Best Topic Word Selection for Topic Labelling

Jey Han Lau,[♠♡] **David Newman,**[♠♢] **Sarvnaz Karimi**[♠] and **Timothy Baldwin**[♠♡]

[♠] NICTA Victoria Research Laboratory, Australia

[♡] Dept of Computer Science and Software Engineering, University of Melbourne, Australia

[♢] Dept of Computer Science, University of California, Irvine, USA

jhlau@csse.unimelb.edu.au, newman@uci.edu, skarimi@unimelb.edu.au, tb@ldwin.net

Abstract

This paper presents the novel task of best topic word selection, that is the selection of the topic word that is the best label for a given topic, as a means of enhancing the interpretation and visualisation of topic models. We propose a number of features intended to capture the best topic word, and show that, in combination as inputs to a reranking model, we are able to consistently achieve results above the baseline of simply selecting the highest-ranked topic word. This is the case both when training in-domain over other labelled topics for that topic model, and cross-domain, using only labellings from independent topic models learned over document collections from different domains and genres.

1 Introduction

In the short time since its inception, topic modelling (Blei et al., 2003) has become a mainstream technique for tasks as diverse as multi-document summarisation (Haghighi and Vanderwende, 2009), word sense discrimination (Brody and Lapata, 2009), sentiment analysis (Titov and McDonald, 2008) and information retrieval (Wei and Croft, 2006). For many of these tasks, the multinomial topics learned by the topic model can be interpreted natively as probabilities, or mapped onto a pre-defined discrete class set. However, for tasks where the learned topics are provided to humans as a first-order output, e.g. for use in document collection analysis/navigation, it can be difficult for the end-user to interpret the rich statistical information encoded in the topics. This research is concerned with making topics more readily human interpretable, by selecting a single term with which to label the topic.

Although topics are formally a multinomial distribution over terms, with every term having finite probability in every topic, topics are usually displayed by printing the top-10 terms (i.e. the 10 most probable terms) in the topic. These top-10 terms typically account for about 30% of the topic mass for reasonable setting of number of topics, and usually provide sufficient information to determine the subject area and interpretation of a topic, and distinguish one topic from another.

Our research task can be illustrated via the top-10 terms in the following topic, learned from a book collection. Terms w_i are presented in descending order of $P(w_i|t_j)$ for the topic t_j:

> trout fish fly fishing water angler stream rod flies salmon

Clearly the topic relates to fishing, and indeed, the fourth term *fishing* is an excellent label for the topic. The task is thus termed *best word* or *most representative word* selection, as we are selecting the label from the closed set of the top-N topic words in that topic.

Naturally, not all topics are equally coherent, however, and the lower the topic coherence, the more difficult the label selection task becomes. For example:

> oct sept nov aug dec july sun lite adv globe

appears to conflate months with newspaper names, and no one of these topic words is able to capture the topic accurately. As such, our methodology presupposes an automatic means of rating topics for coherence. Fortunately, recent research by Newman et al. (2010) has shown that this is achievable at levels approaching human performance, meaning that this is not an unreasonable assumption.

Labelling topics has applications across a diverse range of tasks. Our original interest in the

Lau, Jey Han, David Newman, Sarvnaz Karimi and Timothy Baldwin (to appear) Best Topic
Word Selection for Topic Labelling, In Proceedings of the 23rd International Conference
on Computational Linguistics (COLING 2010), Beijing, China.

problem stems from work in document collection visualisation/navigation, and the realisation that presenting users with topics natively (e.g. as represented by the top-N terms) is ineffective, and would be significantly enhanced if we could automatically predict succinct labels for each topic. Another application area where labelling has been shown to enhance the utility of topic models is selectional preference learning via topic modelling (Ritter et al., to appear). Here, topic labelling via taxonomic classes (e.g. WordNet synsets) can lead to better topic generalisation, in addition to better human readability.

This paper is based around the assumption that an appropriate label for a topic can be found among the high-ranking (high probability) terms in that topic. We assess the suitability of each term by way of comparison with other high-ranking terms in that same topic, using simple pointwise mutual information and conditional probabilities. We first experiment with a simple ranking method based on the component scores, and then move on to using those scores, along with features from WordNet and from the original topic model, in a ranking support vector regression (SVR) framework. Our experiments demonstrate that we are able to perform the task significantly better than the baseline of selecting the topic word of highest marginal probability, including when training the ranking model on labelled topics from other document collections.

