# Improving Topic Evaluation Using Conceptual Knowledge

**Claudiu Cristian Musat[1], Julien Velcin[2], Stefan Trausan-Matu[1] and Marian-Andrei Rizoiu[2]**
[1] Computer Science Department
"Politehnica" University of Bucharest, Romania
[2] ERIC Laboratoire
Université Lumière, Lyon 2, France
{Claudiu.Musat, Trausan}@cs.pub.ro , {Julien.Velcin, Marian-Andrei.Rizoiu}@univ-lyon2.fr

## Abstract

The growing number of statistical topic models led to the need to better evaluate their output. Traditional evaluation means estimate the model's fitness to unseen data. It has recently been proven than the output of human judgment can greatly differ from these measures. Thus the need for methods that better emulate human judgment is stringent. In this paper we present a system that computes the conceptual relevance of individual topics from a given model on the basis of information drawn from a given concept hierarchy, in this case WordNet. The notion of conceptual relevance is regarded as the ability to attribute a concept to each topic and separate words related to the topic from the unrelated ones based on that concept. In multiple experiments we prove the correlation between the automatic evaluation method and the answers received from human evaluators, for various corpora and difficulty levels. By changing the evaluation focus from a statistical one to a conceptual one we were able to detect which topics are conceptually meaningful and rank them accordingly.

## 1 Introduction

Topic models have recently retained a lot of attention in dealing with textual corpora. 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. These mathematical models based on probabilistic Bayesian networks have been designed to address various issues, such as: multi-topics allocation [Blei *et al*., 2003], super and sub-topic hierarchies [Blei *et al*., 2004], temporal evolution of topics [Wang and McCallum, 2006], etc. Plenty of applications can take great benefits from topic models, including information retrieval, database summarization or ontology learning.

However, the comparison of the proposed models remains rather difficult, especially when considering models built on different theoretical basis. A lot of effort is currently put into evaluating topic models. Recent work has proved that using only numerical measures cannot alone solve this issue; human judgment can be of great benefit in the task of topic evaluation [Chang *et al*., 2009a]. Very recent work [Newman *et al*., 2010] uses external resources, especially the Web, as an alternative evaluation measure but deems working with ontologies unfeasible. To our knowledge, no previous work has successfully used ontologies to evaluate the meaning of topic models.

In this paper, our main contribution is to automate the topic model evaluation using an external concept hierarchy – here, WordNet [Miller, 1995] To tackle this problem, we propose the notion of conceptual relevance. The idea behind it is to find the most related concepts to each topic given the concept hierarchy and to evaluate the topic based on these concepts and the strength of the (topic, concept) relations. This semantic approach is very different from that of Newman [2010] which is largely based on statistics.

We prove the correlation between our semantic measure and the human judgment given by 37 external judges. The experiments were made on two different datasets: the first *Suall* [Wang and McCallum, 2006] is a general dataset on American history and the second is a specific dataset containing exclusively economic articles. Although the evaluation correlation shows a similar pattern for both corpora, some quantitative differences exist. Relevant evaluation differences have emerged when confronting the human evaluators with different types of topic word mixes that need to be separated. The influence of these and other external factors on the overall system accuracy is also discussed.

Topic models and their previously used evaluation methods are outlined further in the introduction, the proposed system is detailed in section 2, the experiments and their results are presented in section 3 and our conclusions and future work follow in section 4.

### 1.1 Topic Modeling

Topic models are graphical hierarchical Bayesian networks used to extract the latent meaning of textual datasets. Latent Dirichlet Allocation (LDA) is the prototypic model [Blei *et al*., 2003] following the work of Hofmann [1999]. The main idea of probabilistic models lies in the assumption that the observed texts are derived from a generative model. In such a model, there are unseen latent variables (the topics) from which words and documents are generated. The latent variables are represented as random variables over the set of words or n-grams. Thereby these models attempt to estimate

the probability distributions of the latent variables by using maximum likelihood or Bayesian inference. On top of that basic approach, other generative models have been developed in order to address topics extraction in complex cases, such as for correlated topics [Blei and Lafferty, 2007], n-gram handling [Wang *et al*., 2007], social networks [Chang *et al*., 2009b], opinion mining [Mei *et al*., 2007b],etc.

The English language lexical database WordNet [Miller, 1995] has a long tradition of being used in text classification tasks [Scott and Matwin, 1998]. Also, the idea to mix topic models and ontologies is not new. LDAWN, latent Dirichlet allocation with WordNet [Boyd-Graber *et al*., 2007] is a version of LDA that uses the word sense as a hidden variable and becomes a system for word sense disambiguation.

