Claudiu Musat, Julien Velcin, Stefan Trausan-Matu, Marian-Andrei Rizoiu

To cite this version:

HAL Id: hal-00616245
https://hal.archives-ouvertes.fr/hal-00616245
Submitted on 24 Aug 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.


## Improving Topic Evaluation Using Conceptual Knowledge

Claudiu Cristian Musat¹, Julien Velcin², Stefan Trausan-Matu¹ and Marian-Andrei Rizoiu²

¹Computer Science Department
“Politehnica” University of Bucharest, Romania
²ERIC Laboratoire
Université Lumière, Lyon 2, France

### 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 that 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 multinominal 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 on 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, 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 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.

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 \( T \) and \( V \) the employed vocabulary we aim to evaluate various probability distribution sets over \( T \) with respect to knowledge regarding related concepts. We will use \( z \) to denote a discrete probability distribution function \( \{ p(w|z) \}_{w \in \Theta} \) over \( T \), which we will further refer to as a topic. Each topic \( z \) is a member of \( \Theta = \{ z_1, ..., z_k \} \) a finite set of \( k \) topics extracted using one of the known algorithms given \( T \), with \( \Theta \subseteq Z \), where \( Z \) represents the set of all possible topics over \( T \). Furthermore, \( \Theta \subseteq P(Z) \), the space of all possible probability distribution sets over \( T \).

2.1 Concept Representation

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

2.2 Employed Distances

Within \( C \), 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 \((c_i, c_0)\) or \((c_j, c_0)\) 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 \( C \), \( d(c_i, c_j) \), 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 \( V_C \subseteq V \) where, for each word in \( V_C \) there exists at least one concept in \( C \) that is a sense of the given word. Let \( \delta(w) \subseteq \delta(C) \) be the set of all senses of the word \( w \) within \( C \). 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 \( c \):

\[
\text{dist}(w, c) = \min_{c \in \delta(w)} d(c, c) \forall c \in C, \forall w \in V_C
\]

Let \( c_0(C) \in \delta(C) \) be the root of the \( C \) tree. For instance, in Fig. 1 the distance between \( w_1 \) and the subtree root \( c_0(C) \) is 2, because the distances between its senses and \( c_0(C) \) are
\[ d(w_{1,1}, c_0(C)) = 3 \text{ and } d(w_{1,2}, c_0(C)) = 2, \] where \( w_{1,1} \) and \( w_{1,2} \) are \( w_1 \)'s senses.

\[ C_{z_i} = \{ \delta(z_i, c_i) \}, \] which in the case of \( C_{z_1} \) is outlined with a bold line (dotted or not).

Definition 5. A topical subtree of a concept \( c \) within \( C \) is the reunion of all the \( C_{w,c} \) subtrees of all topical words:

\[ C_{z_i,c} = (\delta(z_i, C_c), r); \delta(C_i, C_c) = \bigcup_{w \in \delta(w, C)} \delta(w, c_0(C)), \forall z_i \in \Theta, \] the reunion of all the \( C_{w,c} \) relevant words.

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 \( cov \): \( \delta(z_i, C) \times \Theta \rightarrow \mathbb{N}_+ \),

\[ c \rightarrow cov(c, z_i) = \frac{card(\{d(w, C)_i \in \delta(z_i, C)\})}{card(d(z_i))}. \] In order to determine a concept’s specificity we rely on two additional features – its height \( h : \delta(C) \rightarrow \mathbb{N}_+ \),

\[ c \rightarrow h(c) = d(c, c_0(C)), \] and depth \( \rho \) with respect to the given topic \( p : \delta(z_i, C), \Theta \rightarrow \mathbb{N}_+ \),

\[ (c, z_i) \rightarrow \rho(c, z_i) = <f >_{w \in \delta(w, C)} (d(w, C) \in \delta(z_i)), \forall z_i \in \Theta, \forall c \in \delta(z_i, C), \] 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:

\[ <f > : \mathbb{R}^k \rightarrow \mathbb{R}_+ ; \cdots \rightarrow <f > \left( \begin{array}{c} x_1 \\ \vdots \\ x_k \end{array} \right) = y \in \mathbb{R}_+. \]

