Automatic Evaluation of Topic Coherence

David Newman,[♠♣] **Jey Han Lau,**[♡] **Karl Grieser**[♢], and **Timothy Baldwin,**[♠♡]
♠ NICTA Victoria Research Laboratory, Australia
♣ Dept of Computer Science, University of California, Irvine
♡ Dept of Computer Science and Software Engineering, University of Melbourne, Australia
♢ Dept of Information Systems, University of Melbourne, Australia
newman@uci.edu, depthchargex@gmail.com,
kgrieser@csse.unimelb.edu.au, tb@ldwin.net

Abstract

This paper introduces the novel task of topic coherence evaluation, whereby a set of words, as generated by a topic model, is rated for coherence or interpretability. We apply a range of topic scoring models to the evaluation task, drawing on WordNet, Wikipedia and the Google search engine, and existing research on lexical similarity/relatedness. In comparison with human scores for a set of learned topics over two distinct datasets, we show a simple co-occurrence measure based on pointwise mutual information over Wikipedia data is able to achieve results for the task at or nearing the level of inter-annotator correlation, and that other Wikipedia-based lexical relatedness methods also achieve strong results. Google produces strong, if less consistent, results, while our results over WordNet are patchy at best.

1 Introduction

There has traditionally been strong interest within computational linguistics in techniques for learning sets of words (aka topics) which capture the latent semantics of a document or document collection, in the form of methods such as latent semantic analysis (Deerwester et al., 1990), probabilistic latent semantic analysis (Hofmann, 2001), random projection (Widdows and Ferraro, 2008), and more recently, latent Dirichlet allocation (Blei et al., 2003; Griffiths and Steyvers, 2004). Such methods have been successfully applied to a myriad of tasks including word sense discrimination (Brody and Lapata, 2009), document summarisation (Haghighi and Vanderwende, 2009), areal linguistic analysis (Daume III, 2009) and text segmentation (Sun et al., 2008). In each case, *extrinsic* evaluation has been used to demonstrate the effectiveness of the learned topics in the application domain, but standardly, no attempt has been made to perform *intrinsic* evaluation of the topics themselves, either qualitatively or quantitatively. In machine learning, on the other hand, researchers have modified and extended topic models in a variety of ways, and evaluated *intrinsically* in terms of model perplexity (Wallach et al., 2009), but there has been less effort on *qualitative* understanding of the semantic nature of the learned topics.

This research seeks to fill the gap between topic evaluation in computational linguistics and machine learning, in developing techniques to perform intrinsic qualitative evaluation of learned topics. That is, we develop methods for evaluating the quality of a given topic, in terms of its coherence to a human. After learning topics from a collection of news articles and a collection of books, we ask humans to decide whether individual learned topics are coherent, in terms of their interpretability and association with a single over-arching semantic concept. We then propose models to predict topic coherence, based on resources such as WordNet, Wikipedia and the Google search engine, and methods ranging from ontological similarity to link overlap and term co-occurrence. Over topics learned from two distinct datasets, we demonstrate that there is remarkable inter-annotator agreement on what is a coherent topic, and additionally that our methods based on Wikipedia are able to achieve nearly perfect agreement with humans over the evaluation of topic coherence.

This research forms part of a larger research agenda on the utility of topic modelling in gisting and visualising document collections, and ultimately enhancing search/discovery interfaces over

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 100–108,
Los Angeles, California, June 2010.

document collections (Newman et al., to appeara). Evaluating topic coherence is a component of the larger question of what are good topics, what characteristics of a document collection make it more amenable to topic modelling, and how can the potential of topic modelling be harnessed for human consumption (Newman et al., to appearb).

2 Related Work

Most earlier work on intrinsically evaluating learned topics has been on the basis of perplexity results, where a model is learned on a collection of training documents, then the log probability of the unseen test documents is computed using that learned model. Usually perplexity is reported, which is the inverse of the geometric mean per-word likelihood. Perplexity is useful for model selection and adjusting parameters (e.g. number of topics T), and is the standard way of demonstrating the advantage of one model over another. Wallach et al. (2009) presented efficient and unbiased methods for computing perplexity and evaluating almost any type of topic model.

