Unsupervised Estimation of Word Usage Similarity

Marco Lui,[♠♡] **Timothy Baldwin,**[♠♡] and **Diana McCarthy** [♣]
♠ NICTA Victoria Research Laboratory
♡ Dept of Computing and Information Systems, The University of Melbourne
♣ Dept of Theoretical and Applied Linguistics, University of Cambridge
mhlui@unimelb.edu.au, tb@ldwin.net, diana@dianamccarthy.co.uk

Abstract

We present a method to estimate word use similarity independent of an external sense inventory. This method utilizes a topic-modelling approach to compute the similarity in usage of a single word across a pair of sentences, and we evaluate our method in terms of its ability to reproduce a human-annotated ranking over sentence pairs. We find that our method outperforms a bag-of-words baseline, and that for certain words there is very strong correlation between our method and human annotators. We also find that lemma-specific models do not outperform general topic models, despite the fact that results with the general model vary substantially by lemma. We provide a detailed analysis of the result, and identify open issues for future research.

1 Introduction

Automated Word Usage Similarity (Usim) is the task of determining the similarity in use of a particular word across a pair of sentences. It is related to the tasks of word sense disambiguation (WSD) and word sense induction (WSI), but differs in that Usim does not pre-suppose a predefined sense inventory. It also captures the fact that word senses may not always be distinct, and that the applicability of word senses is not necessarily mutually exclusive. In Usim, we consider pairs of sentences at a time, and quantify the similarity of the sense of the target word being used in each sentence. An example of a sentence pair (SPAIR) using similar but not identical senses of the word *dry* is given in Figure 1.

Usim is a relatively new NLP task, partly due to the lack of resources for its evaluation. Erk et al. (2009) recently produced a corpus of sentence pairs annotated for usage similarity judgments, allowing Usim to be formulated as a distinct task from the related tasks of word sense disambiguation and word sense induction.

Part c) All this has been a little dry so far: now for some fun.
For people who knew him, it was typical of his dry humor, but some in the audience thought he was tipsy.

Figure 1: Example of an SPAIR judged by annotators to use similar but not identical senses of the word *dry*.

In this work, we propose a method to estimate word usage similarity in an entirely unsupervised fashion through the use of a topic model. We make use of the well-known Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) to model the distribution of topics in a sentence, then examine the similarity between sentences on the basis of the similarity between their topic distributions.

Our main contributions in this work are: (1) we introduce a method to compute word usage similarity in an unsupervised setting based on topic modelling; (2) we show that our method performs better than the bag-of-words modelling approach; (3) we find that each lemma has a distinct optimum parametrization of the approach that does not generalize across parts of speech; and (4) we demonstrate empirically that per-lemma topic models do not perform differently from global topic models.

2 Background

Polysemy is a linguistic phenomenon whereby the same word has different meaning depending on the context it is used it. For example, the use of the word *charge* in the phrase *charge a battery* is different from its use in the phrase *charge a hill*, and also distinct from its use in *charge in court*.

Marco Lui, Timothy Baldwin and Diana McCarthy. 2012. Unsupervised Estimation of Word Usage Similarity. In *Proceedings of Australasian Language Technology Association Workshop*, pages 33−41.

Word sense disambiguation (WSD) is the task of distinguishing between different senses of a word given a particular usage (Agirre and Edmonds, 2006; Navigli, 2009). Word sense disambiguation presupposes the existence of a *sense inventory*, enumerating all possible senses of a word. WSD is the task of selecting the sense of a word being used from the sense inventory given the context of its use. In contrast, word sense induction (WSI) is the task of partitioning uses of a word according to different senses, producing a sense inventory. In most research to date, the applicability of senses has been regarded as binary, in that a sense either entirely applies or entirely does not apply to a particular use of a word, and senses are regarded as mutually exclusive. This does not take into account situations where a word has different but related senses where more than one sense can apply at a time.

