Enforcing Topic Diversity in a Document Recommender for Conversations

Maryam Habibi
Idiap Research Institute and EPFL
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
maryam.habibi@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

This paper addresses the problem of building concise, diverse and relevant lists of documents, which can be recommended to the participants of a conversation to fulfill their information needs without distracting them. These lists are retrieved periodically by submitting multiple implicit queries derived from the pronounced words. Each query is related to one of the topics identified in the conversation fragment preceding the recommendation, and is submitted to a search engine over the English Wikipedia. We propose in this paper an algorithm for diverse merging of these lists, using a submodular reward function that rewards the topical similarity of documents to the conversation words as well as their diversity. We evaluate the proposed method through crowdsourcing. The results show the superiority of the diverse merging technique over several others which not enforce the diversity of topics.

1 Introduction

We present a diverse retrieval technique for ranking documents that are spontaneously retrieved and recommended to people during a conversation. These documents represent potentially useful information for the conversation participants. The information needs of the participants are represented by implicit queries which are built in the background based on their current speech, specifically from keywords obtained from the conversation transcripts. Since people usually mention several topics even during a short conversation span, such keyword sets are made of content words related to different topics. When juxtaposed in an implicit query, these topics may have noisy effects on the retrieval results (Bhogal et al., 2007; Carpineto and Romano, 2012).

The purpose of this paper is to present a method for merging lists of documents retrieved through multiple implicit queries prepared for short conversations spans. Several topically-separated queries are constructed from keywords, and generate several lists of documents. The goal of the method proposed here is to generate a unique and concise list of documents that can be recommended in real time to the conversation participants. The list should cover the maximum number of implicit queries and therefore topics. To merge the lists of documents according to these criteria, we use inspiration from extractive text summarization (Lin and Bilmes, 2011; Li et al., 2012) and from our own previous work on diverse keyword extraction (Habibi and Popescu-Belis, 2013). The method proposed here rewards at the same time topic similarity – to select the most relevant documents to the conversation fragment – and topic diversity – to cover the maximum number of implicit queries and therefore topics in a concise and relevant list of recommendations, if more than one topic is discussed in the conversation fragment.

Several studies have been previously carried out on merging lists of results in information retrieval. Despite the superficial similarity, the problem here is in fact different from distributed information retrieval, where several lists of results from *different* search engines for the *same* query must be merged. Moreover, many studies addressed the topic diversification approach for re-ranking the retrieved results of a single query. However, these approaches are not directly applicable to multiple queries.

The paper is organized as follows. In Section 2 we review existing techniques for merging and re-ranking lists of search results which are applicable here. We then explain the general framework of our document recommender system in Section 3. In Section 4 we describe the proposed algorithm for diverse merging of lists of recommendations. Section 5 presents the data, the parameters setting, and evaluation tasks for comparing document lists. In Section 6 we first demonstrate empirically the benefits, for just-in-time document recommendation, of separating users' information needs into multiple topically-separated queries rather than using a unique query. Then, we compare the proposed diverse merging technique with several alternative ones, showing that it outperforms them according to human judgments of relevance, and also exemplify the results on one conversation fragment given in the Appendix A.

2 Related Work

Just-in-time document retrieval systems have been designed to recommend to their users documents which are potentially relevant to their activities, e.g. individual users authoring documents or browsing various repositories, or small groups holding business or private meetings (Hart and Graham, 1997; Rhodes and Maes, 2000; Popescu-Belis et al., 2008). When using a document recommender system, people are generally unwilling to examine a large number of recommended documents, mainly because this would distract them from their main activity. Several solutions to this problem have been proposed.

For instance, the Watson document recommender system (Budzik and Hammond, 2000), designed for reading or writing activities, clustered the document results and selected from each cluster the best representative to generate a list of recommendations. Clustering results is not suitable for our application where the mixture of topics in a single query will degrade the document results aimed to be clustered (Bhogal et al., 2007; Carpineto and Romano, 2012), and consequently may have a damaging effect on the clusters' representatives. The second part of the method, which selected the best representative of the clusters in the final document list can be helpful; however, its effectiveness relies on having clusters with the same level of importance (Wu and McClean, 2007).

Many studies in information retrieval addressed the problem of diverse ranking, which can be stated as a tradeoff between finding relevant versus diverse information (Robertson, 1997). The existing diverse ranking proposals differ in their diversifying policies and definitions, which can be categorized into implicit methods (Carbonell and Goldstein, 1998; Zhai et al., 2003; Radlinski and Dumais, 2006; Wang and Zhu, 2009) or explicit ones (Agrawal et al., 2009; Carterette and Chandar, 2009; Santos et al., 2010; Vargas et al., 2012). The implicit approaches assume that similar documents will cover similar aspects of a query, and have to be demoted in the ranking to promote relative novelty and reduce overall redundancy. In one of the earliest approaches, Carbonell and Goldstein (1998) introduced Maximal Marginal Relevance (MMR) to re-rank documents based on a tradeoff between the relevance of document results and relative novelty as a measure of diversity. MMR was also used by Radlinski and Dumais (2006) to re-rank results from a query set which is generated for a user query and represents a variety of potential user intents.

