A Hybrid Generative/Discriminative Approach To Citation Prediction

Chris Tanner
Brown University
Providence, RI
christanner@cs.brown.edu

Eugene Charniak
Brown University
Providence, RI
ec@cs.brown.edu

Abstract

Text documents of varying nature (e.g., summary documents written by analysts or published, scientific papers) often cite others as a means of providing evidence to support a claim, attributing credit, or referring the reader to related work. We address the problem of predicting a document's cited sources by introducing a novel, discriminative approach which combines a content-based generative model (LDA) with author-based features. Further, our classifier is able to learn the importance and quality of each topic within our corpus – which can be useful beyond this task – and preliminary results suggest its metric is competitive with other standard metrics (Topic Coherence). Our flagship system, *Logit-Expanded,* provides state-of-the-art performance on the largest corpus ever used for this task.

1 Introduction

The amount of digital documents (both online and offline) continues to grow greatly for several reasons, including the eagerness of users to generate content (e.g., social media, Web 2.0) and the decrease in digital storage costs. Many different types of documents link to or cite other documents (e.g., websites, analyst summary reports, academic research papers), and they do so for various reasons: to provide evidence, attribute credit, refer the reader to related work, etc. Given the plethora of documents, it can be highly useful to have a system which can automatically predict relevant citations, for this could (1) aid authors in citing relevant, useful sources which they may otherwise not know about; and (2) aid readers in finding useful documents which otherwise might not have been discovered, due to the documents' being unpopular or poorly cited by many authors. Specifically, we are interested in *citation prediction* – that is, we aim to predict which *sources* each report document cites. We define a *report* as any document that cites another document in our corpus, and a *source* as a document that is cited by at least one report. Naturally, many documents within a corpus can be both a report and a source. Note, we occasionally refer to *linking* a report and source, which is synonymous with saying the report cites the source.

Citation prediction can be viewed as a special case of the more general, heavily-researched area of *link prediction.* In fact, past research mentioned in Section 2 refers to this exact task as both *citation prediction* and *link prediction*. However, *link prediction* is a commonly used phrase which may be used to describe other problems not concerning documents and citation prediction. In these general cases, a link may be relatively abstract and represent any particular relationship between other objects (such as users' interests or interactions). Traditionally, popular techniques for link prediction and recommendation systems have included feature-based classification, matrix factorization, and other collaborative filtering approaches – all of which typically use meta-data features (e.g., names and interests) as opposed to modelling complete content such as full text documents (Sarwar et al., 2001; Al Hasan and Zaki, 2011). However, starting with Hofmann and Cohn's (2001) seminal work on ci-

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 75–83,
Denver, Colorado, May 31 – June 5, 2015.

tation prediction (PHITS), along with Erosheva et. al.'s (2004) work (LinkLDA), content-based modelling approaches have extensively used generative models – while largely ignoring meta-data features which collaborative filtering approaches often use – thus creating somewhat of a dichotomy between two approaches towards the same problem. We demonstrate that combining (1) a simple, yet effective, generative approach to modelling content with (2) author-based features into a discriminative classifier can improve performance. We show state-of-the-art performance on the largest corpus for this task. Finally, our classifier learns the importance of each topic within our corpus, which can be useful beyond this task.

In the next section, we describe related research. In Section 3 we describe our models and motivations for them. In Section 4 we detail our experiments, including data and results, and compare our work to the current state-of-the-art system. We finally conclude in Section 5.