WSI research to date has been evaluated against fixed sense inventories from resources such as dictionaries or WordNet, since they are the primary resources available. However, WSI is a two-part task, where the first part is to determine the similarity between uses of a word, and the second is to partition the uses based on this similarity. The partitions derived thus divide the usages of a particular word according to its distinct senses. Use of a fixed sense inventory in evaluation makes it impossible to evaluate the similarity comparison independently of the partitioning technique. Furthermore, it prevents us from evaluating a WSI technique's ability to detect novel senses of a word or unusual distributions over common senses, because divergence from the fixed sense inventory is usually penalized.

2.1 Usim

Usim was introduced by Erk et al. (2009) to build a case for a graded notion of word meaning, eschewing the traditional reliance on predefined sense inventories and annotation schemas where words are tagged with the best-fitting sense. They found that the human annotations of word usage similarity correlated with the overlap of paraphrases from the English lexical substitution task. In their study, three annotators were asked to rate the similarity of pairs of usages of a lemma on a 5-point scale, where 1 indicated that the uses were completely different and 5 indicated they were identical. The SPAIRS annotated were drawn from LEXSUB (McCarthy and Navigli, 2007), which comprises open class words with token instances of each word appearing in the context of one sentence taken from the English Internet Corpus (EIC) (Sharoff, 2006). Usim annotations were produced for 34 lemmas spanning nouns, verbs, adjectives and adverbs. Each lemma is the target in 10 LEXSUB sentences, and all pairwise comparisons were presented for annotation, resulting in 45 SPAIRS per lemma, for a total of 1530 comparisons per annotator overall. Erk et al. (2009) provide a detailed analysis of the annotations collected, but do not propose an automated approach to word usage similarity, which is the subject of this work.

2.2 Topic Modelling

Topic models are probabilistic models of latent document structure. In contrast to a standard bag-of-words model, a topic model posits an additional intermediate layer of structure, termed the "topics". Each topic is a distribution over words, and a document is modeled as a finite mixture over topics.

The model that we will be using in this work is the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). In the LDA model, each document is modelled as a mixture of topics. Each topic is a multinomial distribution over words, and LDA places a Dirichlet prior on word distributions in topics. Although exact inference of LDA parameters is intractable, the model has gained prominence due to the availability of computationally efficient approximations, the most popular being based on Gibbs sampling (Griffiths and Steyvers, 2004). For brevity, we do not give a detailed description of the LDA model.

2.3 Related Work

Stevenson (2011) experimented with the use of LDA topic modelling in word sense disambiguation, where he used topic models to provide context for a graph-based WSD system (Agirre and Soroa, 2009), replacing a local context derived from adjacent words. This approach is of limited relevance to our work, as the graph-based approach considered state-of-the-art in unsupervised WSD (De Cao et al., 2010) maps senses to individual nodes in a graph. This presupposes the existence of a fixed sense inventory, and thus does not lend itself to determining unsupervised word usage similarity.

Brody and Lapata (2009) proposed an LDA topic modelling approach to WSI which combines feature sets such as unigram tokens and dependency relations, using a layered feature representation. Yao and Van Durme (2011) extended this work in applying a Hierarchical Dirichlet Process (HDP: Teh et al. (2006)) to the WSI task, whereby the topic model dynamically determines how many topics to model the data with, rather than relying on a preset topic number. Recently, Lau et al. (2012) further extended this work and applied it to the task of novel sense detection.

More broadly, this work is related to the study of distributional semantics of words *in context* (Erk and Padó, 2008). Dinu and Lapata (2010) propose a probabilistic framework for representing word meaning and measuring similarity of words in context. One of the parametrizations of their framework uses LDA to automatically induce latent senses, which is conceptually very similar to our approach. One key difference is that Dinu and Lapata focus on inferring the similarity in use of *different* words given their context, whereas in this work we focus on estimating the similarity of use of a *single* word in a number of different contexts.

