STIF: Semi-Supervised Taxonomy Induction using Term Embeddings and Clustering

Maryam Mousavi
Arizona State University
Tempe, Arizona, USA
maryammousavi@asu.edu

Elena Steiner
Arizona State University
Tempe, Arizona, USA
esteine2@asu.edu

Steven Corman*
Arizona State University
Tempe, Arizona, USA
steve.corman@asu.edu

Scott Ruston
Arizona State University
Tempe, Arizona, USA
scott.ruston@asu.edu

Dylan Weber
Artis Research
St. Michael, Maryland, USA
djweber3@asu.edu

Hasan Davulcu
Arizona State University
Tempe, Arizona, USA
hdavulcu@asu.edu

ABSTRACT

In this paper, we developed a semi-supervised taxonomy induction framework using term embedding and clustering methods for a blog corpus comprising 145,000 posts from 650 Ukraine-related blog domains dated between 2010-2020. We extracted 32,429 noun phrases (NPs) and proceeded to split these NPs into a pair of categories: General/Ambiguous phrases, which might appear under any topic vs. Topical/Non-Ambiguous phrases, which pertain to a topic's specifics. We used term representation and clustering methods to partition the topical/non-ambiguous phrases into 90 groups using the Silhouette method. Next, a team of 10 communications scientists analyzed the NP clusters and inducted a two-level taxonomy alongside its codebook. Upon achieving intercoder reliability of 94%, coders proceeded to map all topical/non-ambiguous phrases into a gold-standard taxonomy. We evaluated a range of term representation and clustering methods using extrinsic and intrinsic measures. We determined that GloVe embeddings with K-Means achieved the highest performance (i.e. 74% purity) for this real-world dataset.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**; *Language models.*

KEYWORDS

Taxonomy induction, Text categorization, Topic detection

ACM Reference Format:
Maryam Mousavi, Elena Steiner, Steven Corman, Scott Ruston, Dylan Weber, and Hasan Davulcu. 2021. STIF: Semi-Supervised Taxonomy Induction using Term Embeddings and Clustering. In *Sanya '21: International Conference on Natural Language Processing and Information Retrieval December 17–20, 2021, Sanya, Hance Provenience, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3508230.3508247

*This research was supported by a grant from the U.S. Office of Naval Research (N00014-18-1-2692)

NLPIR 2021, December 17–20, 2021, Sanya, Hance Provenience, China

ACM ISBN 978-1-4503-8735-4/21/12…$15.00
https://doi.org/10.1145/3508230.3508247

1 INTRODUCTION

A taxonomy is a hierarchical categorization scheme that is an effective way to organize and classify knowledge. Taxonomies play a critical role in boosting the performance of a vast variety of downstream natural language processing (NLP) and information retrieval (IR) applications such as question answering [8, 69], web indexing and search [8, 66], personalized recommendation [18, 71], information management [9, 68] and e-commerce [10, 56]. Many taxonomies have been crafted to organize terms and documents in broad and specific domains such as WordNet [35], Wikidata [61], MeSH [28], and Amazon's Product Taxonomy [23].

The majority of the existing taxonomies are induced manually with the assistance of domain experts or crowdsourcing platforms. Manual induction assures coherence, precision and incorporation of domain expert knowledge into the resulting taxonomy. Although a manually created taxonomy might be incomplete due to the limitations of domain experts' knowledge, it yields high accuracy taxonomies upon resolving term-to-topic mapping conflicts among experts iteratively and through the creation of a codebook documenting coders' decision making and consensus processes [63].

Today's high-volume and fast-streaming applications, such as large volume historical and real-time news, blog, micro-blog applications make manual taxonomy construction and maintenance quite challenging, since without enabling tools and frameworks it is a time-consuming and labor-intensive process. In this paper, we present and evaluate a Semi-supervised Taxonomy Induction Framework (STIF) that yields complete and highly accurate results while reducing the labor needed according to one moderately sized real-world use case.

Our use case comprises 145,000 blog posts from 650 Ukraine related blog domains dated between 2010 and 2020 enveloping the Ukrainian crisis. Ukrainian crises feature the 2013–2014 Euromaidan protests, its emergent social movement for integration of Ukraine into the European Union, followed by the annexation of Crimea by the Russian Federation – which took place between 20 February 2014 and 19 March 2014 – and, the ongoing civil war in the Donbass region since 6 April 2014 until the present day.