### 1.2 Topic Evaluation

Topic models have been proven to be accurate both quantitatively and qualitatively. Quantitatively, it has been shown [Wallach *et al*., 2009], using the perplexity measure, that they possess a high generalization ability on unseen data. This method allows for the estimation of the log-likelihood either on a fraction of the train set (said ``held-out'' data) in a cross-validation way, or on new documents.

Qualitatively, a sample of topics is usually exhibited in order to convince the reader of their usefulness. Each exhibited topic $z$ is a short list of the first terms $w$, from a decreasing probability perspective, associated (sometimes) with their probability values $p(w|z)$, and most of the times with a name given manually by the authors. Here is, for instance, an extract of a topic description illustrating the output of the LDA algorithm [Blei *et al*., 2003]: ``Arts'' (*new, film, show, music, movie, play, musical, best, actor, first, york, opera, theater, actress, love*).

Furthermore, it has been shown [Chang *et al*., *2009*] that human judgment does not coincide with the common automatic evaluation measures. We believe this to be the greatest shortfall of the above measures. This finding has prompted other researchers to look for novel evaluation systems such as that proposed in [Newman *et al*., *2010*]. The latter relies on external resources for the evaluation task and that postulates topic coherence to be at the core of the idea of evaluation. Results obtained by employing Wikipedia or Google were satisfying, while those extracted with WordNet were deemed patchy at best.

The evaluation task is conceptually similar to that of labeling. A "good" label implies an understandable underlying meaning. [Mei *et al*., 2007a] suggested multiple solutions which fell short of being able to generalize concepts through a hypernymy relation, which is usually appropriate, while [Andrzejewski *et al*., *2009*] uses domain knowledge through "must link" or "cannot link" relations in a semi automated model. The latter introduces an important concept – that some words can wrongfully be inserted into a topic. This we believe leads to our conceptual disagreement with Newman's [2010] conclusion on the use of WordNet in topic evaluation. Although some topical words can be outliers, the quality of the topic can remain elevated and the topic itself humanly comprehensible.

## 2 Proposed System

The system we proposed is designed to rank topics according to their relevance to the user, which we defined as a function of their cohesion – how many of the topic words are related and in what degree – and specificity – how general their common hypernym is.

Given a text collection $\boldsymbol{T}$ and $\mathbf{V}$ the employed *vocabulary* we aim to evaluate various probability distribution sets over $\boldsymbol{T}$ with respect to knowledge regarding related concepts. We will use z to denote a discrete probability distribution function $\{p(w|\mathrm{z})\}_{w\in \mathrm{z}}$ over $\boldsymbol{T}$, which we will further refer to as a *topic*. Each topic z is a member of $\Theta = \{\mathrm{z}_1, \dots, \mathrm{z}_k\}$ a finite set of k topics extracted using one of the known algorithms given $\boldsymbol{T}$, with $\Theta \subset Z$ , where $Z$ represents the set of all possible topics over $\boldsymbol{T}$. Furthermore, $\Theta \in P(Z)$, the space of all possible probability distribution sets over $\boldsymbol{T}$.

### 2.1 Concept Representation

As previously said, prior knowledge about related concepts is considered. Let $O = (\boldsymbol{D}, \boldsymbol{\mathcal{R}})$, be the pair of the set of relevant concepts $\boldsymbol{D}$, and a set of relevant relations over $\boldsymbol{D}$, further referred to as $\boldsymbol{\mathcal{R}}$. We address the particular case where there exists a relation $r \in \boldsymbol{\mathcal{R}}$, according to which all concepts in a subset $\delta(\mathcal{C}) \subset \boldsymbol{D}$ form a tree $\mathcal{C} = (\delta(\mathcal{C}), r)$ in which the $\delta(\mathcal{C})$ elements form the tree nodes and $r$ is the relation between them. $\mathcal{C}$ is of course a member of all possible trees over $\delta(\mathcal{C})$, given all possible relations $\boldsymbol{\mathcal{R}}$, a space which we will refer to as $\boldsymbol{\Gamma}$. The $r$ relation can be exemplified as a either hypernymy or hyponymy relations between concepts such as those present in WordNet.

### 2.2 Employed Distances

Within $\mathcal{C}$, we define a *branch* as a path between two concepts $c_i$ and $c_j$ that are either directly or indirectly related, as the pairs $(\mathrm{c}_1, \mathrm{c}_0)$ or $(\mathrm{c}_3, \mathrm{c}_0)$ in Fig.1. In the case of a hypernymy or hyponymy relation, the fact that two concepts are related also implies that one is an ancestor (or a generalization) of the other. To determine the distance between two concepts with respect to $\mathcal{C}$, $d(c_i, c_j)$, we use a slightly modified version of the ancestral path distance with the result being infinity if one is not a direct ancestor of the other. Let $\delta(c_i, c_j)$ be the set containing all the nodes within the branch that connects $c_i$ and $c_j$ .