3 Methodology

Our basic framework is to produce a vector representation for each item in a LEXSUB sentence pair (SPAIR), and then compare the two vectors using a distance measure (Section 3.2). Evaluation is carried out by comparing the per-SPAIR predictions of word usage similarity to the average rating given by human annotators to each SPAIR. The use of the average rating as the goldstandard is consistent with the use of leave-one-out resampling in estimating inter-annotator agreement (Erk et al., 2009). Our evaluation metric is Spearman's ρ with tie-breaking, also consistent with Erk et al. (2009). We compute ρ over the set of all SPAIRS, as well as broken down by part-of-speech and by individual lemma. Positive correlation (higher positive values of ρ) indicates better agreement.

3.1 Background Collections

The data used to learn the parameters of the topic model (henceforth referred to as the *background collection*) has a strong influence on the nature of the topics derived. We investigated learning topic model parameters from 3 global background collections:

SENTENCE The set of 340 sentences used in the Usim annotation

PAGE The set of 340 pages in the EIC from which SENTENCE was extracted

CORPUS The full English Internet Corpus (EIC)

A global background collection is expected to learn word associations ('topics') that are representative of the content of the corpus. Our intuition is that the distribution over topics for similar senses of a word should also be similar, and thus that the distribution over topics can be used to represent a particular use of a word. We discuss how to derive this distribution in Section 3.2.

Prior to learning topic models, we lemmatized the text and eliminated stopwords. In this work, we do not investigate the LDA hyperparameters α and β. We use the common default values of $\alpha = 0.1$ and $\beta = 0.01$.

3.2 Vector-based Representation

Our representation for each usage (each item in an SPAIR) consists of a distribution over topics. We obtain this distribution by mapping each word in the usage context onto a single latent topic using the LDA model. We denote the context in terms of a tuple CONTEXT(a,b). CONTEXT(0,0) indicates that only the annotated sentence was used, whereas CONTEXT(3,3) indicates that three sentences before and three sentences after the annotated sentence were used. Note that in the Usim annotations of Erk et al. (2009), the annotators' judgments were based solely on the sentence pairs, without any additional context. This corresponds exactly to CONTEXT(0,0).

For comparing the vector-based representations of two sentences, we used cosine similarity (`Cosine`). Since the topic vectors can be interpreted as a probability distribution over topics, we also experimented with a similarity metric based on Jensen-Shannon Divergence. We found that cosine similarity provided marginally better results, though the differences were usually minimal.

We also investigated the topic distribution of specific words in the sentence, such as the words

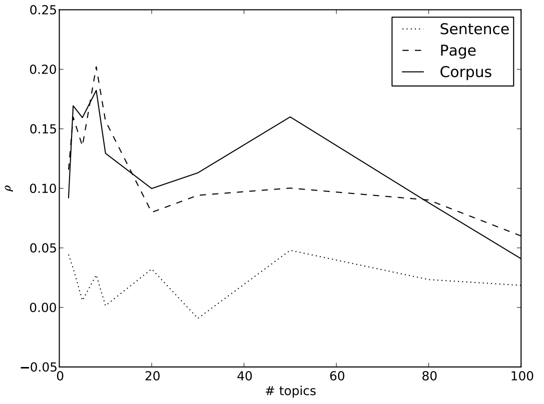

Figure 2: Plot of number of topics against Spearman's ρ per background collection

T_0: ⟨water, plant, area, small, large, fire, tree, tea, food, high⟩
T_1: ⟨quot, http, text, count, amp, book, review, page, language, film⟩
T_2: ⟨war, American, country, force, miltary, government, Iraq, political, United, church⟩
T_3: ⟨service, provide, business, school, cost, need, pay, include, market, information⟩
T_4: ⟨information, system, site, need, computer, number, datum, test, program, find⟩
T_5: ⟨think, question, thing, point, give, want, fact, find, idea, need⟩
T_6: ⟨PM, post, think, Comment, March, Bush, want, thing, write, June⟩
T_7: ⟨look, think, want, find, tell, thing, give, feel⟩

Figure 3: Characteristic terms per topic in the 8-topic model of PAGE

before and after the annotated word, but found that the whole-sentence model outperformed the per-word models, and thus omit the results on per-word models for brevity.