2 Related Work

Hofmann and Cohn's (2001) *PHITS* seminal work on citation prediction included a system that was based on probabilistic latent semantic analysis (PLSA) (Hofmann, 1999). Specifically, they extended PLSA by representing each distinct link to a document as a separate word token – as shown in Equation 1 and represented by s_l. (Note: Table 1 displays common notation that is used consistently throughout this paper.) PHITS assumes both the links and words are generated from the same global topic distributions, and like PLSA, a topic distribution is inferred for each document in the corpus.

$$P(w_i|d_j) = \sum_{k=1}^{K} P(w_i|z_k)P(z_k|d_j), \qquad P(s_l|d_j) = \sum_{k=1}^{K} P(s_l|z_k)P(z_k|d_j) \tag{1}$$

Later, Erosheva et. al.'s (2004) system replaced PLSA with LDA as the fundamental generative process; thus, the topic distributions were assumed to be sampled from a Dirichlet prior, as depicted in the plate notation of Figure 1. We will refer to this model as it is commonly referred, *LinkLDA*, and it is the closest model to our baseline approach (later introduced as *LDA-Bayes*).

M	total # documents in the corpus (both reports and sources)
N	# of words in the particular document
r	a report document
s	a source document
d	a document (report and/or source)
w	a word in a document
K	total # of topics
z	a particular topic
V	corpus' vocabulary size
α, β	concentration parameters to corpus-wide Dirichlet priors
$\triangle(p)$	a simplex of dimension (p-1)
L	number of citations in a particular document
$\Omega_{kd'}$	probability of a link to document d' w.r.t. topic k
s_l	a token representing a link to source s

Table 1: Notation Guide

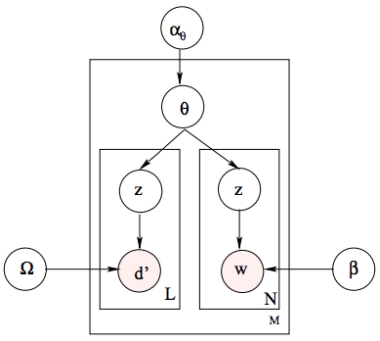

Figure 1: Plate notation of LinkLDA

Others have researched several variants of this LDA-inspired approach, paving the field with promising, generative models. For example, Link-PLSA-LDA is the same as LinkLDA but it treats the generation of the source documents as a separate process inferred by PLSA (Nallapati et al., 2008). Related, Cite-LDA and Cite-PLSA-LDA (Kataria et al., 2010) extend LinkLDA and Link-PLSA-LDA, respectively, by asserting that the existence of a link between a report and source is influenced by the context of where the citation link occurs within the report document. Note, the authors supplemented corpora to include context that surrounds each citation; however, there is currently no freely-available, widely-used corpus which allows one to discern *where* citations appear within each report. Therefore, few systems rely on citation context.

TopicBlock (Ho et al., 2012) models citation prediction with a hierarchical topic model but only uses the first 200 words of each document's abstract. To

our knowledge, *Topic-Link-LDA* (Liu et al., 2009) is the only research which includes both author information and document content into a generative model in order to predict citations. *Topic-Link-LDA* estimates the probability of linking a report-source pair according to the similarity between the documents' (1) author communities and (2) topic distributions – these two latent groups are linearly combined and weighted, and like the aforementioned systems, are inferred by a generative process. PMTLM (Zhu et al., 2013) is reported as the current state-of-the-art system. In short, it is equivalent to PLSA but extended by having a variable associated with each document, which represents that document's propensity to form a link.

As mentioned, although Collaborative Filtering has been used towards citation prediction (McNee et al., 2002), there is little research which includes features based on the entire content (i.e., documents). Very recently, (Wilson et al., 2014) used topic modelling to help predict movie recommendations. Specifically, one feature into their system was the KL-divergence between candidate items' topic distributions, but applying this towards citation prediction has yet to be done. Most similar to our work, (Bethard and Jurafsky, 2010) used a classifier to predict citations, based on meta-data features and compressed topic information (e.g., one feature is the cosine similarity between a report-source pair's topic distribution). As explained in Section 4, we expand the topic information into a vector of length K, which not only improves performance but yields an estimate of the most important, "quality" topics. Further, our system also uses our *LDA-Bayes* baseline as a feature, which by itself yields excellent results compared to other systems on our large corpus. Notably, Bethard and Jurafsky's system (2010) also differs from ours in that (1) their system has an iterative process that alternates between retrieving candidate source documents and learning model weights by training a supervised classifier; and (2) they only assume access to the content of the abstract, not the entire documents. Nonetheless, we use their system's most useful features to construct a comparable system (which we name **WSIC** – "Who Should I Cite"), which we describe in more detail in Section 3.3 and show results for in Section 4.3.

