AUTOMATIC EXTRACTION OF CONCEPTUAL LABELS FROM TOPIC MODELS

Claudiu MUŞAT[1], Ştefan TRĂUŞAN-MATU[2], Julien VELCIN[3], Marian-Andrei RIZOIU[4]

În această lucrare prezentăm un sistem destinat extragerii automate de etichete conceptuale pentru topice obținute prin metode statistice. Realizând o proiecție a unei distribuții peste toate cuvintele din vocabular pe ontologia WordNet reuşim asocierea de concept unor grupuri de cuvinte extrase folosind modele de topice. Contribuțiile cele mai importante ale lucrării sunt legate de validarea rolului acestor concepte ca etichete ale topicelor inițiale şi determinarea corelațiilor care apar între valoarea acestor etichete şi puterea relației dintre concepte şi topice.

This work outlines a novel system that automatically extracts conceptual labels for statistically obtained topics. By creating a projection of the topic, which is a distribution over all the vocabulary words, over the WordNet ontology we succeed in associating concepts to the said groups of words. The most important contributions of this paper are connected to the validation of the role of these concepts as topical labels and the determination of correlations that emerge between the utility of these labels and the strength of the relation between the concepts and the topics.

Keywords: Topic models, labels, WordNet, conceptual processing

1. Introducere

This paper is primarily dedicated to the extraction of statistically meaningful concepts from economic texts

Conceptually, topic modeling falls within the larger study area of generative models, which is itself a part of the even larger research fields of probability and statistics. Starting from the assumption that observable data can be randomly generated following an a priori determined set of rules, topic models seek to detect the abstract ―topics that occur in series of documents. While no generally accepted definition for a topic exists, we will focus on the applications

[1]Post PhD student, Ecole Politechnique Federale de Lausanne, Switzerland, e-mail: Claudiu-cristian.musat@epfl.ch
[2] Prof., Automatic Control and Computers Faculty, University POLITEHNICA of Bucharest, Romania, e-mail: trausan@cs.pub.ro
[3] Prof, ERIC Laboratory, University of Lyon 2, France, e-mail: julien.velcin@eric.univ-lyon2.fr
[4] PhD student, ERIC Laboratory, University of Lyon 2, France, e-mail: marian-andrei.rizoiu@eric.univ-lyon2.fr

of topic models in text mining and we shall use a model to confront it with ontological knowledge. The ontological knowledge we use is contained in the English language ontology WordNet [1].

We create a projection of a trimmed description of the topic word distributions on the noun taxonomy pertaining to the WordNet ontology and obtain the concepts that are most related to the given topics. We then use this projection to evaluate the strength of that relation and based on it we define the internal cohesion and thus importance of the topic itself

This topical evaluation is relevant for obtaining the most important ideas that permeate a text collection and thus the concepts behind them. We go further in our analysis and show that conceptual labels are congruent with human thought and investigate the role of context in the evaluation. The result of all these processing phases is a set of concepts that represent the topics that are truly the most important ones in the given corpus.

The paper begins with an outline of the state of the art regarding opinion mining, topic modeling and ontological knowledge, moving on to the description of the system meant to formalize the pairing of topical concepts and hidden opinions.

2. State of the Art

In our search for better methods of examining the opinions expressed in economic texts, we focus on novel means of extracting what matters within those texts. Although this data retrieval task has many solutions, topic modeling is an elegant one with multiple advantages and applications that will be outlined below.

These mathematical models based on probabilistic Bayesian networks have been designed to address various issues, such as: multi-topics allocation [2], super and sub-topic hierarchies [3], temporal evolution of topics [4], etc. Plenty of applications can take great benefits from topic models, including information retrieval, database summarization or ontology learning. We will however focus on their role in textual processing, both in a general text case and in an economic environment.