As a baseline for comparison, we use a standard bag-of-words representation where frequency vectors of words in context are compared. We use the same contexts for the bag-of-word-model that we used to infer topic distributions, thus allowing for a direct evaluation of topic modelling in contrast to a more conventional text representation. Our baseline results are thus derived by using `Cosine` to quantify the similarity between the bag-of-words of the context of different uses of the same lemma.

4 Results

		8-topic	T-topic	
Lemma/POS	IAA	ρ	T	ρ
bar(n)	0.410	0.244	30	0.306
charge(n)	0.836	**0.394**	10	**0.667**
charge(v)	0.658	**0.342**	30	**0.429**
check(v)	0.448	0.233	8	0.233
clear(v)	0.715	0.224	8	0.224
draw(v)	0.570	0.192	10	**0.606**
dry(a)	0.563	**0.608**	5	**0.756**
execution(n)	0.813	0.174	30	0.277
field(n)	0.267	0.118	3	**0.375**
figure(n)	0.554	0.158	3	**0.356**
flat(a)	0.871	**0.444**	50	**0.684**
fresh(a)	0.260	-0.002	20	**0.408**
function(n)	0.121	0.234	30	0.292
hard(r)	0.432	0.138	5	**0.309**
heavy(a)	0.652	-0.014	5	0.261
investigator(n)	0.299	**0.364**	10	**0.583**
light(a)	0.549	-0.078	20	0.180
match(n)	0.694	-0.228	80	0.227
order(v)	0.740	0.153	10	0.287
paper(n)	0.701	-0.026	3	**0.330**
poor(a)	0.537	0.210	10	**0.353**
post(n)	0.719	**0.482**	8	**0.482**
put(v)	0.414	**0.544**	8	**0.544**
raw(a)	0.386	**0.387**	2	**0.392**
right(r)	0.707	**0.436**	8	**0.436**
rude(a)	0.669	**0.449**	8	**0.449**
softly(r)	0.610	**0.604**	8	**0.604**
solid(a)	0.603	**0.364**	3	**0.417**
special(a)	0.438	0.140	30	**0.393**
stiff(a)	0.386	0.289	8	0.289
strong(a)	0.439	0.163	2	0.292
tap(v)	0.773	0.233	30	0.272
throw(v)	0.401	**0.334**	8	**0.334**
work(v)	0.322	-0.063	80	0.132
adverb	0.585	**0.418**	8	**0.418**
verb	0.634	**0.268**	8	**0.268**
adjective	0.601	**0.171**	50	**0.219**
noun	0.687	**0.109**	3	**0.261**
overall	0.630	**0.202**	8	**0.202**

Table 1: Comparison of mean Spearman's ρ of inter-annotator agreement (IAA), Spearman's ρ for best overall parameter combination for CONTEXT(0,0), and Spearman's ρ for the optimal number of topics, using PAGE as the background collection. ρ values significant at the 0.05 level are presented in **bold**.

For each of the three background collections SENTENCE, PAGE and CORPUS, we considered topic counts between 2 and 100 in pseudo-logarithmic increments. We computed Spearman's ρ between the average human annotator rating for each SPAIR and the output of our method for each combination of background collection and topic count. We analyzed the results in terms of an aggregation of SPAIRs across all lemmas, as well as broken down by lemma and part-of-speech.

We found that the best overall result was obtained using an 8-topic model of PAGE, where the overall Spearman's ρ between the human annota-

T_0: ⟨think, want, thing, look, tell, write, text, find, try, book⟩
T_1: ⟨information, system, need, government, provide, include, service, case, country, number⟩
T_2: ⟨find, give, child, water, place, woman, hand, look, leave, small⟩

Figure 4: Characteristic terms per topic in the 3-topic model of PAGE

Mowing The way that you mow your lawn will also affect how well it survives hot, dry conditions.

