Identifying Sentence-Level Semantic Content Units with Topic Models

Leonhard Hennig, Thomas Strecker, Sascha Narr, Ernesto William De Luca, Sahin Albayrak
Distributed Artificial Intelligence Laboratory (DAI-Lab)
Technische Universität Berlin, Germany
{*leonhard.hennig,thomas.strecker,sascha.narr,ernesto.deluca,sahin.albayrak*}*@dai-labor.de*

***Abstract*—Statistical approaches to document content modeling typically focus either on broad topics or on discourse-level subtopics of a text. We present an analysis of the performance of probabilistic topic models on the task of learning sentence-level topics that are similar to facts. The identification of sentential content with the same meaning is an important task in multi-document summarization and the evaluation of multi-document summaries. In our approach, each sentence is represented as a distribution over topics, and each topic is a distribution over words. We compare the topic-sentence assignments discovered by a topic model to gold-standard assignments that were manually annotated on a set of closely related pairs of news articles. We observe a clear correspondence between automatically identified and annotated topics. The high accuracy of automatically discovered topic-sentence assignments suggests that topic models can be utilized to identify (sub-)sentential semantic content units.**

***Keywords*-text summarization, topic models, latent dirichlet allocation**

I. Introduction

In the field of multi-document summarization (MDS), one of the major challenges is the identification of the core content of a set of thematically related documents. Each document addresses a set of topics related to a main theme, and different documents may address the same or similar topics. Identifying this set of topics and their distribution across documents helps to determine the relative importance of each topic for a summary. Furthermore, as most summarization systems rely on sentence extraction to avoid the problem of having to generate coherent and cohesive natural language text [1], it is desirable to identify topics at the sentence level and not only at the document level.

But what exactly are the subtopics of a document, or a set of documents? In text categorization, documents typically are considered to belong to one or more rather broad categories, such as "Sports" or "Politics". Clearly these document-level types of topics are not very useful in MDS, since input documents in MDS are closely related, and hence would belong to the same categories. A more fine-grained notion of topics is discussed by Barzilay and Lee [2]. They argue that texts from the same domain exhibit an observable structure to facilitate reading comprehension. For example, news articles about earthquakes will usually discuss earthquake magnitude and location, the number of victims, and rescue efforts in the aftermath of the quake. This view of a topic considers information on a sentence level and assigns each sentence to exactly one topic. Multiple subsequent sentences share a common topic label, and a document is represented as a sequence of topics. Various authors have adopted this definition of a topic successfully in single- and multi-document summarization [3]–[5].

But one could take an even more fine-grained perspective, and consider a topic to be a piece of information, similar to a single fact. This definition of a topic relates to that of Summary Content Units [6] or factoids [7]. Each sentence is assumed to relate one or more such facts. For example, the sentence "A small plane carrying John F. Kennedy Jr., son of the former U.S. president, was reported missing early Saturday." gives information that a plane with JFK Jr. on board is missing, and that he was the son of the U.S. president, among others. This type of topic considers information on a subsentential level, and represents each sentence as a mixture of topics. MDS datasets typically consist of news articles that discuss the same or similar facts, but express them using slightly different words and phrases, or by combining facts differently into sentences.

Based on these observations, in this work we examine if a probabilistic topic model of text can be utilized to capture such facts. Topic models are generative latent variable models that can reveal the hidden structure of datasets [8], [9]. Their main claim, when applied to text, is that they can extract a set of meaningful topics, where each topic is represented as a distribution over words, from word-document co-occurrence observations. Our intuition is that we can identify topics similar to facts by modeling patterns of word usage at the sentence level:

- We begin our evaluation by analyzing a set of closely related news article pairs chosen from the MDS task of the 2007 Document Understanding Conference (DUC)[1] to identify sentential clauses with similar word usage that express repeated and unique content. We annotate four different types of topics (Section II) and construct a gold-standard set of topics and topic-sentence assignments for each pair of documents (Section III).
- We then train a Latent Dirichlet Allocation (LDA) topic model [8] on the term-sentence co-occurrence matrix of each pair of documents, and evaluate the quality

[1] http://duc.nist.gov

of the LDA topics and topic-sentence associations using our gold-standard annotations. It turns out that the overall performance of the model is surprisingly good even given only very little contextual information (Section IV). An analysis of the model's performance with respect to the different topic types shows that repeated clauses are the most difficult to identify.

We review related work in Section V and give a conclusion and an outlook on future work in Section VI.