Figure 1 presents of the flowchart of the Semi-supervised Taxonomy

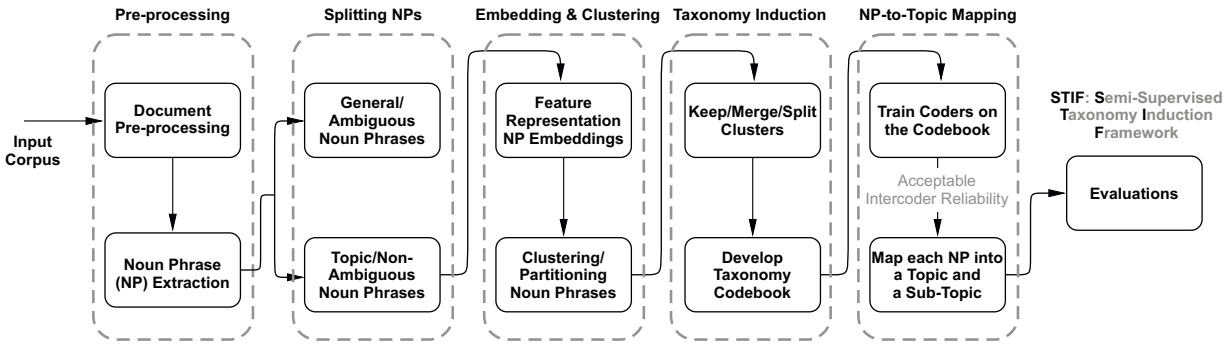

Figure 1: Semi-supervised Taxonomy Induction Framework (STIF)

Induction Framework (STIF) where initially we use noun phrase extraction[24, 31] to extract 32,429 noun phrases (NPs) from the above corpus and proceed to split these NPs into a pair of categories: general/ambiguous phrases, such as "time", "part", "way", "week", "things", etc., which go together with multiple topics vs. topical/non-ambiguous noun phrases which would relate to a single topic in the taxonomy. Next, we use term representation and clustering methods to partition the topical/non-ambiguous phrases. Following that, a team of 10 communications scientists analyze the term frequency-sorted NPs in each of these groups making decisions as a panel about either labeling and keeping a group as a topic as is, splitting a group into a few distinct topics or merging a group with another existing topic. For the Ukraine blog corpus, coding team's joint decision making and consensus processes yielded a hierarchical taxonomy featuring 15 top-level and 85 sub-category level topics alongside a detailed codebook. We present the induced two-level taxonomy in Table 1 below. Upon achieving an intercoder reliability of 94% with Krippendorff's alpha [12] for this mapping task, our coding team proceeded to map all 12,569 topical/non-ambiguous phrases into the gold-standard taxonomy using the codebook developed for this task. We are sharing a link to the codebook guiding the term to topic mappings at following link [1]. Finally, we evaluated a wide range of term representation and clustering methods using intrinsic and extrinsic measures such as Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI) and purity. We determined that GloVe [43] embeddings with K-Means achieved the highest performance for detecting and organizing the topics in this moderately sized real-world corpus.

Although term embedding and clustering methods are well established for taxonomy induction [2, 70], the novel contributions of this paper include the design of the Semi-supervised Taxonomy Induction Framework (STIF) featuring a text processing pipeline that initially splits extracted noun phrases into general/ambiguous vs topical/non-ambiguous ones in order to facilitate taxonomy induction and term-to-topic mapping at high accuracies while reducing the human coding labor needed (i.e. by up to 74%) due to the high coherence, contrast, and purity of its intermediate topic clusters.

Rest of the paper is organized as follows. In Section 2, we present related works aimed for topic taxonomy induction. In Section 3, we present the technical details of the proposed STIF framework. In Section 4, we present the dataset, experimental evaluations, results and findings. Section 5, concludes the paper and discusses future work.

2 RELATED WORK

Conventional taxonomy induction methodologies rely on pattern discovery based approaches. These algorithms perform taxonomy induction by extracting the hyponym (entity) - hypernym (class) term pairs from the input documents and organizing them into a hierarchical tree or directed acyclic graph structure [15]. Hyponym-hyponym term relation extraction runs on predefined textual patterns, such as "is-a" relationship extraction, to acquire child-parent pairs [1, 16, 21]. Researchers incorporated various other lexical patterns [38, 39] and developed bootstrapping methods to automate the process of pattern discovery [37, 55]. These pattern-based methods yield high precision due to their specificity, but yield low recall due to the diversity and complexity of language patterns and underlying relationships which they rely on.

