The Strength of the Weakest Supervision: Topic Classification Using Class Labels

Jiatong Li[1], **Kai Zheng**[2], **Hua Xu**[3], **Qiaozhu Mei**[4], **Yue Wang**[5]
[1]Department of Computer Science, Rutgers University
[2]Department of Informatics, University of California, Irvine
[3]School of Biomedical Informatics, The University of Texas Health Science Center at Houston
[4]School of Information, University of Michigan, Ann Arbor
[5]School of Information and Library Science, University of North Carolina at Chapel Hill
[1]jiatong.li@rutgers.edu, [2]zhengkai@uci.edu, [3]hua.xu@uth.tmc.edu,
[4]qmei@umich.edu, [5]wangyue@email.unc.edu

Abstract

When developing topic classifiers for real-world applications, we begin by defining a set of meaningful topic labels. Ideally, an intelligent classifier can understand these labels right away and start classifying documents. Indeed, a human can confidently tell if a news article is about science, politics, sports, or none of the above, after knowing just the class labels.

We study the problem of training an initial topic classifier using only class labels. We investigate existing techniques for solving this problem and propose a simple but effective approach. Experiments on a variety of topic classification data sets show that learning from class labels can save significant initial labeling effort, essentially providing a "free" warm start to the topic classifier.

1 Introduction

When developing topic classifiers for real-world tasks, such as news categorization, query intent detection, and user-generated content analysis, practitioners often begin by crafting a succinct definition, or a *class label*, to define each class. Unfortunately, these carefully written class labels are completely ignored by supervised topic classification models. Given a new task, these models typically require a significant amount of labeled documents to reach even a modest initial performance. In contrast, a human can readily understand new topic categories by reading the class definitions and making connections to prior knowledge. Labeling initial examples for every new task can be time-consuming and labor-intensive, especially in resource-constrained domains like medicine and law. Therefore it is desirable if a topic classifier can proactively interpret class labels before the training starts, giving itself a "warm start". An imperfect initial model can always be fine-tuned with more labeled documents.

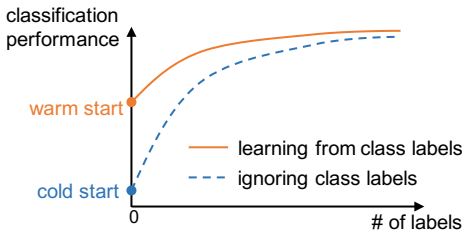

Figure 1: Learning from class labels can give "warm start" to a classifier, accelerating the learning process.

As conceptually shown in Figure 1, a warm start can reduce the total number of training labels for a classifier to reach certain performance level.

In this work, we study algorithms that can initialize a topic classifier using *class labels only*. Since class labels are the starting point of any topic classification task, they can be viewed as the earliest hence weakest supervision signal. We propose a simple and effective approach that combines word embedding and naive Bayes classification. On six topic classification data sets, we evaluate a suite of existing approaches and the proposed approach. Experimental results show that class labels can train a topic classifier that generalizes as well as a classifier trained on hundreds to thousands of labeled documents.

2 Related Work

Text retrieval. Classifying documents by short labels can be viewed as evaluating textual similarity between a document and a label. Baeza-Yates et al. (2011) called this approach "naive text classification". Treating labels as search queries, we can classify a document into a class if it best matches the label of that class. Well-studied text retrieval methods, such as vector space models and probabilistic models (Croft et al., 2010), can produce matching scores. To mitigate vocabulary mismatch, such a classifier can be further enhanced

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 22–28
Minneapolis, Minnesota, June 3 - 5, 2019.

by self-training: the classifier assigns pseudo labels to top-ranked documents as done in pseudo relevance feedback (Rocchio, 1965), and updates itself using those labels.

Semi-supervised learning. Our problem setting can be seen as an *extreme case* of weak supervision: we only use class labels as the (noisy) supervision signal, and nothing else. If we view class labels as "labeled documents", one from each class, and to-be-classified documents as unlabeled documents, then we cast the problem as semi-supervised learning (Zhu, 2006). Self-training is one such technique: a generative classifier is trained using only class labels, and then teaches itself using its own predictions on unlabeled data. If we view class labels as "labeled features", then we expect the classifier to predict a class when a document contains the class label words. For instance, Druck et al. (2008) proposed generalized expectation criteria that uses feature words (class labels) to train a discriminative classifier. Jagarlamudi et al. (2012) and Hingmire and Chakraborti (2014) proposed Seeded LDA to incorporate labeled words/topics into statistical topic modeling. The inferred document-topic mixture probabilities can be used to classify documents.