Surprisingly in such a dry continent as Australia, salt becomes a problem when there is too much water.

If the mixture is too dry, add some water ; if it is too soft, add some flour.

Figure 5: Sentences for *dry(a)* with a strong component of Topic 0 given the 8-topic model illustrated in figure 3

tor averages and the automated word usage similarity computation was a statistically significant 0.202.

A detailed breakdown of the best overall result is given in Table 1. Alongside this breakdown, we also provide: (1) the average inter-annotator agreement (IAA); and (2) the Spearman's ρ for the optimal number of topics for the given lemma.

The IAA is computed using leave-one-out resampling (Lapata, 2006), and is a detailed breakdown of the result reported by Erk et al. (2009). In brief, the IAA reported is the mean Spearman's ρ between the ratings given by each annotator and the average rating given by all other annotators. We also present the Spearman's ρ for the best number of topics in order to illustrate the impact of the number of topics parameter for the model of the background collection. We find that for some lemmas, a lower topic count is optimal, whereas for other lemmas, a higher topic count is preferred. In aggregate terms, we found that verbs, adverbs and nouns performed better with a low topic count, whereas adjectives performed best with a much higher topic count.

On the basis of the best overall result, we examined the effect of the topic count and training collection. These results are shown in Figure 2. We found that aggregated over all lemmas, the topic models learned from the full-page contexts (PAGE) and the whole English Internet Corpus (CORPUS) always do better than those learned from just the single-sentence training collection (SENTENCE). This observation is also true when we examine the aggregation over individual parts-of-speech. In general, the results obtained with topic models of PAGE tend to be similar to those obtained with topics models of CORPUS, and across all lemmas the optimum number of topics is about 8.

The software is the program that sifts through the millions of pages recorded in the index to find match to a search and rank them in order of what it believes is most relevant.

The tag consists of a tiny chip, about the size of a match head that serves as a portable database.

Figure 6: Sentences for *match(n)* with a high concentration of Topic 4

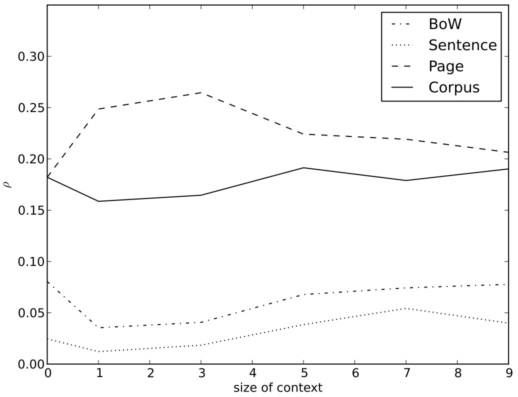

Figure 7: Plot of SPAIR context size against Spearman's ρ (8-topic background collection)

Finally, we also examined our results at a per-lemma level, identifying the optimal topic count for each individual lemma. We found that for all lemmas, there existed a topic count that resulted in a ρ of > 0.4, with the exception of *light(a)*. For some lemmas, their optimal topic count resulted in $\rho > 0.8$ (*check(v)*, *draw(v)*, *softly(r)*). However, the best choice of parameters varied greatly between lemmas, and did not show any observable consistency overall, or between lemmas for a given part-of-speech.

4.1 Topic Modelling vs. Bag-of-words

We computed a baseline result for Usim by using a bag-of-words model for each item in an SPAIR. We examined using only the annotated sentence (CONTEXT(0,0)), as well as varying amounts of context drawn symmetrically around the sentence (CONTEXT(a,b) for $a = b \in \{1, 3, 5, 7, 9\}$).