Instead of implicitly accounting for the aspects covered by each document, another option is to explicitly model these aspects within the diversification approach. Agrawal et al. (2009) introduced a submodular objective function to minimize the probability of average user dissatisfaction by assuming a taxonomy of information and modeling user query aspects at the topical level of this taxonomy. Alternatively, Santos et al. (2010) proposed another submodular objective function to maximize coverage and minimize redundancy with respect to query aspects modeled in a keyword-based representation form instead of a predefined taxonomy.

In our case, the recommender system for conversational environments requires diversity in the results of multiple topically-separated queries, rather than of a single ambiguous query. Therefore, a new approach will be proposed, and will be compared in particular to a version of the explicit diversification approach (Santos et al., 2010) adapted to our problem.

3 Framework of our Document Recommender System

We have designed the Automatic Content Linking Device (ACLD), a speech-based just-in-time document recommender system for business meetings (Popescu-Belis et al., 2008; Popescu-Belis et al., 2011).

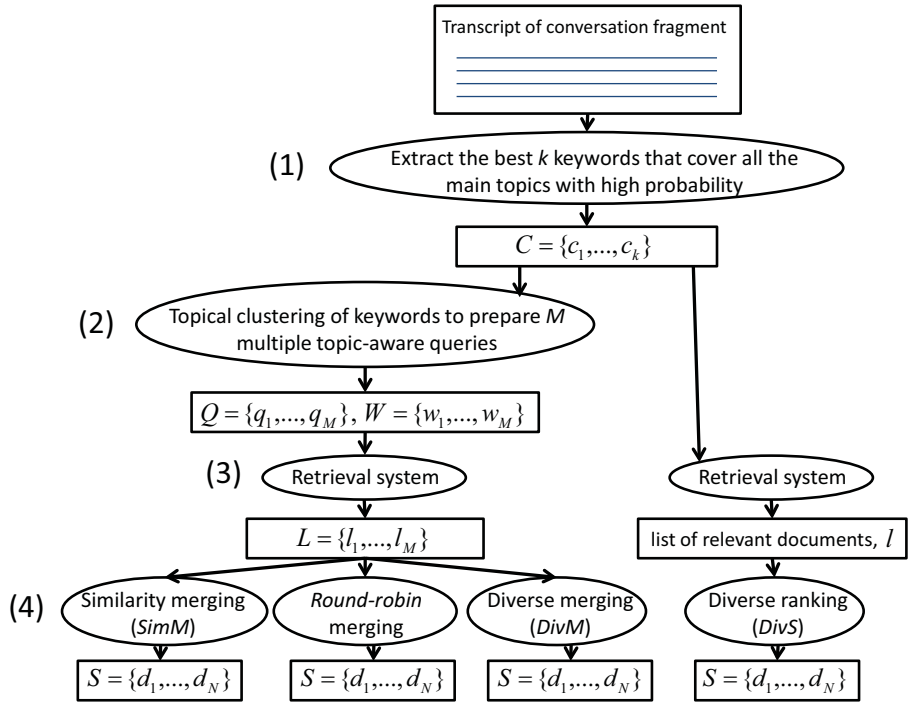

Figure 1: The four stages of our document recommendation approach (shown vertically: 1–4) and the four options considered in this paper (bottom line: *SimM*, *Round-robin*, *DivM*, and *DivS*).

The ACLD monitors the ongoing conversation, and formulates queries based on the words detected by a real-time automatic speech recognition (ASR) system (Garner et al., 2009). The queries are fired periodically to retrieve documents which are then recommended to users by displaying their titles along with relevant excerpts. As these queries are built and triggered in the background, they are referred to as 'implicit queries', as opposed to 'explicit' ones that could be formulated by users. Just-in-time document recommendation in the ACLD system proceeds according to the steps shown in Figure 1, which displays at step 4 the various options for merging lists of results that are the focus of this paper.

Prior to the first processing step outlined in Figure 1, the ACLD must decide when to make a recommendation, and what portion of the conversation prior to that moment should be used. This question is beyond the scope of this paper, and remains to be fully investigated, using verbal and non-verbal criteria. Here, for the reasons explained in Section 5.2, the ACLD recommends documents every two minutes, segmenting the conversation at the end of the nearest utterance and using the entire conversation fragment since the previous recommendation. Although in practice the results of the current recommendation process are merged with the previous ones (using a weighted mechanism that embodies the idea of "persistence" of documents over time), in this paper we will consider the recommendation for each fragment independently of the previous one.