While statistical evaluation of topic models is reasonably well understood, there has been much less work on evaluating the intrinsic *semantic* quality of topics learned by topic models, which could have a far greater impact on the overall value of topic modeling for end-user applications. Some researchers have started to address this problem, including Mei et al. (2007) who presented approaches for automatic labeling of topics (which is core to the question of coherence and semantic interpretability), and Griffiths and Steyvers (2006) who applied topic models to word sense discrimination tasks. Misra et al. (2008) used topic modelling to identify semantically incoherent documents within a document collection (vs. coherent *topics*, as targeted in this research). Chang et al. (2009) presented the first human-evaluation of topic models by creating a task where humans were asked to identify which word in a list of five topic words had been randomly switched with a word from another topic. This work showed some possibly counter-intuitive results, where in some cases humans preferred models with higher perplexity. This type of result shows the need for further exploring measures other than perplexity for evaluating topic models. In earlier work, we carried out preliminary experimentation using pointwise mutual information and Google results to evaluate topic coherence over the same set of topics as used in this research (Newman et al., 2009).

Part of this research takes inspiration from the work on automatic evaluation in machine translation (Papineni et al., 2002) and automatic summarisation (Lin, 2004). Here, the development of automated methods with high correlation with human subjects has opened the door to large-scale automated evaluation of system outputs, revolutionising the respective fields. While our aspirations are more modest, the basic aim is the same: to develop a fully-automated method for evaluating a well-grounded task, which achieves near-human correlation.

3 Topic Modelling

In order to evaluate topic modelling, we require a topic model and set of topics for a given document collection. While the evaluation methodology we describe generalises to any method which generates sets of words, all of our experiments are based on *Latent Dirichlet Allocation* (LDA, aka *Discrete Principal Component Analysis*), on the grounds that it is a state-of-the-art method for generating topics.

LDA is a Bayesian graphical model for text document collections represented by bags-of-words (see Blei et al. (2003), Griffiths and Steyvers (2004), Buntine and Jakulin (2004)). In a topic model, each document in the collection of D documents is modelled as a multinomial distribution over T topics, where each topic is a multinomial distribution over W words. Typically, only a small number of words are important (have high likelihood) in each topic, and only a small number of topics are present in each document.

The collapsed Gibbs sampled topic model simultaneously learns the topics and the mixture of topics in documents by iteratively sampling the topic assignment z to every word in every document, using the Gibbs sampling update:

$$p(z_{id} = t | x_{id} = w, \mathbf{z}^{\neg id}) \propto \frac{N_{wt}^{\neg id} + \beta}{\sum_w N_{wt}^{\neg id} + W\beta} \frac{N_{td}^{\neg id} + \alpha}{\sum_t N_{td}^{\neg id} + T\alpha}$$

where $z_{id} = t$ is the assignment of the i^{th} word in document d to topic t, $x_{id} = w$ indicates that the current observed word is w, and $\mathbf{z}^{\neg id}$ is the vector of all topic assignments not including the current word. N_{wt} represents integer count arrays (with the subscripts denoting what is counted), and α and β are Dirichlet priors.

The maximum a posterior (MAP) estimates of the topics $p(w|t), t = 1 \ldots T$ are given by:

$$p(w|t) = \frac{N_{wt} + \beta}{\sum_w N_{wt} + W\beta}$$

We will follow the convention of representing a topic via its top-n words, ordered by $p(w|t)$. Here, we use the top-ten words, as they usually provide sufficient detail to convey the subject of a topic, and distinguish one topic from another. For the remainder of this paper, we will refer to individual topics by its list of top-ten words, denoted by $\mathbf{w} = (w_1, \ldots, w_{10})$.

