Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality

Jey Han Lau
Dept of Philosophy
King's College London
jeyhan.lau@gmail.com

David Newman
Google
dnewman@google.com

Timothy Baldwin
Dept of Computing and
Information Systems
The University of Melbourne
tb@ldwin.net

Abstract

Topic models based on latent Dirichlet allocation and related methods are used in a range of user-focused tasks including document navigation and trend analysis, but evaluation of the intrinsic quality of the topic model and topics remains an open research area. In this work, we explore the two tasks of automatic evaluation of single topics and automatic evaluation of whole topic models, and provide recommendations on the best strategy for performing the two tasks, in addition to providing an open-source toolkit for topic and topic model evaluation.

1 Introduction

Topic modelling based on Latent Dirichlet Allocation (LDA: Blei et al. (2003)) and related methods is increasingly being used in user-focused tasks, in contexts such as the evaluation of scientific impact (McCallum et al., 2006; Hall et al., 2008), trend analysis (Bolelli et al., 2009; Lau et al., 2012a) and document search (Wang et al., 2007). The LDA model is based on the assumption that document collections have latent topics, in the form of a multinomial distribution of words, which is typically presented to users via its top-N highest-probability words. In NLP, topic models are generally used as a means of preprocessing a document collection, and the topics and per-document topic allocations are fed into downstream applications such as document summarisation (Haghighi and Vanderwende, 2009), novel word sense detection methods (Lau et al., 2012b) and machine translation (Zhao and Xing, 2007). In fields such as the digital humanities, on the other hand, human users interact directly with the output of topic models. It is this context of topic modelling for direct human consumption that we target in this paper.

The topics produced by topic models have a varying degree of human-interpretability. To illustrate this, we present two topics automatically learnt from a collection of news articles:

1. ⟨*farmers, farm, food, rice, agriculture*⟩
2. ⟨*stories, undated, receive, scheduled, clients*⟩

The first topic is clearly related to agriculture. The subject of the second topic, however, is less clear, and may confuse users if presented to them as part of a larger topic model. Measuring the human-interpretability of topics and the overall topic model is the core topic of this paper.

Various methodologies have been proposed for measuring the semantic interpretability of topics. In Chang et al. (2009), the authors proposed an indirect approach based on word intrusion, where "intruder words" are randomly injected into topics and human users are asked to identify the intruder words. The word intrusion task builds on the assumption that the intruder words are more identifiable in coherent topics than in incoherent topics, and thus the interpretability of a topic can be estimated by measuring how readily the intruder words can be manually identified by annotators.

Since its inception, the method of Chang et al. (2009) has been used variously as a means of assessing topic models (Paul and Girju, 2010; Reisinger et al., 2010; Hall et al., 2012). Despite its wide acceptance, the method relies on manual annotation and has never been automated. This is one of the primary contributions of this work: the demonstration that we can automate the method of Chang et al. (2009) at near-human levels of accuracy, as a result of which we can perform automatic evaluation of the human-interpretability of topics, as well as topic models.

There has been prior work to directly estimate the human-interpretability of topics through automatic means. For example, Newman et al.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 530–539,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

(2010) introduced the notion of topic "coherence", and proposed an automatic method for estimating topic coherence based on pairwise pointwise mutual information (PMI) between the topic words. Mimno et al. (2011) similarly introduced a methodology for computing coherence, replacing PMI with log conditional probability. Musat et al. (2011) incorporated the WordNet hierarchy to capture the relevance of topics, and in Aletras and Stevenson (2013a), the authors proposed the use of distributional similarity for computing the pairwise association of the topic words. One application of these methods has been to remove incoherent topics before generating labels for topics (Lau et al., 2011; Aletras and Stevenson, 2013b).

Ultimately, all these methodologies, and also the word intrusion approach, attempt to assess the same quality: the human-interpretability of topics. The relationship between these methodologies, however, is poorly understood, and there is no consensus on what is the best approach for computing the semantic interpretability of topic models. This is a second contribution of this paper: we perform a systematic empirical comparison of the different methods and find appreciable differences between them. We further go on to propose an improved formulation of Newman et al. (2010) based on normalised PMI. Finally, we release a toolkit which implements the topic interpretability measures described in this paper.