3 New Models

3.1 LDA-Bayes

For a baseline system, we first implemented LDA (Blei et al., 2003) topic modelling and ran it on our entire corpus. However, unlike past systems, after our model was trained, we performed citation prediction (i.e., $P(s|r)$) according to Equation 2. Notice, although LDA does not explicitly estimate $P(s|z)$, we can approximate it via Bayes Rule, and we consequently call our baseline *LDA-Bayes*. Doing so allows us to include the prior probability of the given source being cited (i.e., $P(s)$), according to the maximum-likelihood estimate seen during training.

$$P(s|r) = \sum_{k}^{K} P(s|z_k)P(z_k|r), \quad \text{where } P(s|z_j) = \frac{P(z_j|s)P(s)}{\sum_{s'} P(z_j|s')P(s')} \tag{2}$$

Of the past research which uses generative models for citation prediction, we believe LinkLDA is the only other system in which a source's prior citation probability plays any role in training the model. Specifically, in LinkLDA, the prediction metric is identical to ours in that the topics are marginalized over topics (Equation 3). It differs, however, in that their model directly infers $P(s|z_k)$, for it treats each citation link as a word token. Although this does not explicitly factor in each source's prior probability of being cited, it is implicitly influenced by such, for the sources which are more heavily cited during training will tend to have a higher probability of being generated from topics.

$$P(s|r) = \sum_{k}^{K} P(s|z_k)P(z_k|r), \tag{3}$$

Note: the other generative models mentioned in Section 2, after inference, predict citations by sampling from a random variable (typically a Bernoulli or Poisson distribution) which has been conditioned on the topic distributions.

3.2 Logit-Expanded

In attempt to combine the effectiveness of LDA in generating useful topics with the ability of dis-

Table 2: A randomly chosen report and its predicted sources, per LDA-Bayes, illustrating that a report and predicted source may be contextually similar but that their titles may have few words in common.

Report: Japanese Dependency Structure Analysis Based On Support Vector Machines (2000)			
Position	Cited Source?	Year	Source Name
1		1996	A Maximum Entropy Approach To Natural Language Processing Natural Language Processing
2		1993	Building A Large Annotated Corpus Of English: The Penn Treebank
3		1996	A Maximum Entropy Model For Part-Of-Speech Tagging
4		1994	A Syntactic Analysis Method Of Long Japanese Sentences Based On The Detection Of Conjunctive Structures
5		1992	Class-Based N-Gram Models Of Natural Language
...		...	...
11		1996	Three New Probabilistic Models For Dependency Parsing: An Exploration
12		2000	Introduction To The CoNLL-2000 Shared Task: Chunking
13		1995	A Model-Theoretic Coreference Scoring Scheme
14		1988	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
15	✓	1999	Japanese Dependency Structure Analysis Based On Maximum Entropy Models

criminative classifiers to learn important features for classification, we use logistic regression with a linear kernel. Specifically, we train using L2-regularization, which during test time allows us to get a probability estimate for each queried vector (i.e., a report-source pair).

The details of the training and testing data are provided in Section 4.2. However, it is important to understand that each training and testing instance corresponds to a distinct report-source document pair and is represented as a single fixed-length vector. The vector is comprised of the following features, which our experiments illustrate are useful for determining if there exists a link between the associated report and source:

3.2.1 Topic/Content-Based Features

- **LDA-Bayes:** Our baseline system showed strong results by itself, so we include its predictions as a feature (that is, $P(s|r)$).