The recommendation process represented in Figure 1 starts by extracting a set of keywords, C, from the words recognized by the ASR system from the users' conversations. The keywords are extracted using the diverse keyword extraction technique that we proposed (Habibi and Popescu-Belis, 2013), which maximizes the coverage of the topics of a text by the extracted keyword set, as we also target in this paper. Then, implicit queries which express the users' information needs are formulated using the keyword set, following two alternative approaches depicted in step 2 of Figure 1. In a baseline model (right side of the figure), a single query is built for the conversation fragment using the entire keyword list as an implicit query. In the approach we are advocating, multiple topically-separated queries are produced for the conversation fragment (step 2, left side of the figure). This is described in a separate document (Habibi and Popescu-Belis, submitted), but can be outlined as follows. The implicit queries are obtained by clustering the above-mentioned keyword set into several topically-separated subsets, each one corresponding to an abstract topic obtained using topic modeling techniques (similarly to the model

presented in Subsection 4.1). Each subset is an implicit query, and is weighted based on the importance of the topic to which it is associated.

In step 3, we separately submit each implicit query to the Apache Lucene search engine over the English Wikipedia and obtain several lists of relevant articles. Finally, we merge and re-rank these lists before recommendation (step 4). One baseline alternative is the explicit diverse ranking technique proposed by Santos et al. (2010) for diversifying the primary search results retrieved for a single query, shown on the right side of the figure. To compare the methods, we adapted this latter method to make it applicable to our system when a single implicit query is built for a conversation fragment, by defining query aspects using the abstract topics employed for query and document representation. The method is noted *DivS* as it *diversifies* documents from a *single* list.

Our proposal lies at step 4. As represented on the left side of Figure 1, in our system, we merge the lists of documents retrieved for multiple implicit queries. We thus propose a new method noted *DivM* and we compare it with two other merging techniques. The first one, noted *SimM*, ignores the diversity of topics in the list of results and ranks documents only by considering their topic similarity to the conversation fragment. The second one is the merging technique used by the above-mentioned Watson system (Budzik and Hammond, 2000), which uses Round robin merging, hence it is noted *Round-robin*. In contrast, our proposed method, *DivM*, is a *diverse merging* technique which we now proceed to define formally.

4 Diverse Merging of the Results of Multiple Queries

The diverse merging of retrieved document lists is the process of creating a short, diverse and relevant list of recommended documents which covers the maximum number of topics of each conversation fragment. The merging algorithm rewards diversity by decreasing the gain of selecting documents from a list as the number of its previously selected documents increases. The method proceeds in two steps. First, we represent queries and the corresponding list of candidate documents from the Apache Lucene search engine using topic modeling techniques, and then we rank documents by using topical similarity and rewarding the coverage of different lists.

4.1 Document and Query Representation

A topic model represents the abstract topics which occur in a collection of documents – here, preferably, a collection that is representative of the domain of the conversations. Once trained, topic models such as Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA) can be used to determine the distribution of abstract topics in each set of words composing either a conversation fragment, or a query, or a document. LDA implemented in the Mallet toolkit (McCallum, 2002) is used here to train topic models because it does not suffer from the over-fitting of PLSA (Blei et al., 2003).

We first learn a probability model for observing a word v in a document d through the set of abstract topics $T = \{t_1, ..., t_z, ..., t_Z\}$, where Z is the number of topics, using the Mallet toolkit:

$$p(v|d) = \sum_{z=1}^{Z} p(v|t_z) \cdot p(t_z|d) \quad (1)$$

The topic-word distribution $p(v|t_z)$ and the document-topic distribution $p(t_z|d)$, which are obtained using topic modeling, respectively show the contribution of the word v in the construction of the topic t_z, and the distribution of topic t_z in the document d with respect to the other topics.

We represent each new text or fragment A (e.g. from a conversation or document) by a set of probability distributions over all abstract topics T noted as $P(A) = \{p(t_1|A), ..., p(t_z|A), ..., p(t_Z|A)\}$ where $p(t_z|A)$ is inferred using the Gibbs sampling implemented by the Mallet toolkit given the topic models previously learned. We associate to each new document d_i and query q_j a set of topic probabilities according to the above definition noted respectively as $P(d_i) = \{p(t_1|d_i), ..., p(t_z|d_i), ..., p(t_Z|d_i)\}$ and $P(q_j) = \{p(t_1|q_j), ..., p(t_z|q_j), ..., p(t_Z|q_j)\}$.

4.2 Diverse Merging Problem

As stated above, our goal is to recommend a short ranked list of documents answering the users' information needs hypothesized in a conversation fragment, which are modeled by multiple topic-aware implicit

queries as described in Section 3. We build the final list of recommended documents by merging the document lists, one from each implicit query, with the objective of the maximum coverage of the topics of the conversation fragment. Since each document list contains documents found by a search engine given an implicit query, which was prepared for one of the main topics of the conversation fragment, we merge the lists by selecting documents from the maximum number of lists in addition to maximizing their topical similarity to the conversation fragment.

The problem of diverse merging of lists thus amounts to finding a ranked subset of documents $S \subset \cup_{i=1}^{M} l_i$, which are the most representative of all the result lists l_i, and potentially the most informative with respect to the conversation fragment and the information needs that are implicitly stated. This problem is an instance of the maximum coverage problem, which is known to be NP-hard. Our formulation and solution proceed as follows.

