Hierarchical text segmentation for topic modeling
Literature review

Bin Wang BWANG4@STANFORD.EDU
STANFORD UNIVERSITY

Mike Percy MP81@STANFORD.EDU
STANFORD UNIVERSITY

1. Thesis

The major goal of this project is to explore a novel topic extraction method, where topics within a document are explicitly mapped to corresponding document segments. Topics may be nested, i.e. a topic may compose of smaller topics. Another difference between the targeting methods and previous topic extraction methods such as LDA is that the topics are predefined with large text corpus such as Wikipedia.

Our hypothesis is that there exists an universal set of topics in any language. Any document can be decomposed into such a set of the universal topics. And the structure and set of the topics characterizes the document and can be used as features for later tasks such as classification.

This project is divided into three parts. The first part is to extract a set of the universal topics from a large text corpus such as Wikipedia, and then figure out a good model for each topics. The second part is to design a topic identification algorithm for any given document with the predefined set of universal topics. The proposed algorithm should be able to detect the topic transition within the document and assign a proper topic to each document segment. The third part of this project is evaluation. The topics found for each document will be used as input features for the classifier.

2. Literature summary

Here we provide an overview of and references for papers that are included in this review.

Prior work along the lines of what we are investigating was recently done at UIUC involving transformation of documents to a labeled feature space to perform classification without the use of explicit training labels on the original documents (Song and Roth, 2014). The labeled features here are actually the topics we mentioned in our hypothesis. The topics in the work were obtained by running ESA (Gabrilovich and Markovitch, 2007), which we describe here and plan to use as a starting point in our work. An alternative approach to representing topics is Brown clustering (Brown et al., 1992).

Besides representing topics as word frequency vectors (ESA) or word clusters (brown clustering), a generative model as used in LDA (Blei et al., 2003) may also be a good choice. Here, each topic will be represented as a generative model for a word to appear.

Since topic identification in text is very similar to object identification in images, the newly proposed selective search (Uijlings et al., 2013) for image object identification may be adapted to text. A segment proposal algorithm is required by the selective search algorithm. Hence the text segmentation algorithm proposed in (Sun et al., 2008) may be also useful.

There has been recent work involving the classification of text using images as labels (Young et al., 2014), which could potentially have insights applicable to our problem space.

Much recent work has been done on global word vectors, which we could consider using to build up hierarchical similarity clusters of words (Huang et al., 2012). The authors also provide a potentially interesting data set for use in the topic modeling evaluation task. It may also be used to improve Brown's clustering algorithm and then provide a good topic representation scheme.

3. Literature review

3.1 Song and Roth (2014)

This recent paper tries to use a semantic mapping on labels and documents to do classification. The authors use Wikipedia articles as a set of topics and represent each topic with a certain model (TF-IDF model with ESA or word cluster model with Brown clusters). It takes a the description (page content) for each Wikipedia topic page and converts it into a TF-IDF vector (ESA model) or a word cluster vector (Brown clustering) as the semantic "understanding" of the topic label. Upon classification, the authors likewise map a given document into such a vector space and compute the similarity between the document vector and each label vector. The most similar label will be the classification result.

This approach is inspiring because it proposes a simple approach to semantic representation. We can basically represent any text segment as a vector given a set of topics. In this way, it is possible to assign part of the document with a named topic (a Wikipedia topic page). As such, we can represent a document as a hierarchical set of topics corresponding to regions in the document. This is exactly we aim to do in our project.

3.2 Gabrilovich and Markovitch (2007)

Explicit Semantic Analysis (ESA) is an approach to scoring documents by topic using a Wikipedia taxonomy. The topics in the taxonomy are Wikipedia pages, such as `Jaguar (car)`. The first step in the algorithm is to remove topics that likely have little semantic significance, such as extremely short pages and pages whose topics are just lists of things. The next step is to run TF-IDF (term frequency $\cdot$ inverse document frequency) for each non-stopword in each Wikipedia page. We store these TF-IDF scores in an inverted index from word to topic plus the score.

Next, we run the same TF-IDF operation on the document we wish to classify. We perform essentially an "inner join" between the words in the document and the words in the topic index. For each word that matches between the document and the topic index, a score is calculated by multiplying together the TF-IDF score for that word in the document and the topic index. To rank topic relevance for a given document, scores for all matching words are summed. Another possible topic scoring metric is cosine similarity.

As our project needs a good and simple representation of topics, ESA is a perfect candidate. It showed superior performance over other methods in Song and Roth (2014). It can also be applied for text segment representation.