Topic models have recently retained a lot of attention in dealing with textual corpora [reference here]. In brief, topics are multinomial distributions over words or key phrases which aim at capturing the meaning of huge volume of textual data in an unsupervised way. Document clustering is one of the most important areas of natural language processing with applications such as the ability to query large document collections (for instance the Internet). Multiple clustering algorithms, whether supervised or unsupervised, agglomerative or partitional, with a complexity ranging from a simple k-Means to complex kernel based algorithms meant for gigantic data sets have been proposed for this task.

However most of them carry limitations, such as a fixed number of classes or dependence on a multitude of parameters or features which need to be set a priori. While targeted solutions to these inconveniences do exist, such as running through a feature selection phase prior to running the clustering mechanism itself, eliminating the need for such preprocessing becomes a more elegant alternative.

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is the prototypic model [2] following the work of Hofmann [5]. It can be described as a probabilistic generative model designed to extract topics from text corpora.

In brief, the system starts from the bag-of-words model that considers documents as collections of words, without making use of their order. However this assumption can be relaxed, and allow documents to have a Markowian structure, or more easily through the use of bigrams or n-grams. The data is afterwards treated as if being the result of a probabilistic generative process. The intuition behind this reasoning is that documents exhibit multiple topics.

The process of constructing a document can be thought as the process of rolling a die. But it's not a regular die – but one with k facets, where k is also the number of topics in the model. And it's not a fair die either – it's loaded with the probability of each facet being the probability of its associated topic.

After rolling this initial k-facet die, a topic will have been chosen. We now must pick another die to obtain the word. This second die is chosen from k similar dies, each with maybe thousands of facets, a number equal to the number of words in the vocabulary. And all these dies are loaded in a manner similar to the one previously described, except that now the probabilities in question are those of the word given the topic. In just two die throws we have chosen a word for the document. The process is repetitive. The problem with picturing this analogy is the difficulty of picturing (k+1) loaded dies with multiple facets for each document. In a constrained space however the die throwing generative process becomes more intuitive.

However the outcome of this process is not easily humanly understandable. By generating a new document, even with superbly determined probability distributions, one obtains a mere collection of seemingly unrelated words, with no punctuation or structure. Still you could get a sense of what that document is about, just by looking at those words. It does not have to be realistic or contain meaning in the classical understanding case. If this is true, and the document's belonging to certain ideas is easily determinable, than the process that generated the distributions in the first place was correct. Obviously in reality the system is quite the opposite, and we must infer that hidden structure using posterior inference. Using the observations (the documents) and their word

distributions, we contemplate and compute the hidden variables, which become the result of the whole topic modeling task. A document is thus a second order distribution(a distribution over distributions)., and moreover it is a Dirichlet distribution.

2.2. Topic Labeling

The idea of entity labeling is greatly vaster than that of topic model labeling. An interesting analogy is that of topics and clusters, the idea behind both being their ability to separate items that do not share a given set of traits. Thus multiple instances are separated into groups that can be generally described as classes of objects that share a great portion of their relevant traits.

The ability to name these groups is of great interest for their post processing. A good example is that of Osinski [6] , and their Carrot2. Carrot2 is a widely known tool for Web search results clustering and it is relevant to the current discussion because of the fact that it uses frequent patterns for labeling text clusters.

This line of thought has been extended to the field of topic modeling through the work of Rizoiu et al. [7], who use frequent term-based methods to label topics extracted with known topic models. Documents are simplified to feature vectors that encode the presence and absence of certain words, IDF (Inverse Document Frequency) or TFxIDF (Term Frequency - Inverse Document Frequency) data. Then a K-Means variation is used to group the similar documents together and the resulting centroids are being viewed as the resulting labels. Centroid names are then extracted using the data coded in the feature vectors of the documents pertaining to the given class.

Although most of the usual topic models use bag of word approaches, the idea of using the co-occurrence and spatial connections between certain words has been extensively pursuit in the form of n-grams. Wang et al. [8], in their work on topical n-grams even propose n-grams as a useful way of generating topic labels, following the assumption that the bag of words approach can be complemented with spatial data. Moreover their work is interesting because of their ability to limit the context of the said associations. As an example they present the case of the bigram ―white house which has a special meaning in a political environment but less so in a real estate discussion.