2 Related Work

Predictably, there has been significant work on interpreting topics in the context of topic modelling. Topic are conventionally interpreted via the top-N words in each topic (Blei et al., 2003; Griffiths and Steyvers, 2004), or alternatively by post-hoc manual labelling of each topic based on domain knowledge and subjective interpretation of each topic (Wang and McCallum, 2006; Mei et al., 2006).

Mei et al. (2007) proposed various approaches for automatically suggesting phrasal labels for topics, based on first extracting phrases from the document collection, and subsequently ranking the phrases based on KL divergence with a given topic.

Magatti et al. (2009) proposed a method for labelling topics induced by hierarchical topic modelling, based on ontological alignment with the Google Directory (gDir) hierarchy, and optionally expanding topics based on a thesaurus or WordNet. Preliminary experiments suggest the method has promise, but the method crucially relies on both a hierarchical topic model and a pre-existing ontology, so has limited applicability.

Over the general task of labelling a learned semantic class, Pantel and Ravichandran (2004) proposed the use of lexico-semantic patterns involving each member of that class to learn a (usually hypernym) label. The proposed method was shown to perform well over the semantically homogeneous, fine-grained clusters learned by CBC (Pantel and Lin, 2002), but for the coarse-grained, heterogeneous topics learned by topic modelling, it is questionable whether it would work as well.

The first works to report on human scoring of topics were Chang et al. (2009) and Newman et al. (2010). The first study used a novel but synthetic intruder detection task where humans evaluate both topics (that had an intruder word), and assignment of topics to documents (that had an intruder topic). The second study had humans directly score topics learned by a topic model. This latter work introduced the pointwise mutual information (PMI) score to model human scoring. Following this work, we use PMI as features in the ranking SVR model.

3 Methodology

Our task is to predict which words annotators tend to select as most representative or best words when presented with a list of ten words. Since annotators are not generally unanimous in their choice of best word, we formulate this as a ranking task, and treat the top-1, 2 and 3 system-ranked items as the best words, and compare that to the top-1, 2 and 3 words chosen most frequently by annotators. In this section, we describe the features that may be useful for this ranking task. We start with features motivated by word association.

An obvious idea is that the most representative word should be readily evoked by other words in the topic. For example, given a list of words $\langle$*space, earth, moon, nasa, mission*$\rangle$, which is a

Space Exploration topic, *space* could arguably be the most representative word. This is because it is natural to think about the word *space* after seeing the words *earth, moon* and *nasa* individually. A good candidate for best word could be the word that has high average conditional probability given each of the other words. To calculate conditional probability, we use word counts from the entire collection of English Wikipedia articles. Conditional probability is defined as:

$$P(w_i|w_j) = \frac{P(w_i, w_j)}{P(w_j)},$$

where $i \neq j$ and $P(w_i, w_j)$ is the probability of observing both w_i and w_j in the same sliding window, and $P(w_i)$ is the overall probability of word w_i in the corpus. In the above example, *evoked by* means that *space* would fill the slot of w_i. The average conditional probability for word w_i is given by:

$$\text{avg-CP1}(w_i) = \frac{1}{9}\sum_j P(w_i|w_j),$$

for $j = 1 \ldots 10, j \neq i$ (this range of indices applies to all following average quantities).

In other cases, we have the flip situation, where the most representative word may evoke (rather than be evoked by) other words in the list of ten words. Imagine a *NASCAR Racing* topic, which has a list of words $\langle$*race, car, nascar, driver, racing*$\rangle$. Given the word *nascar*, words from the list such as *race, car, racing* and *driver* might come to mind because *nascar* is heavily associated with these words. Therefore, a good candidate, w_i, might also correlate with high $P(w_j|w_i)$. As before, the average conditional probability (here denoted with CP2) for word w_i is given by:

