Diverse Keyword Extraction from 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

A new method for keyword extraction from conversations is introduced, which preserves the diversity of topics that are mentioned. Inspired from summarization, the method maximizes the coverage of topics that are recognized automatically in transcripts of conversation fragments. The method is evaluated on excerpts of the Fisher and AMI corpora, using a crowdsourcing platform to elicit comparative relevance judgments. The results demonstrate that the method outperforms two competitive baselines.

1 Introduction

The goal of keyword extraction from texts is to provide a set of words that are representative of the semantic content of the texts. In the application intended here, keywords are automatically extracted from transcripts of conversation fragments, and are used to formulate queries to a just-in-time document recommender system. It is thus important that the keyword set preserves the diversity of topics from the conversation. While the first keyword extraction methods ignored topicality as they were based on word frequencies, more recent methods have considered topic modeling factors for keyword extraction, but without specifically setting a topic diversity constraint, which is important for naturally-occurring conversations.

In this paper, we propose a new method for keyword extraction that rewards both word similarity, to extract the most representative words, and word diversity, to cover several topics if necessary. The paper is organized as follows. In Section 2 we review existing methods for keyword extraction. In Section 3 we describe our proposal, which relies on topic modeling and a novel topic-aware diverse keyword extraction algorithm. Section 4 presents the data and tasks for comparing sets of keywords. In Section 5 we show that our method outperforms two existing ones.

2 State of the Art in Keyword Extraction

Numerous studies have been conducted to automatically extract keywords from a text or a transcribed conversation. The earliest techniques have used word frequencies (Luhn, 1957), TFIDF values (Salton et al., 1975; Salton and Buckley, 1988), and pairwise word co-occurrence frequencies (Matsuo and Ishizuka, 2004) to rank words for extraction. These approaches do not consider word meaning, so they may ignore low-frequency words which together indicate a highly-salient topic (Nenkova and McKeown, 2012).

To improve over frequency-based methods, several ways to use lexical semantic information have been proposed. Semantic relations between words can be obtained from a manually-constructed thesaurus such as WordNet, or from Wikipedia, or from an automatically-built thesaurus using latent topic modeling techniques. Ye et al. (2007) used the frequency of all words belonging to the same WordNet concept set, while the Wikifier system (Csomai and Mihalcea, 2007) relied on Wikipedia links to compute a substitute to word frequency. Harwath and Hazen (2012) used topic modeling with PLSA to build a thesaurus, which they used to rank words based on topical similarity to the topics of a transcribed conversation. To consider dependencies among selected words, word co-occurrence has been combined with PageRank by Mihalcea and Tarau (2004), and additionally with WordNet by Wang et al. (2007), or with topical information by Z. Liu et al. (2010). However, as shown empirically by Mihalcea and Tarau (2004) and by Z. Liu et al. (2010) with various co-occurrence windows, such approaches have difficulties modeling long-range dependencies between words related to the same

topic. Z. Liu et al. (2009b) used part-of-speech information and word clustering techniques, while F. Liu et al. (2009a) added this information to the TFIDF method so as to consider both word dependency and semantic information. However, although they considered topical similarity, the above methods did not explicitly reward diversity and might miss secondary topics.

Supervised methods have been used to learn a model for extracting keywords with various learning algorithms (Turney, 1999; Frank et al., 1999; Hulth, 2003). These approaches, however, rely on the availability of in-domain training data, and the objective functions they use for learning do not consider yet the diversity of keywords.

3 Diverse Keyword Extraction

We propose to build a topical representation of a conversation fragment, and then to select keywords using topical similarity while also rewarding the diversity of topic coverage, inspired by recent summarization methods (Lin and Bilmes, 2011; Li et al., 2012).

3.1 Representing Topic Information

Topic models such as Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA) can be used to determine the distribution over the topic z of a word w, noted $p(z|w)$, from a large amount of training documents. LDA implemented in the Mallet toolkit (McCallum, 2002) is used in this paper because it does not suffer from the overfitting issue of PLSA (Blei et al., 2003).