4 Topic Evaluation Methods

We experiment with scoring methods based on WordNet (Section 4.1), Wikipedia (Section 4.2) and the Google search engine (Section 4.3). In the case of Google, we query for the entire topic, but with WordNet and Wikipedia, this takes the form of scoring each word-pair in a given topic $\mathbf{w}$ based on the component words $(w_1, \ldots, w_{10})$. Given some (symmetric) word-similarity measure $D(w_i, w_j)$, two straightforward ways of producing a combined score from the 45 (i.e. $\binom{10}{2}$) word-pair scores are: (1) the arithmetic mean, and (2) the median, as follows:

$$\text{Mean-D-Score}(\mathbf{w}) = \text{mean}\{\text{D}(w_i, w_j), ij \in 1 \ldots 10, i < j\}$$

$$\text{Median-D-Score}(\mathbf{w}) = \text{median}\{\text{D}(w_i, w_j), ij \in 1 \ldots 10, i < j\}$$

Intuitively, the median seems the more natural representation, as it is less affected by outlier scores, but we experiment with both, and fall back to empirical verification of which is the better combination method.

4.1 WordNet similarity

WordNet (Fellbaum, 1998) is a lexical ontology that represents word sense via "synsets", which are structured in a hypernym/hyponym hierarchy (nouns) or hypernym/troponym hierarchy (verbs). WordNet additionally links both synsets and words via lexical relations including antonymy, morphological derivation and holonymy/meronym.

In parallel with the development of WordNet, a number of computational methods for calculating the semantic relatedness/similarity between synset pairs (i.e. sense-specified word pairs) have been developed, as we outline below. These methods apply to *synset* rather than word pairs, so to generate a single score for a given word pair, we look up each word in WordNet and exhaustively generate scores for each sense pairing defined by them, and calculate their arithmetic mean.[1]

The majority of the methods (all methods other than HSO, VECTOR and LESK) are restricted to operating strictly over hierarchical links within a single hierarchy. As the verb and noun hierarchies are not connected (other than via derivational links), this means that it is generally not possible to calculate the similarity between noun and verb senses, for example. In such cases, we simply drop the synset pairing in question from our calculation of the mean.

The least common subsumer (LCS) is a common feature to a number of the measures, and is defined as the deepest node in the hierarchy that subsumes both of the synsets under question.

For all our experiments over WordNet, we use the `WordNet::Similarity` package.

Path distance (PATH)

The simplest of the WordNet-based measures is to count the number of nodes visited while going from one word to another via the hypernym hierarchy. The path distance between two nodes is defined as the number of nodes that lie on the shortest path between two words in the hierarchy. This

[1] We also experimented with the median, and trialled filtering the set of senses in a variety of ways, e.g. using only the first sense (the sense with the highest prior) for a given word, or using only the word senses associated with the POS with the highest prior. In all cases, the overall trend was for the correlation with the human scores to drop relative to the mean, so we only present the numbers for the mean in this paper.

count of nodes includes the beginning and ending word nodes.

Leacock-Chodorow (LCH)

The measure of semantic similarity devised by Leacock et al. (1998) finds the shortest path between two WordNet synsets ($sp(c_1, c_2)$) using hypernym and synonym relationships. This path length is then scaled by the maximum depth of WordNet (D), and the log likelihood taken:

$$sim_{lch}(c_1, c_2) = -\log \frac{sp(c_1, c_2)}{2 \cdot D}$$

Wu-Palmer (WUP)

Wu and Palmer (1994) proposed to scale the depth of the two synset nodes ($depth_{c_1}$ and $depth_{c_2}$) by the depth of their LCS ($depth(lcs_{c_1,c_2})$):

$$sim_{wup}(c_1, c_2) = \frac{2 \cdot depth(lcs_{c_1,c_2})}{depth_{c_1} + depth_{c_2} + 2 \cdot depth(lcs_{c_1,c_2})}$$

The scaling means that specific terms (deeper in the hierarchy) that are close together are more semantically similar than more general terms, which have a short path distance between them. Only hypernym relationships are used in this measure, as the LCS is defined by the common member in the concepts' hypernym path.

Hirst-St Onge (HSO)

Hirst and St-Onge (1998) define a measure of semantic similarity based on length and tortuosity of the path between nodes. Hirst and St-Onge attribute directions (up, down and horizontal) to the larger set of WordNet relationships, and identify the path from one word to another utilising all of these relationships. The relatedness score is then computed by the weighted sum of the path length between the two words ($len(c_1, c_2)$) and the number of turns the path makes ($turns(c_1, c_2)$) to take this route:

$$rel_{hso}(c_1, c_2) = C - len(c_1, c_2) - k \times turns(c_1, c_2)$$

where C and k are constants. Additionally, a set of restrictions is placed on the path so that it may not be more than a certain length, may not contain more than a set number of turns, and may only take turns in certain directions.