II. Types of Sentence-Level Topics

The goal of our analysis is to determine the suitability of a probabilistic topic model for identifying (sub-)sentential word usage patterns in a set of related documents. In this section we will introduce the types of patterns we want our model to discover. We assume that each sentence addresses one or multiple topics. Some sentences may repeat information contained in another sentence. Others may repeat only parts, or combine parts of other sentences. The words in many sentences will reflect the main theme of the document, and thus there will be some words that are very common across sentences.

Since a topic model is based on co-occurrence data, it obviously cannot model distinct topics for information occurring in a single sentence only. In addition, if a topic is expressed once as a full sentence s_1, and once as part of second sentence s_2, the remaining words of s_2 will be assigned to the same topic as those of s_1 if they do not occur in any other sentence. The model will therefore only be suitable for learning topics for the following types of content:

- Copies: Sentences that are verbatim copies of a sentence in another document
- Similar: Sentences that express the same content with very similar word usage
- Clause: Sentence parts that are repeated as parts of another sentence or as a full sentence, with similar or identical word usage
- Unique: Sentences that express unique information not repeated in any other sentence

The first type of topic is trivial to identify, as both sentences exhibit the same word pattern. The second type of topic should also be easy to learn, since there is a large overlap in terms of words and phrases, for example:

> (a) *The Supreme Court struck down as unconstitutional a law* giving *the president a line-item veto which lets him* cancel *specific items in tax and spending measures.*
> (b) *The* U.S. *Supreme Court* Thursday *struck down as unconstitutional the line-item veto law that* lets the U.S. *president* strike out *specific items in tax and spending measures.*

The third type is intended for longer phrases that are copied or repeated with similar wording in multiple sentences:

> (a) Germany, Azerbaijan, Greece, France, the Netherlands, Kazakhstan, Ukraine and Russia have been participating in the fight against the blaze that threatened to engulf the entire field of *30 storage tanks containing 1 million tons of crude oil.*
> (b) However, he said the strong fire had destroyed seven storage tanks and damaged two other ones in the refinery which held *30 storage tanks containing 1 million tons of crude oil.*

These two sentences share a longer phrase expressing the fact that there are "30 storage tanks containing 1 million tons of crude oil". Given that the information contained in the remainder of the source sentences also occurs in some other sentences, a topic model could e.g. identify three topics z_1, z_2 and z_3. Topics z_1 and z_2 would be assigned to s_1, topics z_1 and z_3 to s_2, and the word distribution of topic z_1 would correspond to the words of the shared phrase. Finally, sentences with word patterns that are not repeated in any other sentence constitute the fourth type of topic.

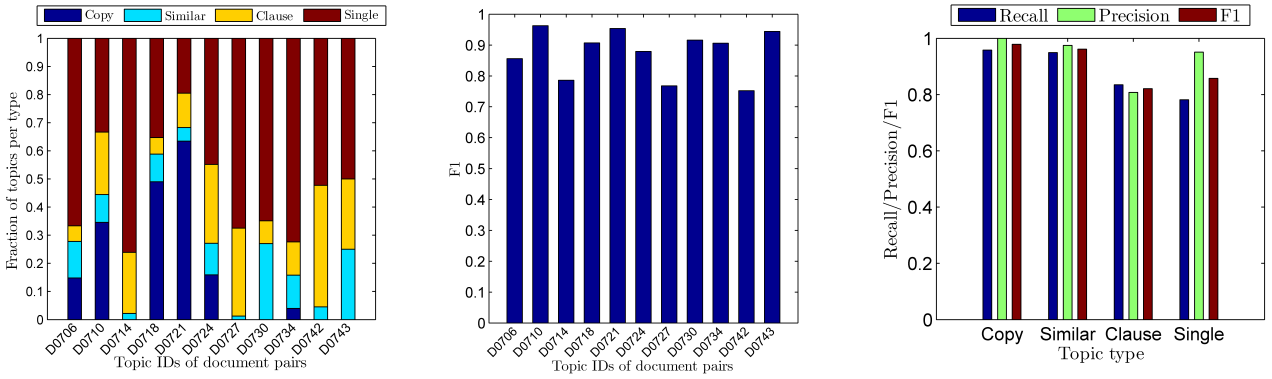
III. Constructing Gold-Standard Topics

Topic models are typically applied to larger text corpora, and on a document level. We, however, intend to model topics on a subsentential level, derived from term-sentence co-occurrence data. This data is very sparse, as most words occur only once per sentence, and the number of words in a sentence is usually very low with respect to the total vocabulary size. We therefore perform our analysis on pairs of input documents which report the same news event, are very similar in terms of word usage, and which were written around the same date. Choosing such closely related documents has the advantage of allowing for the occurrence of all the topic types we are interested in, while at the same type reducing the amount of topics to be discovered.