3.3 Brown et al. (1992)

Brown proposes a method of clustering words by their co-occurrence within documents. The paper proposes two approaches to cluster words.

One way is to group words into clusters that have similar syntactic functions. In a n-gram language model, the Brown et al. assign each word (w_i) a class (c_i) and let a n-gram sequence holding the following property $Pr(w_k|w^{k-1}) = Pr(w_k|c_k)Pr(c_k|c^{k-1})$. It is clear that this assumption implies that each word within a cluster holding similar probabilities to appear in a n-gram sequence which let this clustering make sense. It is worth mentioning that the proposed algorithm gives an hierarchical tree of word clusters, and each word is represented as the path from the root to the leaf where the word resides.

The other way is to cluster words according to their "stickiness". The "stickiness" here refers to whether two words are always close to each other when they show up in a text segment. The words that tends to shows up close to each other are clustered.

This clustering algorithm is relevant to this project since it provides a potential way of representing text segments. In Song and Roth (2014), each text segment was represented as the average of all words in it. Noting that each word in brown clusters is represented as a path vector, the text segments are also represented as vectors whose similarities can be easily calculated.

3.4 Uijlings et al. (2013)

It is interesting to observe that identifying topics within a document with their corresponding locations has a lot of similarity to object identification within images. Therefore, adapting image object identification algorithms to text should be a reasonable approach to consider.

The object identification algorithm proposed in Uijlings et al. (2013) is one of the best choices. It first divides an image into small regions using some heuristics and then groups these regions recursively by their similarity until they all merge into one single region. All of the intermediate regions are recorded as potential regions where an object may exist. After the potential regions are proposed, a pre-trained classifier will be run on these regions to see whether the corresponding region contains an object or not.

We will apply a similar process on text segments in this project, however the similarity measure between text segments, proposal of initial regions, and the algorithms for deciding

whether a topic exists in a text segment still need to be designed. Other related papers are reviewed to address these problems.

3.5 Blei et al. (2003)

Latent Dirichlet Allocation (LDA) has been a popular topic extraction method since its proposal. Many methods which are used in this project deploy LDA as an essential step. Hence, it is quite important to briefly review LDA. Another major motivation for us to review this paper is that both our approach and LDA are topic extraction algorithms, so reviewing how LDA represents topics and is evaluated can be helpful for us to schedule our project and draft our paper.

LDA models the generation of a document as a generative process. For each document, a set of topics with existence probabilities are generated. For each word, a topic is chosen and the word is generated according to the topic. Each topic is modeled by the probability for each word to show up.

This model is constructed in a pretty simple and elegant way, but the number of topics needs to be specified by the user and the topics are proposed according the given training texts. It may be hard for humans to figure out what the topics are, and it is also highly possible that the topics do not make any sense due to the limited size of a training dataset. Moreover, it ignores the fact that topics are not specific to each text corpus – topics may be discussed in different situations with different purposes, but the possible topics are normally the same everywhere.

The original LDA paper evaluated LDA with several different tasks – document modeling, document classification and collaborative filtering. It also provided an example of how LDA models a document so that a human can see whether it makes sense or not. This is quite inspiring. The algorithm proposed in this project may be evaluated in a similar way. Depending on how well the project goes, we may also make some attempt to address more ambitious tasks such as document summarization.

3.6 Sun et al. (2008)

As described earlier, text segmentation is one of the core topics we need to address in this project. Sun et al. (2008) provides a very interesting text segmentation algorithm that may be useful.

In this paper, LDA is run against all proposed segments to extract the potential topics. Then, a Fisher Kernel $K(b_1, b_2)$ is used to measure the similarity between two segments b_1 and b_2. With different segmentation choices, different segments will be generated, and therefore the segments will each have different topic distributions. Clearly, a good topic segmentation will result in adjacent segments having lower similarity than alternative segmentations, which implies a small Fisher Kernel value, and also a smaller segment length variance. The authors define their cost function as $J(\vec{t}; \lambda) = \sum_{m=1}^{M} \lambda F(l_{t_m+1,t_{m+1}}) + K(b_{t_{m-1}+1,t_m}, b_{t_m+1,t_{m+1}})$, where t_m is the position of the mth segmentation, b represents the block on the corresponding position, $F(l_{t_m+1,t_{m+1}}) = \frac{(l_{t_m+1,t_{m+1}}-\mu)^2}{2\sigma^2}$ which represents

the segment length variance, K is the Fisher Kernel, and λ is a constant. The best segmentation can be found by minimizing this cost function. This can be a good heuristic for us to find good candidate regions.