Let us consider a set of implicit queries $Q = \{q_1, ..., q_M\}$, and the corresponding set of document lists $L = \{l_1, ..., l_M\}$ resulting from each query. M is the number of implicit queries of the fragment, and each l_i is a list of documents $\{d_1, ..., d_{N_i}\}$ which are retrieved for query q_i. We define the weight w_i of each query q_i as the importance within the conversation fragment of the topics represented in the query q_i, and compute it as the topical similarity of q_i to the fragment, as shown in Equation 2. In this equation, q is the query made from the whole keyword set, which we call a *collective query*, and includes keywords for all the main topics of the conversation fragment in one query. In turn, we associate to q a set of probabilities over abstract topics, $P(q) = \{p(t_1|q), ..., p(t_Z|q)\}$, similar to the representation of implicit queries explained in Subsection 4.1.

$$w_i = \sum_{z=1}^{Z} p(t_z|q_i) \cdot p(t_z|q) \tag{2}$$

4.3 Defining a Diverse Reward Function

Although the maximum coverage problem is NP-hard, it has been shown that a greedy algorithm can find an approximate solution guaranteed to be within a factor of $(1 - 1/e) \simeq 0.63$ of the optimal one if the coverage function is submodular and monotone non-decreasing[1] (Nemhauser et al., 1978). Several monotone submodular functions have been proposed in various domains for a similar underlying problem, such as explicit diverse re-ranking of retrieval results (Agrawal et al., 2009; Santos et al., 2010; Vargas et al., 2012), extractive summarization of a text (Lin and Bilmes, 2011; Li et al., 2012), or our own model of diverse keyword extraction from a text (Habibi and Popescu-Belis, 2013).

We define a monotone submodular function for diverse merging of document lists inspired by the latter two applications, who proposed a power function with a scaling exponent between 0 and 1 for diverse selection of sentences (or keywords) covering the maximum number of topics of a given document with a fixed number of items. To adapt these techniques to the problem of diverse merging, from the perspective of capturing users' information needs in the set of recommended documents, we define here a reward function enforcing the diverse merging of the lists of document results.

We first estimate the topical similarity of the document subset $S_i = S \cap l_i$ to the collective query q (see Subsection 4.2) as r_{S_i}:

$$r_{S_i} = \sum_{d \in S_i} \sum_{z=1}^{Z} p(t_z|d) \cdot p(t_z|q) \tag{3}$$

We then propose the following reward function f for each S_i containing relevant documents selected from l_i (results of implicit query q_i), where w_i is the topical similarity of q_i to the conversation fragment (see Equation 2), and λ is an exponent parameter between 0 and 1. This reward function is submodular because it has the diminishing returns property when r_{S_i} increases.

$$f : r_{S_i} \rightarrow w_i \cdot r_{S_i}^{\lambda} \tag{4}$$

[1] A function F is *submodular* if $\forall A \subseteq B \subseteq T \setminus t$, $F(A + t) - F(A) \geq F(B + t) - F(B)$ (diminishing returns) and is *monotone non-decreasing* if $\forall A \subseteq B$, $F(A) \leq F(B)$.

The set S is ultimately ranked by maximizing the cumulative reward function $R(S)$ over all the lists, written as follows:

$$R(S) = \sum_{i=1}^{M} w_i \cdot r_{S_i}^{\lambda} \tag{5}$$

The probability of selecting documents from the list of results for q_i thus depends on w_i, the topical similarity of the query to the conversation fragment. This is in contrast to choosing the best representative document from the list of documents relevant to each query, like in the Watson system, which does not select more documents for queries with higher weight before considering lower weight ones. Our model rewards diversity to increase the chance of choosing documents from all the lists of results retrieved for implicit queries.

4.4 Finding the Optimal Document List

Since $R(S)$ is a monotone submodular function, we propose a greedy algorithm (Alg. 1) to maximize $R(S)$. If $\lambda = 1$, the reward function ignores the diversity constraint, because it does not penalize multiple selections from the same list l_i and ranks documents only depending on their similarity to the collective query and on the weights of implicit queries. However, when $0<\lambda<1$, as soon as a document is selected from the list of results of an implicit query, other documents from the same list start having diminishing returns as competitors for selection. Decreasing the value of λ increases the impact of the diversity constraint on ranking documents, which augments the chance of recommending documents from other document lists.

Input : query set Q of size M with probabilities, set of weights W, set of lists of document results L with probabilities, number of recommended documents k
Output: set of recommended documents S
$S \leftarrow \emptyset$;
for $i = 1$ ***to*** M ***step*** 1 **do**
  $S_i \leftarrow \emptyset$;
end
while $|S| \leq k$ **do**
  $S \leftarrow S \cup argmax_{d \in ((\cup_{i=1}^{M} l_i) \setminus S)}(g(d))$ where $g(d) = \sum_{i=1}^{M} w_i \cdot [r_{\{d\} \cap l_i} + r_{S_i}]^{\lambda}$;
  for $i = 1$ ***to*** M ***step*** 1 **do**
    $S_i = l_i \cap S$;
  end
end
return S;

Algorithm 1: Diverse merging of document results for recommendation.