Resnik Information Content (RES)

Resnik (1995) presents a method for weighting edges in WordNet (avoiding the assumption that all edges between nodes have equal importance), by weighting edges between nodes by their frequency of use in textual corpora.

Resnik found that the most effective measure of comparison using this methodology was to measure the Information Content ($IC(c) = -\log p(c)$) of the subsumer with the greatest Information Content from the set of all concepts that subsumed the two initial concepts ($S(c_1, c_2)$) being compared:

$$sim_{res}(c_1, c_2) = \max_{c \in S(c_1,c_2)} [-\log p(c)]$$

Lin (LIN)

Lin (1998) expanded on the Information Theoretic approach presented by Resnik by scaling the Information Content of each node by the information content of their LCS:

$$sim_{lin}(c_1, c_2) = \frac{2 \times \log p(lcs_{c_1,c_2})}{\log p(c_1) + \log p(c_2)}$$

This measure contrasts the joint content of the two concepts with the difference between them.

Jiang-Conrath (JCN)

Jiang and Conrath (1997) define a measure that utilises the components of the information content of the LCS in a different manner:

$$sim_{jcn}(c_1, c_2) = \frac{1}{IC(a) + IC(b) - 2 \times IC(lcs_{a,b})}$$

Instead of defining commonality and difference as with Lin's measure, the key determinant is the specificity of the two nodes compared with their LCS.

Lesk (LESK)

Lesk (1986) proposed a significantly different approach to lexical similarity to that proposed in the methods presented above, using the lexical overlap in dictionary definitions (or *glosses*) to disambiguate word sense. The sense definitions that contain the most words in common indicate the most likely sense of the word given its co-occurrence with similar word senses. Banerjee and Pedersen (2002)

adapted this method to utilise WordNet sense glosses rather than dictionary definitions, and expand the dictionary definitions via ontological links, and it is this method we experiment with in this paper.

Vector (VECTOR)

Schütze (1998) uses the words surrounding a term in a piece of text to form a context vector that describes the context in which the word sense appears. For a set of words associated with a target sense, a context vector is computed as the centroid vector of these words. The centroid context vectors each represent a word sense. To compare word senses, the cosine similarity of the context vectors is used.

4.2 Wikipedia

In the last few years, there has been a surge of interest in using Wikipedia to calculate semantic similarity, using the Wikipedia article content, in-article links and document categories (Strübe and Ponzetto, 2006; Gabrilovich and Markovitch, 2007; Milne and Witten, 2008). We present a selection of such methods below. There are a number of Wikipedia-based scoring methods which we do not present results for here (notably Strübe and Ponzetto (2006) and Gabrilovich and Markovitch (2007)), due to their computational complexity and uncertainty about the full implementation details of the methods.

As with WordNet, a given word will often have multiple entries in Wikipedia, grouped in a disambiguation page. For MIW, RACO and DOCSIM, we apply the same strategy as we did with WordNet, in exhaustively calculating the pairwise scores between the sets of documents associated with each term, and averaging across them.

Milne-Witten (MIW)

Milne and Witten (2008) adapted the Resnik (1995) methodology to utilise the count of links pointing to an article. As Wikipedia is self-referential (articles link to related articles), no external data is needed to find the "referred-to-edness" of a concept. Milne and Witten use an adapted Information Content measure that weights the number of links from one article to another ($c_1 \rightarrow c_2$) by the total number of links to the second article:

$$w(c_1 \rightarrow c_2) = |c_1 \rightarrow c_2| \times \log \sum_{x \in W} \frac{|W|}{|c_1, x)|}$$

where x is an article in W, Wikipedia. This measure provides the similarity of one article to another, however this is asymmetrical. The above metric is used to find the weights of all outlinks from the two articles being compared:

$$\vec{c_1} = (w(c_1 \rightarrow l_1), w(c_1 \rightarrow l_2), \cdots, w(c_1 \rightarrow l_n))$$
$$\vec{c_2} = (w(c_2 \rightarrow l_1), w(c_2 \rightarrow l_2), \cdots, w(c_2 \rightarrow l_n))$$

for the set of links l that is the union of the sets of outlinks from both articles. The overall similarity of the two articles is then calculated by taking the cosine similarity of the two vectors.

