Inference Network-based Indonesian Spoken Query Information Retrieval

Dessi Puji Lestari* and Sadaoki Furui*
* Tokyo Institute of Technology, 2-12-1-W8-77, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: {dessi,furui}@furui.cs.titech.ac.jp Tel: +81-3-5734-3480

***Abstract*—Query term misrecognition caused by the speech recognizer is one of the important issues in the spoken query information retrieval. The misrecognized term in the transcribed query leads to the retrieval of irrelevant documents. To raise the correct ranking of the retrieved documents, we use a speech recognition confidence score based on word posterior probability to weight the term in the inference network-based (IN-based) Indonesian information retrieval system. Our result shows that this technique can improve the mean reciprocal rank (MRR) score of the retrieved documents.**

I. Introduction

Spoken query processing plays an important role in the interactive information retrieval system which enables users to input the query by speech rather than by typing. One of the important issues in the spoken query processing system is the term misrecognition problem caused by incorrect recognition of the terms in the spoken queries by the automatic speech recognizer. When the terms are incorrectly recognized, a number of relevant documents containing the correct terms cannot be retrieved, while a number of irrelevant documents containing the wrong terms are retrieved. The larger the number of misrecognition terms becomes in the transcribed query, the lower the ranking of the retrieved documents becomes.

Inference Network (IN) based information retrieval has an advantage in that it can employ a powerful query language that allows structured query operators and term weighting. Thus, it enables the user to explicitly state the importance of a term comparing to other terms in the query by giving a direct weight to each term in the query. This technique can be used to give more specific information to the query [1].

In the speech recognition system, a confidence score of each recognized word can be used as a measure of how certain the system is about the recognized word [2]. The bigger the confidence score is, the more certain the recognized word is. The confidence score can be used in various ways. For example, in a spoken dialogue system, it allows a dialogue manager to reject uncertain words to avoid unnecessary interactions for utterance verification. One of the popular methods of indicating confidence to a speech recognition result is using a confidence score based on posterior probabilities of the words.

This paper presents the use of a confidence score based on the word posterior probability to explicitly weight each transcribed term in the query for the inference network-based information retrieval. The aim is to give additional information to the query on how certain the recognized word is in the query as a correct term in order to reduce irrelevant documents.

II. Inference Network Model

The Inference Network (IN) model is basically a directed acyclic graph (DAG) of Bayesian Network which is used to model documents, document contents, and queries. Given many sources of evidence, the IN model has the ability to perform a document ranking by combining the evidence. It consists of two sub-networks: the Document Network (DN) and the Query Network (QN) [1] as shown in Fig. 1.

A. Document Network

The DN is produced for documents by indexing and does not change during the retrieval process. It consists of three layers of nodes. The first layer consists of document nodes (d_i nodes) that represent the events for which the documents are observed. Every document in the corpus is represented as a document node. These nodes represent abstract documents rather than the physical representations. The second layer consists of text representation nodes (t_j nodes). In this paper we consider the content of document with a text format only, but it can model not only text, but also image, audio, and video. Since we assume only text representation is available in the information retrieval system, the relationship between the text node and the document node is one-to-one. The last layer in the document network consists of representation nodes (r_k nodes) which represent concepts in the collection. They can be a kind of indexing feature of the document. These nodes can be single terms or proximity representations. A single term corresponds to some term in the corpus, while proximity representations can be phrases, terms appearing ordered or

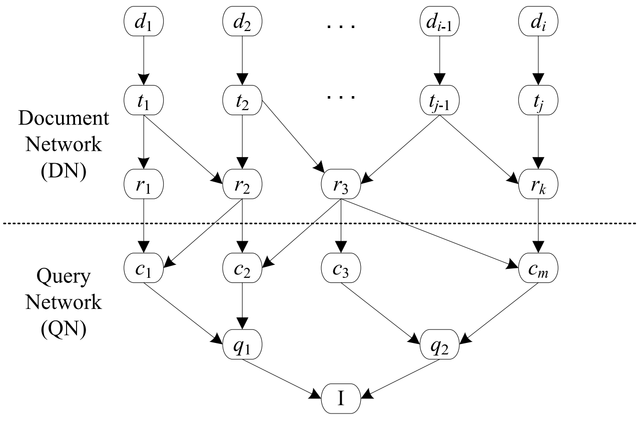

Fig. 1. An Inference Network.

unordered within a fixed length window of words, and other such concepts.