We further assume that there exists a subset of words in the vocabulary $\boldsymbol{V}_{\mathcal{C}} \subset \boldsymbol{V}$ where, for each word in $\boldsymbol{V}_{\mathcal{C}}$ there exists at least one concept in $\mathcal{C}$ that is a sense of the given word. Let $\delta(w) \subset \delta(\mathcal{C})$ be the set of all senses of the word $w$ within $\mathcal{C}$. We define the *distance* between a word and a concept, $dist(w, c)$, as the smallest number of transitions between any concept in $\delta(w)$ and the target concept $c$:

$$d(w, c) = \min_{\mathrm{c_w}\in\delta(w)}\big(d(\mathrm{c_w}, c)\big)\forall \mathrm{c} \in \mathcal{C}, \qquad \forall w \in \boldsymbol{V}_{\mathcal{C}}$$

Let $c_0(\mathcal{C}) \in \delta(\mathcal{C})$ be the root of the $\mathcal{C}$ tree. For instance, in Fig. 1 the distance between $w_1$ and the subtree root $c_0(\mathcal{C})$ is 2, because the distances between its senses and $c_0(\mathcal{C})$ are

$d\big(w_{1,1}, c_0(\mathcal{C})\big) = 3$ and $d\big(w_{1,2}, c_0(\mathcal{C})\big) = 2$, where $w_{1,1}$ and $w_{1,2}$ are $w_1$'s senses.

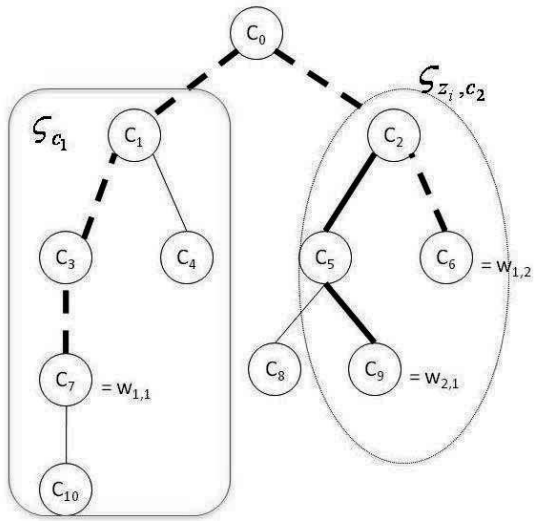

Figure 1. Topical Subtrees

## 2.3 Concept and Topical Subtrees

Definition 1. *A word's subtree within* $\mathcal{C}$ is the reunion of all the branches that relate concepts within $\delta(w)$ to $c_0(\mathcal{C})$, $\mathcal{C}_w = (\delta(w,\mathcal{C}), r);\ \delta(w,\mathcal{C}) = \bigcup_{c_w \in \delta(w)} \delta\big(c_w, c_0(\mathcal{C})\big)$. In Fig. 1 $w_1$'s subtree within $\mathcal{C}$ is marked with a dotted line.

Definition 2. *The subtree of an arbitrary concept c in* $\mathcal{C}$ $\mathcal{C}_c = (\delta(\mathcal{C}_c), r)$ is the subtree of $\mathcal{C}$ whose root concept is $c, \delta(\mathcal{C}_c) \subset \delta(\mathcal{C}), \delta(\mathcal{C}_c) = \{c_i \in \delta(\mathcal{C}) | c_i \overset{r^*}{\rightarrow} c\}$. $\mathcal{C}_{c_1}$ is presented in the figure above in a rectangle.

Definition 3. *A word's w subtree of a concept c within* $\mathcal{C}$, $\mathcal{C}_{\mathrm{w,c}} = (\delta(w, \mathcal{C}_c), r)$ is the subtree of $\mathcal{C}_c = (\delta(\mathcal{C}_c), r)$ that contains all the branches between concepts within $\delta(w)$ and subtree the root c. $\delta(w, \mathcal{C}_c) = \bigcup_{c_w \in \delta(w)} \delta(c_w, c)$. It is noteworthy that $\mathcal{C}_{\mathrm{w,c}}$ is a generalization of $\mathcal{C}_w$. $\mathcal{C}_{\mathrm{w}_1,\mathrm{c}_1}$ is in Fig 1. a reunion of the $(c_1, c_3)$ and $(c_3, c_7)$ arcs.