Related Article Concept Overlap (RACO)

We also determine the category overlap of two articles by examining the outlinks of both articles, in the form of the Related Article Concept Overlap (RACO) measure. The concept overlap of the sets of respective outlinks is given by the union of the two sets of categories from the outlinks from each article:

$$overlap(c_1, c_1) = \left| \left(\bigcup_{l \in ol(c_1)} cat(l) \right) \bigcap \left(\bigcup_{l \in ol(c_2)} cat(l) \right) \right|$$

where $ol(c_1)$ is the set of outlinks from article c_1, and $cat(l)$ is the set of categories of which the article at outlink l is a member. To account for article size (and differing number of outlinks), the Jaccard coefficient is used:

$$rel_{raco}(c_1, c_2) = \frac{\left| \left(\bigcup_{l \in ol(c_1)} cat(l) \right) \bigcap \left(\bigcup_{l \in ol(c_2)} cat(l) \right) \right|}{\left| \bigcup_{l \in ol(c_1)} cat(l) \right| + \left| \bigcup_{l \in ol(c_2)} cat(l) \right|}$$

Document Similarity (DOCSIM)

In addition to these two measures of semantic relatedness, we experiment with simple cosine similarity of the text of Wikipedia articles as a measure of semantic relatedness.

Term Co-occurrence (PMI)

Another variant is to treat Wikipedia as a single meta-document and score word pairs using term co-occurrence. Here, we calculate the pointwise mutual information (PMI) of each word pair, estimated

Selected high-scoring topics (unanimous score=3):
[NEWS] *space earth moon science scientist light nasa mission planet mars ...*
[NEWS] *health disease aids virus vaccine infection hiv cases infected asthma ...*
[BOOKS] *steam engine valve cylinder pressure piston boiler air pump pipe ...*
[BOOKS] *furniture chair table cabinet wood leg mahogany piece oak louis ...*

Selected low-scoring topics (unanimous score=1):
[NEWS] *king bond berry bill ray rate james treas byrd key ...*
[NEWS] *dog moment hand face love self eye turn young character ...*
[BOOKS] *soon short longer carried rest turned raised filled turn allowed ...*
[BOOKS] *act sense adv person ppr plant sax genus applied dis ...*

Table 1: A selection of high-scoring and low-scoring topics

from the entire corpus of over two million English Wikipedia articles (∼1 billion words). PMI has been studied variously in the context of collocation extraction (Pecina, 2008), and is one measure of the statistical independence of observing two words in close proximity. Using a sliding window of 10-words to identify co-occurrence, we computed the PMI of all a given word pair (w_i, w_j) as, following Newman et al. (2009):

$$\text{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

4.3 Search engine-based similarity

Finally, we present two search engine-based scoring methods, based on Newman et al. (2009). In this case the external data source is the entire World Wide Web, via the Google search engine. Unlike the methods presented above, here we query for the topic in its entirety,[2] meaning that we return a topic-level score rather than scores for individual word or word sense pairs. In each case, we mark each search term with the advanced search option + to search for the terms exactly as is and prevent Google from using synonyms or lexical variants of the term. An example query is: *+space +earth +moon +science +scientist +light +nasa +mission +planet +mars*.

Google title matches (TITLES)

Firstly, we score topics by the relative occurrence of their component words in the titles of documents returned by Google:

$$\text{Google-titles-match}(\mathbf{w}) = \mathbf{1}\left[w_i = v_j\right]$$

where $i = 1, \ldots, 10$ and $j = 1, \ldots, |V|$, v_j are all the unique terms mentioned in the titles from the top-100 search results, and **1** is the indicator function to count matches. For example, in the top-100 results for our query above, there are 194 matches with the ten topic words, so Google-titles-match($\mathbf{w}$) = 194.