- **Topics:** LDA-Bayes ranks report-source pairs by marginalizing over all topics (see Equation 2); however, we assert that not all topics are equally important. Allowing each topic to be represented as its own feature, while keeping the value based on the report-source's relationship for that topic (i.e., the absolute value of the difference), can potentially allow the logistic regression to learn both (1) the importance for report-source pairs to be generally similar across most topics and (2) the relative importance of each topic. For all of our experiments (including LDA-Bayes) we used 125 topics to model the corpus; thus, this feature becomes expanded to 125 individual indices within our vector, which is why we name this system Logit-Expanded. Namely, $\forall i \in K$, let feature $f_i = |\theta_{r_i} - \theta_{s_i}|$.

3.2.2 Meta-data Features

- **Report Author Previously Cited Source?:** We believe authors have a tendency to cite documents they have cited in the past

- **Report Author Previously Cited a Source Author?:** Authors also have a tendency to "subscribe" to certain authors and are more familiar with particular people's works, and thus cite those papers more often.

- **Prior Citation Probability:** A distinguishing feature of our LDA-Bayes model is that it factors in the prior probability of a source being cited, based on the maximum likelihood estimate from the training data. So, we explicitly include this as a feature.

- **Number of Overlapping Authors:** Authors have a tendency to cite their co-authors, in part because their co-authors' past work has an increased chance of being relevant.

- **Number of Years between Report and Source:** Authors tend to cite more recent papers.

- **Title Similarity between Report and Source:** As shown in Table 2, some sources erroneously returned by our baseline system could have been discarded had we judged them by how dissimilar their titles are from the report's title. In Table 2's example, the one correct source to find (within ∼12,000) was returned at position 15 and has many words in common with the report (namely, "Japanese Dependency Structure Analysis Based On" appears in the titles of both the report and correctly predicted source).

3.3 WSIC (Who Should I Cite?)

In attempt to compare our systems against Bethard and Jurafsky's system (2010), we implemented the features they concluded to be most useful for retrieval, and like our Logit-Expanded system, used logistic regression as the mechanism for learning citation prediction. Instead of using only the text from the abstracts, like in their research, to make the comparison more fair we used text from the entire documents – just like we did for the rest of our systems. Specifically, adhering to their naming convention, the features from their system that we used are: *citation-count, venue-citation-count, author-citation-count, author-h-index, age (# years between report and source), terms-citing, topics, authors, authors-cited-article, and authors-cited-author.*

4 Experiments

4.1 Corpora

The past research mentioned in Section 2 primarily makes use of three corpora: Cora, CiteSeer, and WebKB. As shown in Table 3, these corpora are relatively small with ~3,000 documents, an average of less than three links per document, and a modest number of unique word types.

We wanted to use a corpus which was larger, provided the complete text of the original documents, and included meta-data such as author information. Thus, we used the ACL Anthology (Radev et al., 2013) (the December 2013 release), which provides author and year information for each paper, and the corpus details are listed in Table 3. For the task of citation prediction, we are the first to use full content information from a corpus this large.

4.2 Training/Testing Data

The research listed in Section 2 commonly uses 90% of all positive links (i.e., a distinct report-to-source instance) for training purposes and 10% for testing. LDA-based topic modelling approaches, which are standard for this task, require that at testing time each report and candidate source has already been observed during training. This is because at test time the topic distribution for each document must have already been inferred. Additionally, it is common to make the assumption that the corpus is split into a bipartite graph: a priori we know which documents are reports and which are sources, with most being both. At testing time, one then predicts sources from the large set of candidate sources, all of which were seen at some point during training (as either a report or a source document).

We follow suit with the past research and randomly split the ACL Anthology's report-to-source links (citations) into 90% for training and 10% for testing, with the requirement that every candidate source document during testing was seen during training as either a report or a source – ensuring we have a topic distribution for each document. On average, each report has 6.8 sources, meaning typically at test time each report has just a few (e.g., 1-5) sources which we hope to predict from our 12,265 candidate sources. For all of our experiments, the systems (e.g., LDA-Bayes, LinkLDA, Logit-Expanded, etc) were evaluated on the exact same randomly chosen split of training/testing data.