5 Data, Settings and Evaluation Method

The experiments were performed on conversational data from the ELEA Corpus (Emergent LEader Analysis, Sanchez-Cortes et al. (2012)). Implicit queries were formulated as presented above in Figure 1 using keywords extracted from each conversation fragment, defined as below (Subsection 5.1). Each subset of keywords obtained by topical clustering of the keyword set resulted in an implicit query. The lists of document results for each implicit query were obtained by submitting the query to the Apache Lucene search engine[2] over the English Wikipedia[3]. These initial lists of results were ultimately merged into final recommendation lists of documents using the four alternative methods from Figure 1, including the one we proposed. This section presents the data, system parameters, and evaluation methods used in our experiments.

[2]Available from `http://lucene.apache.org`.
[3]A local copy was downloaded from `http://dumps.wikimedia.org`.

5.1 Conversational Corpus

The ELEA Corpus comprises nearly ten hours of recorded meetings in English and French. Each meeting consists in a role play game in which participants play survivors of an airplane crash in a mountainous region. They must rank a list of 12 items with respect to their utility for surviving until they are rescued. We used from the ELEA corpus four English conversations of around fifteen minutes each, which have been manually transcribed and segmented at the speaker turn level.

One of the most important issues for a just-in-time document recommender system is to determine the appropriate timing of the recommendations, and the size of the context to use for computing them. Here, awaiting future investigations[4], we decided to make recommendations approximately every two minutes, at the end of an ongoing speaker turn, and consider as input the words uttered since the previous recommendation. A segment size of two minutes enables us to collect an appropriate number of words (neither too small nor too large) in order to extract keywords, model the topics, and formulate implicit queries. Based on our experience with the ACLD, it also corresponds to an acceptable frequency for receiving suggestions.

Therefore, our test data comprises 26 two-minute segments, each of them ending at a speaker change. On average, segments contain 278 words (including stop words). Once topic modeling is applied, the average number of topics per fragment is 5, with an observed minimum of 3 and a maximum of 9.

5.2 Parameter Settings for Experimentation

As document search is performed over the English Wikipedia, we trained our topic models on this corpus as well. We used only a subset of it for tractability reasons, i.e. about 125,000 articles as in other studies (Hoffman et al., 2010). The subset is randomly selected from the entire English Wikipedia. As in previous studies, we fixed the number of topics at 100 (Boyd-Graber et al., 2009; Hoffman et al., 2010).

The exponent of the submodular function was set to $\lambda = 0.75$, as in our diverse keyword extraction study (Habibi and Popescu-Belis, 2013). This was found to be the best value for diverse merging of lists of results, as it leads to a reasonable balance between relevance and diversity in the aggregated list of documents. Of course, if sufficient training data were available, this could be used to optimize λ.

The number of recommended documents was fixed at five in our experiments. This value was selected again based on user preferences observed with the ACLD. Moreover, this is also the value of the average number of topics in a conversation fragment, which allows the system to cover on average one result per topic. Experiments with other values were not carried out due to the cost of evaluation.

5.3 Evaluation Protocol and Metrics

We designed a task that measures the relevance of recommended document lists for each of the test conversation fragment. Based on validation experiments in our previous work (Habibi and Popescu-Belis, 2012), the task requires subjects to compare two lists obtained by two different methods. Using a web browser, the subjects had to read the conversation transcript, answer several control questions about its content, and then decide which of the two lists provides more relevant documents, with the following options: the first list is better than the second one; the second is better than the first; both are equally relevant; or both are equally irrelevant. The position of each system (first or second) was randomized across the tasks.

The 26 comparison tasks (one for each ELEA fragment) were crowdsourced via Amazon's Mechanical Turk as "human intelligence tasks" (HITs). For each HIT we recruited ten workers, only accepting those with greater than 95% approval rate and more than 1000 previously approved HITs (qualification control). We only kept answers from the workers who answered correctly our control questions about each HIT. Each worker could answer the entire set of 26 HITs, or part of it. We observed that the average time spent per HIT was around 90 seconds.

[4]For instance, they could combine an analysis of non-verbal information to detect "interruptibility" and of verbal information to detect topic changes and perform online segmentation (Mohri et al., 2010). Topic changes, however, are not appropriate moments to make recommendations because it would be useless to recommend documents about a topic that the users no longer discuss (Jones and Brown, 2004).

To consolidate the comparative judgments over a large number of subjects and conversation fragments, and compute an aggregated score, we applied a qualification control factor to the human judgments (to reduce the effect of judgments which disagree with the majority vote) and another one to the HITs (to reduce the impact of undecided HITs on the global scores). This was done by using the PCC-H metric, defined and validated in our previous work (Habibi and Popescu-Belis, 2012), which provides two scores, one for each document list, summing up to 100%; a higher value indicates a better list. In addition to PCC-H, we also provide below (Table 1) the raw preference scores for each comparison, i.e. the number of times a system was preferred over another one, although PCC-H was shown to be a more reliable indicator of quality.