Google log hits matches (LOGHITS)

Second, we issue queries as above, but return the log number of hits for our query:

$$\text{Google-log-hits}(\mathbf{w}) = \log_{10}(\text{\# results from search for } \mathbf{w})$$

where $\mathbf{w}$ is the search string $+w_1$ $+w_2$ $+w_3$ ... $+w_{10}$. For example, our query above returns 171,000 results, so Google-log-hits($\mathbf{w}$) = 5.2. and the URL titles from the top-100 results include a total of 194 matches with the ten topic words, so for this topic Google-titles-match($\mathbf{w}$)=194.

5 Experimental Setup

We learned topics for two document collections: a collection of news articles, and a collection of books. These collections were chosen to produce sets of topics that have more variable quality than one typically observes when topic modeling highly uniform content. The collection of $D = 55,000$ news articles was selected from English Gigaword, and the collection of $D = 12,000$ books was downloaded from the Internet Archive. We refer to these collections as NEWS and BOOKS, respectively.

Standard procedures were used to tokenize each collection and create the bags-of-words. We learned

[2] All queries were run on 15/09/2009.

Resource	Method	Median	Mean
WordNet	HSO	−0.29	0.34
	JCN	0.08	0.22
	LCH	−0.18	−0.07
	LESK	0.38	0.37
	LIN	0.18	0.25
	PATH	0.19	0.11
	RES	−0.10	0.13
	VECTOR	0.07	0.20
	WUP	0.03	0.10
Wikipedia	RACO	0.61	0.63
	MIW	0.69	0.60
	DOCSIM	0.45	0.50
	PMI	0.78	0.77
Google	TITLES	**0.80**	
	LOGHITS	0.46	
Gold-standard	IAA	0.79	0.73

Table 2: Spearman rank correlation ρ values for the different scoring methods over the NEWS dataset (best-performing method for each resource underlined; best-performing method overall in **boldface**)

Resource	Method	Median	Mean
WordNet	HSO	0.15	0.59
	JCN	−0.20	0.19
	LCH	−0.31	−0.15
	LESK	0.53	0.53
	LIN	0.09	0.28
	PATH	0.29	0.12
	RES	0.57	0.66
	VECTOR	−0.08	0.27
	WUP	0.41	0.26
Wikipedia	RACO	0.62	0.69
	MIW	0.68	0.70
	DOCSIM	0.59	0.60
	PMI	**0.74**	**0.77**
Google	TITLES	0.51	
	LOGHITS	−0.19	
Gold-standard	IAA	0.82	0.78

Table 3: Spearman rank correlation ρ values for the different scoring methods over the BOOKS dataset (best-performing method for each resource underlined; best-performing method overall in **boldface**)

topic models of NEWS and BOOKS using $T = 200$ and $T = 400$ topics respectively. We randomly selected a total of 237 topics from the two collections for user scoring. We asked $N = 9$ users to score each of the 237 topics on a 3-point scale where 3="useful" (coherent) and 1="useless" (less coherent).

We provided annotators with a rubric and guidelines on how to judge whether a topic was useful or useless. In addition to showing several examples of useful and useless topics, we instructed users to decide whether the topic was to some extent coherent, meaningful, interpretable, subject-heading-like, and something-you-could-easily-label. For our purposes, the usefulness of a topic can be thought of as whether one could imagine using the topic in a search interface to retrieve documents about a particular subject. One indicator of usefulness is the ease by which one could think of a short label to describe a topic.

Table 1 shows a selection of high- and low-scoring topics, as scored by the $N = 9$ users. The first topic illustrates the notion of labelling coherence, as *space exploration*, e.g., would be an obvious label for the topic. The low-scoring topics display little coherence, and one would not expect them to be useful as categories or facets in a search interface. Note that the useless topics from both collections are not chance artifacts produced by the models, but are in fact stable and robust statistical features in the data sets.

6 Results

The results for the different topic scoring methods over the NEWS and BOOKS collections are presented in Tables 2 and 3, respectively. In each table, we separate out the scoring methods into those based on WordNet (from Section 4.1), those based on Wikipedia (from Section 4.2), and those based on Google (from Section 4.3).