Zero-shot learning aims to classify visual objects from a new class using only word descriptions of that class (Socher et al., 2013). It first learns visual features and their correspondence with word descriptions, and then constructs a new classifier by composing learned features. Most research on zero-shot learning focuses on image classification, but the same principle applies to text classification as well (Pushp and Srivastava, 2017). Our proposed method constructs a new classifier by composing learned word embeddings in a probabilistic manner. Since the new classifier transfers semantic knowledge in word embedding to topic classification tasks, it is broadly related to **transfer learning** (Pan and Yang, 2010). The main difference is that in transfer learning the information about the new task is in the form of labeled data, not class definition words.

3 Proposed Method

Let a test document x be a sequence of words $(w_1, \cdots, w_j, \cdots)$, and a class topic description y be a sequence of words $d_y = (w_1, \cdots, w_y, \cdots)$. All words are in vocabulary V. We propose a generative approach, where the predictive probability $p(y|x) \propto p(x|y)p(y)$. Generative approaches tends to perform well when training data is scarce, which is the case in our setting.

We assume there exists weak prior knowledge on which classes are popular and which are rare. We can then construct rough estimates $\hat{p}(y)$ using simple heuristics as described in (Schapire et al., 2002). It distributes probability mass q evenly among majority classes, and $1-q$ evenly among minority classes. We treat the most frequent class as the majority class, the rest as minority classes, and $q = 0.7$ in our experiments.

By interpreting class topic description as words, we obtain $\hat{p}(x|y) = p(x|d_y)$. We assume that the d_y expresses a noisy-OR relation of the words it contains (Oniśko et al., 2001). Up to first-order approximation:

$$\begin{aligned} p(x|d_y) &= 1 - \prod_{w_y \in d_y} (1 - p(x|w_y)) \\ &\approx \sum_{w_y \in d_y} p(x|w_y), \end{aligned} \tag{1}$$

where each w_y is a word in the class topic description d_y. Further, we assume that words in document x are conditionally independent given a label word w_y (naïve Bayes assumption):

$$p(x|w_y) = \prod_{w_j \in x} p(w_j|w_y). \tag{2}$$

Combining (1) and (2), the document likelihood is

$$\hat{p}(x|y) = \sum_{w_y \in d_y} \prod_{w_j \in x} p(w_j|w_y). \tag{3}$$

To this end, we need a word association model $p(w_1|w_2), \forall w_1, w_2 \in V$. It can be efficiently learned by word embedding algorithms. The skip-gram algorithm (Mikolov et al., 2013) learns vector representations of words, such that for words w_1, w_2, their vectors $\mathbf{u}_{w_1}, \mathbf{v}_{w_2}$ approximate the conditional probability[1]

$$p(w_1|w_2) = \frac{\exp\left(\mathbf{u}_{w_1}^\top \mathbf{v}_{w_2}\right)}{\sum_{w \in V} \exp\left(\mathbf{u}_w^\top \mathbf{v}_{w_2}\right)}. \tag{4}$$

[1]The two sets of word vectors $\{\mathbf{u}_w : w \in V\}$ and $\{\mathbf{v}_w : w \in V\}$ produced by skip-gram correspond to the input and output parameters of a two-layer neural network. Typically, only the output parameters are used as the "learned word vectors". Here we need both input and output parameters to compute $p(w_1|w_2)$.

Combining (3) with (4), the document likelihood becomes

$$\hat{p}(x|y) = \sum_{w_y \in d_y} \exp\left(\sum_{w_j \in x} \left(\mathbf{u}_{w_j}^\top \mathbf{v}_{w_y} - C_{w_y}\right)\right) ,$$

where $C_{w_y} = \log \sum_{w \in V} \exp\left(\mathbf{u}_w^\top \mathbf{v}_{w_y}\right)$ is independent of document x and only related to label word w_y, therefore can be precomputed and stored to save computation.

Finally, we construct an generative classifier as $\hat{p}(y|x) \propto \hat{p}(x|y)\hat{p}(y)$. We call this method *word embedding naïve Bayes* (WENB).