6 Experimental Results

We merged and re-ranked the document lists intended to be recommended during a conversation by the four methods presented above in Section 3 and Figure 1. Three methods merge lists of results from topically-separated queries: *SimM* only considers their similarity with the fragment; *Round-robin* picks the best document in each list; and our proposal, *DivM*, considers the diversity and importance of topics. A fourth method, *DivS*, uses one query made of all keywords extracted from the conversation fragment, and ranks the documents using the diverse re-ranking technique proposed by Santos et al. (2010).

Binary comparisons were performed between pairs of techniques, using crowdsourcing over 26 conversation fragments of the ELEA Corpus, and aiming to minimize the number of binary comparisons while still ordering completely the methods according to their perceived quality.

6.1 Diverse Re-ranking vs. Similarity Merging

We first performed a comparison between the top five documents generated by two recommendation strategies, *DivS* and *SimM*, over 26 conversation fragments of the ELEA Corpus. The consolidated relevance score (PCC-H) is 75% for *SimM* vs. 25% for *DivS*, as shown in Table 1. These scores indicate the superiority of *SimM* over *DivS*. In other words, separating the mixture of topics of a fragment into multiple topically-separated queries mitigates the negative effect of the mixture of topics on the suggestions.

6.2 Comparison across Merging Techniques

Binary comparisons were then performed between pairs of merging techniques (*SimM*, *Round-robin*, and *DivM*), using the same experimental settings. The PCC-H scores are 62% for *DivM* vs. 38% for *Round-robin*, 59% for *DivM* vs. 41% for *SimM*, and 56% for *Round-robin* vs. 44% for *SimM*, as shown in Table 1. The scores show that the diverse merging of lists of documents improves recommendations, and indicate the following high to low ranking: *DivM* $>$ *Round-robin* $>$ *SimM*.

SimM ranks lowest in this ordering, likely because of the ignorance of diversity in the list of results. *Round-robin* is second, likely because it disregards the major differences of importance among implicit queries in a conversation fragment. The results of the comparisons confirm that the *DivM* technique, which merges lists of documents by considering the diversity of topics in the list of recommendations, in proportion to their importance in the conversation, is the most satisfying to the majority of human subjects.

6.3 Impact of the Topical Diversity of Fragments

To further examine the benefits of our method, we studied its sensitivity to the number of topics in the conversation fragments. For this purpose, we divided the set of test fragments into two subsets. The first one (noted 'A' in Table 1) gathers the fragments for which fewer than or exactly five main topics (and therefore implicit queries) have been computed. The other fragments, with more than five main topics, form the second subset (noted 'B'). The value of five corresponds to the average number of main topics per fragment as well as to the number of recommended documents in our experiments.

As shown in Table 1, although there is an improvement in the comparison scores of *DivS* over *SimM* when the number of conveyed topics in the fragments is higher than the number of recommended documents (subset B), the comparison scores indicate the superiority of *SimM* over *DivS* in both cases, and

Compared methods (m_1 vs. m_2)	PCC-H relevance score (%)						Raw preferences (%)	
	A		B		A ∪ B		A ∪ B	
	m_1	m_2	m_1	m_2	m_1	m_2	m_1	m_2
SimM* vs. *DivS	**80**	20	**70**	30	**75**	25	**70**	30
Round-robin* vs. *SimM	33	**67**	**68**	32	**56**	44	**52**	48
DivM* vs. *Round-robin	**64**	36	**60**	40	**62**	38	**58**	42
DivM* vs. *SimM	**54**	46	**60**	40	**59**	41	**58**	42

Table 1: Comparative scores of the recommended document lists from four methods: *DivS*, *SimM*, *Round-robin*, and *DivM*, evaluated by human judges over the ELEA Corpus. Subset A gathers fragments with fewer than or exactly five topics, while subset B gathers all the other fragments. The results imply the following ranking: *DivM* > *Round-robin* > *SimM* > *DivS*.

confirm the benefit of the diverse merging techniques. When comparing *Round-robin* versus *SimM*, the scores show the superiority of the former method when the number of conveyed topics in fragments is higher than the number of recommended documents, because it provides a diverse lists of documents in which documents relevant to less important topics are not displayed. However, when the number of topics is smaller than the number of recommendations, *SimM* provides better results. The reason of the decrease in the scores of *Round-robin* is likely the ignorance of the actual importance of the main topics when ranking documents. Overall, as shown in Table 1, regardless of the number of topics conveyed in the fragments, *DivM* always outperforms *Round-robin* and *SimM*.

6.4 Example of Document Results

To illustrate how *DivM* surpasses the other techniques, we consider an example from one of the conversation fragments of the ELEA Corpus. The manual transcript of this conversation fragment is given in the Appendix A. As described in Section 5, the conversation participants had to select a list of 12 items vital to survive in winter while waiting to be rescued. The keywords extracted from the manual transcript of this fragment by our method (Habibi and Popescu-Belis, 2013) are: *fire, lighter, cloth, shoe, cold, die, igloo, walking*. As our keyword extraction method was shown to be robust to ASR noise, we only use here the reference transcripts (Habibi and Popescu-Belis, submitted).

