INTENT FOCUSED SUMMARIZATION OF CALLER-AGENT CONVERSATIONS

Shajith Ikbal[1] *Ashish Verma*[1] *Prasanta Ghosh*[2] *Kenneth Church*[3] *Jeffrey Marcus*[4]

[1] IBM Research, India. [2] Indian Institute of Science (IISc), Bangalore, India.
[3] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA.
[4] Nuance Communications Inc., Burlington, MA, USA.

{shajmoha, vashish}@in.ibm.com, prasantg@ee.iisc.ernet.in, kwchurch@us.ibm.com, jeffrey.marcus@nuance.com

ABSTRACT

In this paper, we propose a conditional random field (CRF) based approach to identify segments within call center conversations that convey caller intent. A distinguishing aspect of our approach is the use of context information of the intent bearing segments to predict the presence or absence of intents within various segments. The context is represented through a set of phrase features that are frequently present in and around the intent bearing segments. These phrases, identified in a data-driven manner, are used along with conventional word features in a CRF based sequence labeling framework to assign intent/non-intent labels to each utterance in a conversation. Another distinguishing aspect of our approach is that instead of using 1-best label alignment, we extract N-best label alignments at the output of CRF and combine evidences from them to rank the utterances according to their intent bearing potential, so that top ranked utterances can be chosen as the intent summary. To demonstrate the effectiveness of our approach and to evaluate the influence of automatic speech recognition (ASR) errors we evaluated our approach using manually transcribed and ASR transcribed conversations. Experimental results show improved summarization accuracy using our approach. Specifically, in 92% of the manually transcribed conversations accurate summaries of just one utterance length can be extracted using the proposed approach.

Index Terms— caller intent, intent focused summarization, phrase features, conditional random field, n-best alignment

1. INTRODUCTION

Conversations involving call center agents of enterprises and their customers contain information about various issues faced by the customers and hence can be used to gain insights about how to improve the business. Advancements in conversational speech recognition systems and text mining techniques lead to several approaches for automatically extracting such business insights through analysis of a large collection of caller-agent conversations[1, 2]. One crucial information that enterprises would like to extract from these conversations is the list of caller intents, i.e., various reasons why the customers are calling into the call centers such as problems faced, information needed, and so on. This information could potentially be used for various tasks including: optimization of call flow dialog for interactive voice response (IVR) systems, identification of a set of product or service related issues faced by the customers, identification of leads for new products or features, improvement of operational efficiency, and decision about the best sequence of actions to be performed by the agents during a conversations [3].

Our aim in this paper is to extract segments within the conversations where the caller intents are expressed, i.e., *intent focused summarization*. Figure 1 illustrates a typical conversation, where after exchanging greetings, the customer conveys to the agent the reason why he/she called (i.e., the caller intent) which the agent first acknowledges and further tries to resolve during the remaining part of the conversation. In this conversation, the caller intent is conveyed in the highlighted utterances, where the customer's service request of *password reset for voicemail access* is expressed. The aim of intent focused summarization is to identify these highlighted utterances automatically. Caller intent is typically conveyed within small segments of the conversations. Hence, extracting intent summaries could potentially serve as a useful preprocessing step for further processing of the conversations to extract the list of caller intents. In such a scenario, summarization would play the role of a feature selection module to choose essential intent conveying part useful for further processing, discarding the irrelevant parts. A similar approach of using shorter segments of the conversations has been shown to be beneficial in a call classification task [4] where simply the initial part of the conversation is used. However, it still used large initial parts of the conversations assuming the relevant part is captured there.

Although utterances conveying caller intent typically occur in the initial part of the call, their exact location may vary due to reasons such as variability in the length of the greetings part of the call, authentication of the customer at the start of the call and so on. In addition, the intent is typically expressed within small segments. These factors make the accurate extraction of short intent summaries a challenging task. In this paper we propose a conditional random field (CRF) based approach that uses the knowledge of context of the intent bearing parts of the conversations to accurately locate the intent bearing segments. A set of phrase features representing the context are extracted and used to classify each utterance in the call as to whether it is intent bearing or not. In addition, to improve the robustness of such classification, an evidence combination using N-best classification outputs of CRF is performed. As will be shown later in the paper, in the absence of speech recognition errors, this method is able to extract accurate intent summaries of just one utterance length in 92% of the conversations.