2 Related Work

Chang et al. (2009) challenged the conventional wisdom that held-out likelihood — often computed as the perplexity of test data or unseen documents — is the only way to evaluate topic models. To measure the human-interpretability of topics, the authors proposed a word intrusion task and conducted experiments using three topic models: Latent Dirichlet Allocation (LDA: Blei et al. (2003)), Probabilistic Latent Semantic Indexing (PLSI: Hofmann (1999)) and the Correlated Topic Model (CTM: Blei and Lafferty (2005)). Contrary to expectation, they found that perplexity correlates negatively with topic interpretability.

In the word intrusion task, each topic is presented as a list of six words — the five most probable topic words and a randomly-selected "intruder word", which has low probability in the topic of interest, but high probability in other topics — and human users are asked to identify the intruder word that does not belong to the topic in question.

Newman et al. (2010) capture topic interpretability using a more direct approach, by asking human users to rate topics (represented by their top-10 topic words) on a 3-point scale based on how coherent the topic words are (i.e. their observed coherence). They proposed several ways of automating the estimation of the observed coherence, and ultimately found that a simple method based on PMI term co-occurrence within a sliding context window over English Wikipedia produces the consistently best result, nearing levels of inter-annotator agreement over topics learnt from two distinct document collections.

Mimno et al. (2011) proposed a closely-related method for evaluating semantic coherence, replacing PMI with log conditional probability. Rather than using Wikipedia for sampling the word co-occurrence counts, Mimno et al. (2011) used the topic-modelled documents, and found that their measure correlates well with human judgements of observed coherence (where topics were rated in the same manner as Newman et al. (2010), based on a 3-point ordinal scale). To incorporate the evaluation of semantic coherence into the topic model, the authors proposed to record words that co-occur together frequently, and update the counts of all associated words before and after the sampling of a new topic assignment in the Gibbs sampler. This variant of topic model was shown to produce more coherent topics than LDA based on the log conditional probability coherence measure.

Aletras and Stevenson (2013a) introduced distributional semantic similarity methods for computing coherence, calculating the distributional similarity between semantic vectors for the top-N topic words using a range of distributional similarity measures such as cosine similarity and the Dice coefficient. To construct the semantic vector space for the topic words, they used English Wikipedia as the reference corpus, and collected words that co-occur in a window of ± 5 words. They showed that their method correlates well with the observed coherence rated by human judges.

3 Dataset

As one of the primary foci of this paper is the automation of the intruder word task of Chang et al. (2009), our primary dataset is that used in the original paper by Chang et al. (2009), which provides topics and human annotations for a range of

domains and topic model types. In the dataset, two text collections were used: (1) 10,000 articles from English Wikipedia (WIKI); and (2) 8,447 articles from the New York Times dating from 1987 to 2007 (NEWS). For each document collection, topics were generated by three topic modelling methods: LDA, PLSI and CTM (see Section 2). For each topic model, three settings of T (the number of topics) were used: $T = 50$, $T = 100$ and $T = 150$. In total, there were 9 topic models (3 models $\times$ 3 T) and 900 topics (3 models $\times$ $(50 + 100 + 150)$) for each dataset.[1]

For some of topic interpretability estimation methods, we require a reference corpus to sample lexical probabilities. We use two reference corpora: (1) NEWS-FULL, which contains 1.2 million New York Times articles from 1994 to 2004 (from the English Gigaword); and (2) WIKI-FULL, which contains 3.3 million English Wikipedia articles (retrieved November 28th 2009).[2] The rationale for choosing the New York Times and English Wikipedia as the reference corpora is to ensure domain consistency with the word intrusion dataset; the full collections are used to more robustly estimate lexical probabilities.

4 Human-Interpretability at the Model Level

In this section, we evaluate measures for estimating human-interpretability at the *model* level. That is, for a measure — human-judged or automated — we first aggregate its coherence/interpretability scores for all topics from a given topic model to obtain the topic model's average coherence score. We then calculate the Pearson correlation coefficients between the two measures using the topic models' average coherence scores. In summary, the correlation is computed over nine sets of topics (3 topic modellers $\times$ 3 settings of T) for each of WIKI and NEWS.