There are more than one possible definitions for the specificity, one of which is a weighted average of the concept’s height and depth:

\[ spec : C_{z_i} \times \Theta \rightarrow \mathbb{R}; c, z_i \rightarrow spec(c, z_i) = \omega_h \cdot h(c) + \omega_p \cdot \rho(c, z_i), \] 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 \( \Theta \),

\[ \eta : P(Z) \rightarrow \mathbb{R}_+ ; \Theta \rightarrow \eta(\Theta) = <f > (\eta(z)), \forall z \in \Theta. \] The evaluation of the model as a whole depends on the individual evaluation of each of its topics \( \eta : \Theta \rightarrow \mathbb{R}_+ , z_i \rightarrow \eta(z_i) = <f > (\phi(C_{z_i}), \forall z_i \in \Theta, \) in which \( C_{z_i} \) is the topical subtree of \( z_i \) given \( C \). Each node from \( C_{z_i} \) is assigned a positive fitness value based on its relevance to the given topic \( z_i \).

\[ \phi : C_{z_i} \times \Theta \rightarrow \mathbb{R}_+ \]. We express this relevance as a weighted average of its coverage and specificity, with \( \omega_{cov} \) and \( \omega_{spec} \) set a priori,

\[ c \rightarrow \phi(c) = \omega_{cov} \cdot cov(c, z_i) + \omega_{spec} \cdot spec(c, z_i) \]

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 – \(hit_{+}\) and \(hit_{-}\). The difference between the two, as a percentage of \(hit_{-}\) is shown as the gain, in the last column of Table 1.

<table>
<thead>
<tr>
<th>Data</th>
<th>Type</th>
<th>(hit_{+})</th>
<th>(\sigma(hit_{+}))</th>
<th>(hit_{-})</th>
<th>(\sigma(hit_{-}))</th>
<th>(+)(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Close</td>
<td>0.37</td>
<td>0.29</td>
<td>0.27</td>
<td>0.23</td>
<td>39.33</td>
</tr>
<tr>
<td></td>
<td>Far</td>
<td>0.69</td>
<td>0.29</td>
<td>0.65</td>
<td>0.23</td>
<td>6.93</td>
</tr>
<tr>
<td>Suall</td>
<td>Close</td>
<td>0.51</td>
<td>0.23</td>
<td>0.3</td>
<td>0.24</td>
<td>66.76</td>
</tr>
<tr>
<td></td>
<td>Far</td>
<td>0.75</td>
<td>0.24</td>
<td>0.59</td>
<td>0.33</td>
<td>28.55</td>
</tr>
</tbody>
</table>

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. \(hit_{+}\). 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.

<table>
<thead>
<tr>
<th>Data</th>
<th>Type</th>
<th>(hit_{+})</th>
<th>(\sigma(hit_{+}))</th>
<th>(hit_{-})</th>
<th>(\sigma(hit_{-}))</th>
<th>(+)(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Close</td>
<td>0.27</td>
<td>0.26</td>
<td>0.21</td>
<td>0.2</td>
<td>32.08</td>
</tr>
<tr>
<td></td>
<td>Far</td>
<td>0.6</td>
<td>0.32</td>
<td>0.53</td>
<td>0.24</td>
<td>13.92</td>
</tr>
<tr>
<td>Suall</td>
<td>Close</td>
<td>0.48</td>
<td>0.23</td>
<td>0.23</td>
<td>0.18</td>
<td>105.4</td>
</tr>
<tr>
<td></td>
<td>Far</td>
<td>0.71</td>
<td>0.23</td>
<td>0.44</td>
<td>0.32</td>
<td>60.35</td>
</tr>
</tbody>
</table>

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 | \(\bar{w}_c\) | \(\sigma(w_c)\) | \(\bar{w}_r\) | \(\sigma(w_r)\) | +(%)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Close</td>
<td>1.35</td>
<td>0.38</td>
<td>1.4</td>
<td>0.43</td>
<td>3.45</td>
</tr>
<tr>
<td></td>
<td>Far</td>
<td>1.31</td>
<td>0.38</td>
<td>1.54</td>
<td>0.39</td>
<td>15.19</td>
</tr>
<tr>
<td>Suall</td>
<td>Close</td>
<td>1.24</td>
<td>0.32</td>
<td>1.37</td>
<td>0.37</td>
<td>9.05</td>
</tr>
<tr>
<td></td>
<td>Far</td>
<td>1.28</td>
<td>0.31</td>
<td>1.43</td>
<td>0.38</td>
<td>10.93</td>
</tr>
</tbody>
</table>

Table 3. Chosen Words Number

Results show that the number of chosen words \(\bar{w}\) 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.