As stated in Section 4, we experiment with two methods for combining the word-pair scores (for all methods other than the two Google methods, which operate natively over a word set), namely the arithmetic mean and median. We present the numbers for these two methods in each table. In each case, we evaluate via Spearman rank correlation, reversing the sign of the calculated ρ value for PATH (as it is the only instance of a distance metric, where the gold-standard is made up of similarity values).

We include the inter-annotator agreement (IAA) in the final row of each table, which we consider

to be the upper bound for the task. This is calculated as the average Spearman rank correlation between each annotator and the mean/median of the remaining annotators for that topic. Encouragingly, there is relatively little difference in the IAA between the two datasets; the median-based calculation produces slightly higher ρ values and is empirically the method of choice.[3]

Of all the topic scoring methods tested, PMI (term co-occurrence via simple pointwise mutual information) is the most consistent performer, achieving the best or near-best results over both datasets, and approaching or surpassing the inter-annotator agreement. This indicates both that the task of topic evaluation as defined in this paper is computationally tractable, and that word-pair based co-occurrence is highly successful at modelling topic coherence.

Comparing the different resources, Wikipedia is far and away the most consistent performing, with PMI producing the best results, followed by MIW and RACO, and finally DOCSIM. There is relatively little difference in results between NEWS and BOOKS for the Wikipedia methods. Google achieves the best results over NEWS, for TITLES (actually slightly above the IAA), but the results fall away sharply over BOOKS. The reason for this can be seen in the sample topics in Table 1: the topics for BOOKS tend to be more varied in word class than for NEWS, and contain less proper names; also, the genre of BOOKS is less well represented on the web. We hypothesise that Wikipedia's encyclopedic nature means that it has good coverage over both domains, and thus more robust.

Turning to WordNet, the overall results are markedly better over BOOKS, again largely because of the relative sparsity of proper names in the resource. The results for individual methods are somewhat surprising. Whereas JCN and LCH have been shown to be two of the best-performing methods over lexical similarity tasks (Budanitsky and Hirst, 2005; Agirre et al., 2009), they perform abysmally at the topic scoring task. Indeed, the spread of results across the WordNet similarity methods (notably HSO, JCN, LCH, LIN, RES and WUP) is much greater than we had expected. The single most consistent method is LESK, which is based on lexical overlap in definition sentences and makes relatively modest use of the WordNet hierarchy. Supplementary evaluation where we filtered out all proper nouns from the topics (based on simple POS priors for each word learned from an automatically-tagged version of the British National Corpus) led to a slight increase in results for the WordNet methods; the full results are omitted for reasons of space. In future work, we intend to carry out error analysis to determine why some of the methods performed so badly, or inconsistently across the two datasets.

There is no clear answer to the question of whether the mean or median is the best method for combining the pair-wise scores.

[3]Note that the choice of mean or median for IAA is independent of that for the scoring methods, as they are combining different things: annotator scores in the one hand, and word/concept pair scores on the other.

7 Conclusions

We have proposed the novel task of topic coherence evaluation as a form of intrinsic topic evaluation with relevance in document search/discovery and visualisation applications. We constructed a gold-standard dataset of topic coherence scores over the output of a topic model for two distinct datasets, and evaluated a wide range of topic scoring methods over this dataset, drawing on WordNet, Wikipedia and the Google search engine. The single best-performing method was term co-occurrence within Wikipedia based on pointwise mutual information, which achieve results very close to the inter-annotator agreement for the task. Google was also found to perform well over one of the two datasets, while the results for the WordNet-based methods were overall surprisingly low.

Acknowledgements

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence programme. DN has also been supported by a grant from the Institute of Museum and Library Services, and a Google Research Award.

References

E Agirre, E Alfonseca, K Hall, J Kravalova, M Paşca, and A Soroa. 2009. A study on similarity and re-

latedness using distributional and WordNet-based approaches. In *Proc. of HLT: NAACL 2009*, pages 19–27, Boulder, Colorado.