4.1 Indirect Approach: Word Intrusion

The word intrusion task measures topic interpretability indirectly, by computing the fraction of annotators who successfully identify the intruder word. A limitation of the word intrusion task is that it requires human annotations, therefore preventing large-scale evaluation. We begin by proposing a methodology to fully automate the word intrusion task.

Lau et al. (2010) proposed a methodology that learns the *most representative* or *best* topic word that summarises the semantics of the topic. Observing that the word intrusion task — the task of detecting the *least representative* word — is the converse of the best topic word selection task, we adapt their methodology to automatically identify the intruder word for the word intrusion task, based on the knowledge that there is a unique intruder word per topic.

The methodology works as follows: given a set of topics (including intruder words), we compute the word association features for each of the top-N topic words of a topic,[3] and combine the features in a ranking support vector regression model (SVM^{rank}: Joachims (2006)) to learn the intruder words. Following Lau et al. (2010), we use three word association measures:

$$\text{PMI}(w_i) = \sum_{j}^{N-1} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

$$\text{CP1}(w_i) = \sum_{j}^{N-1} \frac{P(w_i, w_j)}{P(w_j)}$$

$$\text{CP2}(w_i) = \sum_{j}^{N-1} \frac{P(w_i, w_j)}{P(w_i)}$$

We additionally experiment with normalised pointwise mutual information (NPMI: Bouma (2009)):

$$\text{NPMI}(w_i) = \sum_{j}^{N-1} \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}$$

In the dataset of Chang et al. (2009) (see Section 3), each topic was presented to 8 annotators, with small variations in the displayed topic words (including the intruder word) for each annotator. That is, each topic has essentially 8 subtly different representations. To measure topic interpretability, the authors defined "model precision": the relative success of human annotators at identifying the intruder word, across all representations of the different topics. The model precision scores produced by human judges are henceforth referred to as WI-Human, and the scores produced by our

[1] In the WIKI topics there were corrupted symbols in the topic words for 24 topics. We removed these topics, reducing the total number of topics to 876.

[2] For both corpora we perform tokenisation and POS tagging using OpenNLP and lemmatisation using Morpha (Minnen et al., 2001).

[3] N is the number of topic words displayed to the human users in the word intrusion task, including the intruder word.

Topic Domain	Ref. Corpus	Pearson's r with WI-Human	
		WI-Auto-PMI	WI-Auto-NPMI
WIKI	WIKI-FULL	0.947	0.936
	NEWS-FULL	0.801	0.835
NEWS	NEWS-FULL	0.913	0.831
	WIKI-FULL	0.811	0.750

Table 1: Pearson correlation of WI-Human and WI-Auto-PMI/WI-Auto-NPMI at the model level.

automated method for the PMI and NPMI variants as WI-Auto-PMI and WI-Auto-NPMI respectively.[4]

The Pearson correlation coefficients between WI-Human and WI-Auto-PMI/WI-Auto-NPMI at the model level are presented in Table 1. Note that our two reference corpora are used to independently sample the lexical probabilities for the word association features.

We see very strong correlation for in-domain pairings (i.e. WIKI+WIKI-FULL and NEWS+NEWS-FULL), achieving $r > 0.9$ in most cases for both WI-Auto-PMI or WI-Auto-NPMI, demonstrating the effectiveness of our methodology at automating the word intrusion task for estimating human-interpretability at the model level. Overall, WI-Auto-PMI outperforms WI-Auto-NPMI.

Note that although our proposed methodology is supervised, as intruder words are synthetically generated and no annotation is needed for the supervised learning, the whole process of computing topic coherence via word intrusion is fully automatic, without the need for hand-labelled training data.

4.2 Direct Approach: Observed Coherence

Newman et al. (2010) defined topic interpretability based on a more direct approach, by asking human judges to rate topics based on the observed coherence of the top-N topic words, and various methodologies have since been proposed to automate the computation of the observed coherence. In this section, we present all these methods and compare them.