Mei et al. [9] propose another probabilistic alternative to manually labeling topics, starting from the observation that single words tend to be too general labels, while whole sentences tend to be too specific. They treat the labeling task as an optimization problem that aims at minimizing the Kullback – Leibler (KL) divergence between word distributions. The two compared distributions in this case are the topic itself and the candidate label word

distribution. The obtained distance is an indicator of the meaningfulness of that particular label to the analyzed topic. A similar track is pursuit by Chen et al [10] in their algorithm that extracts diverse topic phrases as summaries for large corpora.

2.3. WordNet

The English language ontology WordNet [1] has a long tradition of being used in text classification tasks [11]. WordNet may be considered a general ontology or a lexical database. It is in fact a huge semantic network linking the majority of usual words in English through a fixed set of relations like: synonymy, hypernomy/hyponymy (super/sub concept), meronymy/holonymy (part/whole), antonymy, etc. Each word may have several senses and for each sense it has a set of synonyms (a synset). Each synset represent a distinct concept and semantic distances between pairs of words may be computed [12]. Consequently, sets of words may be grouped in their respective semantic neighborhoods.

There are several differences between WordNet semantic neighborhoods and semantic spaces of LSA or topics discovered with LDA. First of all, the former are obtained from the word networks built explicitly by humans, starting from psycholinguistics data, while LSA and LDA word grouping is determined statistically from text corpora. The advantage of using WordNet is precision while the disadvantage is the lack of dynamics and of the possibility to handle very specific domains. Even if it has more than 200,000 word-sense pairs, WordNet cannot cope with very specialized terms or neologisms. A second difference is that in WordNet words are not only grouped by similarity, they are also related by various relations, as mentioned above and thirdly, each word in WordNet has a gloss. The latter two features may be exploited for further semantic processing. The idea to mix topic models and ontologies is not new. LDAWN, latent Dirichlet allocation with WordNet [13] is a version of LDA that uses the word sense as a hidden variable and becomes a system for word sense disambiguation.

3. Proposed System

In the presented system, we propose a new framework for generating topical labels, by passing the Rubicon from the word space to the conceptual space. The system is based on the previously presented method [14] of attaching concepts to topical relevant words via WordNet projection.

The primary aim of the approach is to determine whether a concept constitutes a better label than a regular bag of words approach that uses the top scoring word as the main idea of the topic and thus its label. We then determine which factors affect the quality of the conceptual labels and calculate correlations

between the labels' usefulness and various parameters such as the topic relevance scores. We also discuss the apparent role of other topics in labeling a single instance and determine whether this apparently exterior influence is itself correlated with the topical individual score.

We test this hypothesis in multiple usage cases. The first factor we want to assess is the relevance of the system on different corpora. We thus use both the corpora already presented (Suall [4] and Associated Press [14]) and test whether differences emerge. The second factor we want to assess is actually a parameter of the system – the number of concepts that best describe the analyzed topics. We test the system for both the top concept and top three concepts cases and compare the results.

4. Experiments

The topic label scoring results are divided into two conceptual zones – the first contains model oriented data, while the second is centered on the human evaluator. By contrasting the two areas and starting from the premise that a good automatic system must remain congruent with human judgment, we assess the quality of our approach from the evaluator point of view.

4.1. Methodology

The questions regarding topic labeling represent a second batch of questions each evaluator was asked to respond to, each having at most 20 questions. We define a model by selecting the topics drawn from the pool obtained by applying LDA on a given corpus, and having k (the number of topics) set. In our case this translates into dividing the label evaluations into groups such as ―topics obtained from the economic corpus in the 200 topic set.