The intent focused summarization approach discussed in this paper falls under the category of extractive targeted summarization techniques discussed in the past literature, since the aim is to extract segments that would summarize a specific aspect of the conversation, not the entire call. Although a majority of the prior work on summarization have focused on written text [5, 6, 7], recently there is a growing interest in the spoken content summarization. However, majority of the work on spoken content summarization focus mainly on summarization of the entire call such as broadcast news summarization [8, 9], voicemail summarization [10] and meeting sum-

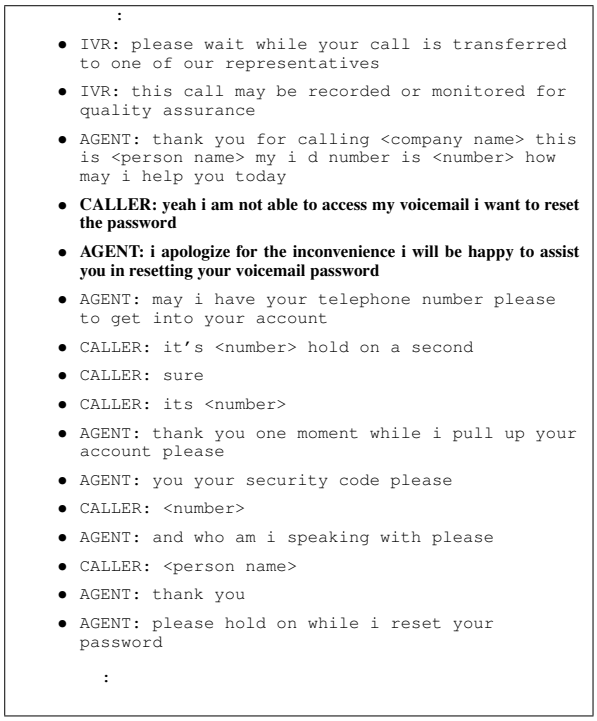

Fig. 1. Transcript of a typical caller-agent conversation.

marization [11, 12]. Query-focused summarization [13, 14] is one specific case of the targeted summarization where the aim is to summarize multiple documents relevant to a particular query. Various approaches in the literature for extractive summarization first use a metric to rank various parts of the calls/documents according to their relevance and then extract top ranking parts as the summary. Some of the metrics used include, maximum marginal relevance [7], term weighting measure from information retrieval literature [15], and semantic similarity measure [16].

The organization of the rest of this paper: Section 2 describes the extraction of context representing phrase features. Section 3 describes a CRF based method to extract intent summaries using those phrase features and a combination of evidences from N-best alignments out of CRF to rank the utterances. Section 4 describes the experimental setup for evaluation. Section 5 presents and discusses the results. Section 6 concludes and discusses future directions.

2. PHRASE FEATURES REPRESENTING INTENT CONTEXT

As can be seen from the example call given in section 1, the intent bearing utterances are typically surrounded by their context denoting key words. For example, the agents typically say *how may I help you?* or *how may I assist you today?* before the intent bearing utterance. Similarly after the intent bearing utterance agents typically say phrases like *apologize for the inconvenience* or *happy to assist you*. Also within the intent bearing utterance itself the callers typically mention key phrases such as *not able to*, *i am faced with*, *i have problem in*, etc. Hence it should be possible to use these key words, or in general key phrases, to identify the context of the intent bearing utterances, in order to further identify the intent bearing utterances themselves. However, the question is how do we find a set of all these phrases automatically instead of listing them manually. Interestingly, phrases that occur consistently in the context of the intent would emerge as top frequent phrases in and around true intent utterances in a large collection of conversations. Hence we can identify them automatically from training data containing conversations marked with intent utterances in a data-driven manner.