The distribution of each topic z in a given conversation fragment t, noted $p(z|t)$, can be computed by summing over all probabilities $p(z|w)$ of the N words w spoken in the fragment:

$$p(z|t) = \frac{1}{N} \sum_{w \in t} p(z|w).$$

3.2 Selecting Keywords

The problem of keyword extraction with maximal topic coverage is formulated as follows. If a conversation fragment t mentions a set of topics Z, and each word w from the fragment t can evoke a subset of the topics in Z, then the goal is to find a subset of unique words $S \subseteq t$, with $|S| \leq k$, which maximzes the number of covered topics for each number of keywords k.

This problem is an instance of the maximum coverage problem, which is NP-hard. Nemhauser et al. (1978) showed that a greedy algorithm can find an approximate solution guaranteed to be within $(1 - \frac{1}{e}) \simeq 0.63$ of the optimal solution if the coverage function is submodular and monotone nondecreasing[1].

To find a monotone submodular function for keyword extraction, we used inspiration from recent work on extractive summarization methods (Lin and Bilmes, 2011; Li et al., 2012), which proposed a square root function for diverse selection of sentences to cover the maximum number of key concepts of a given document. The function rewards diversity by increasing the gain of selecting a sentence including a concept that was not yet covered by a previously selected sentence. This must be adapted for keyword extraction by defining an appropriate reward function.

We first introduce $r_{S,z}$, the topical similarity with respect to topic z of the keyword set S selected from the fragment t, defined as follows:

$$r_{S,z} = \sum_{w \in S} p(z|w) \cdot p(z|t).$$

We then propose the following reward function for each topic, where $p(z|t)$ is the importance of the topic and λ is a parameter between 0 and 1:

$$f : r_{S,z} \rightarrow p(z|t) \cdot r_{S,z}^{\lambda}\,.$$

This is clearly a submodular function with diminishing returns as $r_{S,z}$ increases.

Finally, the keywords $S \subseteq t$, with $|S| \leq k$, are chosen by maximizing the cumulative reward function over all the topics, formulated as follows:

$$R(S) = \sum_{z \in Z} p(z|t) \cdot r_{S,z}^{\lambda}\,.$$

Since $R(S)$ is submodular, the greedy algorithm for maximizing $R(S)$ is shown as Algorithm 1 on the next page, with $r_{\{w\},z}$ being similar to $r_{S,z}$ with $S = \{w\}$. If $\lambda = 1$, the reward function is linear and only measures the topical similarity of words with the main topics of t. However, when $0 < \lambda < 1$, as soon as a word is selected from a topic, other words from the same topic start having diminishing gains.

4 Data and Evaluation Method

The proposed keyword extraction method was tested on two conversational corpora, the Fisher

[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 nondecreasing* if $\forall A \subseteq B$, $F(A) \leq F(B)$.

(a) (b)

Please select one of the following options:
1. Image (a) represents the conversation fragment better than (b).
2. Image (b) represents the conversation fragment better than (a).
3. Both (a) and (b) offer a good representation of the conversation.
4. None of (a) and (b) offer a good representation of the conversation.

Figure 1: Example of a HIT based on an AMI discussion about the impact on sales of some features of remote controls (the conversation transcript is given in the Appendix). The word cloud was generated using Wordle™ from the list produced by the diverse keyword extraction method with $\lambda = 0.75$ (noted D(.75)) for image (a) and by a topic similarity method (TS) for image (b). TS over-represents the topic "color" by selecting three words related to it, but misses other topics such as "remote control", "losing a device" and "buying a device" which are also representative of the fragment.

Input : a given text t, a set of topics Z, the number of keywords k
Output: a set of keywords S
$S \leftarrow \emptyset$;
while $|S| \leq k$ **do**
  $S \leftarrow S \cup \{argmax_{w \in t \setminus S}(h(w))$ where $h(w) = \sum_{z \in Z} p(z|t)[r_{\{w\},z} + r_{S,z}]^{\lambda}\}$;
end
return S;

Algorithm 1: Diverse keyword extraction.