Figure 7 shows the result of varying the size of the context used. On the x-axis, a value of 0 indicates no additional context was used (i.e. only the annotated sentence was used). A value of 3 indicates that CONTEXT(3,3) was used (i.e. 3 sentences before and after, in addition to the annotated sentence). Based on earlier results, we only considered 8-topic models for each background collection. In general, we found that the page-level(PAGE) and corpus-level(CORPUS) topic models perform better than the bag-of-words (BoW) model and the sentence-level topic model(SENTENCE).

For each context, we used Welch's t-test to determine if the difference between background collections was statistically significant. We found that at the 5% level, for all contexts, CORPUS and PAGE are different from BoW. We also found that at the 5% level, for all contexts, CORPUS and PAGE are different. Overall, the best performance was observed on the 8-topic PAGE model, using CONTEXT(3,3). This yielded a Spearman's ρ of 0.264 with respect to the gold standard annotations.

4.2 Global vs. Per-lemma Topic Models

We have already demonstrated that the topic modelling approach yields improvements over the bag of words model for estimating word usage similarity, provided that that PAGE or CORPUS background collections are used. However, performance on individual lemmas varies widely. [1] One possible reason for this is that the topics being learned are too general, and thus the latent semantics that they capture are not useful for estimating the similarity in word use. To address this issue, we experiment with learning topic models *per-lemma*, learning topics that are specific to each target lemma.[2]

In the per-lemma approach, instead of a single global topic model, we learn a distinct set of topics for each lemma. The per-lemma models use only sentences in which the target lemma occurs, plus one sentence before and one sentence after (CONTEXT(1,1)). Thus, the background collections for each lemma are a (small) subset of CORPUS, and have some overlap with PAGE, although they also include uses of the lemmas that were not annotated and therefore not present in PAGE. We assembled the background collection for each lemma before part-of-speech tagging, so for *charge(n)* and *charge(v)* a single topic model was used. This gave us 33 different topic models for the set of 34 lemmas.

We compare the use of a global topic model to the use of per-lemma topic models in estimating the similarity in word usage of a given lemma L in each sentence in an SPAIR. Given a topic count T and a symmetric usage context CONTEXT(k, k), we map each word in the context into the corresponding topic space. For the global model, we use a topic space of T topics in PAGE, and for the per-lemma model we use the T topic model of all the occurrences of L in CORPUS. Note that the context extracted for each occurrence of L in CORPUS is kept constant (CONTEXT(1,1)); it is the context k used to represent each sentence in the annotated SPAIR that is varied. Thus, the overall result we compute for the global topic models is based on inference on a single global topic model, whereas the overall result reported for the per-lemma approach is based on inference in each of the per-lemma topic models.

Disappointingly, the per-lemma topic models fail to improve on the performance of the global topic models. Across a range of context sizes k and number of topics T in the topic model, we find that the global PAGE model and the per-lemma topic models have nearly indistinguishable performance. The per-lemma models are actually worse than the global model at a high topic count ($T = 50$ and beyond). The overall best result remains the 8-topic PAGE model using CONTEXT(3,3), which a Spearman's ρ of 0.264 with respect to the gold standard annotations. The best result for the per-lemma topic models is 0.209, obtained with an 8-topic model using CONTEXT(5,5).

5 Discussion & Future Work

From the breakdown given in Table 1, we observe that the effectiveness of our approach varies significantly between parts-of-speech and between individual lemmas. For example, for *dry(a)*, we see a fairly strong correlation between the calculated similarity and the human annotations. This

[1] Human performance also varies by lemma as shown by the range in IAA scores. System performance would be increased if we could focus on those lemmas with higher IAA but since we would have no way of predicting IAA in advance we include all lemmas in our overall figures.

[2] This also addresses the unlikely situation where an SPAIR shares two target lemmas, where the uses of one are very similar and the uses of the other are very different.

correlation is much stronger than that observed across all the adjectives.