We display the topically-aware implicit queries prepared by our method from this keyword list along with their weights in Table 2. Then, in Table 3 we show the retrieval results (five highest-ranked Wikipedia pages) obtained by the four methods using the reference transcript of this fragment.

As shown in Table 2, each implicit query corresponds to one of the main topics of the fragment with a specific weight. In this example, the main topics spoken in the fragment are about making an igloo, lightening a fire, having warm clothes, and suitable shoes for walking.

As shown in Table 3, *DivS* provides two irrelevant documents likely because the single (collective) query does not separate the mixture of topics in the conversation fragment, and leads to some poor results (Wikipedia pages) such as "Cold Fire (Koontz novel)". *SimM* slightly improves the results by separating the discussed topics of the conversation fragment into multiple queries. However, it does not cover all the

Implicit queries	Weights
q_1 = {fire, cold, igloo, lighter}	w_1 = 0.110
q_2 = {shoe, lighter, walking}	w_2 = 0.097
q_3 = {cloth}	w_3 = 0.058
q_4 = {die}	w_4 = 0.040
q_5 = {igloo}	w_5 = 0.026

Table 2: Example of implicit queries built from the keyword list extracted from a sample fragment of the ELEA Corpus. Each query covers one of the main topics of the fragment and has a different weight.

DivS	***SimM***	***Round-robin***	***DivM***
Flint spark lighter	Igloo	Igloo	Igloo
Extended Cold Weather Clothing System	Flint spark lighter	Shoe	Shoe
Cold Fire (Koontz novel)	Lighter	Jersey (clothing)	Flint spark lighter
Igloo	Lighter (barge)	Die Hard	Jersey (clothing)
Walking	Worcester Cold Storage Warehouse fire	Flint spark lighter	Lighter

Table 3: Example of retrieved Wikipedia pages from the four different methods tested in this paper. Results of diverse merging (*DivM*) appear to cover more topics relevant to the conversation fragment than other methods. The average ranking (*DivM* > *Round-robin* > *SimM* > *DivS*) is also observed in this example.

topics mentioned in the fragment due to mostly focusing on the single topic represented by q_1. *Round-robin* further enhances the results by adding diversity, but as it gives the same level of importance to all topics, it provides a poor result like "Die Hard" from a topic of the conversation fragment with a small weight. The results of *DivM* appear to be the most useful ones, as they include other articles relevant to q_1, q_2, and q_3 before showing results relevant to the low weight queries q_4 and q_5. Therefore, in this example, *DivM* provides better ranking of documents by covering the largest number of main topics mentioned in the fragment.

7 Conclusion

We proposed a diverse merging technique for combining lists of documents from multiple topically-separated implicit queries, prepared using keyword lists obtained from the transcripts of conversation fragments. Our *diverse merging* method *DivM* provides a short, diverse, and relevant list of recommendations, which avoids distracting participants that would consider it during the conversation. We also compared *DivM* to existing merging techniques, in terms of comprehensiveness and relevance of the final recommended list of documents to the conversation fragment. The human judgments collected via Amazon Mechanical Turk showed that *DivM* outperforms all other methods.

Moreover, these results emphasized the benefit of splitting the keyword set into multiple topically-separated queries: the suggested lists of documents from *DivS* (which accounts for the diversity of results by re-ranking the documents of a single list) were indeed found less relevant than those from *SimM* and the other two methods, which merged results from multiple queries.

In the future, the diverse merging method *DivM* will be integrated in the ACLD just-in-time retrieval system for conversational environments, with implicit queries that are prepared from the ASR transcript of users' conversation. User-oriented evaluation experiments will be conducted. We will also enable the system to answer explicit queries asked by users, considering contextual factors to improve the relevance of the answers, which will complement the recommendation functionality based on implicit queries.

Acknowledgments

The authors are grateful to the Swiss National Science Foundation for its support through the IM2 NCCR on Interactive Multimodal Information Management (2002-2013, see `http://www.im2.ch`), and to the Hasler Foundation for its support through the REMUS project (Re-ranking Multiple Search Results for Just-in-Time Document Recommendation, 2014). The authors also thank the anonymous reviewers for their helpful suggestions.

References

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM.

Jagdev Bhogal, Andy Macfarlane, and Peter Smith. 2007. A review of ontology based query expansion. *Information and Processing Management*, 43:866–886.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jonathan Boyd-Graber, Jordan Chang, Sean Gerrish, Chong Wang, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1–9.

Jay Budzik and Kristian J. Hammond. 2000. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 5th International Conference on Intelligent User Interfaces (IUI)*, pages 44–51. ACM.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1–56.

Ben Carterette and Praveen Chandar. 2009. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1287–1296.