We selected document pairs from 11 different DUC 2007 document clusters by first computing the pair-wise cosine similarity $sim(d_i, d_j)$ of all documents of a cluster[2], and then choosing the most similar document pair with $sim(d_i, d_j) \leq 0.85$. The upper bound on the similarity was introduced to avoid selecting documents which are copies or minor revisions of each other. On average, each document pair contained 34.3 sentences and had a vocabulary of 169 words, which occurred a total of 393 times.

Three different annotators identified topics for these document pairs, with at least two annotators per document pair. Six document pairs were processed by all human annotators. The output of each annotation is a matrix $\widehat{\Theta}^{(a)}$ of topic-sentence assignments, where $\widehat{\Theta}_{ij} = 1$ if topic i

[2]We removed stop words and performed stemming with the NLTK toolkit; http://www.nltk.org



(a) Annotated topics by type (b) Topic model performance on document pairs (c) Performance for different topic types

Figure 1. (a) Distribution of number of annotated topics per type for selected document pairs, averaged over annotators. (b) F_1 scores of the topic model per document pair, averaged over matching topics and annotators. The mean F_1 score across all document pairs is 0.88. (c) Performance of LDA model for different types of gold-standard topics.

is associated with sentence j, and (a) is the annotator. Sentences may belong to multiple topics, and topics are defined as a combination of sentences. For each topic $\hat{z}_k$, we estimate its distribution over words $p(w|\hat{z}_k)$. We compute $p(w_i|\hat{z}_k)$ as the relative frequency of word w_i in the union of topic $\hat{z}_k$'s sentences. We denote the resulting matrix as $\widehat{\Phi}^{(a)}$. This simplified approach will not result in completely adequate word distributions for topics of type *Clause*, since all words – and not only the words of the clause – will be included in the distribution.

Interannotator agreement is high: The mean pairwise Pearson correlation of annotators on the number of topics is 0.97. The mean percentage of full topic matches between annotators is 0.69, i.e. the annotators agreed fully on 69% of the topics[3].

Figure 1(a) shows the distribution of topics per type for all document pairs, averaged over annotators. We see that for some document pairs, such as for D0718 and D0721, the majority of topics are of type *Copy* or *Similar*. On the other hand, some document pairs do not share any or only very few similar or copied sentences, e.g. D0727 and D0742.

IV. Learning Sentence-level Topic Models

A topic model is a generative latent variable model that views each topic as a distribution over words. Each document is represented as a mixture of topics. For our analysis, we use the LDA model introduced by [8]. In this model, each document is generated by first choosing a distribution over topics $\theta^{(d)}$, parametrized by a conjugate Dirichlet prior α. Subsequently, each word of this document is generated by drawing a topic z_k from $\theta^{(d)}$, and then drawing a word w_i from topic z_k's distribution over words $\phi^{(k)}$. ϕ is parametrized by a conjugate Dirichlet prior β. For T topics, the matrix Φ specifies the probability $p(w|z)$ of words given topics, and the matrix Θ specifies the probability $p(z|d)$ of topics given documents. To estimate Φ and Θ from a set of documents, we employ Gibbs sampling [9].

Since we are interested in modeling topics for sentences, we treat each sentence as a document. We construct a matrix A for each document pair, using word-sentence co-occurrence observations and preprocessing sentences as described before. Each entry A_{ij} corresponds to the frequency of word i in sentence j. In our input data, the majority of these frequencies is 1. We run the Gibbs sampling algorithm on A, setting the parameter T, the number of latent topics to learn, equal to the number of manually annotated topics.

Since we want to learn a topic model with a structure that reflects the type of topics defined in Section II, the topic distribution for each sentence should be peaked toward a single or only very few topics. To ensure that the topic-specific word distributions $p(w|z)$ as well as the sentence-specific topic distributions $p(z|d)$ behave as intended, we set $\alpha = 0.01$ and $\beta = 0.01$. This enforces a bias toward sparsity, resulting in more peaked distributions [10]. A low value of β also favors more fine-grained topics [9]. The exact values of the parameters were determined experimentally on the D0742 document pair. We run the Gibbs sampler for 2000 iterations, and collect a single sample from the resulting posterior distribution over topic assignments for words. From this sample, we compute the conditional distributions $p(w|z)$ and $p(z|d)$.

A. Matching annotated and LDA topics

In order to compare the topic-sentence associations computed by the LDA topic model with the gold-standard assignments, we have to match the LDA topics to the manually annotated topics. We consider topics as similar if their word distributions are similar. We therefore compute the pair-wise Jensen-Shannon (JS) divergence between columns of Φ and

[3]Computing a more sophisticated agreement measure such as Krippendorf's alpha is desirable, but problematic due to the varying total number of topics identified by each annotator and the problem of partial matches.