Starting from a topic's most important words, we construct a substructure of the concept representation – a *topical subtree* that comprises all the concepts related to the selected topic words. The idea of using the topic's most important words is conceptually similar to the method employed in [Chang *et al*., 2009a] where those words alone are sufficient to lead to a satisfying assessment of a topic's quality. For a given topic model $\Theta = \{\mathrm{z}_1, \ldots, \mathrm{z}_k\}$ we define a term importance function $\tau: \Theta \mathrm{x} \boldsymbol{V'} \rightarrow \mathbb{R}_+;\ \mathrm{z_i}, w \mapsto \tau(\mathrm{z_i}, w), \forall \mathrm{z_i} \in \Theta, \forall w \in \boldsymbol{V'}$**.** Thus, for each $\mathrm{z_i} \in \Theta$ we extract a set of relevant words $\mathrm{P(z_i)}$ as the words that have $\tau(\mathrm{z_i}, w) > \alpha$, with $\alpha \in \mathbb{R}_+$ previously determined.

Let a topic's relevant concepts $\delta(\mathrm{z_i})$ within $\delta(\mathcal{C})$ be the reunion of the $\mathcal{C}$ senses of the topic's relevant words**:** $\delta(\mathrm{z_i}) = \bigcup_{w \in \mathrm{P(z_i)}} \delta(\mathrm{w}),\ \forall \mathrm{z_i} \in \Theta$

Definition 4. *A topical subtree within* $\mathcal{C}$ is the reunion of all the $\mathcal{C}$ subtrees of all the topic's relevant words:

$$\mathcal{C}_{\mathrm{z_i}} = (\delta(\mathrm{z_i}, \mathcal{C}), r);\ \delta(\mathrm{z_i}, \mathcal{C}) = \bigcup_{w \in \mathrm{P(z_i)}} \bigcup_{c_w \in \delta(\mathrm{w})} \delta(c_w, c_0(\mathcal{C}))\ ,\ \forall \mathrm{z_i} \in \Theta$$

In Fig 1. $\mathcal{C}_{\mathrm{z_i}}$ is outlined with a bold line (dotted or not).

Definition 5. *A topical subtree of a concept c within* $\mathcal{C}$ is the reunion of all the $\mathcal{C}_{\mathrm{w,c}}$ subtrees of all topical words:

$$\mathcal{C}_{\mathrm{z_i,c}} = (\delta(\mathrm{z_i}, \mathcal{C}_c), r);\ \delta(\mathrm{z_i}, \mathcal{C}_c) = \bigcup_{w \in \mathrm{P(z_i)}} \bigcup_{c_w \in \delta(\mathrm{w})} \delta(c_w, c)\ , \forall \mathrm{z_i} \in \Theta,$$

which in the case of $\mathcal{C}_{\mathrm{z_i,c_2}}$ is shown in the Fig 1. ellipse.

## 2.4 Concept Metrics

We aim to identify the topical subtrees that include at least one sense for as many of the topic's words as possible while at the same time having a root concept as specific as possible. We are thus trapped in the age old tradeoff between coverage and specificity.

We define a subtrees' coverage $cov: \delta(\mathrm{z_i}, \mathcal{C}) \times \Theta \rightarrow \mathbb{N}_+;\ c \mapsto cov(c, \mathrm{z_i}) = \frac{card\{\cap\, \delta(\mathrm{z_i}), \delta(\mathrm{z_i}, \mathcal{C}_c)\}}{card\{\delta(\mathrm{z_i})\}}$. In order to determine a concept's specificity we rely on two additional features – its height $h: \delta(\mathcal{C}) \rightarrow \mathbb{N}_+;\ c \mapsto h(c) = d(c, c_0(\mathcal{C}))$, and depth $\rho$ with respect to the given topic: $\rho: \delta(\mathrm{z_i}, \mathcal{C}), \Theta \rightarrow \mathbb{N}_+;\ (c, \mathrm{z_i}) \mapsto \rho(c, \mathrm{z_i}) = <f>_{w \in \delta(\mathrm{z_i})} \big(d(w, c) | c \in \delta(\mathrm{z_i})\big)$, $\forall \mathrm{z_i} \in \Theta, \forall c \in \delta(\mathrm{z_i}, \mathcal{C})$, where $<f>$ is a member of a given family of functions such as the minimum, maximum or average that obtain a single scalar value for an input vector: $<f>: \mathbb{R}_+^{\mathrm{k}} \rightarrow \mathbb{R}_+;\ \begin{matrix} x_1 \\ \cdot\cdot \\ x_k \end{matrix} \mapsto <f> \begin{pmatrix} x_1 \\ \cdot\cdot \\ x_k \end{pmatrix} = y \in \mathbb{R}_+$.