3.1 Continued Training

The proposed method produces pseudo labels $\hat{p}(y|x_j)$ for unlabeled documents $\{x_j\}_{j=1}^m$. When true labels $\{(x_i, y_i)\}_{i=1}^n$ are available, we can train a new discriminative logistic regression classifier $p_\theta(y|x)$ using both true and pseudo labels (θ is the model parameter):

$$J(\theta) = \sum_{i=1}^{n} \sum_{y \in Y} -\mathbf{1}_{\{y_i = y\}} \log p_\theta(y|x_i) + \lambda \|\theta\|^2 + \mu \sum_{j=1}^{m} \sum_{y \in Y} -\hat{p}(y|x_j) \log p_\theta(y|x_j) . \quad (5)$$

To find the balance of pseudo vs. true labels in (5), we search the hyperparameter μ on a 5-point grid $\{10^{-2}, 10^{-1}, 0.4, 0.7, 1\}$. We expect pseudo labels to have comparable importance as true labels when n is small (fine granularity for $\mu \in [10^{-1}, 1]$), and their importance will diminish as n gets large ($\mu = 10^{-2}$). μ is automatically selected such that it gives the best 5-fold cross-validation accuracy on n true labels.

4 Experiments

We compare a variety of methods on six topic classification data sets. The goals are (1) to study the best classification performance achievable using class labels only, and (2) to estimate the equivalent amount of true labels needed to achieve the same warm-start performance.

4.1 Compared Methods

Retrieval-based methods. We use language modeling retrieval function with Dirichlet smoothing (Zhai and Lafferty, 2001) ($\mu = 2500$) to match a document to class labels (**IR**). The top 10 results are then used as pseudo-labeled documents to retrain three classifiers: **IR+Roc**: a Rocchio classifier ($\alpha = 1, \beta = 0.5, \gamma = 0$); **IR+NB**: a multinomial naive Bayes classifier (Laplace smoothing, $\alpha = 0.01$); **IR+LR** a logistic regression classifier (linear kernel, $C = 1$).

Semi-supervised methods. **ST-0**: the initial self-training classifier using class labels as "training documents" (multinomial naïve Bayes, Laplace smoothing $\alpha = 0.01$). **ST-1**: ST-0 retrained on 10 most confident documents predicted by itself. **GE**: a logistic regression classifier trained using generalized expectation criteria (Druck et al., 2008). Class labels are used as labeled features. **sLDA**: a supervised topic model trained using seeded LDA (Jagarlamudi et al., 2012). Besides k seeded topics (k is the number of classes), we use an extra topic to account for other content in the corpus.

Word embedding-based methods. **Cosine**: a centroid-based classifier, where class definitions and documents are represented as average of word vectors. **WENB**: The proposed method (Section 3). **WENB+LR**: a logistic regression classifier trained only on pseudo labels produced by WENB (Section 3.1, $n = 0$).

For general domain tasks, we take raw text from English Wikipedia, English news crawl (WMT, 2014), and 1 billion word news corpus (Chelba et al., 2013) to train word vectors. For medical domain tasks, we take raw text from MEDLINE abstracts (NLM, 2018) to train word vectors. We find 50-dimensional skip-gram word vectors perform reasonably well in the experiments.

4.2 Data Sets

We consider six topic classification data sets with different document lengths and application domains. Table 1 summarizes basic statistics of these data sets. Table 4 and 5 in the appendix show actual class labels used in each data set.

Data set	Avg word/doc	# classes	# docs
Wiki Titles	3.1 (1.1)	15	30,000
News Titles	6.7 (9.5)	4	422,937
Y Questions	5.0 (2.6)	10	1,460,000
20 News	101.6 (438.5)	20	18,846
Reuters	76.5 (117.3)	10	8,246
Med WSD	202.8 (46.6)	2/task	190/task

Table 1: Statistics of topic classification data sets. Numbers in column "Avg word/doc" are "mean (standard deviation)".