As for training Logit-Expanded, naturally there are vastly more negative examples (i.e., no link between the given report-source pair) than positive examples; most sources are not cited for a given report. This represents a large class-imbalance problem, which could make it difficult for the classifier to learn our task. Consequently, we downsampled the negative examples. Specifically, for each report, we included all positive examples (the cited sources), and for each positive example, we included 5 randomly selected negative examples (sources). Note, for testing our system, we still need to evaluate every possible candidate report-source pair – that is $\sim$12,265 candidate sources per tested report.

Table 3: Report-to-Source Citation Prediction Corpora

	Cora	CiteSeer	WebKB	ACL
# docs	2,708	3,312	3,453	17,298
# links	5,429	4,608	1,733	106,992
vocab size	1,433	3,703	24,182	137,885
# authors	-	-	-	14,407

4.3 Results

4.3.1 Report-To-Source Citation Prediction

First, we tested our LDA-Bayes baseline system and compared it to LinkLDA and PMTLM (Zhu et

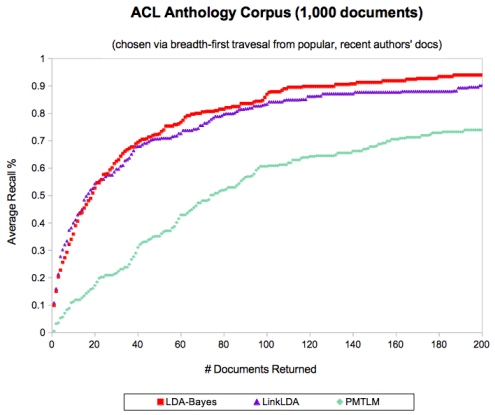

Figure 2: Average Recall Performance across all Reports from a 1,000 document subset of the ACL Anthology

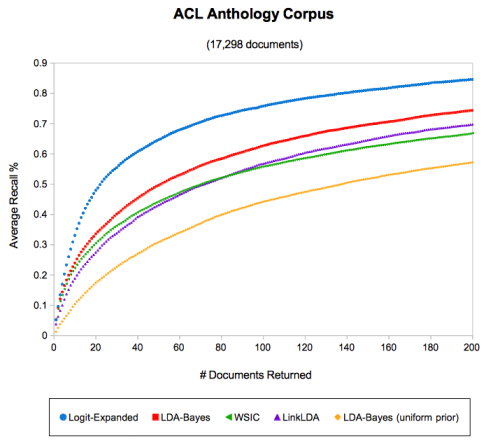

Figure 3: Average Recall Performance across all reports from the full ACL Anthology

al., 2013) – the current state-of-the-art system. Due to the slow running time of PMTLM, we restricted our preliminary experiment to just 1,000 documents of the ACL Anthology, and Figure 2 shows the average recall performance across all reports. Surprisingly, PMTLM performed worst. Note: the authors of PMTLM compared their system to LinkLDA for a different task (predicting research area) but did not compare to LinkLDA during their analysis of citation prediction performance. Thus, it was not previously asserted that PMTLM would outperform LinkLDA.

As we can see, LDA-Bayes, despite being simple, performs well. As mentioned, LDA-Bayes explicitly captures the prior probability of each source being cited (via maximum-likelihood estimate), whereas LinkLDA and PMTLM approximates this during inference. We believe this contributes towards the performance differences.

It was expected that when run on the entire ACL corpus, WSIC and our Logic-Expanded systems would have sufficient data to learn authors' citing preferences and would outperform the other generative models. As shown in Figure 3 and 4, our flagship Logit-Expanded system greatly outperformed all other systems, while our baseline LDA-Bayes continued to offer strong results. Note, the full recall performance results include returning 12,265 sources, but we only show the performance for returning the first 200 returned sources. Further, Table 4 shows the same experimental results but for the performance when returning just the first 50 predicted sources per report.