There are more than one possible definitions for the specificity, one of which is a weighted average of the concept's height and depth: $spec: \mathcal{C}_{\mathrm{z_i}} \times \Theta \rightarrow \mathbb{R};\ c, \mathrm{z_i} \mapsto spec(c, \mathrm{z_i}) = \omega_h \cdot h(c) + \omega_\rho \cdot \rho(c, \mathrm{z_i})$, with the weights set a priori.

## 2.5 Topic Evaluation

We define the evaluation of the whole topic model as an aggregate evaluation function: for each distribution set $\Theta$, $\eta: P(Z) \rightarrow \mathbb{R}_+;\ \Theta \mapsto \eta(\Theta) = <f>\big(\eta(\mathrm{z})\big), \forall \mathrm{z} \in \Theta$. The evaluation of the model as a whole depends on the individual evaluation of each of its topics $\eta: \Theta \rightarrow \mathbb{R}_+, \mathrm{z_i} \mapsto \eta(\mathrm{z_i}) = <f>(\phi(\mathcal{C}_{\mathrm{z_i}}), \forall \mathrm{z_i} \in \Theta$, in which $\mathcal{C}_{\mathrm{z_i}}$ is the topical subtree of $\mathrm{z_i}$ given $\mathcal{C}$. Each node from $\mathcal{C}_{\mathrm{z_i}}$ is assigned a positive fitness value based on its relevance to the given topic $\mathrm{z_i}$, $\phi: \mathcal{C}_{\mathrm{z_i}} \times \Theta \rightarrow \mathbb{R}_+$. We express this relevance as a weighted average of its coverage and specificity, with $\omega_{cov}$ and $\omega_{spec}$ set a priori,

$$c \mapsto \phi(c) = \omega_{cov} \cdot cov(\mathrm{c}, \mathrm{z_i}) + \omega_{spec} \cdot spec(\mathrm{c}, \mathrm{z_i})$$

The higher the coverage of the concept with the highest fitness value, the higher the topic's cohesion given that concept. Furthermore, more specific concepts are attached to more specific topics. The total relevance of an individual topic is thus a function of its cohesion and specificity.

## 3 Experiments

Three types of experiments were devised, in order to capture the correlation between the human verdict and the calculated topic relevance. The LDA [Blei *et al*., 2003] model built into the Mallet suite [McCallum, 2002] was used to generate the topics and all automatically determined relevancies were calculated using the framework above while human evaluations similar to those employed by [Chang *et al*., 2009a] were gathered using a binary question answering system. Evaluators were asked to extract the unrelated words from a group containing one topic and an additional spurious word. One or more unrelated words were chosen for each group.

In the first experiment, the analyzed topics were separated into relevant or not from an algorithmic point of view and the aim is to see whether an improvement of the spurious word detection is visible from one category to the other. The second experiment shows how the improvement gains and the confidence in the experiment vary when modifying k (the number of topics in the model). These variations are shown for three different metrics – the chance to hit the spurious word, the same limited to the evaluator's first choice word and the total number of chosen words. The third experiment regarded the correlation between evaluator agreement, accuracy and ontological topic relevance.

Two corpora were used to find whether results would differ greatly if the focus is changed from a general purpose corpus such as the Suall [Wang and McCallum, 2006] to a more specific one, in our case an economic corpus. We built the second corpus from publicly available Associated Press articles published in the Yahoo! Finance section. A total of 23986 news broadcasts which had originally appeared between July and October 2010 were gathered.

### 3.1 Spurious Word Types

As previously said, human evaluation of a topic depends on the chance that the evaluator correctly detects a spurious word that is mixed with the topic's words. The choice of the spurious word is not obvious as it influences the outcome of the experiment. While [Chang *et al*. 2009] use a *random* word from those relevant to the other topics but not relevant to the current one, we believe that a discussion is necessary. Within each model, we compute all the inter-topic Kullback – Leibler divergences and for each topic we select a word from both the closest and the farthest remaining topics which shall serve as spurious words. The aim is to detect differences in evaluator agreement and evaluation performance depending on the spurious word choice. For instance, one of the topics obtained from the AP corpus was {*drug, treatment, company, patient, cancer*}. The word chosen from the closest KL neighbor was *hospital* while the choice from the farthest topic was *pound*.

A total of 37 evaluators were each given 40 groups of six words in a randomized order containing the five most important words for a topic (the ones carrying the highest probability in the LDA model) and a spurious word. In the example above, one examiner will be asked to choose a spurious word from the {*cancer, drug, pound, treatment, company, patient*} group while {*company, patient, cancer, hospital, drug, treatment*} will be shown to another person.

The questions were balanced to have an equal number of topics evaluated for the two corpora, for each topic number $k \in \{30, 50, 100, 200, 300\}$ and for each of the two spurious word types.