3.7 Young et al. (2014)

Young, et al. propose a method to measure semantic similarity between two images based on the text captions for the images. In their data set, each image is captioned multiple times by different writers. They first build a subsumption graph on all of the images and their associated strings based on a semantic analysis of the captions, after multiple preprocessing steps, including POS tagging and hypernym expansion based on WordNet. The subsumption graph here refers to a graph where the parents are more general than their children.

With this graph, evaluation was performed using two tasks: caption entailment and text semantic similarity. Only the text semantic similarity is related to our project. For any caption, they are able to propose a set of images that may be described by the caption. The similarity between two captions is measured using two methods. One is based on the intersection size of the two image sets that can be described by each caption. The other similarity measurement is a form of PMI that is also based on images described by each caption.

This paper proposes a novel method for similarity calculation, which is also a core part of our project when we need to propose potential regions within a document for a topic to raise in the second part of our project. The fact that they use a form of semantic similarity is particularly appealing for topic classification.

3.8 Huang et al. (2012)

Word representation and sense disambiguation are also concerns of ours. With better word representation methods, each document or text segment can be more accurately represented, leading to a better topic identification result. It may also be a good substitute approach for Brown clustering.

Huang et al. (2012) proposed a good improvement that deals with sense disambiguation. In Young's paper, a neural network which composes two sub networks was proposed. One neural network takes an n-gram word sequence, s, as input, where each word is represented as a vector and the whole sequence is represented as word vector concatenations. The other network takes the whole document word sequence, d, as input. The final output score of the whole network is the sum of the two sub-networks. Then, for each n-gram sequence, the last word may be replaced randomly, where the new n-gram sequence is s^w and the corresponding document is d^w, so that there are more training inputs. The word vectors and the neural network weights are updated at the same time so that the final score for the original s is at least larger than the final score of mutated sequence s^w by 1 for any w.

The n-gram sequence is noted as local context and the whole document is noted as global context in the paper. This method deals with sense disambiguation by clustering the local contexts of each word and assigns each context with a different sense.

GloVe (Pennington et al., 2014) is the most successful global word representation algorithm, and has similarities to Huang et al. (2012) in terms of approaching the problem of representing words using a global feature vector.

The proposed methods outperform other algorithms, hence could be potentially good word representation schemes in this project.

4. Comparison Between ESA, Brown Clustering, and LDA

Since we are trying to invent a novel algorithm, many of the papers we have reviewed are only tangentially related to each other, and are therefore difficult to compare. They serve our project in different ways. However, ESA and Brown clustering are similar to each other since they both provide a statistical vector representation of a topic.

The primary difference is that ESA represents a topic with a word frequency vector, while Brown clustering is a word clustering algorithm and thus the final outputs are averages of the word vectors within the document. It is apparent that the word frequency approach should work better since it takes aspects of the entire corpus (IDF) into account. The Brown word average approach considers each word to be equally important, which is a false assumption.

The actual result in Song and Roth (2014) is consistent with this analysis. ESA outperforms Brown clustering in their evaluation.

Our third model, the LDA model, differs from the previous two models in that it is a generative model. Generative models are typically less accurate than discriminative models due to their strong assumptions on independence for each word they have generated. However, all three of these models have a bag-of-words (BoW) assumption, where word ordering is ignored. Thus, there is no significant drawback to using LDA to generate a semantic representation of a topic and it is a worth trying this approach as well. We may also compare "vanilla" LDA to our hierarchical model in our experiments.

5. Conclusions

To sum up, our project aims to represent a document with a hierarchical topic tree where the topics are obtained from a large third party text corpus such as Wikipedia. The first step is to model each topic using its associated text description. Brown et al. (1992), Gabrilovich and Markovitch (2007) and Blei et al. (2003) can be used to model topics. Next, we need to propose text regions in the document that may potentially contain different topics. Uijlings et al. (2013) and Sun et al. (2008) are good references for this task. While proposing the regions, we need to combine finer regions by the similarity between them. Young et al. (2014) proposes an inspiring approach. Finally, we need to evaluate our method. Document classification and document summarization are good tasks to test with. Blei et al. (2003) also provides a great example of how to evaluate a topic extraction algorithm.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4): 467–479, 1992.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12, 2014.

Yangqiu Song and Dan Roth. On dataless hierarchical text classification. In *Proceedings of AAAI*, 2014.

Qi Sun, Runxin Li, Dingsheng Luo, and Xihong Wu. Text segmentation with lda-based fisher kernel. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 269–272. Association for Computational Linguistics, 2008.

Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104 (2):154–171, 2013.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.