	Wiki Titles	News Titles	Y Questions	20 News	Reuters	Med WSD
Majority guess	.83	13.26	1.82	.48	6.47	34.20
IR	3.14 (.25)	14.20 (.06)	6.15 (.06)	19.57 (.95)	8.37 (.55)	52.99 (.64)
IR+Roc	2.93 (.24)	14.20 (.06)	8.35 (1.12)	25.09 (.93)	19.33 (1.87)	59.89 (.54)
IR+NB	5.44 (.53)	32.98 (2.13)	14.45 (.45)	30.45 (1.46)	**62.59 (2.43)**	**82.12 (.41)**
IR+LR	3.26 (.30)	13.44 (.10)	7.38 (2.08)	**34.76 (1.50)**	6.48 (.07)	68.35 (.38)
ST-0	3.16 (.32)	16.03 (.16)	6.15 (.02)	19.49 (.98)	6.79 (.17)	69.11 (.26)
ST-1	5.62 (.29)	24.34 (.36)	10.02 (.49)	22.91 (1.29)	**55.77 (1.62)**	**82.97 (.56)**
GE	9.55 (.90)	14.54 (.08)	31.72 (.05)	**48.71 (.41)**	21.65 (27.36)	62.63 (.37)
sLDA	7.07 (0.97)	51.16 (8.10)	40.98 (2.61)	24.80 (4.98)	30.61 (4.80)	69.81 (1.09)
Cosine	**27.67 (.59)**	33.49 (.11)	31.16 (.03)	26.19 (.75)	6.56 (.16)	32.65 (.19)
WENB	**26.70 (.48)**	**63.02 (.10)**	**44.89 (.06)**	32.23 (.48)	34.99 (1.99)	68.27 (.20)
WENB+LR	24.88 (.39)	**63.76 (.11)**	**45.69 (.09)**	30.57 (.71)	32.04 (1.44)	62.57 (.19)

Table 2: Macro-averaged F_1 (%) of compared methods on different data sets. The numbers are "mean (standard deviation)" of 5-fold cross validation. Top two numbers in each column are highlighted in **boldface**.

Data set	# of labels
Wiki Titles	1500
News Titles	200
Y Questions	1500-2000
20 News	100-200
Reuters	100-200
Med WSD	20/task $\times$ 198 tasks

Table 3: Number of true labels needed for a logistic regression classifier to achieve the same performance as "WENB+LR".

Three short text data sets are (1) **Wiki Titles**: Wikipedia article titles sampled from 15 main categories (Wikipedia Main Topic). (2) **News Titles**: The UCI news title data set (Lichman, 2013). (3) **Y Questions**: User-posted questions in Yahoo Answers (Yahoo Language Data, 2007).

Three long text data sets are (1) **20 News**: The well-known 20 newsgroup data set. (2) **Reuters**. The Reuters-21578 data set (Lewis). We take the articles from the 10 largest topics. (3) **Med WSD**: The MeSH word sense disambiguation (WSD) data set (Jimeno-Yepes et al., 2011).

Each WSD task aims to tell the sense (meaning) of an ambiguous term in a MEDLINE abstract. For instance, the term "cold" may refer to *Low Temperature, Common Cold*, or *Chronic Obstructive Lung Disease*, depending on its context. These senses are used as the class labels. We use 198 ambiguous words with at least 100 labeled abstracts in the data set, and report the average statistics over 198 independent classification tasks.

Although no true labels are used for training, some methods require unlabeled data for retrieval, pseudo-labeling, and re-training. We split unlabeled data into 5 folds, using 4 folds to "train" a classifier and 1 fold for test. We use macro-averaged F_1 as the performance metric because not all data sets have a balanced class distribution.

4.3 Results and Discussion

Label savings. Table 2 shows that overall, class labels can train text classifiers remarkably better than majority guess. This is no small feat considering that the classifier has not seen any labeled documents yet. Such performance gain essentially comes "for free", as any text classification task has to start by defining classes. In Table 3, we report the number of true labels needed for a logistic regression model to achieve the same performance as WENB+LR. The most significant savings happen on short documents: class labels are equivalent to hundreds to thousands of labeled documents at the beginning of the training process.

Effect of document length. On short documents (Wiki Titles, News Titles, Y Questions), leveraging unlabeled data does not help with most semi-supervised methods due to severe vocabulary mismatch. The proposed methods (WENB and WENB+LR) show robust performance, because pretrained word vectors can capture semantic similarity even without any word overlap between a class label and a document. This prior knowledge is essential when documents are short. On long documents (20 News, Reuters, Med WSD), leveraging unlabeled data helps, since long documents have richer content and are more likely to contain not only label words themselves, but also other topic-specific words. Retrieval-based and semi-supervised methods are able to learn these words by exploiting intra-document word co-occurrences.