As with the first experiment, regarding topic quality, the best and worst ten topics from each experiment were considered and evaluated. They were shuffled and showed in a random order to the evaluators, with the condition that no topic is shown twice to the same person.

The structure of the question lot is as follows: the evaluator is initially asked to rate the fitness of the most relevant word in the topic (with the highest probability) as the topical label. The scale is from 1 to 5, where 1 means that the proposed label has absolutely no connection with the topic, 2 implies some vague connection, 3signals that the topic and label are probably related, at least in the same area, 4 underlines a strong connection while 5 is the highest grade and is only fit for perfect matches. Since the first question regards the most important word, which is visible to the evaluator (being in the top 5 topical words), it is rarely graded as a 1, usually the lowest grade is a 2 – vaguely connected.

The second question regards the perceived fitness as a topical label of the most relevant concept attached to the topic following the presented topical sub tree method. Similarly, the third question garners the evaluator's opinion on the fitness of a three concept label, consisting of the three most relevant concepts – with the highest fitness values – for that topic.

Questions 4 and 5 in the second lot were asked to determine the separability capacity of the given 3 concept label. If the label obtains a high score in conjunction with the current topic but lower scores with others, than it succeeded in separating the current ones from its conceptual neighbors. If not, then probably the concept set is not specific enough to constitute a relevant label or the competing topics are extremely close from a conceptual standpoint to the initial topic.

Topic 41												
Topic:												
percent		stock		market		investor		bond				
a) Please rate how well the following labels - formed by one or three concepts - represent the underlying meaning of the topic: [1 to 5]												
	(percent)							1 = absolutely no connection				
->	2	<-	[1 to 5]					2 = Some vague connection				
								3 = At least in the same area				
	(person)							4=Well related				
->	1	<-	[1 to 5]					5= perfect match				
(person + organism + artifact)												
->	1	<-	[1 to 5]									
b) How well would the 3 concept label (person + organism + artifact) also apply to the following topics?												
(plant		energy		power		project		company)				
	->	1	<-	[1 to 5]								
(share		percent		revenue		cent		company)				
	->	1	<-	[1 to 5]								

Fig. 1 Topic Label Evaluator Question Format

4.2. Results

We used two perspectives to interpret the results obtained. The first is model oriented, and its usage enables the detection of model specific correlations, such as the one between concept topical fitness and label quality from a human

point of view. In the second, all results are centered on the evaluator, and the topics are only separated by their fitness type – good or bad from an algorithmic point of view. The latter permits the detection of general trends that separate the method from other known ones.

In the first experiment we test whether using a concept as a label is significantly better viewed by humans than using the highest scoring concept within that topic. As previously said, the topics obtained from the two corpora (Suall and Associated Press) are divided into the best and lowest scoring, according to their most important concept scores. Each class (for instance good topics obtained from the economic corpus) is then divided into groups according to the number of classes from the experiment they were drawn from (for instance good topics obtained from the economic corpus running LDA with *k*=50 topics).

For each verdict v and for each topic number k we compute the differences between the average evaluator answers for questions 2 and 1 for the selected topic set, $dif_{21_{v,k}}$and in an analogue manner the difference between the answers for questions 3 and 1, dif_{31}. We are interested in the cases where the concept evaluation is greater than the top word evaluation and we separate those cases as $card^{+}_{21_{vk}}$and $card^{+}_{31_{vk}}$ – the number of cases in which the criteria is met. We then computed the ratio between the number found and the total number of instances within that topic set, $total_{vk}$:

$$r^{+}_{21} = \frac{card^{+}_{21}}{total_{vk}};\ r^{+}_{31} = \frac{card^{+}_{31}}{total_{vk}} \tag{1}$$

Table 1

Bottom scoring topic

Corpus	*k*	$\overline{fit_k}$	$r^{+}_{21_{0k}}$	$r^{+}_{31_{0k}}$
Associated Press	30	1.18	0.23	0.29
	50	1.13	0.26	0.26
	100	1.11	0.27	0.24
	200	1.23	0.29	0.42
	300	1.15	0.37	0.49
Suall	30	1.51	0.33	0.5
	50	1.36	0.29	0.37
	100	1.22	0.49	0.56
	200	1.26	0.36	0.47
	300	1.22	0.43	0.51
	Average		0.332	0.411

These ratios symbolize the proportion of cases in which the one or three concept label was better than the default top probability word case, for the given verdict (top or bottom topics) and for the considered number of topics within the model.