The method we use to find these phrases first identifies a larger list of candidate phrases and scores each of them based on frequency of occurrences at different locations within the conversation, namely within the intent utterances, in neighboring utterances before and after the intent utterances and elsewhere in the conversation. Based on the scores a subset that distinctively represent the context of the intent bearing segment is chosen. For this, we divide the training data into 3 different sets based on the location of intent marked utterances in each conversation:

- *Set-int*: The set of all intent bearing utterances from the whole training set.
- *Set-pre*: The set of utterances preceding all the intent bearing utterances in the training set. For each conversation if i denote the index of the intent bearing utterance then utterances with indices from $i-1$ to $i-N$ are included in this set. N denotes the context length.
- *Set-fol*: The set of utterances following all the intent bearing utterances in the training set, i.e., utterances with indices from $i+1$ to $i+N$ are included in this set.

From each of these data sets top frequent phrases of various lengths L occurring within window lengths W_L are identified. These frequent phrases are collected together as the larger list of candidate phrases to score them using their frequency counts within different data sets. The following subsections describe these steps in detail.

2.1. Identification of Frequent Phrases

We use a frequent phrase identification algorithm to extract top frequent phrases within the utterances of each data set *set-int*, *set-pre* and *set-fol*, described above. The phrases we target to extract are not only the simple n-grams but many variations of them, specifically, phrases with gaps and phrases with same set of words but in different order. This is expected to help in improving the recall of these phrases making them robust to the variability of their expressions in the actual call transcripts. For example, different variations of the phrase *bill pay* that it can capture include: *pay bill*, *pay my bill*, *pay the bill*, *pay this month bill* and *bill pay*.

The method we used to extract the frequent phrases is motivated from the a-priori algorithm [17] which is well known for market-basket analysis. It starts with extraction of frequent unigrams and then goes on to discover word groups, i.e., phrases of increasing length such that all the words in a group occur within a pre-specified window length. As like a-priori algorithm, our method assumes that a higher length word group will only be frequent if all its constituent word sub-groups are also frequent. This way, it can discover frequent phrases of interest quite efficiently. In the work for this paper, we considered phrases of lengths $L = 1$, 2 and 3, for which respective window lengths of $W_L = 1$, 3 and 7 are used.

2.2. Selection of Phrases Representing Intent Context

A larger list of all the candidate phrases extracted, as explained above, is used to further choose a subset based on the potential of each phrase to represent the intent context. Let f^p_{int}, f^p_{pre}, and f^p_{fol} respectively denote the frequencies of a phrase p within the data sets

set-int, *set-pre* and *set-fol*, and let f^p_{all} denote its frequency in the entire training set. Then,

- phrase p identified from *set-int* is chosen as an intent context representing phrase if $\frac{f^p_{int}}{f^p_{all}} \geq T$, where T denote frequency ratio threshold.
- phrase p identified from either *set-pre* or *set-fol* is chosen if $\left|\frac{f^p_{pre}-f^p_{fol}}{f^p_{all}}\right| \geq T$.

Table 1 shows some of the top scored phrases representing the intent context, found using data from telecom domain. Phrases such as *my internet service*, *send text*, *wondering you* are typically uttered by the caller while conveying the intent. Phrases such as *apologize any inconvenience*, *glad assist* are uttered by the agent after hearing the caller intent, while *great day*, *problem sir* are typically uttered before hearing the caller intent. The subset of phrases identified in this manner are further used in the CRF based approach to extract intent summaries as explained in the next section.

Table 1. Some of the top intent context representing phrases found.

great day, glad assist, verification, apologize any inconvenience, happy assist you, problem sir, my internet service, send text, my voicemail get, my number transfer, paid phone, phone bill, wondering you, bill my, inconvenience, blocked, pay dollars

3. UTTERANCE CLASSIFICATION AND RANKING USING CRF

Intent summaries of length M are extracted as the top-M ranked utterances from the conversations after ranking of the utterances according to their intent bearing potential. The intent bearing potential of the utterances is measured through a classifier trained to discriminate the intent utterances from the rest. We use conditional random field (CRF) for such classification of utterances. CRFs are probabilistic discriminative models for the task of sequence labeling [18]. They have been shown to achieve state-of-the-art performance in various natural language processing tasks such as segmentation, parts-of-speech tagging, named entity recognition and information extraction.