### 3.2 Spurious Word Detection Variation

In the first experimental setup the analyzed topics were separated into relevant and irrelevant from an algorithmic point of view. The first lot contained the top ten topics for each *k* above and each corpus, ranked by conceptual relevance while the second lot contained the bottom ten. The aim was to see whether an improvement of the spurious word detection is visible from one category to the other. The setup was duplicated for the close or far spurious word poison type.

For each examiner's answers we computed the average ratio of topics where the spurious word was detected in the two situations – top and bottom topics – $\overline{hit_+}$ and $\overline{hit_-}$. The difference between the two, as a percentage of $\overline{hit_-}$ is shown as the gain, in the last column of Table 1.

| Data | Type | $\overline{hit_+}$ | $\sigma(hit_+)$ | $\overline{hit_-}$ | $\sigma(hit_-)$ | +(%) |
|---|---|---|---|---|---|---|
| AP | Close | 0.37 | 0.29 | 0.27 | 0.23 | 39.33 |
| | Far | 0.69 | 0.29 | 0.65 | 0.23 | 6.93 |
| Suall | Close | 0.51 | 0.23 | 0.3 | 0.24 | 66.76 |
| | Far | 0.75 | 0.24 | 0.59 | 0.33 | 28.55 |

Table 1. Spurious Word Detection Ratios

The influence of the spurious word choice on the outcome of the experiment is noteworthy. On average, as shown in Table 1., the detection rate was 92.7% higher when the spurious words had been taken from far topics than in the close topic scenario.

The standard deviations for each topic type $\sigma(hit_+)$ and $\sigma(hit_-)$ are also shown and the evolution is mixed, for reasons which will be discussed in section 3.4.

Another metric is whether the spurious word was detected from the first word in the evaluator's answer rather than the others, $\overline{fhit}$. For instance, in the second example presented earlier, an evaluator answered *(company, hospital)*, which means that, although the spurious word has been detected, the initial bias was towards *company*. The gains obtained when passing from the low quality topics to high ranking ones are even more pronounced in this case.

This implies that if the evaluators were forced to only give one answer they would have detected the inserted words in even twice as many cases for the good topics than for the lower quality ones.

| Data | Type | $\overline{fhit_+}$ | $\sigma(fhit_+)$ | $\overline{fhit_-}$ | $\sigma(fhit_-)$ | +(%) |
|---|---|---|---|---|---|---|
| AP | Close | 0.27 | 0.26 | 0.21 | 0.2 | 32.08 |
| | Far | 0.6 | 0.32 | 0.53 | 0.24 | 13.92 |
| Suall | Close | 0.48 | 0.23 | 0.23 | 0.18 | 105.4 |
| | Far | 0.71 | 0.23 | 0.44 | 0.32 | 60.35 |

Table 2. Spurious Word Hit from the First Chosen Word

A third metric is the number of chosen words to answer one question. A higher number signals a lower quality topic

where the evaluator is unable to choose an outlying word with enough confidence. On average for each corpus and for each spurious word type 23 out of the 37 evaluators gave at least an answer containing a minimum of two words.

| Data | Type | $\overline{w_+}$ | $\sigma(w_+)$ | $\overline{w_-}$ | $\sigma(w_-)$ | +(%) |
|---|---|---|---|---|---|---|
| AP | Close | 1.35 | 0.38 | 1.4 | 0.43 | 3.45 |
| | Far | 1.31 | 0.38 | 1.54 | 0.39 | 15.19 |
| Suall | Close | 1.24 | 0.32 | 1.37 | 0.37 | 9.05 |
| | Far | 1.28 | 0.31 | 1.43 | 0.38 | 10.93 |

Table 3. Chosen Words Number

Results show that the number of chosen words $\overline{w}$ and the standard deviation are always smaller for good topics in all cases shown. The better the quality of the topic is, the lesser the need to also insert one of the real topic words alongside with the poison.

## 3.4 Metric and Topic Number Correlation

The second experiment setup shows the improvement gains presented in the previous experiment divided according to the number of LDA topics. Results obtained using the two spurious word types are presented separately.

| | $(\overline{hit_+} - \overline{hit_-})/\overline{hit_-}$ (%) | | $(\overline{fhit_+} - \overline{fhit_-})/\overline{fhit_-}$ (%) | | $(\overline{w_-} - \overline{w_+})/\overline{w_+}$ (%) | |
|---|---|---|---|---|---|---|
| k | Close | Far | Close | Far | Close | Far |
| 30 | -9.55 | -2.52 | -29.17 | -10.26 | 27.3 | 11.91 |
| 50 | -37.09 | -20.27 | -49.02 | -28.68 | 45.9 | 41.16 |
| 100 | 35.29 | 32.8 | 83.75 | 111.52 | 0.17 | 10.25 |
| 200 | 80.06 | 42 | 183.33 | 139.79 | 23.19 | 35.24 |
| 300 | 100 | 65.42 | 160 | 665.71 | 14.8 | 24.9 |