$$\text{avg-CP2}(w_i) = \frac{1}{9}\sum_j P(w_j|w_i).$$

Another approach to measuring word association is by calculating pointwise mutual information (PMI) between word pairs. Unlike conditional probability, PMI is symmetric and thus the order of words in a pair does not matter. We calculate PMI using word counts from English Wikipedia as follows:

$$\text{PMI}(w_i, w_j) \quad = \quad log\frac{P(w_i, w_j)}{P(w_i)P(w_j)}.$$

The average PMI for word w_i is given by:

$$\text{avg-PMI}(w_i) = \frac{1}{9}\sum_j \text{PMI}(w_i, w_j).$$

The topic model produces an ordered list of words for each topic, and the ordering is given by the marginal probability of each word given that topic, $P(w_i|t_j)$. The ranking of words based on these probabilities indicates the importance of a word in a topic, and it is also a feature that we use for predicting the most representative word.

We also observe that sometimes the most representative words are generalized concepts of other words. As such, hypernym relations could be another feature that may be relevant to predicting the best word. To this end, we use WordNet to find hypernym relations between pairs of words in a topic and obtain a set of boolean-valued relationships for each topic word.

Our last feature is the distributional similarity scores of Pantel et al. (2009), as trained over Wikipedia.[1] This takes the form of representing the distributional similarity between each pairing of terms $\text{sim}(w_i|w_j)$; if w_i is not in the top-200 most similar terms for a given w_j, we assume it to have a similarity of 0.

While the above features can be used alone to get a ranking on the ten topic words, we can also use various combinations of features in a reranking model such as support vector regression (SVM^{rank}: Joachims (2006)). Applying the features described above — conditional probabilities, PMI, WordNet hypernym relations, the topic model word rank, and Pantel's distributional similarity score — as features for SVM^{rank}, a ranking of words is produced and candidates for the most representative word are selected by choosing the top-ranked words.

NEWS	stock market investor fund trading investment firm exchange ... police gun officer crime shooting death killed street victim ... food restaurant chef recipe cooking meat meal kitchen eat... patient doctor medical cancer hospital heart blood surgery ...
BOOKS	loom cloth thread warp weaving machine wool cotton yarn ... god worship religion sacred ancient image temple sun earth ... crop land wheat corn cattle acre grain farmer manure plough ... sentence verb noun adjective grammar speech pronoun ...

Figure 1: Selected topics from the two collections (each line is one topic, with fewer than ten topic words displayed because of limited space)

Features	Description
PMI	Pointwise mutual information
CP1	Conditional probability $P(w_i\|*)$
CP2	Conditional probability $P(*\|w_i)$
TM Rank	Original topic model word rank
Hypernym	WordNet hypernym relationships
PDS	Pantel distributional similarity score

Table 1: Description of feature sets

4 Datasets

We used two collections of text documents from different genres for our experiments. The first collection (NEWS) was created by selecting 55,000 news articles from the LDC Gigaword corpus. The second collection (BOOKS) was 12,000 English language books selected from the Internet Archive American Libraries collection. The NEWS and BOOKS collections provide a diverse range of content for topic modeling. In the first case – news articles from the past decade written by journalists — each article usually attempts to clearly and concisely convey information to the reader, and hence the learned topics tend to be fairly interpretable. For BOOKS (with publication dates spanning more than a century), the writing style often uses lengthy and descriptive prose, so one sees a different style to the learned topics.

The input to the topic model is a bag-of-words representation of the collection of text documents, where word counts are preserved, but word order is lost. After performing fairly standard tokenization and limited lemmatisation, and creating a vocabulary of terms that occurred at least ten times, each corpus was converted into its bag-of-words representation. We learned topic models for the two collections, choosing a setting of $T = 200$ topics for NEWS and $T = 400$ topics for BOOKS. After computing the PMI-score for each topic (according to Newman et al. (2010)), we selected 60 topics with high PMI-score, and 60 topics with low PMI-score, from both corpora, resulting in a total of 240 topics for human evaluation.

The 240 topics selected for human scoring were each evaluated by between 10 and 20 users. For the two topic models, we used the conventional approach of displaying each topic with its top-10 terms. In a typical survey, a user was asked to evaluate anywhere from 60 to 120 topics. The instructions asked the user to perform the following tasks, for each topic in the survey: (a) score the topic for "usefulness" or "coherence" on a scale of 1 to 3; and (b) select the single best word that exemplifies the topic (when score=3).