Performance of other methods. Learning from class labels themselves provides very limited help (IR and ST-0). Using class labels as search queries and labeled documents are closely related: IR and ST-0 perform similarly; so do IR+NB and ST-1. When using class labels as search queries,

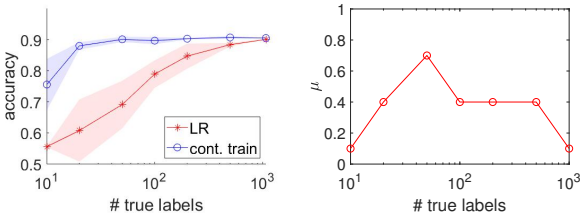

Figure 2: Continued training behavior: *Atheism* vs. *Autos*. Colored band: ± 1 standard deviation.

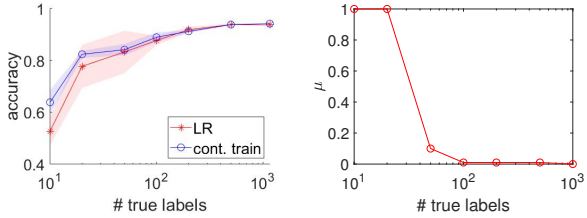

Figure 3: Continued training behavior: *Medical* vs. *Mideast*. Colored band: ± 1 standard deviation.

re-ranking (IR+Roc) is less useful than training classifiers (IR+NB and IR+LR). After initial retrieval, training a naïve Bayes classifier is almost always better than a logistic regression classifier (IR+NB vs. IR+LR), demonstrating the power of generative models when supervision signal is sparse. Using class labels as labeled features (GE and sLDA) performs well occasionally (GE on 20 News; sLDA on Y Questions), but not consistently. The Cosine method performs well only on Wiki Titles, the shortest documents, because without supervision, representing a long document as an average of word vectors causes significant information loss. Finally, it is encouraging to see WENB+LR sometimes outperform WENB, as WENB+LR is much smaller than WENB+LR in terms of model size.

4.4 Continued Training and Error Analysis

Figure 2 and 3 compare logistic regression classifiers trained with and without pseudo labels generated by WENB. Note that the classifier trained with pseudo labels (cont. train) has a much lower performance variance than the logistic regression classifier trained only on true labels (LR).

The warm-started classifier can serve as a good starting point for further training. Figure 2 shows a salient warm-start effect on a balanced binary classification task in 20 News. The weight μ of pseudo labels increases when true labels are few (initial classifier as an informative prior). As expected, μ decreases when true labels become abundant.

Figure 3 shows another binary classification task in 20 News where the warm-start effect is limited. Correspondingly, μ quickly diminishes as more true labels are available. With 100 or more true labels, pseudo labels have a negligible weight ($\mu = 10^{-2}$). In machine learning terms, these pseudo labels specify an incorrect prior that the model should quickly forget, so that it will not hinder the overall learning process.

A closer investigation reveals that the word vector for *mideast* (the class label of one topic in Figure 3) is not well-trained. This is because in general text corpus, the word *mideast* is rather infrequent compared to commonly used alternatives, such as *middle_east*. The word vector of *mideast* is surrounded by other infrequent words or misspellings (such as *hizballah*, *jubeir*, *saudis*, *isreal*) as opposed to more frequent and relevant ones (such as *israel*, *israeli*, *saudi*, *arab*). Since WENB uses the semantic knowledge in word vectors to infer pseudo labels, the quality of class label word vectors will affect the pseudo label accuracy.

5 Conclusion and Future Directions

We studied the problem of training topic classifiers using only class labels. Experiments on six data sets show that class labels can save a significant amount of labeled examples in the beginning. Retrieval-based and semi-supervised methods tend to perform better on long documents, while the proposed method performs better on short documents.

This study opens up many interesting avenues for future work. First, we introduce a new perspective on text classification: can we build a text classifier by just providing a short description of each class? This is a more challenging (but more user-friendly) setup than standard supervised classification. Second, future work can investigate tasks such as sentiment and emotion classification, which are more challenging than topic classification tasks. Third, the two approaches – leveraging unlabeled data (retrieval-based and semi-supervised methods) and leveraging pretrained models (the proposed method) – could be combined to give robust performance on both short and long documents. Finally, we can invite users into the training loop: in addition to labeling documents, users can also revise the class definitions to improve the classifier.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was in part supported by the National Library of Medicine under grant number 2R01LM010681-05. Qiaozhu Mei's work was supported in part by the National Science Foundation under grant numbers 1633370 and 1620319. Yue Wang would like to thank the support of the Eleanor M. and Frederick G. Kilgour Research Grant Award by the UNC-CH School of Information and Library Science.