We then compute the average fitness of the topics from that given experiment, $\overline{fit_k}$, and plot the results of the experiment in tables 1 (for the lower scoring topics) and 2 (for the top scoring ones). In table 2 the differences between the results obtained for the best and worst scoring topics in the one concept label's case are expressed as

$$dif^{+}_{21_k} = \frac{r^{+}_{21_{1k}} - r^{+}_{21_{0k}}}{r^{+}_{21_{0k}}} * 100\,, \tag{2}$$

Whereas in a similar manner the difference for the three concept label, $dif^{+}_{31_k}$ is expressed as:

$$dif^{+}_{21_k} = \frac{r^{+}_{21_{1k}} - r^{+}_{21_{0k}}}{r^{+}_{21_{0k}}} * 100\,, \tag{3}$$

Table 2

Top scoring topics and relations to bottom scoring ones

Corpus	$id,	$\overline{fit_k}$	$r^{+}_{21_{1k}}$	$r^{+}_{31_{1k}}$	$dif^{+}_{21_k}$	$dif^{+}_{31_k}$
Associated Press	30	1.74	0.56	0.53	144.19	87.21
	50	1.91	0.49	0.56	89.92	117.05
	100	2.02	0.43	0.59	61.18	143.78
	200	2.14	0.78	0.92	168.2	118.25
	300	2.22	0.62	0.74	66.93	50.06
Suall	30	2.07	0.81	0.67	142.86	33.33
	50	2.14	0.88	0.75	202.27	103.57
	100	2.26	0.54	0.7	10.81	25.26
	200	2.17	0.71	0.71	97.8	51.26
	300	2.18	0.6	0.74	41.09	44.7
	Average		0.642	0.691	102.525	77.447

The fact that the averages of the two differences, $dif^{+}_{21_k}$ and $dif^{+}_{31_k}$, which we denoted as $\overline{dif^{+}_{21_k}}$ and $\overline{dif^{+}_{31_k}}$ are both positive and have values close to or exceeding 100% is a clear indication that a conceptual label has roughly twice as many chances of being considered helpful by a human if the connection between that concept and the topic itself is a close one. Furthermore this validates our

method of evaluating the relation between concepts and topics as the presented fitness score.

We observe that in both cases topics' relevance scores are correlated with the probability of the concept label being considered better than the first word label. This is more visible in the one concept case – an observation that can be explained through the fact that we considered the most relevant score to determine the topical fitness in the first place.

It is noteworthy that for good and bad topics alike the three concept label performed better than the one concept one. However, for poorly extracted topics, the top concept does not represent the 5 words better than the top word. For the top scoring topics on the other hand, it does. While the same observation is true for the three concept sets, the difference between $r^{+}_{31_{1k}}$ and $r^{+}_{31_{0k}}$is much less pronounced than in the one concept case.

4.4. Individual Correlation

In a second experiment we crossed the borders of individual models and bundled all the topics that were shown to a given evaluator. As in the experiment above, we are interested in the proportion of the cases where the one or three concept label was considered better than the first topical word.

For each evaluator we computed the correlation between the array of topical fitnesses (containing a numeric value for each topic shown) - fit_n and the array of ones or zeros $r^{+}_{21_n}$ containing the same number of values corresponding to whether for the given topic the one or three concept label was better than the benchmark.

As in the topic evaluation phase, we used Pearson's coefficient to establish the degree in which the analyzed values are correlated. The average $\bar{r}$, maximum r_{max} and standard deviation $\sigma(r)$ of the set of correlations determined are shown in table 3.