The word intrusion dataset is not annotated with human ratings of observed coherence. To create gold-standard coherence judgements, we used Amazon Mechanical Turk:[5] we presented the topics (with intruder words removed) to the Turkers and asked them to rate the topics using on a 3-point ordinal scale, following Newman et al. (2010). In total, we collected six to fourteen annotations per topic (an average of 8.4 annotations per topic). The observed coherence of a topic is computed as the arithmetic mean of the annotators' ratings, once again following Newman et al. (2010). The human-judged observed topic coherence is henceforth referred to as OC-Human.

For the automated methods, we experimented with the following methods for estimating the human-interpretability of a topic t:

1. **OC-Auto-PMI:** Pairwise PMI of top-N topic words (Newman et al., 2010):
$$\text{OC-Auto-PMI}(\mathbf{t}) = \sum_{j=2}^{N}\sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)P(w_j)}$$
2. **OC-Auto-NPMI:** NPMI variant of OC-Auto-PMI:
$$\text{OC-Auto-NPMI}(\mathbf{t}) = \sum_{j=2}^{N}\sum_{i=1}^{j-1} \frac{\log \frac{P(w_j,w_i)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}$$
3. **OC-Auto-LCP:** Pairwise log conditional probability of top-N topic words (Mimno et al., 2011):[6]
$$\text{OC-Auto-LCP}(\mathbf{t}) = \sum_{j=2}^{N}\sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)}$$
4. **OC-Auto-DS:** Pairwise distributional similarity of the top-N topic words, as described in Aletras and Stevenson (2013a).

For OC-Auto-PMI, OC-Auto-NPMI and OC-Auto-LCP, all topics are lemmatised and intruder words are removed before coherence is computed.[7] In-domain and cross-domain pairings of

[4] Note that both variants use CP1 and CP2 features, i.e. WI-Auto-PMI uses PMI+CP1+C2 while WI-Auto-NPMI uses NPMI+CP1+C2 features.

[5] https://www.mturk.com/mturk/

[6] Although the original method uses the topic-modelled document collection and document co-occurrence for sampling word counts, for a fairer comparison we use log conditional probability only as a replacement to the PMI component of the coherence computation (i.e. words are still sampled using a reference corpus and a sliding window). For additional evidence that the original method performs at a subpar level, see Lau et al. (2013) and Aletras and Stevenson (2013a).

[7] We once again use Morpha to do the lemmatisation, and determine POS via the majority POS for a given word, aggregated over all its occurrences in English Wikipedia.

Topic Domain	Ref. Corpus	Pearson's r with OC-Human			
		OC-Auto-PMI	OC-Auto-NPMI	OC-Auto-LCP	OC-Auto-DS
WIKI	WIKI-FULL	0.490	0.903	0.959	0.859
	NEWS-FULL	0.696	0.844	0.913	
NEWS	NEWS-FULL	0.965	0.979	0.887	0.941
	WIKI-FULL	0.931	0.964	0.872	

Table 2: Pearson correlation of OC-Human and the automated methods — OC-Auto-PMI, OC-Auto-NPMI, OC-Auto-LCP and OC-Auto-DS — at the model level.

the topic domain and reference corpus are experimented with for these measures.

For OC-Auto-DS, all topics are lemmatised, intruder words are removed and English Wikipedia is used to generate the vector space for the topic words. The size of the context window is set to ± 5 word (i.e. 5 words to either side of the target word). We use PMI to weight the vectors, cosine similarity for measuring the distributional similarity between the top-N topic words, and the "Topic Word Space" approach to reduce the dimensionality of the vector space. A complete description of the parameters can be found in Aletras and Stevenson (2013a). Note that cross-domain pairings of the topic domain and reference corpus are not tested: in line with the original paper, we use only English Wikipedia to generate the vector space before distributional similarity.

We present the Pearson correlation coefficient of OC-Human and the four automated methods at the model level in Table 2. For OC-Auto-NPMI, OC-Auto-LCP and OC-Auto-DS, we see that they correlate strongly with the human-judged coherence. Overall, OC-Auto-NPMI has the best performance among the methods, and in-domain pairings generally produce the best results for OC-Auto-NPMI and OC-Auto-LCP. The results are comparable to those for the automated intruder word detection method in Section 4.1.