<table>
<thead>
<tr>
<th>(\frac{hit_c - hit_r}{hit_r})</th>
<th>(\frac{fhit_c - fhit_r}{fhit_r})</th>
<th>(\frac{w_c - w_r}{w_r})</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>Close</td>
<td>Far</td>
</tr>
<tr>
<td>30</td>
<td>-9.55</td>
<td>-2.52</td>
</tr>
<tr>
<td>50</td>
<td>-37.09</td>
<td>-20.27</td>
</tr>
<tr>
<td>100</td>
<td>35.29</td>
<td>32.8</td>
</tr>
<tr>
<td>200</td>
<td>80.06</td>
<td>42</td>
</tr>
<tr>
<td>300</td>
<td>100</td>
<td>65.42</td>
</tr>
</tbody>
</table>

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.

3.4.1 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 \(hit_{+i}\) or \(hit_{-i}\) 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, \(hit_{+/-k}\) and \(\sigma(hit_{+/-k})\) and subtracted the average value for all \(k\) values, \(hit_{+/-} = \sum_k hit_{+/-k}/k\) and \(\sigma(hit_{+/-}) = \sum_k \sigma(hit_{+/-k})/k\):

- \(\partial hit_{+/-k} = \frac{hit_{+/-k} - hit_{+/-}}{hit_{+/-}}\)
- \(\partial \sigma_{+/-k} = \frac{\sigma(hit_{+/-k}) - \sigma(hit_{+/-})}{\sigma(hit_{+/-})}\)

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

- \(\partial rel_{+/-k} = rel_{+/-k} - rel_{+/-}\)

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.

<table>
<thead>
<tr>
<th>(k)</th>
<th>(\partial hit_{-k})</th>
<th>(\partial \sigma_{-k})</th>
<th>(\partial rel_{-k})</th>
<th>(\partial hit_{+k})</th>
<th>(\partial \sigma_{+k})</th>
<th>(\partial rel_{+k})</th>
<th>(\partial hit_{+/-k} - \partial hit_{-/-k})</th>
<th>(\partial \sigma_{+/-k} - \partial \sigma_{-/-k})</th>
<th>(\partial rel_{+/-k} - \partial rel_{-/-k})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close Spurious Word</td>
<td>30</td>
<td>0.08</td>
<td>0.04</td>
<td>0.22</td>
<td>-0.02</td>
<td>0.05</td>
<td>-0.21</td>
<td>0.01</td>
<td>-0.43</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.02</td>
<td>0.06</td>
<td>0.06</td>
<td>-0.16</td>
<td>0.0</td>
<td>0.03</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.05</td>
<td>-0.06</td>
<td>0</td>
<td>-0.05</td>
<td>0.03</td>
<td>0.08</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.23</td>
<td>-0.04</td>
<td>0.1</td>
<td>0.23</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.17</td>
<td>0</td>
<td>-0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.704</td>
<td>-0.016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distant Spurious Word</td>
<td>30</td>
<td>-0.11</td>
<td>-0.01</td>
<td>0.22</td>
<td>-0.18</td>
<td>0.11</td>
<td>-0.21</td>
<td>-0.07</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>-0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>-0.08</td>
<td>0.04</td>
<td>-0.03</td>
<td>-0.05</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.02</td>
<td>0.01</td>
<td>0.13</td>
<td>-0.03</td>
<td>0.08</td>
<td>0.11</td>
<td>-0.03</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.04</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.02</td>
<td>-0.04</td>
<td>0.1</td>
<td>-0.02</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.07</td>
<td>-0.04</td>
<td>-0.17</td>
<td>0.1</td>
<td>-0.08</td>
<td>0.06</td>
<td>0.03</td>
<td>-0.04</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.557</td>
<td>-0.979</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 P-SDRU/6/1.5/S/19 7713

References