A causal link represented as down arrow between nodes indicates that the parent nodes are represented by the children node. Each link contains a conditional probability, or weight, to indicate the strength of the relationship. The evaluation of a node is performed using the value of the parent nodes and the conditional probabilities. Each document node has a prior probability associated with it that describes the probability of observing that document. This prior probability will generally be set to 1/(collection size). Each representation node contains a specification of the conditional probability associated with the node given its set of parent text nodes. This specification is basically any indexing weight such as the term frequency for each parent text or term weights such as the inverse document frequency associated with the representation concept. In practical use, the canonical representations are implemented that allow us to compute the conditional probabilities when required.

B. Query Network

The QN is produced during the retrieval process. It consists of nodes that represent the required concepts of the query called the query concept nodes (c_m nodes) and the nodes that represent the query (q nodes). The final leaf of the QN is the leaf node that represents the user information need (I node). Each query node contains a specification in the form of the link matrices to describe the dependence of the query on its parent query concepts. The form of the link matrix is determined by the dependance type or the query type; a link matrix simulating a query contains a boolean operator such as AND, OR, and NOT operator is different than a matrix simulating a weighted query such as WSUM (weighted sum) operator.

In the retrieval process, to form the complete IN, the QN is attached to the DN if the concepts in both networks are the same. After the attachment phase, the complete IN is evaluated for each document node to form the probability of the relevance to the query. The evaluation is initialized by setting the output of one document node to true (1) and all the other document nodes to false (0). This is applied to each document node in turn. The probability of document relevance is taken from the final node I and is used to produce the ranking [3].

For all non-root nodes in the inference network, probability of each node needs to be calculated using its parent values. If a node A has a set of parents $\pi = \ p_1, \ . \ . \ . \ , p_n$, we need to estimate P(A|p_1, . . . ,p_n). It usually uses a link matrix to provide diagnostic information to the set of parents based on belief with A. In practical use, a canonical link matrix form is implemented. This link matrix can be used to implement a variety of weighting schemes, including familiar term weighting schemes based on the frequency of a term in a single document (tf), and inverse value of the frequency of documents including the term (idf), often combining together (tf.idf) [1]. Some researchers used smoothed language model estimates [4].

C. The tf.idf Observation Estimates

Belief of node r_k (bel(r_k)) is computed when the concept r_k is true given a single document d_i i.e. $bel(r_k) = P(r_k = true|d_i = true, d_{j\neq i} = false)$. Using the tf.idf weight, belief of node r_k can be computed as follows:

$$bel(r_k) = \lambda + (1-\lambda)\overline{tf_{rk,di}}idf_{rk} \qquad (1)$$

where λ is an arbitrary default belief. This ensures that every representation is allocated a non-zero belief for the observed document, even if it is not present in the document. The $\overline{tf_{rk,di}}idf_{rk}$ value can be calculated using any standard method for estimating tf.idf weights for the representation r_k. Here we use the Okapi tf score [5], and a standard idf score. The okapi tf score for representation r_k and document d_i and idf score can be written as follows:

$$\overline{tf_{rk,di}} = \frac{tf_{rk,di}}{tf_{rk,di} + 0.5 + 1.5\frac{|d_i|}{|D_{avg}|}} \qquad (2)$$

$$idf_{rk} = \frac{\log(\frac{|C|+0.5}{tf_{rk,di}})}{\log(|C|+1)} \qquad (3)$$

where $tf_{rk,di}$ is the number of times that representation r_k is matched in a document d_i, $|d_i|$ is the length of document i, $|D_{avg}|$ is the average of document length in the collection, and $|C|$ is the number of words in the collection.