The CRF treats the problem of identifying intent bearing part of the conversation as a sequence labeling problem, trying to assign a sequence of intent flags to the utterances based on the sequence of input features derived from those utterances. As described in the previous section, the presence or absence of a context denoting phrase feature in the neighborhood of an utterance is an important cue for deciding its intent bearing potential. Hence we use the set of context denoting phrase features as extracted in the previous section in addition to the conventional word features during the intent labeling process using CRF. The process of constructing input feature vectors for CRF using these phrase features is described in the next subsection.

3.1. Features for Classification

For each utterance i in the conversation, a basic feature vector F_i is constructed first using the conventional set of word features present in the utterance appended with two additional features that may be useful for the current task namely: speaker of the utterance (agent/caller) and the position of the utterance in terms of its index within the conversation. Then F_i is further concatenated with a set of indicator features denoting the presence or absence of the context denoting phrase features in the neighborhood of the current utterance. Note that the true intent utterance is expected to be surrounded by some of the context denoting phrase features. Hence, some of the indicator features are expected to be switched on for potential intent bearing utterances. The initial feature vector F_i is updated using the set of phrase features by repeating the following steps for each phrase p in the set of phrases as identified in Section 2.2:

- $F_i = \{F_i, p^{flag}_{cur}\}$, where p^{flag}_{cur} is an indicator to denote the presence or absence of the phrase p in the current utterance. Subscript cur is to denote that this feature is related to the current utterance. Superscript $flag \in \{0, 1\}$ denotes the presence or absence of the phrase.
- $F_i = \{F_i, p^{flag}_{pre}\}$, where p^{flag}_{pre} is an indicator to denote the presence or absence of p in the utterances preceding i, i.e., from $i-1$ to $i-N$, as specified by the subscript pre.
- $F_i = \{F_i, p^{flag}_{fol}\}$, where p^{flag}_{fol} is an indicator to denote the presence or absence of p in the utterances following i, i.e., from $i+1$ to $i+N$, as specified by the subscript fol.

In this paper, we have used $N = 1$, chosen experimentally.

3.2. Utterance Ranking and Intent Summary Extraction

The feature vectors as computed above from the utterances of conversations in the training set are used along with the corresponding intent labels to train the CRF models. The trained models are then used to assign intent flags to the utterances of the test conversation based on feature vectors extracted from those utterances. The intent flags assigned to the utterances are based on the best alignment obtained during Viterbi decoding using the CRF models for the feature vectors derived from those utterances. In our approach, instead of using only the best alignment, we generate multiple alignments using CRF, i.e., N-best alignments. The intent bearing potential of each utterance is scored based on the count of intent flags assigned to that utterance in all the alignments. Let L^j_i, represent the intent flag assigned to i^{th} utterance in the j^{th} alignment, $L^j_i \in \{0, 1\}$ where 1 means intent bearing and 0 means otherwise. The rank of an utterance is computed using scores:

$$N_i = \sum_j L^j_i \tag{1}$$

Higher scoring utterances get higher ranks. In cases where more than one utterances get the same score, utterance occurring earlier in the call is given a higher rank. This is based on our prior knowledge that the intent bearing utterances typically occur early in the call, as discussed earlier in section 1. After the ranking of the utterances, we choose top-M ranked utterances as the M-length intent summary of the conversation.

4. EXPERIMENTAL SETUP

The data set, baseline and metrics used to evaluate the proposed approach for intent summarization are explained in this section.

4.1. Data

We used two datasets consisting of real-life call center conversations to evaluate the proposed approach for intent discovery: 1) 581 manually transcribed conversations (*data-man*) to estimate the performance in the absence of ASR errors, and 2) 3676 automatically transcribed conversations (*data-asr*). Average length of each conversation is between 4-5 minutes. All the conversations are manually marked with the location of intent conveying utterances within them. In case of *data-man*, 300 calls are used for training and the remaining 281 for testing. In case of *data-asr*, to evaluate the influence of the amount of training data used for training CRF, we repeated the

experiments for two cases: 1) using 300 calls for training and 3376 for testing and 2) using 1000 calls for training and 2676 for testing. As explained in section 1, the number of intent bearing utterances could vary from call to call. Hence the analysts were instructed to mark all of intent bearing utterances they find within the call, not limiting to any particular number. The manual summary size varies from 1 to 10 with 85% calls containing summary size ≤ 3.