From both NEWS and BOOKS, the 40 topics with the highest average human scores had relatively complete data for the 'best word' selection task (i.e. every time a user gave a topics score=3, they also selected a 'best word'). The remainder of this paper is concerned with the 40 NEWS topics and 40 BOOKS topics where we had 'best word' data from the annotators. Sample topics from these two sets are given in Figure 1.

To measure presentational bias (i.e. the extent to which annotators tend to choose a word seen earlier rather than later, particularly when armed with the knowledge that words are presented in order of probability), we reissued a survey using the 40 NEWS topics to ten additional annotators, but this time the top-10 topic words were presented in random order. Again, these ten new annotators were asked to select the best word.

5 Experiments

We used average PMI and conditional probabilities, CP1 and CP2, to rank the ten words in each topic. Candidates for the best words were selected by choosing the top-1, 2 and 3 ranked words.

We used the following weighted scoring function for evaluation:

$$\text{Best-N score} = \frac{\sum_{i=1}^{N} n(w_{\text{rev}_i})}{\sum_{i=1}^{N} n(w_i)}$$

[1] Accessed from http://demo.patrickpantel.com/Content/LexSem/thesaurus.htm.

Features	Best-1	Best-2	Best-3
Baseline	0.35	0.50	0.59
PMI	0.25	0.38	0.49
CP1	0.30	0.42	0.51
CP2	0.15	0.27	0.45
Upper bound	0.48	—	—

Table 2: Best-1,2,3 scores for ranking with single feature sets (PMI and both conditional probabilities) for NEWS

Features	Best-1	Best-2	Best-3
Baseline	0.38	0.48	0.60
PMI	0.25	0.38	0.49
CP1	0.30	0.38	0.47
CP2	0.15	0.30	0.49
Upper bound	0.64	—	—

Table 3: Best-1,2,3 scores for ranking with single feature sets (PMI and both conditional probabilities) for BOOKS

Feature Set	Best-1	Best-2	Best-3
Baseline	0.35	0.50	0.59
All Features	0.43	0.56	0.62
−PMI	0.45 (+0.02)	0.52 (−0.04)	0.62 (±0.00)
−CP1	0.35 (−0.08)	0.49 (−0.07)	0.57 (−0.05)
−CP2	0.40 (−0.03)	0.50 (−0.06)	0.61 (−0.01)
−TM Rank	0.40 (−0.03)	0.52 (−0.04)	0.57 (−0.05)
−Hypernym	0.43 (±0.00)	0.57 (+0.01)	0.62 (±0.00)
−PDS	0.43 (±0.00)	0.53 (−0.03)	0.62 (±0.00)
Upper bound	0.48	—	—

Table 4: SVR-based best topic word results for NEWS for all six feature types, and feature ablation over each (numbers in brackets show the relative change over the full feature set)

Feature Set	Best-1	Best-2	Best-3
Baseline	0.38	0.48	0.60
All Features	0.40	0.51	0.62
−PMI	0.38 (−0.02)	0.51 (±0.00)	0.63 (+0.01)
−CP1	0.33 (−0.07)	0.47 (−0.04)	0.56 (−0.06)
−CP2	0.40 (±0.00)	0.50 (−0.01)	0.64 (+0.02)
−TM Rank	0.35 (−0.05)	0.49 (−0.02)	0.63 (+0.01)
−Hypernym	0.40 (±0.00)	0.50 (−0.01)	0.61 (−0.01)
−PDS	0.45 (+0.05)	0.48 (−0.03)	0.67 (+0.05)
Upper bound	0.64	—	—

Table 5: SVR-based best topic word results for BOOKS for all six feature types, and feature ablation over each (numbers in brackets show the relative change over the full feature set)

where w_{rev_i} is the i^{th} term ranked by the system and w_i is the i^{th} most popular term selected by annotators; rev$_i$ gives the index of the word w_i in the annotator's list; and $n(w)$ is the number of votes given by annotators for word w.