Corpus (Cieri et al., 2004), and the AMI Meeting Corpus (Carletta, 2007). The former corpus contains about 11,000 topic-labeled telephone conversations, on 40 pre-selected topics (one per conversation). We created a topic model using Mallet over two thirds of the Fisher Corpus, given its large number of single-topic documents, with 40 topics. The remaining data is used to build 11 artificial "conversations" (1-2 minutes long) for testing, by concatenating 11 times three fragments about three different topics.

The AMI Corpus contains 171 half-hour meetings about remote control design, which include several topics each – so they cannot be directly used for learning topic models. While selecting for testing 8 conversation fragments of 2-3 minutes each, we trained topic models on a subset of the English Wikipedia (10% or 124,684 articles). Following several previous studies, the number of topics was set to 100 (Boyd-Graber et al., 2009; Hoffman et al., 2010).

To evaluate the relevance (or representativeness) of extracted keywords with respect to a conversation fragment, we designed comparison tasks. In each task, a fragment is shown, followed by three control questions about its content, and then by two lists of nine keywords each, from two different extraction methods. To improve readability, the keyword lists are presented to the judges using a word cloud representation generated by Wordle™ (http://www.wordle.net), in which the words ranked higher are emphasized in the word cloud (see example in Figure 1). The judges had to read the conversation transcript, answer the control questions, and then decide which word cloud better represents the content of the conversation.

The tasks were crowdsourced via Amazon's Mechanical Turk (AMT) as "human intelligence tasks" (HITs). One of them is exemplified in Figure 1, without the control questions, and the respective conversation transcript is given in the Appendix. Ten workers were recruited for each corpus. An example of judgment counts for each of the 8 AMI HITs comparing two methods is shown in Table 1. After collecting judgments, the comparative relevance values were computed by first applying a qualification control factor to the human judgments, and then averaging results over all judgments (Habibi and Popescu-Belis, 2012).

Moreover, to verify the diversity of the key-

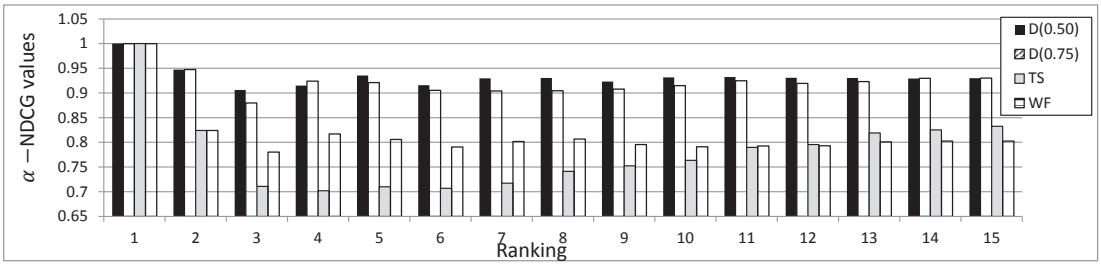

Figure 2: Average α-NDCG over the 11 conversations from the Fisher Corpus, for 1 to 15 extracted keywords.

word set, we use the α-NDCG measure (Clarke et al., 2008) proposed for information retrieval, which rewards a mixture of relevance and diversity – with equal weights when $\alpha = .5$ as set here. We only apply α-NDCG to the three-topic conversation fragments from the Fisher Corpus, relevance of a keyword being set to 1 when it belongs to the fragment corresponding to the topic. A higher value indicates that keywords are more uniformly distributed across the three topics.

Corpus	Compared methods (m_1 vs. m_2)	Relevance (%)	
		m_1	m_2
Fisher	D(.75) vs. TS	68	32
	TS vs. WF	82	18
	WF vs. D(.5)	95	5
AMI	D(.75) vs. TS	78	22
	TS vs. WF	60	40
	WF vs. D(.5)	78	22

Table 2: Comparative relevance scores of keyword extraction methods based on human judgments.

