 score assuming a=0.5.

The optimal k value for KNN is specified from the interval [1, 30] for each training set. It is natural that we do not use any information from the test sets during the process of parametric optimization in order to keep the results as representative as possible. Searching for k, we split each training set into a new training and a validating set in order to perform exhaustive search among all possible k using this pair of sets.

We tested all possible combinations of the term weighting methods (with and without the feature transformation method) and the machine learning algorithms for both classification tasks:

Table 1: Results of the classification algorithms with different term weighting methods tested for the *two-class* task with the *speaker-independent* validation.

KNN		SVM-FLM	
TM2+FT	**0.907**	NTW	**0.910**
RF	0.896	RF	**0.907**
TRR	0.892	IDF	**0.905**
NTW+FT	0.892	TM2+FT	0.903
CW	0.887	TRR	0.901
IDF	0.881	NTW+FT	0.891
CW+FT	0.880	RF+FT	0.880
TRR+FT	0.879	CW+FT	0.876
RF+FT	0.879	TM2	0.871
NTW	0.879	TRR+FT	0.867
TM2	0.879	IDF+FT	0.859
GR	0.877	CW	0.836
IDF+FT	0.857	GR+FT	0.731
GR+FT	0.772	GR	0.638

Table 2: Results of the classification algorithms with different term weighting methods tested for the *three-class* task with the *speaker-independent* validation.

KNN		SVM-FLM	
NTW	**0.869**	RF	**0.893**
NTW+FT	**0.866**	NTW	**0.887**
TRR	**0.864**	TRR	0.887
RF+FT	**0.863**	IDF	0.886
CW	**0.863**	TM2+FT	0.863
TM2+FT	**0.863**	TRR+FT	0.850
CW+FT	**0.862**	NTW+FT	0.849
RF	**0.862**	CW	0.849
TRR+FT	**0.857**	CW+FT	0.848
TM2	0.856	RF+FT	0.844
IDF	0.846	TM2	0.837
IDF+FT	0.843	IDF+FT	0.827
GR	0.843	GR	0.739
GR+FT	0.606	GR+FT	0.387

The tables contain unweighted mean recall values. The mark '+FT' means that the feature transformation method is implemented. The best term weighting methods are emphasized and have no statistically significant difference between each other within one column. The variable distributions are close to normal that allows us to use t-test. The given results are relevant with the confidential probability 0.95.

According to t-test, SVM-FLM works significantly better than KNN for the task with three classes. For the two-class task, there is no significant difference between the classification algorithms.

5 CONCLUSIONS

The off-talk detection task has been solved with the proposed approach based on text classification, which shows effective results for both problem statements. There is a set of the best term weighting methods with no significant difference in their effectiveness within each problem statement. The highest unweighted mean recall for the two-class task equals 0.910 and is reached with NTW and SVM-FLM, for the three-class task - 0.893 (with RF and SVM-FLM).

The proposed approach outperforms the baseline one: R equals 0.681 and 0.600 at the two- and three-class task respectively for the baseline approach, while the proposed approach demonstrates the values 0.910 and 0.893 at the same tasks. A possible reason of this is the assumption that the proposed approach uses more effective data pre-processing methods and machine learning algorithms for off-talk detection. Another possible reason is a conceptual contradiction in the baseline approach; the fact that it uses prosodic features for off-talk detection means that users are supposed to change their normal manner of speech once they start talking to a computer. Such an approach can work now, since modern dialogue systems are still far from perfection, and users have to adapt their behaviour talking to them. However, it does not correspond to the main direction of automatic dialogue system development – to make the interaction between a user and a system as natural as possible.

Any additional data processing (speech recognition, text pre-processing, etc.) causes an information loss. Deep learning neural networks possess some features, which could improve classification effectiveness: due to their ability to work with entities of different abstraction levels, they do not require additional data processing and are able to make the system more effective and flexible. Moreover, the works on off-talk detection (Shriberg et al., 2012) and (Batliner et al., 2006) state that using more than one group of features significantly improves classification effectiveness. The choice of relevant features can also be delegated to a system based on deep learning neural networks. Therefore, as a future direction, we propose the research of a deep learning neural network-based approach to off-talk detection.

REFERENCES

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1-47.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513-523.

Debole, F. and Sebastiani, F. 2004. Supervised term weighting for automated text categorization. *Text mining and its applications*:81-97. Springer Berlin Heidelberg.

Soucy P. and Mineau G. W. 2005. Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*:1130-1135.

Xu, H. and Li, C. 2007. A Novel term weighting scheme for automated text Categorization. *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*:759-764. IEEE.

Lan, M., Tan, C. L., Su, J., and Lu, Y. 2009. Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):721-735.

Ko, Y. 2012. A study of term weighting schemes using class information for text classification. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*:1029-1030. ACM.

Gasanova, T., Sergienko, R., Akhmedova, S., Semenkin, E., and Minker, W. 2014. Opinion Mining and Topic Categorization with Novel Term Weighting. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Baltimore, Maryland, USA*, 84–89.

Fan, R. E., Chang, K. W., Hsieh C. J., Wang X. R., Lin C. J. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9, 1871–1874.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. *ICML*, vol. 9:412-420.

Batliner, A., Hacker, C., and Noth, E. 2006. To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk. *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, 79-100.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E. 2003. How to Find Trouble in Communication. *Speech Communication*, 40, 117–143.

Batliner, A., Nutt, M., Warnke, V., Noth, E., Buckow, J., Huber, R., Niemann, H. 1999. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. *Proc. of Eurospeech99*, 519–522.

Zhou, Y., Li, Y., and Xia, S. 2009. An improved KNN text classification algorithm based on clustering. *Journal of computers*, 4(3), 230-237.

Sergienko, R., Muhammad, S., and Minker, W. 2016. A comparative study of text preprocessing approaches for topic detection of user utterances. *In Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Baeza-Yates, R; Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. New York, NY: ACM Press, Addison-Wesley, 75.

Shriberg, E., Stolcke, A., Hakkani-Tur, D., Heck, L. 2012. Learning When to Listen: Detecting System-Addressed Speech in Human-Human-Computer Dialog. *Proceedings of Interspeech 2012*, 334-337.

Shafait, F., Reif, M., Kofler, C., and Breuel, T. M. 2010. Pattern recognition engineering. In: *RapidMiner Community Meeting and Conference*, Citeseer, vol 9.