One merit of IN based information retrieval systems comparing to other traditional IR approach is the possibility to explicitly weight the term in the query. The query operator that is used for weighting the query is the #WSUM operator. Let assume that we have a simple query of the form #WSUM (w_1q_1, . . . , w_nq_n). By propagating beliefs from the nodes corresponding to the q to I nodes for an observed document d_j, the belief for #WSUM becomes:

$$bel_{wsum}(I) = \frac{\sum_i w_i p_i}{\sum_i w_i} = \sum_i w_i \overline{tf_{q_i,d_j}} idf_{q_i} \qquad (4)$$

D. Smoothed Language Model Estimates

In the language modeling for information retrieval, smoothing plays the same role as idf weighting in tf.idf based systems [6]. Here, by using Jelinek-Mercer smoothing, which is a method of interpolating between the document and collection language model, the belief r_k given a single document d_i can be computed as follows:

$$bel(r_k) = \lambda\frac{tf_{rk,di}}{|d_i|} + (1-\lambda)\frac{cf_{rk}}{|C|} \qquad (5)$$

where λ is the smoothing parameter, $tf_{rk,di}$ is the number of times that representation r_k is matched in document d_i, $|d_i|$ is the number of words in the document, cf_{rk} is the number of times that word r_k appears in the entire collection, and $|C|$ is the number of words in the collection.

This method also allows us to give explicit weighting to the term in the query. Instead of the #WSUM operator, the #Weight operator is typically used to calculate the belief of the query. Suppose that we have a simple query of the form #Weight $(w_1q_1, \ldots, w_nq_n)$, belief of I can be calculated by propagating beliefs from the nodes corresponding to the q to I node for an observed document d_j as follows:

$$bel_{weight}(I) = \prod_i (\lambda \frac{tf_{q_i,d_j}}{|d_j|} + (1-\lambda)\frac{cf_{q_i}}{|C|})^{w_i} \tag{6}$$

III. Confidence Scoring using Word Posterior Probability

A recognition decoder typically outputs results as a form of N-best list or word lattice [2]. The word lattice contains a large number of alternative word hypotheses and their associated likelihoods. By parsing the hypotheses, the posterior probability of each word hypothesis is computed. Since it reflects the relative distribution of word likelihoods among many alternatives, the posterior probability works well as the confidence measure [7].

Let τ denote the starting time and t the ending time of word w. $W[w;\tau,t]$ denotes all sentences that contain the hypothesis $[w;\tau,t]$. Given an N-best list or a word lattice from recognition decoder, the posterior probability $p([w;\tau,t]|X)$ of a specific word hypothesis $[w;\tau,t]$ over the acoustic observation X can be computed by summing up the posterior probabilities of all paths which contain the hypothesis $[w;\tau,t]$.

$$p([w;\tau,t]|X) = \sum_{W \epsilon W_{w;\tau,t}} \frac{p(X|W)p(W)}{p(X)} = \sum_{W \epsilon W_{w;\tau,t}} \frac{e^{g(W)}}{p(X)} \tag{7}$$

where $g(W)$ is log likelihood of a sentence hypothesis W derived from the recognition decoder defined as follows:

$$g(W) = \log p(X|W)p(W) \tag{8}$$

$p(X)$ is approximated by summation over all paths through the lattice. The word posterior probability can be used directly as the confidence score of the word hypothesis as follows:

$$C([w;\tau,t]) = \sum_{W \epsilon W_{w;\tau,t}} \frac{e^{\alpha . g(W)}}{p(X)} \tag{9}$$

where α is a scaling parameter $(\alpha < 1)$ to avoid only a few words to dominate the sums in these equations due to a large dynamic range of acoustic likelihoods[7]. The sum of probabilities of paths $W[w;\tau,t]$ and probabilities of all paths for $p(X)$ is computed to estimate a word confidence score of $W[w;\tau,t]$. If the hypotheses are given as N-best list, the computation process is a summation of scores over the sentences containing $w[w;\tau,t]$. On the word lattice, the forward-backward algorithm is usually applied.

IV. Methods

We use both the tf.idf estimates and the Jelinek-mercer smoothed language model estimates in the inference network-based Indonesian IR. We also use the speech recognition confidence score to explicitly weight each term in the query.