5 Experimental Results

We have compared several versions of the diverse keyword extraction method, noted D(λ), for $\lambda \in \{.5, .75, 1\}$, with two other methods. The first one uses only word frequency (not including stopwords) and is noted WF. We did not use TFIDF because it sets low weights on keywords that are repeated in many fragments but which are nevertheless important to extract. The second method is based on topical similarity (noted TS) but does not specifically enforce diversity (Harwath and Hazen, 2012). In fact TS coincides with D(1), so it is noted TS. As the relevance of keywords for D(.5) was already quite low, we did not test lower values of λ. Similarly, we did not test additional values of λ above .5 because the resulting word lists were very similar to tested values.

First of all, we compared the four methods with respect to the diversity constraint over the concatenated fragments of the Fisher Corpus, by using α-NDCG to measure how evenly the extracted keywords were distributed across the three topics. Figure 2 shows results averaged over 11 conversations for various sizes of the keyword set (1–15). The average α-NDCG values for D(.75) and D(.5) are similar, and clearly higher than WF and TS for all ranks (except, of course, for a single keyword). The values for TS are quite low, and only increase for a large number of keywords, demonstrating that TS does not cope well with topic diversity, but on the contrary first selects keywords from the dominant topic. The values for WF are more uniform as it does not consider topics at all.

HIT	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
TS more relevant	4	1	1	1	2	2	1	1
D(.75) more rel.	4	1	8	9	6	6	6	8
Both relevant	2	5	1	0	2	2	3	1
Both irrelevant	0	3	0	0	0	0	0	0

Table 1: Number of answers for each of the four options of the comparative evaluation task, from ten human judges. The 8 HITs compare the D(.75) and TS methods on 8 AMI HITs.

To measure the overall representativeness of keywords, we performed binary comparisons between the outputs of each method, using crowdsourcing, over 11 fragments from the Fisher Corpus and 8 fragments from AMI. The goal is to rank the methods, so we only report here on the comparisons required for complete ordering. AMT workers compared two lists of nine keywords each, with four options: X more representative or relevant than Y, or vice-versa, or both relevant, or both irrelevant. Table 1 shows the judgments collected when comparing the output of D(.75) with TS on the AMI Corpus. Workers disagreed for the first two HITs, but then found that the keywords extracted by D(.75) were more representative compared to TS. The consolidated rel-

evance (Habibi and Popescu-Belis, 2012) is 78% for D(.75) vs. 22% for TS.

The averaged relevance values for all comparisons needed to rank the four methods are shown in Table 2 separately for the Fisher and AMI Corpora. Although the exact differences vary, the human judgments over the two corpora both indicate the following ranking: D(.75) $>$ TS $>$ WF $>$ D(.5). The optimal value of λ is thus around .75, and with this value, our diversity-aware method extracts more representative keyword sets than TS and WF. The differences between methods are larger for the Fisher Corpus, due to the artificial fragments that concatenate three topics, but they are still visible on the natural fragments of the AMI Corpus. The low scores of D(.5) are found to be due, upon inspection, to the low relevance of keywords. In particular, the comparative relevance of D(.75) vs. D(.5) on the Fisher Corpus is very large (96% vs. 4%).

6 Conclusion

The diverse keyword extraction method with $\lambda = .75$ provides the keyword sets that are judged most representative of the conversation fragments (two conversational datasets) by a large number of human judges recruited via AMT, and has the highest α-NDCG value. Therefore, enforcing both relevance and diversity brings an effective improvement to keyword extraction.

Setting λ for a new dataset remains an issue, and requires a small development data set. However, preliminary experiments with a third dataset showed that $\lambda = .75$ remains a good value.

In the future, we will use keywords to retrieve documents from a repository and recommend them to conversation participants by formulating topically-separate queries.