The non-normalised variant OC-Auto-PMI correlates well for NEWS but performs poorly for WIKI, producing a correlation of only 0.490 for the in-domain pairing. We revisit this in Section 6, and provide a qualitative analysis to explain the discrepancy in results between OC-Auto-PMI and OC-Auto-NPMI.

4.3 Word Intrusion vs. Observed Coherence

In the previous sections, we showed for both the direct and indirect approaches that the automated methods correlate strongly with the manually-annotated human-interpretability of topics at the model level (with the exception of OC-Auto-PMI). One question that remains unanswered, however, is whether word intrusion measures topic interpretability differently to observed coherence. This is the focus of this section.

From the results in Table 3 for the intruder word model vs. observed coherence, we see a strong correlation between WI-Human and OC-Human. This observation is insightful: it shows that the topic interpretability estimated by the two approaches is almost identical at the model level.

Between WI-Human and the observed coherence methods automated methods, overall we see a strong correlation for the OC-Auto-NPMI, OC-Auto-LCP and OC-Auto-DS methods. OC-Auto-PMI once again performs poorly over WIKI, but this is unsurprising given its previous results (i.e. its poor correlation with OC-Human). In-domain pairings tend to perform better, and the performance of OC-Auto-NPMI, OC-Auto-LCP and OC-Auto-DS is comparable, with no one clearly best method.

5 Human-Interpretability at the Topic Level

In this section, we evaluate the various methods at the *topic* level. We group together all topics for each dataset (without distinguishing the topic models that produce them) and calculate the correlation of one measure against another. That is, the correlation coefficient is computed for 900 topics/data points in the case of each of WIKI and NEWS.

5.1 Indirect Approach: Word Intrusion

In Section 4.1, we proposed a novel methodology to automate the word intrusion task (WI-Auto-PMI and WI-Auto-NPMI). We now evaluate its performance at the topic level, and present its correlation with the human gold standard (WI-Human) in Table 4.

The correlation of WI-Human and WI-Auto-PMI/WI-Auto-NPMI at the topic level is considerably worse, compared to its results at the model

Topic Domain	Ref. Corpus	Pearson's r with WI-Human				
		OC-Human	OC-Auto-PMI	OC-Auto-NPMI	OC-Auto-LCP	OC-Auto-DS
WIKI	WIKI-FULL	0.900	0.638	0.927	0.911	0.907
	NEWS-FULL		0.614	0.757	0.821	
NEWS	NEWS-FULL	0.915	0.865	0.866	0.867	0.925
	WIKI-FULL		0.838	0.874	0.893	

Table 3: Word intrusion vs. observed coherence: Pearson correlation coefficient at the model level.

Topic Domain	Ref. Corpus	Pearson's r with WI-Human		Human Agreement
		WI-Auto-PMI	WI-Auto-NPMI	
WIKI	WIKI-FULL	0.554	0.573	0.735
	NEWS-FULL	0.622	0.592	
NEWS	NEWS-FULL	0.602	0.612	0.770
	WIKI-FULL	0.638	0.648	

Table 4: Pearson correlation coefficient of WI-Human and WI-Auto-PMI/WI-Auto-NPMI at the topic level.

level (Table 1). The performance between WI-Auto-PMI and WI-Auto-NPMI is not very different, and the cross-domain pairing slightly outperforms the in-domain pairing.

To better understand the difficulty of the task, we compute the agreement between human annotators by calculating the Pearson correlation coefficient of model precisions produced by randomised sub-group pairs in the topics.[8] That is, for each topic, we randomly split the annotations into two sub-groups, and compute the Pearson correlation coefficient of the model precisions produced by the first sub-group and that of the second sub-group.

The original dataset has 8 annotations per topic. Splitting the annotations into two sub-groups reduces the number of annotations to 4 per group, which is not ideal for computing model precision. We thus chose to expand the number of annotations by sampling 300 random topics from each domain (for a total of 600 topics) and following the same process as Chang et al. (2009) to get intruder word annotations using Amazon Mechanical Turk. On average, we obtained 11.7 *additional* annotations per topic for these 600 topics. The human agreement scores (i.e. the Pearson correlation coefficient of randomised sub-group pairs) for the sampled 600 topics are presented in the last column of Table 4.