4.2. Baseline

Baseline for the proposed intent summarization approach is a simple CRF based approach that uses only the basic features, as explained in section 3.1, namely the word features, speaker id and the position of the utterance within the conversation. This baseline is to examine the effectiveness of the proposed context denoting phrase features and the usefulness of the additional evidence gained from multiple alignment outputs of the CRF.

In addition, since the intent is typically conveyed in the early part of the call we found that simply choosing the initial $x\%$ of the call from the first caller utterance after both agent and caller come into the conversation is also an effective and competitive baseline for this particular dataset. Before agent comes into the conversation the customer might be interacting with automated system and after agent comes in the first utterance from the customer is likely to express the intent, as illustrated in Figure 1. Hence we assign top ranks to the initial $x\%$ utterances of the call starting from first caller utterance after agent comes into the conversation. We call this an *initial-N* approach.

4.3. Evaluation Metric

For evaluation, we use mean reciprocal rank (MRR) [19] computed as an average of the inverse rank of the first ground truth intent bearing utterance appearing in the ranked list generated by intent summarization approach. This metric measures the ability of the algorithm to assign high ranks to the true intent bearing utterances. Higher MRR value denote higher accuracy.

5. EXPERIMENTAL RESULTS AND DISCUSSION

Table 2 shows results of experimental evaluation of the proposed intent summarization approach using data sets *data-man* and *data-asr*. In case of *data-asr*, the experiments were repeated for two different sizes of training set, one with 300 calls and the other with 1000 calls. As can be seen from the table, the use of ASR transcripts leads to drop in MRR values indicating a relatively poor summarization accuracy as a result of the ASR errors. In case of *data-asr*, use of bigger training set for the CRF training resulted in improved MRR values. The use of *phrase features* and *N-best alignments* result in improved MRR values over the baseline both independently as well as in combination, in most of the cases. The use of phrase features result in milder improvement in comparison to the improvements observed using N-best alignments. In case of *data-man* the simultaneous use of both the phrase features and the N-best alignments result in a significant improvement. This achieves the best MRR value of 0.93 which means the proposed method is able to assign highest rank to the correct intent utterances most of the time. In fact, in a separate analysis we found that, in 92% of the conversations the proposed approach is able to assign first rank to one of the true intent bearing utterances. This effectively means that the proposed method is able to extract summaries of just one utterance length accurately in 92% of the conversations. On the other hand, in case of *data-asr* the use of just N-best alignments from CRF seems to result in larger improvements. Erroneous ASR transcripts seems to result in extraction of a few noisy phrase features.

Another observation from the table is that in most of cases the proposed method is able to achieve better accuracy than the relatively simple but effective method of *initial-N*, explained in Section 4.2. Note that the rule of choosing initial $x\%$ of utterances in the conversation would be effective only in data similar to the data used in this paper, where intent is typically conveyed early in the call. On the other hand, the proposed method do not have this constraint. Improvements over *initial-N* method mean the proposed method is able to accurately locate the intent conveying utterances occurring in the later parts of the call too. Note that the MRR value for *initial-N* is lower for *data-asr* than *data-man*. This is partly due to the inaccurate segmentation during ASR transcription and due to the difficulty faced during manual marking of the ground truth intent utterances due to the ASR errors. Both these factors resulted in marking of the ground truth intent utterances later in the call.

Table 2. Comparison of MRR (mean reciprocal rank) achieved using various approaches for intent focused summarization

Method	MRR
Dataset: *data-man*, train size=300	
1) CRF: word+speaker+position features (baseline)	0.88
2) CRF: baseline+N-best	0.89
3) CRF: baseline+phrase features	0.88
4) CRF: baseline+phrase features+N-best	**0.93**
5) Choosing initial-N (another baseline)	0.88
Dataset: *data-asr*, train size=300	
1) CRF: word+speaker+position features (baseline)	0.56
2) CRF: baseline+N-best	**0.63**
3) CRF: baseline+phrase features	0.57
4) CRF: baseline+phrase features+N-best	**0.63**
5) Choosing initial-N (another baseline)	0.50
Dataset: *data-asr*, train size=1000	
1) CRF: word+speaker+position features (baseline)	0.57
2) CRF: baseline+N-best	**0.70**
3) CRF: baseline+phrase features	0.59
4) CRF: baseline+phrase features+N-best	0.67
5) Choosing initial-N (another baseline)	0.50

6. CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the problem of intent focused summarization to extract segments within call center conversations that convey the caller intents. We described a CRF based method that uses intent context denoting phrases as features to label the utterances of the conversation with intent/non-intent labels. Instead of using just the 1-best label alignment at the output of CRF we extracted N-best alignments and combined evidences in them to rank the utterances according to their intent bearing potential. In the absence of ASR errors, the proposed approach is able to achieve accurate ranking of the utterances. In 92% of the manually transcribed conversations, our method is able to assign highest rank to one of the ground truth summary utterance, thus demonstrating its ability to extract accurate summaries of just one utterance length. Errors in transcripts as a result of ASR degrade the accuracy of summarization. In case of ASR transcribed conversations, the use of N-best alignment outputs from CRF leads to best improvement in summarization accuracy over the baseline, although the use of phrase features also result in milder improvements. One potential future direction of work is to improve the robustness of the context denoting phrases features in the presence of ASR errors. In addition, we are also planning to use the intent summaries instead of the entire conversation to extract the actual caller intents using methods such as clustering.

7. REFERENCES

[1] L. V. Subramaniam, T. A. Faruquie, S. Ikbal, S. Godbole, and M. K. Mohania, "Business intelligence from voice of customer," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE'09)*, 2009.

[2] H. Takeuchi, L. V. Subramaniam, T. Nasukawa, and S. Roy, "Getting insights from the voice of customers: Conversation mining at a contact center," *ACM Trans. on Information Sciences*, vol. 179, no. 11, pp. 1584–1591, May 2009.

[3] U. Nambiar, T. Faruquie, L. Subramaniam, S. Negi, and G. Ramakrishnan, "Discovering customer intent in real-time for streamlining service desk conversations," in *Proceedings of CIKM*, 2011, pp. 2349–2352.

[4] Y. Park, "Automatic call section segmentation for contact-center calls," in *Proceedings of Conference on Information and Knowledge Management*, 2007, pp. 117–126.

[5] I. Mani and T. Maybury (Editors), *Advances in Automatic Text Summarization*, MIT Press, 1999.

[6] J. Kupiec, J. Pedersoen, and F. Chen, "A trainable document summarizer," in *Proceedings of ACM-SIGIR*, 1995, pp. 68–73.

[7] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of 21st Annual International ACM SIGIR Conference*, 1998.

[8] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic summarization of english broadcast news speech," in *Proceedings of HLT*, 2002.

[9] Sameer Raj Maskey and Julia Hirschberg, "Automatic summarization of broadcast news using structural features," in *Proceedings of EUROSPEECH*, 2003.

[10] K. Koumpis and S. Renals, "Transcription and summarization of voicemail speech," in *Proceedings of ICSLP*, 2000.

[11] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. of Interspeech'05*, 2005.

[12] J. J. Zhang and P. Fung, "Automatic parliamentary meeting minute generation using rhetorical structure modeling," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 9, pp. 2492–2504, 2012.

[13] S. Fisher, B. Roark, J. Yank, and B. Hersh, "Ogi/ohsu baseline query-directed multi-document summarization system for duc-2005," in *Proceedings of the Document Understanding Workshop (DUC)*, 2005.

[14] W. Li, B. Li, and M. Wu, "Query focus guided sentence selection strategy for duc," in *Proc. of the Document Understanding Workshop at the HLT/NAACL Annual Meeting*, Brooklyn, New York, 2006.

[15] G. Murray and S. Renals, "Extractive summarization of meeting recordings," in *Proc. of MLMI*, 2007.

[16] Iryna Gurevych and Michael Strube, "Semantic similarity applied to spoken dialogue summarization," in *Proceedings of COLING*, 2004.

[17] R. Agarwal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of 20th International Conference on Very Large Data Bases*, 1994.

[18] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of International Conference on Machine Learning*, 2001, pp. 282–289.

[19] E. M. Voorhees, "Trec-8 question answering track report," in *Proc. of the 8th Text Retrieval Conference*, 1999, pp. 77–82.