Let assume that we have a simple spoken query q. The spoken query q is first transcribed using a speech recognizer and converted to $q(q_1...q_n)$ as its representation. By adding the confidence score, representation of the weighted query becomes $\#WSUM(c_1q_1...c_nq_n)$ for the case of tf.idf based IN, and $\#Weight(c_1q_1...c_nq_n)$ for the case of smoothed language model based IN. The confidence score is calculated using the posterior probabilities on the generated word lattice as shown in Equation (9).

The belief of query q given a document d_j using the tf.idf based IN is then calculated by incorporating Equations (4) and (9) as follows:

$$bel_{wsum}(q) = \sum_i Cs([q_i;\tau,t])\overline{tf_{q_i,d_j}}idf_{q_i} \tag{10}$$

The belief of query q given a document d_j using the smoothed language model estimates is calculated by incorporating Equations (6) and (9) as follows:

$$bel_{weight}(q) = \prod_i (\lambda \frac{tf_{q_i,d_j}}{|d_j|} + (1-\lambda)\frac{cf_{q_i}}{|C|})^{Cs([q_i;\tau,t])} \tag{11}$$

where $Cs([q_i;\tau,t]$ is the confidence score of q_i in the query q. As mentioned above, here we use Jelinek-Mercer smoothing. The probability of document relevance is taken from the node q and is used to produce document ranking.

V. Experiments

The spoken query is first transcribed using the Bahasa Indonesia LVCSR system that we built previously [8]. We used Julius 4.0 as the speech decoder[1]. After removing the stop words in Bahasa Indonesia [9], the transcribed query with a speech recognition confidence score for each term in the query is fed into the IR system. We used the Lemur toolkit[2] provided by Carnegie Mellon University and the University of Massachusetts, Amherst, to build the Indonesian IR system.

A. Experimental Data

Since there is no standard evaluation corpus for spoken query IR in Bahasa Indonesia, we recorded spoken queries by 20 native Indonesian speakers (11 males, 9 females), each uttering 35 queries with different topics. The queries were derived from the Bahasa Indonesia IR collection developed by the ILPS [10]. The articles in the corpus were taken from the two popular Indonesian newspaper[3] and magazine[4] sites. There are 35 query topics available for the magazine corpus and the newspaper corpus in the ILPS corpus. In the experiment in this paper, we only used the corpus taken from the magazine. For each of the 35 topics of the query, we developed three kinds of spoken queries in terms of the length: short query (2-4 words), medium-length query (4-8 words), and long query (8-16 words). The aim was to analyze

[1] http://julius.sourceforge.jp/index.php

[2] http://www.lemurproject.org

[3] http://www.kompas.com

[4] http://www.tempointeraktif.com

TABLE I
WORD CORRECTNESS AVERAGE FOR DIFFERENT LENGTH OF SPOKEN QUERY

Query length	Word correctness (%)
short	86.30
medium	87.60
long	84.41
all	86.10

the effect of the word length in the future to the retrieval performance. There are 2100 Indonesian spoken queries in total. The Indonesian text corpus provided by ILPS was divided into two parts. The first part was used to train the language model of Bahasa Indonesia LVCSR, and the second part was used as the document collection for the IR system.

B. Evaluation

We compared the results between the text query and the spoken query information retrieval. We used the mean reciprocal rank (MRR) as the measure for IR and the word correctness instead of the word accuracy for the speech recognition performance measure. In our experiment, we have found that the word correctness has more influence to the IR performance comparing to the word accuracy, i.e. a query with low accuracy but high word correctness can give a better retrieval result even in the worst case where the word accuracy is 0%. For example, if query 4 in the testing data is uttered by speaker 17 the accuracy was 0%, while if it is uttered by speaker 18 the accuracy was 25% higher than speaker 17, however the MRR score given by speaker 17 was 1.0 while the MRR score given by speaker 18 was only 0.25. This happened since the word correctness of query 4 uttered by speaker 17 was 75% and the word correctness of query 4 spoken by speaker 18 was only 50%.