Table 3

Correlations between topical fitness and label utility

Correlation	$r^{+}_{21_n}$: fit_n	$r^{+}_{31_n}$: fit_n
$\bar{r}$	0.390	0.414
r_{max}	0.663	0.704
$\sigma(r)$	0.156	0.199

We notice that in all cases the correlation between the odds that the label is considered better than its competitor is a positive one. Confirming the findings in the experiment above, the three concept solution is slightly better than the single concept one both as an average and as a maximum value. However this also leads

to a higher level of uncertainty – with the standard deviation of the correlations in the three concept case being mode than 20% higher than in the primary case.

5. Conclusions

The results presented within this paper show the need to progress from word only topical labels to ones consisting of upper level concepts. Our experiments show that not only are conceptual labels more accurate and more consistent with human thought than word only definitions, but they also indicate the presence of a significant topic. The correlations that emmerge between label fitness and topical fitness show the need for advanced labeling techniques to further develop the analysis of topic models.

Furthermore our experiments show that label quality is consistent with evaluator agreement and that conceptually cohesive topics are easily labeled and understandable.

R E F E R E N C E S

[1] *G.A Miller*, "WordNet: a lexical database for English". Journal Communications of the ACM. 38(11), pp. 39-41, ACM (1995)

[2] *D.M. Blei, A. Ng, M. Jordan*, "Latent Dirichlet Allocation". The Journal of Machine Learning Research. 3, pp. 993 – 1022 (2003)

[3] *D.M., Blei, T. Griffiths, M. Jordan, J. Tenenbaum*, "Hierarchical topic models and the nested Chinese restaurant process". Advances in neural information processing systems, 16 , 106 – 114 (2004)

[4] *X. Wang and A. McCallum*, "Topics over time: a non-Markov continuous-time model of topical trends". Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 424-433. 2006. ACM

[5] *T. Hofmann*, "Probabilistic latent semantic indexing". In Proceedings of the 22th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. 50-57, 1999.

[6] *S. Osinski, D. Weiss*, "Carrot2: Design of a Flexible and Efficient Web Information Retrieval Framework", Advances in web intelligence - Third International Atlantic Web Intelligence Conference, AWIC 2005, Lodz, Poland, June 6-9, 2005

[7] *M.-A Rizoiu, J. Velcin, J-H. Chauchat*, "Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes". 10éme conférence Extraction et Gestion des Connaissances (EGC 2010), Hammamet, Tunisie, 2010.

[8] *X. Wang, A. McCallum, X. Wei,* "Topical n-grams: Phrase and topic discovery, with an application to information retrieval". In Proceedings of the 7th ieee International Conference on Data Mining, pp. 697–702 (2007)

[9] *Q. Mei, X. Shen, and C. Zhai,* "Automatic labeling of multinomial topic models". In Proceedings of KDD 2007. 490–499. 2007

[10] *J. Chen, J. Yan, B. Zhang*, "Diverse topic phrase extraction through latent semantic analysis Data Mining, 2006. ICDM'06. Sixth International Conference on Data Modeling

[11] *S. Scott, S. Matwin*. "Text Classification using WordNet Hypernyms". In Proceedings of the Association for Computational Linguistics Conference. 38-44. 1998
[12] *A. Budanitsky, G. Hirs*t "Evaluating WordNet-based measures of semantic distance." Computational Linguistics, 32(1), pp. 13—47 (2006)
[13] *J. Boyd-Graber, D.M. Blei, and X. Zhu*, "A Topic Model for Word Sense Disambiguation". In Empirical Methods in Natural Language Processing. 1024-1033. 2007
[14] *C. Musat, J. Velcin, S. Trausan-Matu, M-A. Rizoiu*, "Improving Topic Evaluation Using Contextual Knowledge", proceedings of IJCAI 2011.