The baseline is obtained using the original word rank produced by the topic model based on topic word probabilities $P(w_i|t_j)$. An upperbound is calculated by evaluating the decision of an annotator against others for each topic. This upperbound signifies the maximum accuracy for human annotators on average; since the annotators were asked to pick a single best word in the survey, only the Best-1 upperbound can be obtained.

The Best-1/2/3 results are summarized in Table 2 for NEWS and Table 3 for BOOKS. These Best-N scores are computed just using the single feature of PMI, CP1 and CP2 (each in turn) to rank the words in each topic. None of these features alone produces a result that exceeds baseline performance.

To make better use of all the features described in Section 3, namely the PMI score, conditional probabilities (both directions), topic model word rank, WordNet Hypernym relationships and Pantel's distributional similarity score, we build a ranking classifier using SVM^{rank} and evaluating using 10-fold cross validation. Our first approach is to use the entire set of features to train the classifier. Following this, we also measure the effect of each feature by ablating (removing) one feature at a time. The drop in Best-N score indicates which features are the strongest predictors of the best words (a larger drop in score indicates that feature is more important). The results for Best-1, Best-2 and Best-3 scores are summarized in Table 4 for NEWS, and Table 5 for BOOKS (averaged across the 10 iterations of cross validation).

We then produced a condensed set of features, consisting of the conditional probabilities, the original topic model word rank and the WordNet hypernym relationships. This "best" set of features is used to make predictions of best words. Results are improved in most cases, and are summarized in Table 6 for both NEWS and BOOKS.

Dataset		Best-1	Best-2	Best-3
NEWS	Baseline	0.35	0.50	0.59
	Best Feat. Set	0.45	0.50	0.65
	Upper bound	0.48	—	—
BOOKS	Baseline	0.38	0.48	0.60
	Best Feat. Set	0.48	0.56	0.66
	Upper bound	0.64	—	—

Table 6: Results with the best feature set compared to the baseline

Dataset	Best-1	Best-2	Best-3
NEWS baseline	0.35	0.50	0.59
BOOKS → NEWS	0.38	0.56	0.62
NEWS upper bound	0.48	—	—
BOOKS baseline	0.38	0.48	0.60
NEWS → BOOKS	0.48	0.56	0.65
BOOKS upper bound	0.64	—	—

Table 7: Results for cross-domain learning

We also tested whether the SVM classifier could be trained using data from one domain, and run on data from another domain. Using our two datasets as these different domains, we trained a model using BOOKS data and made predictions for NEWS, and then we trained a model using NEWS data and made predictions for BOOKS.

The results, shown in Table 7, indicate that we are still able to outperform the baseline, even when the ranking classifier is trained on a different domain. In fact, when we trained a model using NEWS, we saw almost no drop in performance for predicting best words for BOOKS, and improvement is seen for Best-2 score from NEWS. This implies that the SVM classifier generalizes well across domains and suggests the possibility of having a fixed training model to predict best words for any data.

In these experiments, topic words are presented in the original order that the topic model produces, i.e. in descending order of probability of a word under a topic $P(w_i|t_j)$. We noticed that the first words of the topics are frequently selected as the best words by annotators, and suspected that this was introducing a bias towards the first word. As our baseline scores are derived from this topic word ordering, such a bias could give rise to an artificially high baseline.

To investigate this effect, we ran a second annotation exercise over the same set of topics (but different annotators), to obtain a new set of best word annotations for NEWS, with the topic words presented in random order. In Figure 2, we plot the cumulative proportion of words selected as best word by the annotators across the topics, in the case of the random topic word order, mapping the topic words back onto their original ranks in the topic model. A slight drop can be observed in the proportion of first- and second-ranked topic words being selected when we randomise the topic word order. When we recalculate the baseline accuracy for NEWS on the basis of the new set of annotations, we observe an appreciable drop in the scores (see Table 8).

Word Order	Best-1	Best-2	Best-3
Original	0.35	0.50	0.59
Randomized	0.23	0.33	0.46

Table 8: Reduction of baseline scores for NEWS when words are presented in random order to annotators.