For the lemmas *dry(a)*, *investigator(n)* and *put(v)*, the correlation between our method and the human annotations exceeds the average inter-annotator agreement. This is due to variation in the inter-annotator agreement by lemma. Often, two annotators produce rankings that are similar, with the third being very different. In this case our system output may be more similar to the average of the three annotators than the mean similarity of each annotator to the average of the other two. In absence of additional annotations,[3] it may be possible to correct for systematic differences in the annotators use of the gradings to achieve a more consistent ordering over SPAIRS. We leave this investigation to future work.

For part-of-speech aggregation, the highest correlation is seen in adverbs, which is somewhat surprising since adverbs are not normally thought of as being strongly topical in nature. In order to gain further insight into the sources of the correlations, we examined the data in greater detail. In particular, we manually inspected the characteristic terms of each topic learned by the topic model. These terms are reproduced in Figure 3. For contrast, we include the best terms for each topic in the 3-topic model of PAGE, which was the best overall for nouns (Figure 4).

We examined the topic distribution for all the sentences for *dry(a)*. We found that in the 8-topic model of PAGE, Topic 0 clusters terms associated with water, food and plants. The sentences with a strong component of Topic 0 are reproduced in Figure 5. We found that sentences with strong components of Topic 0 were likely to use *dry* in the sense of "lacking in water", thus this particular topic was well suited to measuring the similarity in the use of the word *dry*; uses of *dry* in relation to water had a strong component of Topic 0, whereas uses of *dry* not related to water did not.

Although a topic count of 2 is unusually low for LDA modelling of text, we found that for some lemmas this was the optimum topic count, and for *raw(a)* the correlation between annotations and our usage similarity estimation was statistically significant. A possible explanation is that the topics in the 2-topic model aligned with variation in the senses of raw found in different genres of text.[4]

The use of topic distribution is not a panacea for usage similarity. An example of how topic modelling can be misleading is given in Figure 6. Here, we find two sentences with a high concentration of Topic 4, which is related to computers. Both sentences do indeed talk about concepts related to computers; however the use of *match(n)* in the two sentences is completely different. In the first instance, the use of match is topical to the concepts of searching and ranking, whereas in the second instance the term match is used for an analogy about size, and thus this usage of match has little to no topical relation with the rest of the sentence.

Overall, we find that the use of topic models provides a statistically significant improvement in estimating word usage similarity with respect to a bag-of-words model. We observe that for both BoW and topic-modelling approaches, modelling a usage using only the sentence that it occurs in provides inferior results to using a larger context. For the BoW model, the results kept improving as the context was increased, though at larger contexts the comparison essentially becomes one of word distributions in the entire document rather than in the particular usage of a word. This illustrates a key issue with a bag-of-words model of a single sentence: the resulting vectors are very sparse, which makes judging their similarity very difficult.[5]

We observed that the per-lemma topic models did not perform any better than the global topic models, which suggests that the performance increase of automated estimation of word usage similarity may be simply due to dimensionality reduction rather than the latent semantic properties of the topic model. However, we found that the PAGE models outperformed the CORPUS models. This indicates that the actual data used for topic modelling has an impact on per-

[3] In more recent work (Erk et al., 2012) judgments are collected from eight annotators which increases inter-annotator agreement overall, although agreement per-lemma will still vary depending on the semantics of the lemma in question.

[4] Inspection of the top terms for each of the two topics suggested a rough division between "news" and "lay language", but it is not clear exactly how these align with the uses or *raw(a)*. We leave further analysis of this for future work.