Table I
EXAMPLE MATCHES OF LDA AND GOLD-STANDARD TOPICS

	LDA Topic 5	Topic 2	LDA Topic 8	Topic 4
Top Terms	blaze	30	oil	storag
	engulf	azerbaijan	crude	tank
	entir	blaze	tank	1
	field	engulf	ton	30
	fight	entir	storag	crude

$\widehat{\Phi}^{(a)}$. Topics from $\widehat{\Phi}^{(a)}$ are matched to topics of Φ on the basis of this dissimilarity using a simple greedy approach. We reorder the rows of Θ according to the computed matching. Table I shows the most likely terms for some example topic matches. The first topic captures the fact that different countries helped to fight the blaze that threatened to engulf the entire field of 30 storage tanks, the second lists words related to the fact that the storage tanks contained 1 million tons of crude oil.

During our experiments, we observed that the Gibbs Sampler did not always use all the topics available. Instead, some topics had a uniform distribution over words, i.e. no words were assigned to these topics during the sampling process. Since we set T, the number of topics, to the number of manually annotated topics, this indicates that some of the manually annotated topics cannot be discovered from the available data. On average, 15.2% of the sampled topics had a uniform word distribution, and were therefore discarded before the matching step.

B. Evaluation

To compare the topic distributions $p(z|d)$ with the gold-standard topic-sentence assignments $\widehat{\Theta}^{(a)}$, we binarize Θ to give $\Theta^{'}$ by setting all entries $\Theta^{'}_{ij} = 1$ if $\Theta_{ij} > 0.1$, and 0 otherwise[4]. We can now compute precision, recall and F_1-scores for each topic. Averaged over all assignments, these measures give us an indication of how well the LDA model captured the topics we are interested in.

Figure 1(b) shows the F_1 scores for each of the 11 document pairs. All values are averaged over topics and annotators. We see that LDA topics correspond quite well to the manually annotated topics. The mean F_1 score, calculated over all document pairs and annotators, is 0.88. The precision of the topic-sentence assignments is consistently higher than recall for all document pairs, with the average precision being 0.92, and average recall 0.83 (not shown here). The results of our analysis suggest that a probabilistic topic model can successfully discover sentence- and clause-level topics.

Figure 2 shows some example topic matchings. Each cell displays the JS divergence of the word distributions of an LDA topic (columns) compared to a gold-standard topic (rows). On the diagonal, the best-matching topics are ordered by increasing JS divergence. Multiple points with low JS divergence in a single row, observed e.g. in row 4 of document pair D0706, indicate that more than one LDA topic was very similar to this gold-standard topic. The graphs show a clear correspondence between many LDA and gold-standard topics.

For some document pairs, the LDA topics have a precision of close to 1 (D0718, D0721, D0734, D0743, not shown). Topic models for document pairs that contain many *Clause* topics seem to be more difficult to learn. This is indicated by the relatively low F_1 scores for document pairs D0727 and D0742. On the other hand, the F_1 score for document pair D0743, for which approximately a quarter of the annotated topics is based on shared clauses, is among the best of all models.

An evaluation of the performance of topic models with respect to the different types of gold-standard topics confirms our intuition that topics of type *Clause* are the most difficult to identify. Figure 1(c) shows that the F_1 scores of topics corresponding to copied or similar sentences is very high. For topic type *Clause*, however, the F_1 score is only about 0.8. There are two main reasons for this lower score: First, this type of topic must deal with 'noise' in the form extra words in the enclosing sentences. Second, the word distributions of the gold-standard topics of this type are not adequate, as explained in Section III. The greedy matching process intuitively prefers matching clearly defined topic-word distributions, and the noise introduced by the extra words may well dilute the word distribution of the topics too much in order for them to be matched correctly. A better modeling of the gold-standard word distributions is therefore necessary to show the real performance of this topic type.