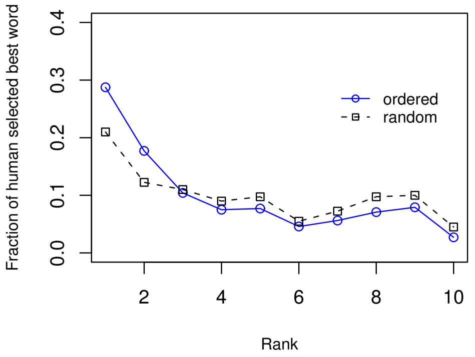

Figure 2: Bias for humans selecting the best word, when the topic words are presented in their original ordering (*ordered*) or randomised (*random*)

6 Discussion

From the experiments in Section 5, perhaps the first thing to observe is: (a) the high performance of the baseline, and (b) the relatively low (Best-1) upper bound accuracy for the task. The first is perhaps unsurprising, given that it represents the topic model's own interpretation of the word(s) which are most representative of that topic. In this sense, we have set our sights high in attempting to better the baseline. The upper bound accuracy is a reflection of both the inter-annotator agreement, and the best that we can meaningfully expect to do for the task. That is, any result higher than this would paradoxically suggest that we are able to do better at a task than humans, where we are evaluating ourselves relative to the labellings of those humans. The upper bound for News was slightly less than 0.5, indicating that humans agree on the best topic word only 50% of the time. To better understand what is happening here, consider the following topic from Figure 1:

health drug patient medical doctor hospital care cancer treatment disease

This is clearly a coherent topic, but at least two topic words suggest themselves as labels: *health* and *medical*. By way of having between 10 and 20 annotators (uniquely) label a given topic, and interpreting the multiple labellings probabilistically, we are side-stepping the inter-annotator agreement issue, but ultimately, for the Best-1 evaluation, we are forced to select one term only, and consider any alternative to be wrong. Because annotators selected only one best topic word, we unfortunately have no way of performing Best-2 or Best-3 upper bound evaluation and deal with topics such as this, but would expect the numbers to rise appreciably.

Looking at the original feature rankings in Tables 2 and 3, no clear picture emerges as to which of the three methods (PMI, CP1 and CP2) was most successful, but there were certainly clear differences in the relative numbers for each, pointing to possible complementarity in the scoring. This expectation was born out in the results for the reranking model in Tables 4 and 5, where the combined feature set surpassed the baseline in all cases, and feature ablation tended to lead to a drop in results, with the single most effective feature set being CP1 ($P(w_i|*)$), followed by CP2 ($P(*|w_i)$) and topic model rank. The lexical semantic features of WordNet hypernymy and PDS (Pantel's distributional similarity) were the worst performers, often having no or negative impact on the results.

Comparing the best results for the SVR-based reranking model and the upper bound Best-1 score, we approach the upper bound performance for News, but are still quite a way off with Books when training in-domain. This is encouraging, but a slightly artificial result in terms of the broader applicability of this research, as what it means in practical terms is that if we can access multi-annotator best word labelling for the majority of topics in a given topic model, we can use those annotations to predict the best word for the remainder of the topics with reasonably success. When we look to the cross-domain results, however, we see that we almost perfectly replicate the best-achieved Best-1, Best-2 and Best-3 in-domain results for Books by training on News (making no use of the annotations for Books). Applying the annotations for Books to News is less successful in terms of Best-1 accuracy, but we actually achieve higher Best-2, and largely mirror the Best-3 results as compared to the best of the in-domain results in Table 6. This leads to the much more compelling conclusion that we can take annotations from an independent topic model (based on a completely unrelated document collection), and apply them to successfully model the best topic word for a new topic model, without requiring any additional annotation. As we now have two sets of topics multiply-annotated for best words, this result suggests that we can perform the best topic word selection task with high success over novel topic models.

We carried out manual analysis of topics where the model did particularly poorly, to get a sense for how and where our model is being led astray. One such example is the topic:

race car nascar driver racing cup winston team gordon season

where the following topic words were selected by our annotators: *nascar* (8 people), *race* (2 people), and *racing* (2 people). First, we observe the split between *race* and *racing*, where more judicious lemmatisation/stemming would make both the annotation easier and the evaluation cleaner. The SVR model tends to select more common, general terms, so in this case chose *race* as the best word, and ranked *nascar* third. This is one instance were *nascar* evokes all of the other words effectively, but not conversely (*racing* is associated with many events/sports beyond *nascar*, e.g.).