Comparing to the simple query, the weighted query using a confidence score gives a slightly better performance both for tf.idf estimates and smoothed language model estimates as shown in Tables II and III. Between these two estimate methods, the tf.idf estimate gives better MRR score than the smoothed language model estimate.

We have found that in the case where almost all the terms in the query are recognized correctly, the confidence score does not necessarily provide a good certainty measure in the word recognition, i.e. two words that recognized correctly may have a big difference in the confidence score value. In the case where the terms are recognized correctly, similar weighing values should be given to avoid biases in the document retrieval process.

TABLE II
COMPARISON OF MRR SCORE (%) FOR DIFFERENT LENGTH OF QUERY FOR TEXT QUERY (T-QUERY), SIMPLE SPOKEN QUERY (S-QUERY), AND WEIGHTED SPOKEN QUERY (W-QUERY) USING TF.IDF IN

Query Length	t-query	s-query	w-query
short	89.28	79.63	80.29
medium	89.28	81.01	81.63
long	90.1	82.42	84.68
all	89.56	81.02	82.20

TABLE III
COMPARISON OF MRR SCORE (%) FOR DIFFERENT LENGTH OF QUERY FOR TEXT QUERY (T-QUERY), SIMPLE SPOKEN QUERY (S-QUERY), AND WEIGHTED SPOKEN QUERY (W-QUERY) USING SMOOTHED LM ESTIMATES

Query Length	t-query	s-query	w-query
short	89.14	77.79	79.38
medium	83.10	76.19	77.06
long	85.48	76.02	77.18
all	85.90	76.67	77.87

VI. CONCLUSIONS

In this paper we have compared two inference network (IN) based IR with Indonesian spoken queries, using whether the tf-idf estimates or the smoothed language model estimates. The term weighing strategy using the speech recognition confidence score has shown its great potential to improve document ranking in the inference network-based IR. It works well especially for the spoken query with low correctness of recognition. However, further work using confidence score weighting needs to be conducted to improve the performance of the spoken query with high correctness of recognition.

ACKNOWLEDGMENT

The authors would like to thank F. Z. Tala for providing a Bahasa Indonesia text corpus for information retrieval.

REFERENCES

[1] H. R. Turtle and W. B. Croft, "Inference networks for document retrieval," in *ACM Transactions on Information Systems,* vol. 9, no 3, pp. 187-222, July 1991.

[2] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proc. EUROSPEECH,* vol. 2, pp. 827-830, Greece, September 1997.

[3] H. R. Turtle and W. B. Croft, "Efficient probabilistic inference for text retrieval," in *RIAO 3,* pp. 644-661, Spain, 1991.

[4] D. Metzler and W. B. Croft, "Combining the language model and inference network approaches to retrieval," in *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval,* vol. 40(5), pp.735-750, 2004.

[5] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," In *Proc. the 17th annual international ACM SIGIR conference on research and development in information retrieval,* pp. 232-241, New York, 1994.

[6] D. Hiemstra,"Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term." In *Proc. of the 25th annual international ACM SIGIR conference on research and development in information retrieval,* pp. 35-41, Finland, 2002.

[7] F. Wessel, R. Schluter, K. Macherey, and H. Ney,, "Confidence measures for large vocabulary continuous speech recognition," in *IEEE Trans. Speech and Audio Process.,* pp. 288-298, vol. 9, no. 3, March 2001.

[8] D. P . Lestari, K . Iwano, and S . Furui, "A large vocabulary continuous speech recognition system for indonesian language," in *Proc. 15th Indonesian Scientific Conference in Japan (ISA-Japan),* pp.17-22, Hiroshima, Japan, 2006.

[9] F .Z . Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis,* Appendix D, pp. 39-46, University of Amsterdam, 2003.

[10] F .Z Tala, J . Kamps, K . Muller, and M . de Rijke, 'The Impact of Stemming on Information Retrieval in Bahasa Indonesia," *14th Meeting of Computational Linguistics in the Netherlands (CLIN-2003),* Netherland, 2003.