References

2014. WMT 2014 English News Crawl. http://www.statmt.org/wmt14/training-monolingual-news-crawl.

Ricardo Baeza-Yates, Berthier de Araújo Neto Ribeiro, et al. 2011. *8.3.2 Naive Text Classification*, chapter 8. New York: ACM Press; Harlow, England: Addison-Wesley,.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

W Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Retrieval Models*, chapter 7. Addison-Wesley Reading.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602. ACM.

Swapnil Hingmire and Sutanu Chakraborti. 2014. Topic labeled text classification: a weakly supervised approach. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 385–394. ACM.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.

Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223.

David D. Lewis. Reuters-21578. http://www.daviddlewis.com/resources/testcollections/reuters21578.

M Lichman. 2013. Uci machine learning repository [http://archive. ics. uci. edu/ml]. university of california, school of information and computer science. *Irvine, CA*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

NLM. 2018. MEDLINE/PubMed Data. ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline.

Agnieszka Oniśko, Marek J Druzdzel, and Hanna Wasyluk. 2001. Learning bayesian network parameters from small data sets: Application of noisy-or gates. *International Journal of Approximate Reasoning*, 27(2):165–182.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*.

J. J. Rocchio. 1965. Relevance feedback in information retrieval, report no. *ISR-9 to the National Science Foundation, The Computation Laboratory of Harvard University, to appear August*.

Robert E Schapire, Marie Rochery, Mazin Rahim, and Narendra Gupta. 2002. Incorporating prior knowledge into boosting. In *ICML*, volume 2, pages 538–545.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Wikipedia Main Topic. Category:Main topic classifications. https://en.wikipedia.org/wiki/Category:Main_topic_classifications.

Yahoo Language Data. 2007. Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part). https://webscope.sandbox.yahoo.com/catalog.php?datatype=l.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of 24th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342.

Xiaojin Zhu. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3):4.

A Class Labels Used in Each Data Set

Data set	Class labels
Wiki Titles	*Arts, Games, Geography, Health, History, Industry, Law, Life, Mathematics, Matter, Nature, People, Religion, Science/Technology, Society*
News Titles	*Business, Technology, Entertainment, Health*
Y-Questions	*Society/Culture, Science/Mathematics, Health, Education/Reference, Computers/Internet, Sports, Business/Finance, Entertainment/Music, Family Relationships, Politics/Government*
20 News	*Atheism, Graphics, Microsoft, IBM, Mac,Windows, Sale, Autos, Baseball, Motorcycles, Hockey, Encrypt, Electronics, Medical, Space, Christian Guns, Mideast, Politics, Religion*
Reuters	*Earnings/Forecasts, Mergers/Acquisitions, Crude Oil, Trade, Foreign Exchange, Interest Rates, Money Supply, Shipping, Sugar, Coffee*

Table 4: Class labels in 5 topic classification data sets.

Task (ambiguous term)	Class labels (senses)
AA	*Amino Acids, Alcoholics Anonymous*
ADA	*Adenosine Deaminase, American Dental Association*
ADH	*Alcohol dehydrogenase, Argipressin*
ADP	*Adenosine Diphosphate, Automatic Data Processing*
Adrenal	*Adrenal Glands, Epinephrine*
Ala	*Alanine, Alpha-Linolenic Acid, Aminolevulinic Acid*
ALS	*Antilymphocyte Serum, Amyotrophic Lateral Sclerosis*
ANA	*American Nurses' Association, Antibodies, Antinuclear*
Arteriovenous Anastomoses	*Arteriovenous anastomosis procedure, Structure of anatomic-arteriovenous anastomosis*
Astragalus	*Talus, Astragalus Plant*
B-Cell Leukemia	*B-Cell Leukemia, Chronic Lymphocytic Leukemia*
BAT	*Chiroptera, Brown Fat*
BLM	*Bloom Syndrome, Bleomycin*
Borrelia	*Lyme Disease, Borrelia bacteria*
BPD	*Bronchopulmonary Dysplasia, Borderline Personality-Disorder*
BR	*Brazil, Bromides*
Brucella abortus	*Brucella abortus infection, Brucella abortus bacterium*
BSA	*Body Surface Area, Bovine Serum Albumin*
BSE	*Bovine Spongiform-Encephalopathy, Breast Self-Examination*
Ca	*Hippocampus (Brain), Calcium, California, Canada*

Table 5: The first 20 ambiguous terms/tasks in Med WSD data set.