The sub-group correlation is around $r = 0.75$ for the topics from both datasets. As such, estimating topic interpretability at the topic level is a much harder task than model-level evaluation. Our automated methods perform at a highly credible $r = 0.6$, but there is certainly room for improvement. Note that the correlation values reported in Newman et al. (2010) are markedly higher than ours, as they evaluated based on Spearman rank correlation, which isn't attuned to the relative differences in coherence values and returns higher values for the task.

[8]To counter for the fact that annotators labelled varying numbers of topics.

5.2 Direct Approach: Observed Coherence

We repeat the experiments of observed coherence in Section 4.2, and evaluate the correlation of the automated methods (OC-Auto-PMI, OC-Auto-NPMI, OC-Auto-LCP and OC-Auto-DS) on the human gold standard (OC-Human) at the topic level. Results are summarised in Table 5.

OC-Auto-PMI performs poorly at the topic level in the WIKI domain, similar to what was seen at the model level in Section 4.2. Overall, both OC-Auto-NPMI and OC-Auto-DS are the most consistent methods. OC-Auto-LCP performs markedly worse than these two methods.

To get a better understanding of how well human annotators perform at the task, we compute the one-vs-rest Pearson correlation coefficient using the gold standard annotations. That is, for each topic, we single out each rating/annotation and compare it to the average of all other ratings/annotations. The one-vs-rest correlation result is displayed in the last column (titled "Human Agreement") in Table 5. The best automated methods surpass the single-annotator performance, indicating that they are able to perform the task as well as human annotators (unlike the topic-level results for the word intrusion task where humans were markedly better at the task than the automated methods).

Topic Domain	Ref. Corpus	Pearson's r with OC-Human				Human Agreement
		OC-Auto-PMI	OC-Auto-NPMI	OC-Auto-LCP	OC-Auto-DS	
WIKI	WIKI-FULL	0.533	0.638	0.579	0.682	0.624
	NEWS-FULL	0.582	0.667	0.496		
NEWS	NEWS-FULL	0.719	0.741	0.471	0.682	0.634
	WIKI-FULL	0.671	0.722	0.452		

Table 5: Pearson correlation of OC-Human and the automated methods at the topic level.

Topic Domain	Ref. Corpus	Pearson's r with WI-Human				
		OC-Human	OC-Auto-PMI	OC-Auto-NPMI	OC-Auto-LCP	OC-Auto-DS
WIKI	WIKI-FULL	0.665	0.472	0.557	0.547	0.639
	NEWS-FULL		0.504	0.571	0.455	
NEWS	NEWS-FULL	0.641	0.629	0.634	0.407	0.649
	WIKI-FULL		0.604	0.633	0.390	

Table 6: Word intrusion vs. observed coherence: pearson correlation results at the topic level.

5.3 Word Intrusion vs. Observed Coherence

In this section, we bring together the indirect approach of word intrusion and the direct approach of observed coherence, and evaluate them against each other at the topic level. Results are summarised in Table 6.

We see that the correlation between the human ratings of intruder words and observed coherence is only modest, implying that there are topic-level differences in the output of the two approaches. In Section 6, we provide a qualitative analysis and explanation as to what constitutes the differences between the approaches.

For the automated methods, OC-Auto-DS has the best performance, with OC-Auto-NPMI performing relatively well (in particularly in the NEWS domain).

6 Discussion

Normalised PMI (NPMI) was first introduced by Bouma (2009) as a means of reducing the bias for PMI towards words of lower frequency, in addition to providing a standardised range of $[-1, 1]$ for the calculated values.

We introduced NPMI to the automated methods of word intrusion (WI-Auto-NPMI) and observed coherence (OC-Auto-NPMI) to explore its suitability for the task. For the latter, we saw that NPMI achieves markedly higher correlation than OC-Human (in particular, at the model level). To better understand the impact of normalisation, we inspected a list of WIKI topics that have similar scores for OC-Human and OC-Auto-NPMI but very different OC-Auto-PMI scores. A sample of these topics is presented in Table 7. WIKI-FULL is used as the reference corpus for computing the scores. Note that the presented OC-Auto-NPMI* and OC-Auto-PMI* scores are post-normalised to the range $[0, 1]$ for ease of interpretation. To give a sense of how readily these topic words occur in the reference corpus, we additionally display the frequency of the first topic word in the reference corpus (last column).