Distributional methods mitigate the problem of "dictionary learning" and sparsity by representing words in a low-dimensional vector space, by computing "semantic similarity" of the words' representations and, inferring the hyponym-hypernym relationships [14, 33, 65]. The term pairs observed in a hypernym relation can be turned into a taxonomy by using graph-based methods presented in [25, 40, 59]. Aforementioned algorithms lack completeness since they consider each keyword individually and independently from others using brittle patterns, instead of taking into account various types of their local and global contextual relationships.

Taxonomy learning can be modeled as a clustering problem where similar terms that are grouped together may share the same hypernym [63]. Both word2vec [34] and GloVe [43] enable us to represent a word in the form of a vector (often called embedding). They are the two most popular algorithms for word embeddings that bring out the semantic similarity of words that captures different facets of the meaning of a word by taking into account various types of local and global contextual correlations. Using the term embeddings, clustering is performed to group terms into various topics [62, 67].

[1] https://rb.gy/gplfek

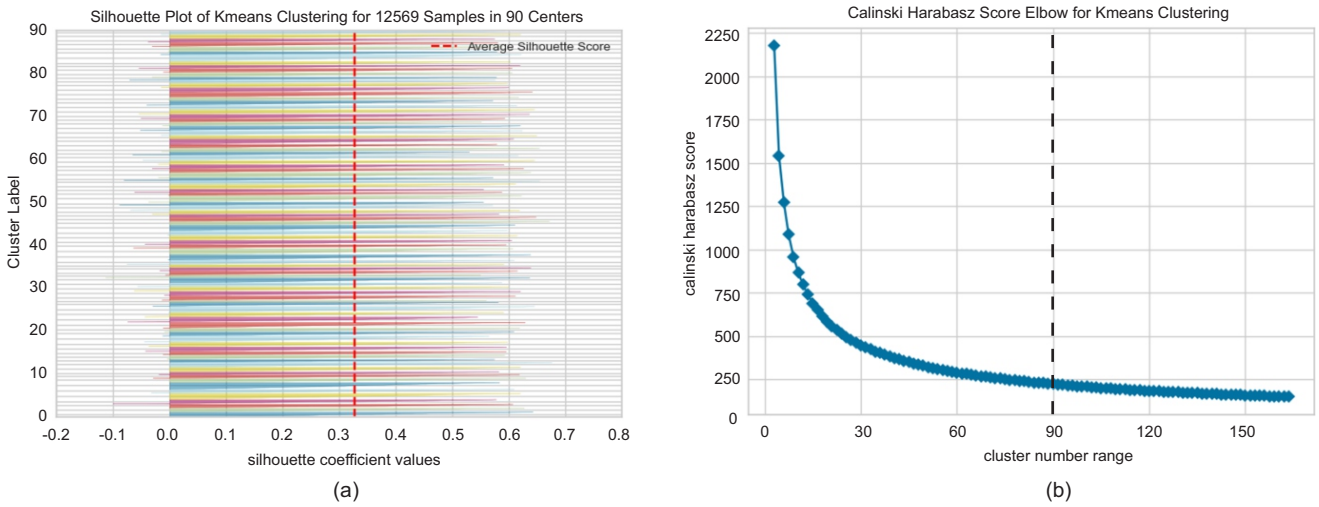

Figure 2: Elbow and Silhouette Charts indicating K=90 as the Close-to-Optimal Clusters