Another topic where our model had difficulty was:

> window nave aisle transept chapel tower arch pointed arches roof

where our best model selected *nave*, while the human annotators selected *chapel* (6 people), *arch* (2 people), *nave*, *roof*, *tower* and *transept* (1 person each). Clearly, our annotators struggled to come up with a best word here, despite the topic again being coherent. This is an obvious candidate for labelling with a hypernym/holonym of the topic words (e.g. *church* or *church architecture*), and points to the limitations of best word labelling — there are certainly many topics where best word labelling works, as our upper bound analysis demonstrated, but there are equally many topics where the most natural label is not found in the top-ranked topic words. While this points to slight naivety in the current task set up — we are forcing annotators to label words with topic words, where we know that this is sub-optimal for a significant number of topics — we contend that our numbers suggest that: (a) consistent best topic word labelling is possible at least 50% of the time; and (b) we have developed a method which is highly adept at labelling these topics. As a way forward, we intend to relax the constraint on the topic label needing to be based on a topic word, and explore the possibility of predicting which topics are best labelled with topic words, and which require independent labels. For topics which can be labelled with topic words, we can use the methodology developed here, and for topics where this is predicted to be sub-optimal, we intend to build on the work of Mei et al. (2007), Pantel and Ravichandran (2004) and others in selecting phrasal/hypernym labels for topics. We are also interested in applying the methodology proposed herein to the closely-related task of intruder word, or *worst* topic word, detection, as proposed by Chang et al. (2009).

Finally, looking to the question of the impact of the presentation order of the topic words on best word selection, it would appear that our baseline is possibly an over-estimate (based on Table 8). Having said that, the flipside of the bias is that it leads to more consistency in the annotations, and tends to help in tie-breaking of examples such as *race* and *racing* from above, for example. In support of this claim, the upper bound Best-1 accuracy of the randomised annotations, relative to the original gold-standard is 0.44, slightly below the original upper bound for NEWS. More work is needed to determine the real impact of this bias on the overall task setup and evaluation.

7 Conclusion

This paper has presented the novel task of best topic word selection, that is the selection of the topic word that is the best label for a given topic. We proposed a number of features intended to capture the best topic word, and demonstrated that, while they were relatively unsuccessful in isolation, in combination as inputs to a reranking model, we were able to consistently achieve results above the baseline of simply selecting the highest-ranked topic word, both when training in-domain over other labelled topics for that topic model, and cross-domain, using only labellings from independent topic models learned over document collections from different domains and genres.

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

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

Brody, S. and M. Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, pages 103–111, Athens, Greece.

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd*

Annual Conference on Neural Information Processing Systems (NIPS 2009), pages 288–296, Vancouver, Canada.

Griffiths, T. and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Haghighi, A. 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, Boulder, USA.

Joachims, T. 2006. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 217–226, Philadelphia, USA.

Magatti, D., S. Calegari, D. Ciucci, and F. Stella. 2009. Automatic labeling of topics. In *Proceedings of the International Conference on Intelligent Systems Design and Applications*, pages 1227–1232, Pisa, Italy.

Mei, Q., C. Liu, H. Su, and C. Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th International World Wide Web Conference (WWW 2006)*, pages 533–542.

Mei, Q., X. Shen, and C. Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 490–499, San Jose, USA.

Newman, D., 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.

Pantel, P. and D. Lin. 2002. Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.

Pantel, P. and D. Ravichandran. 2004. Automatically labeling semantic classes. In *Proceedings of the 4th International Conference on Human Language Technology Research and 5th Annual Meeting of the NAACL (HLT-NAACL 2004)*, pages 321–328, Boston, USA.

Pantel, P., E. Crestan, A. Borkovsky, A-M. Popescu, and V. Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 938–947, Singapore.

Ritter, A, Mausam, and O Etzioni. to appear. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the ACL (ACL 2010)*, Uppsala, Sweden.

Titov, I. and R. McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, pages 111–120, Beijing, China.

Wang, X. and A. McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 424–433, Philadelphia, USA.

Wei, S. and W.B. Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 178–185, Seattle, USA.