All topics presented have an OC-Human score of 3.0 (i.e. these topics are rated as being very coherent by human judges) and similar OC-Auto-NPMI values. Their OC-Auto-PMI scores, however, are very different between the top-3 and bottom-3 topics. The bias of PMI towards lower frequency words is clear: topic words that occur frequently in the corpus receive a lower OC-Auto-PMI score compared to those that occur less frequently, even though the human-judged observed coherence is the same. OC-Auto-NPMI on the other hand, correctly estimates the coherence.

We observed, however, that the impact of normalising PMI is less in the word intrusion task. One possible explanation is that for the automated methods WI-Auto-PMI and WI-Auto-NPMI, the PMI/NPMI scores are used indirectly as a feature to a machine learning framework, and the bias could be reduced/compensated by other features.

On the subject of the difference between observed coherence and word intrusion in estimating topic interpretability, we observed that WI-Human and OC-Human correlate only moderately ($r \approx 0.6$) at the topic level (Table 6). To better understand this effect, we manually analysed topics that have differing WI-Human and OC-Human scores. A sample of topics with high divergence in estimated coherence score is given in Table 8. As before, the presented the OC-Human* and WI-

Topic	OC-Human	OC-Auto-NPMI*	OC-Auto-PMI*	Word Count
cell hormone insulin muscle receptor	3.0	0.59	0.61	#(cell) = 1.1M
electron laser magnetic voltage wavelength	3.0	0.52	0.54	#(electron) = 0.3M
magnetic neutrino particle quantum universe	3.0	0.55	0.55	#(magnetic) = 0.4M
album band music release song	3.0	0.56	0.37	#(album) = 12.5M
college education school student university	3.0	0.57	0.38	#(college) = 9.8M
city county district population town	3.0	0.52	0.34	#(city) = 22.0M

Table 7: A list of WIKI topics to illustrate the impact of NPMI.

Topic #	Topic	OC-Human*	WI-Human*
1	business company corporation **cluster** [loch] shareholder	0.94	0.25
2	**song** actor [clown] play role theatre	1.00	0.50
3	census [ethnic] female male population **village**	0.92	0.25
4	composer **singer** jazz music opera [piano]	1.00	0.63
5	choice count give i.e. simply [**unionist**]	0.14	1.00
6	[**digital**] clown friend love mother wife	0.17	1.00

Table 8: A list of WIKI topics to illustrate the difference between observed coherence and word intrusion. Boxes denote human chosen intruder words, and boldface denotes true intruder words.

Human* scores in the table are post-normalised to the range $[0, 1]$ for ease of comparison.

In general, there are two reasons for topics to have high OC-Human and low WI-Human scores. First, if a topic has an outlier word that is mildly related to the topic, users tend to choose this word as the intruder word in the word intrusion task, yielding a low WI-Human score. If they are asked to rate the observed coherence, however, the single outlier word often does not affect its overall coherence, resulting in a high OC-Human score. This is observed in topics 1 and 2 in Table 8, where *loch* and *clown* are chosen by annotators in the word intrusion task, as they detract from the semantics of the topic. This results in low WI-Human scores, but high observed coherence scores (OC-Human).

The second reason is the random selection of intruder words related to the original topic. We see this in topics 3 and 4, where related intruder words (*village* and *singer*) were selected.

For topics with low OC-Human and high WI-Human scores, the true intruder words are often very different to the domain/focus of other topic words. As such, annotators are consistently able to single them out to yield high WI-Human scores, even though the topic as a whole is not coherent. Topics 5 and 6 in Table 8 exhibit this.