A recent method, TaxoGen [70] applies an adaptive spherical clustering on a local-embedding of the terms to identify fine-grained topics. They reported relation accuracy of 52% and term coherency of 59% in their experimental evaluations with 94,476 abstracts from signal processing area publications. Another recent method [2] inducted a business domain taxonomy by clustering the terms extracted from an annual reports corpus comprising 20,040 documents. They benchmarked the validity of the braching structure of their induced taxonomy qualitatively against the official Chinese National Equities Exchange and Quotations(NEEQ) classification guide via human evaluations. Bhardwaj et al. [4] utilized term embedding and clustering techniques to induce a taxonomy in order to extract insights for improved decision-making. They utilized maintenance records collected from in-field operations and a standard industrial taxonomy. They report the Adjusted Rank Index (ARI) [44] as a similarity metric between their induced taxonomy and the industrial one in two experimental settings. In the first setting, failure mechanism taxonomy, the authors aim to identify clusters of maintenance records that describe maintenance events caused by similar failure mechanisms. In the second setting, on basis of mud-pump equipment taxonomy, they aim to identify the clusters of maintenance records which describe activities concerned with constituent components. They reported ARI values of 0.34 for the first setting and 0.42 for the second.

Hierarchical topic modeling algorithms [5, 36] organize terms to form a taxonomy of topics, where each topic is represented by a word distribution. Automatically induced taxonomies are prone to error due to multiple challenges such as word sense ambiguities, idiomatic expressions, parsing errors and uninformative extractions [27, 63]. It is also essential to incorporate domain knowledge in domain-specific corpora to ensure accurate topic separation, completeness and high purity. However incorporation of domain-specific knowledge remains challenging in unsupervised and automated approaches leading to taxonomies that are noisy and untrustworthy.

Another taxonomy related task that researchers focused on is ongoing maintenance of an existing taxonomy by expanding, completing, and enhancing it. In HiExpan [54] and CoRel [19], the problem of seed-guided topical taxonomy construction has been explored. HiExpan method expands an existing taxonomy vertically and horizontally by adding sibling nodes using the set expansion algorithm and analogical reasoning to capture additional relationships between parent-child pairs [53]. CoRel method [19] aims to complete a seed taxonomy in depth and breadth by utilizing its modules of relation transferring, which learns the user-expressed relationships in seed parent-child pairs and transfers them along multiple paths to expand the taxonomy, as well as concept learning, that finds discriminative topical clusters for each concept during the process of jointly embedding concepts and words.

There are also semi-supervised learning methods, such as our STIF, developed for taxonomy induction [25, 26]. These methods train a classifier based on either a sampling of hyponym-hypernym pairs [29, 52] or by incorporating other syntactic contextual information [32, 66]. Next, they use a classifier to extract additional term pairs for incorporation into the hyponym-hypernym relationships. Recent techniques, such as [33], use word embeddings to curate the training dataset and then learn a classifier. Semi-supervised methods provide more control and supervision over the learned taxonomy. They guide the learning process to follow a more precisely formulated and step-by-step verifiable pathway. They also provide opportunities for incorporating domain experts' knowledge. In this work, we demonstrate an effective semi-supervised topic taxonomy induction framework which incorporates domain experts' knowledge and decision making judiciously and assiduously on a moderately sized real-world dataset in order to induce a high quality taxonomy. STIF also provide assistance with term-to-topic mappings due to the high purity and other high quality characteristics (i.e. Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI)) of its intermediate results.

3 FRAMEWORK OVERVIEW

An overview of the STIF framework is illustrated in Figure 1 and we will describe the details of each block in following subsections.

3.1 Document Pre-processing

This unit applies a pair of pre-processing tasks on the input corpus: (1) Document cleaning, such as the removal of HTML tags, stop words, URL's, followed by tokenization and lemmatized keywords using the *nltk* package. We also perform Noun Phrase (NP) extraction by using the *Textblob* Library's ConllExtractor extractor. For our Ukraine blogs corpus, this step yields 32,429 noun phrases occurring with a minimum frequency of 50.

3.2 Splitting Noun Phrases

The next block aims to split the extracted noun phrases into two categories: (1) General/Ambiguous Noun Phrases such as "part", "time", "last years", etc., which might correspond to more than one topic. (2) Topical/Non-Ambiguous noun phrases that pertain to a specific topic only; including its "whats" (i.e. facts, aspects and specifics) and "whys" that make a particular topic a subject of interest. A comprehensive list of NP's in both categories (i.e. 19,860 in General/Ambiguous category and 12,569 in Topical/Non-Ambiguous category) is available online at following link [2]. Initially NP splitting is done manually with the help of a team of communication scientists. We also trained a term embedding based classifier, using Glove embeddings [43] which yielded a classifier with F1 score of 71% for splitting these noun phrases automatically. Table 2 shows the final noun phrase frequency statistics obtained from our Ukraine blogs corpus.