Appendix: Conversation transcript of AMI ES2005a meeting (00:00:5-00:01:52)

The following transcript of a four-party conversations (speakers noted A through D) was submitted to our keyword extraction method and a baseline one, generating respectively the two word clouds shown in Figure 1.

A: The only the only remote controls I've used usually come with the television, and they're fairly basic. So uh
D: Yeah. Yeah.
C: Mm-hmm.
D: Yeah, I was thinking that as well, I think the the only ones that I've seen that you buy are the sort of one for all type things where they're, yeah. So presumably that might be an idea to
C: Yeah the universal ones. Yeah.
A: Mm. But but to sell it for twenty five you need a lot of neat features. For sure.
D: put into.
C: Yeah.
D: Yeah, yeah. Uh 'cause I mean, what uh twenty five Euros, that's about I dunno, fifteen Pounds or so?
C: Mm-hmm, it's about that.
D: And that's quite a lot for a remote control.
A: Yeah, yeah.
C: Mm. Um well my first thoughts would be most remote controls are grey or black. As you said they come with the TV so it's normally just your basic grey black remote control functions, so maybe we could think about colour? Make that might make it a bit different from the rest at least. Um, and as you say, we need to have some kind of gimmick, so um I thought maybe something like if you lose it and you can whistle, you know those things?
D: Uh-huh. Mm-hmm. Okay. The the keyrings, yeah yeah. Okay, that's cool.
C: Because we always lose our remote control.
B: Uh yeah uh, being as a Marketing Expert I will like to say like before deciding the cost of this remote control or any other things we must see the market potential for this product like what is the competition in the market? What are the available prices of the other remote controls in the prices? What speciality other remote controls are having and how complicated it is to use these remote controls as compared to other remote controls available in the market.
D: Okay.
B: So before deciding or before finalising this project, we must discuss all these things, like and apart from this, it should be having a good look also, because people really uh like to play with it when they are watching movies or playing with or playing with their CD player, MP three player like any electronic devices. They really want to have something good, having a good design in their hands, so, yes, all this.

Acknowledgments

The authors are grateful to the Swiss National Science Foundation for its financial support through the IM2 NCCR on Interactive Multimodal Information Management (see www.im2.ch).

References

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)*.

Jean Carletta. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2):181–190.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher Corpus: a resource for the next generations of speech-to-text. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*, pages 69–71.

Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666.

Andras Csomai and Rada Mihalcea. 2007. Linking educational materials to encyclopedic knowledge. *Frontiers in Artificial Intelligence and Applications*, 158:557.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI 1999)*, pages 668–673, Stockholm, Sweden.

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)*, page 15.

David Harwath and Timothy J. Hazen. 2012. Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5073–5076. IEEE.

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.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 216–223, Sapporo, Japan.

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 49th Annual Meeting of the ACL*.

Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. 2009a. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the ACL (HLT-NAACL)*, pages 620–628.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2009b. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 257–266.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 366–376.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Yutaka Matsuo and Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1):157–169.

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

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 404–411, Barcelona.

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.

Ani Nenkova and Kathleen McKeown, 2012. *A Survey of Text Summarization Techniques*, chapter 3, pages 43–76. Springer.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management Journal*, 24(5):513–523.

Gerard Salton, Chung-Shu Yang, and Clement T. Yu. 1975. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.

Peter Turney. 1999. Learning to extract keyphrases from text. Technical Report ERB-1057, National Research Council Canada (NRC).

Jinghua Wang, Jianyi Liu, and Cong Wang. 2007. Keyword extraction based on PageRank. In *Advances in Knowledge Discovery and Data Mining (Proceedings of PAKDD 2007)*, LNAI 4426, pages 857–864. Springer-Verlag, Berlin.

Shiren Ye, Tat-Seng Chua, Min-Yen Kan, and Long Qiu. 2007. Document concept lattice for text understanding and summarization. *Information Processing and Management*, 43(6):1643–1662.