All topic evaluation measures described in this paper are implemented in an open-source toolkit.[9]

[9] https://github.com/jhlau/topic_interpretability

7 Conclusion

In this paper, we examined various methodologies that estimate the semantic interpretability of topics, at two levels: the model level and the topic level. We looked first at the word intrusion task proposed by Chang et al. (2009), and proposed a method that fully automates the task. Next we turned to observed coherence, a more direct approach to estimate topic interpretability. At the model level, results were very positive for both the word intrusion and observed coherence methods. At the topic level, however, the results were more mixed. For observed coherence, our best methods (OC-Auto-NPMI and OC-Auto-DS) were able to emulate human performance. For word intrusion, the automated methods were slightly below human performance, with some room for improvement. We finally observed that there are systematic differences in the topic-level scores derived from the two task formulations.

Acknowledgements

This work was supported in part by the Australian Research Council, and for author JHL, also partly funded by grant ES/J022969/1 from the Economic and Social Research Council of the UK. The authors acknowledge the generosity of Nikos Aletras and Mark Stevenson in providing their code for OC-Auto-DS, and Jordan Boyd-Graber in providing the data used in Chang et al. (2009).

References

N. Aletras and M. Stevenson. 2013a. Evaluating topic coherence using distributional semantics. In *Proceedings of the Tenth International Workshop on Computational Semantics (IWCS-10)*, pages 13–22, Potsdam, Germany.

N. Aletras and M. Stevenson. 2013b. Representing topics using images. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 158–167, Atlanta, USA.

D. Blei and J. Lafferty. 2005. Correlated topic models. In *Advances in Neural Information Processing Systems 17 (NIPS-05)*, pages 147–154, Vancouver, Canada.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

L. Bolelli, Ş. Ertekin, and C.L. Giles. 2009. Topic and trend detection in text collections using Latent Dirichlet Allocation. In *Proceedings of ECIR 2009*, pages 776–780, Toulouse, France.

G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, pages 31–40, Potsdam, Germany.

J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 21 (NIPS-09)*, pages 288–296, Vancouver, Canada.

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies 2009 (NAACL HLT 2009)*, pages 362–370.

D. Hall, D. Jurafsky, and C.D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 363–371, Honolulu, USA.

M. Hall, P. Clough, and M. Stevenson. 2012. Evaluating the use of clustering for automatically organising digital library collections. In *Proceedings of the Second International Conference on Theory and Practice of Digital Libraries*, pages 323–334, Paphos, Cyprus.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of 22nd International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 50–57, Berkeley, USA.

T. Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, USA.

J.H. Lau, D. Newman, S. Karimi, and T. Baldwin. 2010. Best topic word selection for topic labelling. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Posters Volume*, pages 605–613, Beijing, China.

J.H. Lau, K. Grieser, D. Newman, and T. Baldwin. 2011. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 1536–1545, Portland, USA.

J.H. Lau, N. Collier, and T. Baldwin. 2012a. Online trend analysis with topic models: #twitter trends detection topic model online. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1519–1534, Mumbai, India.

J.H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. 2012b. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the EACL (EACL 2012)*, pages 591–601, Avignon, France.

J.H. Lau, T. Baldwin, and D. Newman. 2013. On collocations and topic models. *ACM Transactions on Speech and Language Processing*, 10(3):10:1–10:14.

A McCallum, G.S. Mann, and D Mimno. 2006. Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries 2006 (JCDL'06)*, pages 65–74, Chapel Hill, USA.

D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 262–272, Edinburgh, UK.

G. Minnen, J. Carroll, and D. Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

C. Musat, J. Velcin, S. Trausan-Matu, and M.A. Rizoiu. 2011. Improving topic evaluation using conceptual knowledge. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-2011)*, pages 1866–1871, Barcelona, Spain.

D. Newman, J.H. Lau, K. Grieser, and T. Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 100–108, Los Angeles, USA.

M. Paul and R. Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the 24th Annual Conference on Artificial Intelligence (AAAI-10)*, Atlanta, USA.

J. Reisinger, A. Waters, B. Silverthorn, and R.J. Mooney. 2010. Spherical topic models. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 903–910, Haifa, Israel.

X. Wang, A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 697–702, Omaha, USA.

B. Zhao and E.P. Xing. 2007. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems (NIPS 2007)*, pages 1689–1696, Vancouver, Canada.