Table 4: Performance of each system, averaged across all reports while returning the top 50 predicted sources for each. 125 topics were used for every system.

	recall	precision	fscore
Logit-Expanded	.647	.016	.031
LDA-Bayes	.496	.012	.024
WSIC	.442	.011	.021
LinkLDA	.431	.011	.021
LDA-Bayes (uniform prior)	.309	.007	.014

Again, we further see how effective it is to have a model influenced by a source's prior probability, for when we change LDA-Bayes such that $P(SourceCited)$ is uniform for all sources, performance falls greatly – represented as *LDA (uniform prior)*.

We analyzed the benefits of each feature of Logit-Expanded in 2 ways: (1) starting with the full-feature set experiment (whose results we showed), we evaluate each feature by running an experiment whereby the said feature is removed; and (2) starting with our LDA-Bayes baseline as the only feature for our Logit-Expanded system, we evaluate each feature by running an experiment whereby the said feature is paired with LDA-Bayes as the only two features used. For both of these approaches, we measure performance by looking at recall, precision, and f-score when returning the first 50 predicted sources. The results are shown in Table 5; technique (1) is shown in column *removal*, and (2)

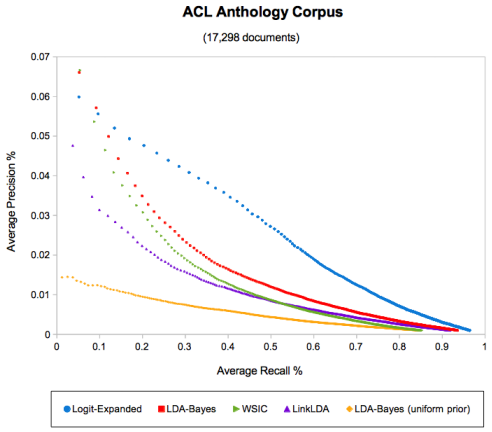

Figure 4: Recall vs Precision Performance across all Reports from the full ACL Anthology. Logit-Expanded's slight blips at recall = 0.25, 0.33, and 0.5 is due to the truth set having many reports with only 4, 3, or 2 golden sources, respectively.

is in column *addage*.

Table 5 reveals insightful results: it is clear that LDA-Bayes is a strong baseline and useful feature to include in our system, for removing it from our feature list causes performance to decrease more than removing any other feature. *PrevCitedSource* and *Topics Expanded* are the second and third strongest features, respectively. We suspect that *PrevCitedSource* was a good feature because our corpus was sufficiently large; had our corpus been much smaller, there might not have been enough data for this feature to provide any benefit. Next, *Title Similarity* and *# Shared Authors* were comparably good features. *PrevCitedAuthor* and *# Years Between* were the worst features, as we see negligible performance difference when we (1) pair either with LDA-Bayes, or (2) remove either from our full feature list. An explanation for the former feature's poor performance could be that authors vary in (1) how often they repeatedly cite authors, and most likely (2) many authors have small publication histories within training, so it might be unwise to base prediction on this limited information. Last, it is worth noting that when we pair Topics Expanded with LDA-Bayes, that alone is not enough to give the best performance from a pair. An explanation is that it dominates the system with too much content-based (i.e., topic) information, overshadowing the prior-citation-probability that plays a role in LDA-Bayes. Supporting this idea, we see the biggest performance increase when we pair LDA-Bayes with the PrevCitedSource feature – a non-topic-based feature, which provides the system with a different *type* of data to leverage.

Table 5: Analysis of each feature used in Logit-Expanded. Results based on the first 50 sources returned, averaged over all reports. Our *Starting Point** system listed within the "Addage" columns used LDA-Bayes as the only feature. Our *Starting Point** system within the "Removal" columns used every feature.