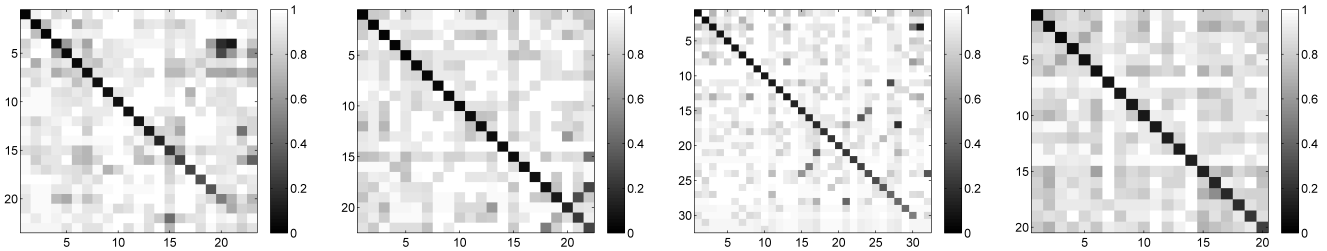
V. RELATED WORK

Topic models have been successfully applied to a variety of tasks [8], [9], [11]. In text summarization, most topic modeling approaches utilize a term-sentence co-occurrence matrix, but learn topics at the document level. Each sentence is assigned to exactly one topic, and a topic is a cluster of multiple sentences [3]–[5].

The authors of [12] propose to identify shared phrases in a set of similar sentences. The model uses manually constructed paraphrasing rules to merge sentences which express the same or similar content. Paraphrasing rules consider linguistic information, such as sentence constituent ordering. In contrast, our approach does not rely on manually constructed rules, and also models topics for semantically distinct content.

Marcu introduces another linguistically motivated approach for the identification of subsentential content units in [13]. He shows that the nuclei of parse trees that are based on Rhetorical Structure Theory can be used to construct a

[4]Since the LDA algorithm learns very peaked distributions, the actual value of this threshold does not have a large impact on the resulting binary matrix and subsequent evaluation results.



(a) D0706 (b) D0710 (c) D0724 (d) D0734

Figure 2. Pairwise Jensen-Shannon divergence of word distributions of manually annotated topics and LDA topics. Matching topics are ordered by increasing divergence along the diagonal, using a simple greedy algorithm. The examples show a clear correspondence of LDA topics to the gold-standard topics.

summary. The work focuses on an evaluation of the correlation between the importance of parse tree elements and their salience for a summary. An identification of semantically similar parse tree elements, e.g. to avoid redundancy, is not performed.

Some summarization evaluation methods, such as the Pyramid method [6], or the factoid approach proposed by [7], also identify subsentential content units with the same meaning. Due to the difficulty of this task, it is currently performed manually.

VI. Conclusion

We have presented an analysis of the learnability of subsentential and sentence-level topics that represent content with the same meaning. Our analysis assumes that each sentence relates one or more facts, and different sentences utilize similar words and phrases to express the same facts. We have evaluated our approach on a set of closely related pairs of news articles for which we manually identified gold-standard topics. We showed that a probabilistic topic model can learn topics with word distributions that are similar to the word distributions of manually identified topics. The topics are discovered in a completely unsupervised fashion, using no information except the distribution of the words themselves. We observed a clear correspondence between automatically derived and annotated topics. An evaluation of the discovered topic-sentence assignments revealed a surprisingly high agreement with gold-standard assignments.

The results of our analysis suggest that topic models, with their shallow statistical approach to semantics, can be utilized to identify sentence-level topics which are similar to facts. Our approach has many interesting applications. For example, it can be seen as a step toward the automatic acquisition of Summary Content Units (SCU) used in the Pyramid summarization evaluation method. In future work, we therefore intend to investigate the correspondence of the topics of an topic model trained on human summaries with manually annotated SCUs.

References

[1] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, "Multi-document summarization by sentence extraction," in *NAACL-ANLP WS on Automatic summarization*, 2000.

[2] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," in *Proceedings of HLT-NAACL '04*, 2004.

[3] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-Document summarization using sentence-based topic models," in *Proceedings of ACL-IJCNLP '09*, 2009, pp. 297–300.

[4] J. Tang, L. Yao, and D. Chen, "Multi-topic based query-oriented summarization," in *Proceedings of SDM '09*, 2009.

[5] L. Hennig, "Topic-based multi-document summarization with probabilistic latent semantic analysis," in *Proceedings of RANLP '09*, 2009.

[6] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The Pyramid method," in *Proceedings of HLT-NAACL*, 2004.

[7] S. Teufel and H. V. Halteren, "Evaluating information content by factoid analysis: human annotation and stability," in *Proceedings of EMNLP*, 2004.

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *J. of Machine Learning Research*, vol. 3, 2003.

[9] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl 1, 2004.

[10] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning*, T. Landauer, S. D. McNamara, and W. Kintsch, Eds., 2007.

[11] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of EMNLP '08*, 2008, pp. 363–371.

[12] R. Barzilay, K. McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in *Proceedings of the ACL '99*, 1999, pp. 550–557.

[13] D. Marcu, "The rhetorical parsing of natural language texts," in *Proceedings of the ACL '97*, 1997, pp. 96–103.