Table 4. Metric and Topic Number Correlation

The wide differences between the results obtained show the need for a finer granularity. Although the ratios are computed based on the same top-bottom topics dichotomy, we observe that, for these two corpora, with $k = \{30,50\}$ the algorithmically best topics perform poorer than their counterparts while for the upper $k$ values the scales turn pronouncedly. As we will show in the following experiment, this is due to the fact that the relevance difference between the better and the poorer quality topics varies with $k$. We underline the fact that the presented ratios move in tandem when varying $k$, regardless of the type of poisoning.

## 3.4 Evaluator Agreement

In the third experimental framework we divided the question instances by their spurious word type and again by the number of classes of the model the contained topic came from. For each topic $i$ we bundled all the examiners' answers and we computed their average $\overline{hit_{+_i}}$ or $\overline{hit_{-_i}}$ and standard deviation $\sigma(hit_-)_i$ or $\sigma(hit_+)_i$ respectively.

We then computed the average of these individual instance averages for each $k$ value, $\overline{hit_{+/-_k}}$ and $\overline{\sigma(hit_{+/-_k})}$ and subtracted the average value for all $k$ values, $\overline{hit_{+/-}} = \sum_k \overline{hit_{+/-_k}}/k$ and $\overline{\sigma(hit_{+/-})} = \sum_k \overline{\sigma(hit_{+/-_k})}/k$:

- $\partial hit_{+/-\ k} = \overline{hit_{+/-_K}} - \overline{hit_{+/-}}$
- $\partial\sigma_{+/-_k} = \overline{\sigma(hit_{+/-_k})} - \overline{\sigma(hit_{+/-})}$

Depending on whether it is part of the good or bottom topics, each topic $i$ has an automatically determined relevance $\overline{rel_{+_i}}$ or $\overline{rel_{-_i}}$. Based on these values we can compute for each $k$ value $\overline{rel_{+/-_k}}$ and $\overline{\sigma(rel_{+/-_k})}$ in an analogous manner as above. By subtracting the average relevance value for the whole experiment we obtain:

- $\partial rel_{+/-_k} = \overline{rel_{+/-_K}} - \overline{rel_{+/-}}$

The difference $\partial hit_{+_k} - \partial hit_{-_k}$ shows how much easier the evaluators were able to find the spurious words relative to the whole experiment average, given a fixed $k$ value. Also $\partial\sigma_{+_k} - \partial\sigma_{-_k}$ shows the variation of the degree of uncertainty in the previous improvement, above the experiment average. Results are shown in Table 5.

The goal is to find the correlations between the accuracy increases, uncertainty decreases and relevance variations for different $k$ values. We computed the Pearson correlation between the $(\partial hit_{+_k} - \partial hit_{-_k})$ and $(\partial rel_{+_k} - \partial rel_{-_k})$ and also between $(\partial\sigma_{+_k} - \partial\sigma_{-_k})$ and $(\partial rel_{+_k} - \partial rel_{-_k})$ for the two spurious word cases. A value close to 1 or -1 implies strong positive or negative correlation while values close to 0 show a lack of linear correlation.

| | $k$ | $\partial hit_{-_k}$ | $\partial\sigma_{-_k}$ | $\partial rel_{-_k}$ | $\partial hit_{+_k}$ | $\partial\sigma_{+_k}$ | $\partial rel_{+_k}$ | $\partial hit_{+_k} - \partial hit_{-_k}$ | $\partial\sigma_{+_k} - \partial\sigma_{-_k}$ | $\partial rel_{+_k} - \partial rel_{-_k}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Close Spurious Word | *30* | 0.08 | 0.04 | 0.22 | -0.02 | 0.05 | -0.21 | -0.1 | 0.01 | -0.43 |
| | *50* | 0.02 | 0.06 | 0.06 | -0.16 | 0 | -0.03 | -0.18 | -0.06 | -0.08 |
| | *100* | -0.05 | -0.06 | 0 | -0.05 | 0.03 | 0.08 | -0.01 | 0.09 | 0.08 |
| | *200* | 0 | 0.01 | -0.11 | 0.23 | -0.04 | 0.1 | 0.23 | -0.05 | 0.21 |
| | *300* | -0.06 | -0.05 | -0.17 | 0 | -0.04 | 0.06 | 0.06 | 0.01 | 0.22 |
| Correlation | | | | | | | | **0.704** | **-0.016** | |
| Distant Spurious Word | *30* | -0.11 | -0.01 | 0.22 | -0.18 | 0.11 | -0.21 | -0.07 | 0.12 | -0.43 |
| | *50* | -0.02 | 0.04 | 0.06 | -0.08 | 0.04 | -0.03 | -0.05 | 0 | -0.08 |
| | *100* | 0.02 | -0.01 | 0 | 0.13 | -0.03 | 0.08 | 0.11 | -0.03 | 0.08 |
| | *200* | 0.04 | 0.01 | -0.11 | 0.02 | -0.04 | 0.1 | -0.02 | -0.05 | 0.21 |
| | *300* | 0.07 | -0.04 | -0.17 | 0.1 | -0.08 | 0.06 | 0.03 | -0.04 | 0.22 |
| Correlation | | | | | | | | **0.557** | **-0.979** | |