	Addage			Removal		
	recall	precision	fscore	recall	precision	fscore
Starting Point*	.496	.012	.024	.647	.016	.031
LDA-Bayes	-	-	-	**.583**	**.014**	**.028**
Topics Expanded	**.564**	**.014**	**.027**	**.606**	**.015**	**.028**
PrevCitedSource	**.581**	**.014**	**.028**	**.599**	**.014**	**.028**
PrevCitedAuthor	.484	.012	.023	.641	.016	.030
# Shared Authors	.543	.013	.026	.636	.015	.029
Prior Prob. Cited	.501	.012	.023	.639	.015	.030
Title Similarity	.513	.012	.023	.623	.015	.029
# Years Between	.498	.012	.023	.645	.016	.030

Additionally, when using only the metadata features (i.e., not LDA-Bayes or Topics-Expanded), performance for returning 50 sources averaged 0.403, 0.010, and 0.019 for recall, precision, and fscore, respectively – demonstrating that the metadata features alone do not yield strong results but that they complement the LDA-Bayes and Topics-Expanded features.

4.3.2 Topic Importance

Although Report-to-Source citation prediction was our primary objective, our feature representation of topics allows logistic regression to appropriately learn which *topics* are most useful for predicting citations. In turn, these topics are arguably the most cohesive; thus, our system, as a byproduct, provides a metric for measuring the "quality" of each topic. Namely, the weight associated with each topic feature indicates the topic's importance – the lower the weight the better.

Table 6 shows our system's ranking of the most important topics, signified by "Logit-weight." We did not prompt humans to evaluate the quality of the topics, so in attempt to offer a comparison, we also rank each topic according to two popular metrics: Pointwise Mutual Information (PMI) and Topic Coherence (TC) (Mimno et al., 2011). For a topic k,

let $V^{(k)}$ represent the top M words for K; where $V^{(k)} = (v_i^{(k)}, ..., v_M^{(k)})$ and $D(v)$ represents the document frequency of word type v. Then, $PMI(k)$ is defined by Equation 4 and $TC(k)$ is defined by Equation 5.

In Table 6, we see that our most useful topic (Topic 49) concerns vision research, and since our corpus is heavily filled with research concerning (non-vision-related) natural language processing, it makes sense for this topic to be highly important for predicting citations. Similarly, we see the other top-ranking topics all represent a well-defined, subfield of natural language processing research, including parsing, text generation, and Japanese-English machine translation.

$$PMI(k; V^{(k)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{p(V_m^{(k)}, V_l^{(k)})}{p(V_m^{(k)}) p(V_l^{(k)})} \quad (4)$$

$$TC(k; V^{(k)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(V_m^{(k)}, V_l^{(k)})}{D(V_m^{(k)})} \quad (5)$$

Table 7 shows the worst 5 topics according to Logit-Expanded. Topic 96 concerns Wikipedia as a corpus, which naturally encompasses many areas of research, and as we would expect, the mention of such is probably a poor indicator for predicting citations. Topic 77 concerns artifacts from the OCR-rendering of our corpus, which offers no meaningful information. In general, the worst-ranking topics concern words that span many documents and do not represent cohesive, well-defined areas of research. Additionally, in both Table 6 and 7 we see that Pointwise Mutual Information (PMI) disagrees quite a bit with our Logit-Expanded's ranking, and from this initial result, it appears Logit-Expanded's ranking might be a better metric than PMI – at least in terms of quantifying relevance towards documents being related and linked via a citation.

This cursory, qualitative critique of the metrics warrants more research, ideally with human-evaluation. However, one can see how these metrics differ: TC and PMI are both entirely concerned with just the co-occurrence of terms, normalized by the general popularity of the said terms. Therefore, words could highly co-occur together but otherwise represent nothing special about the corpus at large. On the other hand, Logit-Expanded's ranking is mainly concerned with quantifying how well each topic represents discriminatively useful content within a document.

Table 6: The highest quality topics (out of 125), sorted according to Logit-Expanded's estimate. Topics are also ranked according to Pointwise Mutual Information (PMI) and Topic Coherence (TC).