Philip N. Garner, John Dines, Thomas Hain, Asmaa El Hannani, Martin Karafiat, Danil Korchagin, Mike Lincoln, Vincent Wan, and Le Zhang. 2009. Real-time ASR from meetings. In *Proceedings of Interspeech 2009 (10th Annual Conference of the International Speech Communication Association)*, pages 2119–2122.

Maryam Habibi and Andrei Popescu-Belis. 2012. Using crowdsourcing to compare document recommendation strategies for conversations. In *Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2011)*, pages 15–20.

Maryam Habibi and Andrei Popescu-Belis. 2013. Diverse keyword extraction from conversations. In *Proceedings of the ACL 2013 (51th Annual Meeting of the Association for Computational Linguistics)*, pages 651–657.

Maryam Habibi and Andrei Popescu-Belis. submitted. Keyword extraction and clustering for document recommendation in conversations. Manuscript submitted for publication.

Peter E. Hart and Jamey Graham. 1997. Query-free information retrieval. *International Journal of Intelligent Systems Technologies and Applications*, 12(5):32–37.

Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for Latent Dirichlet Allocation. *Proceedings of 24th Annual Conference on Neural Information Processing Systems*, 23:856–864.

Gareth J.F. Jones and Peter J. Brown. 2004. Context-aware retrieval for ubiquitous computing environments. In *Mobile and ubiquitous information access*, pages 227–243. Springer.

Jingxuan Li, Lei Li, and Tao Li. 2012. Multi-document summarization via submodularity. *Applied Intelligence*, 37(3):420–430.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the ACL 2011 (49th Annual Meeting of the Association for Computational Linguistics)*, pages 510–520.

Andrew K. McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. 2010. Discriminative topic segmentation of text and speech. In *International Conference on Artificial Intelligence and Statistics*, pages 533–540.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming Journal*, 14(1):265–294.

Andrei Popescu-Belis, Erik Boertjes, Jonathan Kilgour, Peter Poller, Sandro Castronovo, Theresa Wilson, Alejandro Jaimes, and Jean Carletta. 2008. The AMIDA Automatic Content Linking Device: Just-in-time document retrieval in meetings. In *Proceedings of MLMI 2008 (Machine Learning for Multimodal Interaction)*, LNCS 5237, pages 272–283.

Andrei Popescu-Belis, Majid Yazdani, Alexandre Nanchen, and Philip N. Garner. 2011. A speech-based just-in-time retrieval system using semantic search. In *Proceedings of 49th Annual Meeting of the ACL*, pages 80–85.

Filip Radlinski and Susan Dumais. 2006. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 691–692. ACM.

Bradley J. Rhodes and Pattie Maes. 2000. Just-in-time information retrieval agents. *IBM Systems Journal*, 39(3.4):685–704.

Stephen E. Robertson. 1997. The probability ranking principle in IR. In Karen Sparck Jones and Peter Willett, editors, *Readings in information retrieval*, pages 281–286. Morgan Kaufmann Publishers Inc.

Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. on Multimedia*, 14(3):816–832.

Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th Int. Conf. on the World Wide Web*, pages 881–890. ACM.

Saúl Vargas, Pablo Castells, and David Vallet. 2012. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 75–84. ACM.

Jun Wang and Jianhan Zhu. 2009. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122. ACM.

Shengli Wu and Sally McClean. 2007. Result merging methods in distributed information retrieval with overlapping databases. *Information Retrieval*, 10(3):297–319.

Cheng Xiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–17. ACM.

Appendix A. Transcript of a Conversation Fragment from the ELEA Corpus

The following transcript of a conversation fragment (speakers noted A through C) was submitted to the document recommender system and is exemplified in Section 6.4. The corresponding implicit queries and recommendations are respectively shown in Tables 2 and 3.

```
  A: okay I start.
  B: how  how do you want to proceed?
  A: I guess -
  C: yes what is the most important?
  A: I guess fire light.
  B: fire lighter?
  A: fire, yes.  I would say if we had something we can fire with -- I guess that
the lighter is useful in getting some sparks.
  B: hopefully.
  A: so we can use either newspaper or -- something like that.
  C: but again - first it is more important to have enough err clothes.
  A: and for me, more important to know where to go.  I would say that the compass.
  C: I mean -- if you don't have enough clothes so -- at one point you can --
  B: you can die.
  C: yes you can -- you will die.  so first issue, try to keep yourself alive and
then you can --
  A: but -- but you already have some --
  B: basics.  you everything.  you have enormous which is and so is no shoes here.
  C: okay that we have shoes so -- okay.
  B: because seventy kilometers will take you how many days?  err in the snow --
what do you think?
  A: two or three.
  B: it can be two or three days?
  C: yes, but okay you cannot always have fire with you -- but you need always
have clothes with you.  I mean it is the only thing that protects you when you are
walking.
  B: oh yes.  and erm you can make an igloo during the evening.  not that cold.
only about five degrees.  so lighting a fire is not so important.
  C: I guess fire is an extra.  I mean it is important but err for me first it is
important that when you keep walking you should be protected.
```