S Banerjee and T Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. *Proc. of CICLing'02*, pages 136–145.

DM Blei, AY Ng, and MI Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

S Brody and M Lapata. 2009. Bayesian word sense induction. In *Proc. of EACL 2009*, pages 103–111, Athens, Greece.

A Budanitsky and G Hirst. 2005. Evaluating WordNet-based Measures of Lexical Sematic Relatedness. *Computational Linguistics*, 32(1):13–47.

WL Buntine and A Jakulin. 2004. Applying discrete PCA in data analysis. In *Proc. of UAI 2004*, pages 59–66.

J Chang, J Boyd-Graber, S Gerris, C Wang, and D Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proc. of NIPS 2009*.

H Daume III. 2009. Non-parametric bayesian areal linguistics. In *Proc. of HLT: NAACL 2009*, pages 593–601, Boulder, USA.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6).

C Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

E Gabrilovich and S Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proc. of IJCAI'07*, pages 1606–1611, Hyderabad, India.

T Griffiths and M Steyvers. 2004. Finding scientific topics. In *Proc. of the National Academy of Sciences*, volume 101, pages 5228–5235.

T Griffiths and M Steyvers. 2006. Probabilistic topic models. In *Latent Semantic Analysis: A Road to Meaning*.

A Haghighi and L Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proc. of HLT: NAACL 2009*, pages 362–370, Boulder, USA.

G Hirst and D St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropism. In Fellbaum (Fellbaum, 1998), pages 305–332.

T Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

JJ Jiang and DW Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of COLING'97*, pages 19–33, Taipei, Taiwan.

C Leacock, G A Miller, and M Chodorow. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–65.

M Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proc. of SIGDOC'86*, pages 24–26, Toronto, Canada.

D Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING/ACL'98*, pages 768–774, Montreal, Canada.

C-Y Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proc. of the ACL 2004 Workshop on Text Summarization Branches Out (WAS 2004)*, pages 74–81, Barcelona, Spain.

Q Mei, X Shen, and CX Zhai. 2007. Automatic labeling of multinomial topic models. In *Proc. of KDD 2007*, pages 490–499.

D Milne and IH Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proc. of AAAI Workshop on Wikipedia and Artificial Intelligence*, pages 25–30, Chicago, USA.

H Misra, O Cappe, and F Yvon. 2008. Using LDA to detect semantically incoherent documents. In *Proc. of CoNLL 2008*, pages 41–48, Manchester, England.

D Newman, S Karimi, and L Cavedon. 2009. External evaluation of topic models. In *Proc. of ADCS 2009*, pages 11–18, Sydney, Australia.

D Newman, T Baldwin, L Cavedon, S Karimi, D Martinez, and J Zobel. to appeara. Visualizing document collections and search results using topic mapping. *Journal of Web Semantics*.

D Newman, Y Noh, E Talley, S Karimi, and T Baldwin. to appearb. Evaluating topic models for digital libraries. In *Proc. of JCDL/ICADL 2010*, Gold Coast, Australia.

K Papineni, S Roukos, T Ward, and W-J Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL 2002*, pages 311–318, Philadelphia, USA.

P Pecina. 2008. *Lexical Association Measures: Collocation Extraction*. Ph.D. thesis, Charles University.

P Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of IJCAI'95*, pages 448–453, Montreal, Canada.

H Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

M Strübe and SP Ponzetto. 2006. WikiRelate! computing semantic relateness using Wikipedia. In *Proc. of AAAI'06*, pages 1419–1424, Boston, USA.

Q Sun, R Li, D Luo, and X Wu. 2008. Text segmentation with LDA-based Fisher kernel. In *Proc. of ACL-08: HLT*, pages 269–272.

HM Wallach, I Murray, R Salakhutdinov, and DM Mimno. 2009. Evaluation methods for topic models. In *Proc. of ICML 2009*, page 139.

D Widdows and K Ferraro. 2008. Semantic Vectors: A scalable open source package and online technology management application. In *Proc. of LREC 2008*, Marrakech, Morocco.

Z Wu and M Palmer. 1994. Verb selection and lexical selection. In *Proc. of ACL'94*, pages 133–138, Las Cruces, USA.