Logit's Rank	PMI Rank	TC Rank	Logit Weight	Topic #	Top Words
1	116	103	-5.50	49	image, visual, multimodal, images, spatial, gesture, objects, object, video, scene, instructions, pointing
2	33	44	-4.76	25	grammar, parsing, grammars, left, derivation, terminal, nonterminal, items, free, string, item, derivations, cfg
3	68	37	-4.71	65	generation, generator, generated, realization, content, planning, choice, nlg, surface, generate
4	49	27	-4.28	32	noun, nouns, phrases, adjectives, adjective, compound, verb, head, compounds, preposition
5	107	61	-4.24	0	japanese, ga, expressions, wo, accuracy, bunsetsu, ni, dictionary, wa, kanji, noun, expression

Table 7: The lowest quality topics (out of 125), sorted by Logit-Expanded's estimate. Topics are also ranked according to Pointwise Mutual Information (PMI) and Topic Coherence (TC).

Logit's Rank	PMI Rank	TC Rank	Logit Weight	Topic #	Top Words
121	13	110	-1.45	96	wikipedia, links, link, articles, article, title, page, anchor, pages, wiki, category, attributes
122	83	122	-1.20	77	x1, x2, c1, c2, p2, a1, p1, a2, r1, l1, xf, fi
123	42	36	-1.09	91	annotation, agreement, annotated, annotators, annotator, scheme, inter, annotate, gold, kappa
124	10	34	-0.75	43	selection, learning, active, selected, random, confidence, sample, sampling, cost, size, select
125	65	115	-0.33	30	region, location, texts, city, regions, weather, locations, map, place, geographic, country

5 Conclusions

We have provided a strong baseline, LDA-Bayes, which when run on the largest corpus for this task, offers compelling performance. We have demonstrated that modelling the prior probability of each candidate source being cited is simple yet important, for it allows all of our systems to outperform the previous state-of-the-art – our large corpus helps towards making this a useful feature, too.

Our biggest contribution is our new system, Logit-Expanded, which combines both the effectiveness of the generative model LDA with the power of logistic regression to discriminately learn important features for classification. By representing each topic as its own feature, while still modelling the re-

lationship between the candidate report-source pair, we allow our system to learn (1) that having similar topic distributions between reports and sources is indicative of a link, and (2) which topics are most important for predicting a link. Because we used a linear kernel, we are able to discern exactly how important it ranks each topic. A cursory, qualitative assessment of its metric shows promising and competitive performance with that of Pointwise Mutual Information and Topic Coherence.

References

Mohammad Al Hasan and Mohammed J Zaki. 2011. A survey of link prediction in social networks. In *Social network data analytics*, pages 243–275. Springer.

Steven Bethard and Dan Jurafsky. 2010. Who should i cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 609–618. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Elena Erosheva, Stephen Fienberg, and John Lafferty. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220–5227.

Qirong Ho, Jacob Eisenstein, and Eric P Xing. 2012. Document hierarchies from text and links. In *Proceedings of the 21st international conference on World Wide Web*, pages 739–748. ACM.

David Hofmann and Thomas Cohn. 2001. The missing link-a probabilistic model of document content and hypertext connectivity. In *Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems. The MIT Press*, pages 430–436.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. 2010. Utilizing context in generative bayesian models for linked corpus. In *AAAI*, volume 10, page 1.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. Topic-link lda: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning*, pages 665–672. ACM.

Sean M McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K Lam, Al Mamunur Rashid, Joseph A Konstan, and John Riedl. 2002. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 116–125. ACM.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics.

Ramesh M Nallapati, Amr Ahmed, Eric P Xing, and William W Cohen. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550. ACM.

DragomirR. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, pages 1–26.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM.

Jobin Wilson, Santanu Chaudhury, Brejesh Lall, and Prateek Kapadia. 2014. Improving collaborative filtering based recommenders using topic modelling. *arXiv preprint arXiv:1402.6238*.

Yaojia Zhu, Xiaoran Yan, Lise Getoor, and Cristopher Moore. 2013. Scalable text and link analysis with mixed-topic link models. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 473–481. ACM.