3.3 Term Embedding and Clustering

This block applies modern algorithms for term embeddings on the topical/non-ambiguous noun phrases. The algorithms that we experiment with include (1) traditional TF-IDF (as the baseline), (2) a local word vector embedding algorithm, Word2Vec [34] and (3) a global vector embedding algorithm, Glove [43].

TF-IDF [46], stands for term frequency-inverse document frequency, is a term-document mapping representation which provides a practical approach for topic analysis in a text corpus. This approach assigns a score (weight) to each word such that the words with low frequency gets higher scores and lower scores are assigned to more frequently occurring words such as "a", 'the", "as" and etc. We applied TF-IDF vectorizer from scikit-learn python package [41] as our baseline feature representation model. This vectorizer ranks the importance of the terms in a document based on their distribution throughout the entire corpus.

Word2Vec embedding technique, introduced in [34], is a neural network-based model which calculates a distributed representation of words and phrases in a vector space by inspecting the local information and surrounding context of words. In our work, a pre-trained bag of word (CBOW) version of word2vec has been fine-tuned on our corpus using the Gensim package [45]. This pre-trained model itself relies on a 300-dimensional term representations that was trained on a large Google News dataset comprising about 100 billion words.

GloVe [42], global vector representation is an unsupervised method which provides a vector representation of words and phrases based on aggregated global word-word co-occurrence statistics derived from a corpus. In this work we fine tuned and used a 25-dimensional pre-trained GloVe model, which was initially trained on 2 billion tweets from Twitter platform comprising 27 billion tokens, and 1.2 million vocabulary.

After obtaining the term embeddings, we experimented with three different clustering algorithms to partition the NPs into K-clusters. The number of clusters is identified by using the Elbow [22] and the Silhouette [50] methods – both indicating K=90 as the close-to-optimal number of clusters based on the charts shown in Figure 2. A Silhouette plot is a graphical representation indicating that a data point is well matched to its own cluster and poorly matched to its neighboring clusters [11, 48]. In the Silhouette plot, the ideal number of clusters is identified when majority of clusters have a silhouette score above their average silhouette score and the thickness of the silhouette plots for the clusters do not have wide fluctuations. In the Calinksi Harabasz Score Elbow plot [7, 58], the score is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. The "elbow" or "knee of a curve" is selected as the ideal cutoff point indicating that additional clusters do not improve the quality of the fit.

We designed experiments to find the best combination of feature representation and clustering method that would yield the highest quality topic taxonomy agreeing with the gold standard taxonomy and its term-to-topic mappings. The clustering methods that we evaluated include: (1) K-means, (2) K-medoids and (3) Agglomerative (a bottom-up hierarchical clustering technique) which are all commonly used in text analysis.

We used K-means and K-mediods algorithm implementations of scikit-learn package [41], with Euclidean distance measure and K (number of clusters) set to 90. K-Medoids is known to be more robust and less sensitive to outliers compared to the K-means. The third clustering algorithm applied is "agglomerative" from the scikit-learn package [41]. This clustering method works in a bottom-up fashion where initially each object is considered a single member cluster. Most similar clusters are progressively merged together thereby reducing the total number of clusters until K=90 clusters remain. In our experiments we used the Euclidean distance measure with Ward's method [64], in order to minimize the within-cluster variance.

3.4 Taxonomy Induction

In the taxonomy induction block of the STIF framework, a team of 10 social scientists examined the 90 frequency sorted resultant NP clusters and decided on one of these three actions for each cluster: (1) Keep the cluster "as is" as a topic and label the topic by selecting one of the cluster's most frequently occurring terms; (2) Split the cluster into a few distinct clusters if there are more than one topics within it; (3) Merge the cluster with another one if they share a topic together. The result of this effort is the induction of a gold-standard topic taxonomy. The gold-standard taxonomy induced from the Ukraine blog corpus is presented below in Table 1. The codebook

[2] https://rb.gy/bttrbd

Table 1: Gold-standard Hierarchical Taxonomy for the Ukraine Blogs (2010 - 2020)