Table 5. Evaluator Agreement

Results prove our assumption that the chance of the spurious word being detected is directly correlated with the quality of the topic, with values of 0.557 for the distant spurious words and 0.704 for the second case. Moreover the inverse correlation of -0.979 is extremely strong for the distant spurious word uncertainty variation. The better the topics, compared to the average value, the better the evaluator's answers and the greater their agreement.

## 4 Conclusions and Future Work

We have successfully proven that there is a strong correlation between the ontological evaluation of topics and the way humans interpret them. We have outlined the important impact corpus choice or question formation have on that correlation. By shifting to a conceptual perspective we believe that we will be better equipped to rank topics and their respective models in a manner congruent with user needs. Topic labeling and corpus summarization are just two of the applications that could benefit from the model.

A natural follow up of the evaluation task is the attempt to improve the given models. We have already developed a system based on a similar WordNet topical subtree framework to detect conceptual outliers within the topic relevant words, which should be functional in the near future. moreover we will analyze the quantitative impact of automatically labeling topics from a conceptual standpoint rather than a statistical one. Another inciting application of the framework is to create conceptual neighborhoods within a topic model and detect the importance of the context created by the other topics when labeling or evaluating a single one.

## Acknowledgments

This work is supported by the European Union Grant PO-SDRU/6/1.5/S/19 7713

## References

[Andrzejewski *et al., 2009*] David Andrzejewski, Xiaojin Zhu and Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. *Proceedings of the 26th Annual International Conference on Machine Learning*. 382-286, 2009.

[Blei *et al*., 2003] David M. Blei, Andrew Y. Ng, Michael I. Jordan and John Lafferty. Latent Dirichlet Allocation. *In Journal of Machine Learning Research*. 3: 993-1022, 2003.

[Blei *et al*., 2004] David M. Blei, Tom. L. Griffiths and Michael I. Jordan. Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems 16*. 2004.

[Blei and Lafferty, 2007]. David M. Blei and John Lafferty. A correlated topic model of Science. In *Annal of Applied Statistics*. 1(1). 17-35. 2007

[Boyd-Graber *et al*., 2007] Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A Topic Model for Word Sense Disambiguation. In *Empirical Methods in Natural Language Processing*. 1024-1033. 2007.

[Chang *et al*., 2009a] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei, Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information Processing Systems*, 2009

[Chang *et al*., 2009b] Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. Connections between the Lines: Augmenting Social Networks with Text. *Conference on Knowledge Discovery and Data Mining*, Paris, France, 2009

[Hofmann, 1999] Thomas 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.

[McCallum, 2002] Andrew K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002.

[Mei *et al*., 2007a] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of KDD 2007*. 490–499. 2007

[Mei *et al*., 2007b], Qiaozhu Mei Xu Ling, Matthew Wondra, Hang Su and ChengXiang Zhai,. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*. 171-180. 2007

[Miller, 1995] George A. Miller. WordNet: a lexical database for English. In *Communications of the ACM*. 38(11): 39-41, 1995. ACM.

[Newman *et al*., 2010] David Newman, Jey Han Lau, Karl Grieser and Timothy Baldwin. Automatic Evaluation of Topic Coherence. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 100-108. June 2010.

[Scott and Matwin, 1998] Sam Scott and Stan Matwin. Text Classification using WordNet Hypernyms. In *Proceedings of the Association for Computational Linguistics Conference*. 38-44. 1998

[Wallach *et al*., 2009] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinovn and David Mimno. Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*. 1105-1112, 2009.

[Wang and McCallum, 2006] Xuerui Wang and Andrew 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

[Wang *et al*., 2007] Xuerui Wang, Andrew McCallum and Xing Wei. Topical n-grams. Phrase and topic discovery with an application to information retrieval. In *Proceedings of ICDM*. 697-702. 2007