[5] It may be possible to use second order co-occurrences to alleviate this to some extent by using the centroid of vectors of the words in context where those vectors are taken from a whole corpus.

formance, suggesting that some latent semantic properties are being recovered. CORPUS is a very much larger background collection than PAGE. In this respect we would expect a much larger diversity of topics in CORPUS than in PAGE, and to some extent this is supported by the results presented in Figure 2. Here, we see a peak in performance at $T = 50$ topics for the CORPUS model that is not present in the PAGE model. However, this is a local optimum. The best topic count for both PAGE and CORPUS was at $T = 8$ topics. The reasons for this are not fully clear, but perhaps may be again attributable to sparsity. Where a large number of topics is used, only a very small number of words may be assigned to each topic. This is supported by the results in Figure 7, where we see an initial increase in performance as we increase the size of the context. However, this increase due to increased context is counteracted by a decreased topical coherence in the larger context, thus for the PAGE model we see that performance decreases after CONTEXT(3,3). Interestingly, for the CORPUS model there is no corresponding decrease in performance. However, at larger context sizes we are reaching a limit in the context in that the entire document is being used, and thus this increase cannot be extended indefinitely.

Overall, this work has shown promising results for a topic modelling approach to estimating word usage similarity. We have found that topic distributions under a topic model can be effective in determining similarity between word usages with respect to those determined by human annotators. One problem that we faced was that the optimal parameters varied for each lemma, and there was no obvious way of predicting them in an unsupervised context. We found that although the globally-optimal approach produced a statistically significant correlation with human annotators for many of the lemmas, most lemmas had a different locally-optimal parametrization. This suggests that a promising avenue for future research is a semi-supervised approach to estimating word usage similarity. Given a small amount of training data, it may be possible to determine the optimal parameters for topic-modelling-based estimation of word usage similarity, which can then be applied to word usage similarity estimation in much larger text collections. The HDP model of Teh et al. (2006) would be an alternative approach to resolving this issue.

We also have not fully explored the effect of the background collection. We found that topic models of background collections drawn at the document level performed better than those drawn at the corpus level, but that those drawn at the per-lemma sentence level were not measurably different from those drawn at the document level. Two additional background collections could be investigated: (1) at the per-lemma document level, where entire documents containing a given lemma are used; and (2) at a cross-corpus level. The former would give insight on whether there is an issue of data sparsity at the parameter estimation phase, since we found that for global models, the document-level background collection outperformed the sentence-level background collection. For the latter, including data from additional corpora may result in a better correspondence between topics and senses, allowing for better estimation of word usage similarity.

6 Conclusion

In this work, we examined automated estimation of word usage similarity via vector similarity over topic vectors inferred from LDA topic models. We found that such topic vectors outperform a bag-of-words baseline, with the globally optimal parametrization attaining Spearman's ρ of 0.264 with the average annotation given by 3 human annotators across 1530 SPAIRS. We also found that each lemma has a different optimum topic count. In some cases, the correlation between our method and the average of human annotations exceeds the inter-annotator agreement. However, the optimum topic count is difficult to predict, and is not consistent within parts of speech. Finally, we found that per-lemma topic models do not significantly improve results with respect to global topic models. Overall, we have shown that a topic modelling approach has potential for automated estimation of word usage similarity, but there remain a number of open issues to investigate which may lead to even better performance.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, Netherlands.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, (April):33–41.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

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

Diego De Cao, Roberto Basili, Matteo Luciani, Francesco Mesiano, and Riccardo Rossi. 2010. Robust and efficient page rank for word sense disambiguation. In *Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 24–32, Uppsala, Sweden, July. Association for Computational Linguistics.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172, Cambridge, MA, October. Association for Computational Linguistics.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii, October. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - ACL-IJCNLP '09*, page 10, Morristown, NJ, USA. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nick Gaylord. 2012. Measuring word meaning in context. *to appear in Computational Linguistics*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the EACL (EACL 2012)*, pages 591–601, Avignon, France.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2).

Serge Sharoff. 2006. Open-source Corpora Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 4:435–462.

Mark Stevenson. 2011. Disambiguation of medline abstracts using topic models. In *Proceedings of the ACM fifth international workshop on Data and text mining in biomedical informatics*, pages 59–62. ACM.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Xuchen Yao and Benjamin Van Durme. 2011. Nonparametric bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14, Portland, USA.