Topic	Sub-Topics
Agreement	Dialogue, Diplomacy, Negotiations, Peace, Cooperation
Conflict	Immigration, Threat, Dissent, Protest, Violence, Terrorism, Insurgency
Defense	Weapons, Warfare, Intelligence, Coast Guard, Army, Navy, Air Force, Bases, Drones, Nuclear, Proxy-Wars, Missiles, Veterans
Economy	Banks, Debt, Energy, Food, Housing, Labor, Metals, Oligarchs, Recession, Stock Market, Taxes, Trade, Investments, Sanctions
Science and Education	History, Sciences, Education
Government	Elections, Border, Impeachment, Politics, NGO
Health	COVID-19, Vaccine, Diseases
Identity	Religious, Racial, Ethnic, Gender, Generational, Minorities
Ideology	National, Political, Conspiracy
Infrastructure	Housing, Transportation
Justice System	Crime, Corruption, Narcotics, Policing, Prisons, Genocide, Abortion, Gun Debate
Media	Propaganda/Disinformation, Mainstream, Alternative, Social Media
Morality	eMFD Vices and Virtues[a]
Sports	Olympics, World Cup, NHL, Soccer, Hockey
Planet	Environment, Climate Change, Water, Disasters, Animals

[a] we used the provided list in [17]

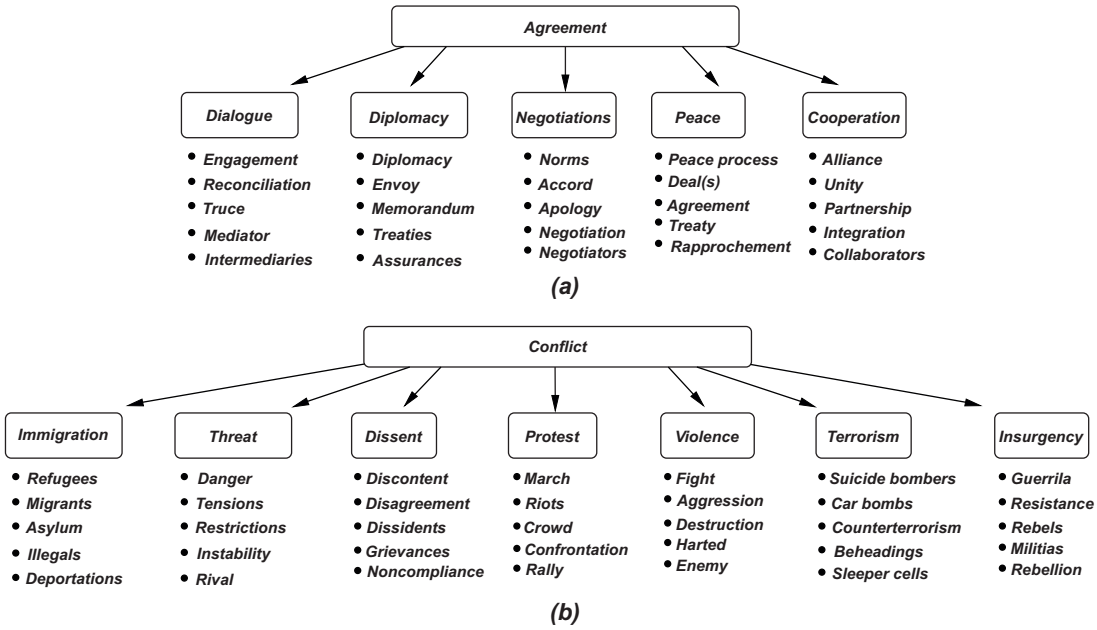

Figure 3: Sample topics and their sub-topics

providing term-to-topic mapping instructions can also be found at this link below[3].

3.5 Noun Phrase to Topic Mapping

In the next block 12,569 topical/non-ambiguous NPs are mapped to the topic taxonomy according to the instructions and examples provided in the gold-standard taxonomy's codebook. A team of 10 communication scientists were trained on the codebook and took part in multiple rounds of individually mapping 500 random NPs into the gold-standard taxonomy. Krippendorff's alpha (α) [12] is a reliability coefficient developed to measure the inter-coder agreements drawing distinctions among typically unstructured phenomena or assign computable values to them with an application of the same coding scheme. The minimum acceptable alpha coefficient should be chosen according to the importance of the conclusions to

[3] https://rb.gy/gplfek

Table 2: Noun Phrase Statistics

Noun phrase category	Count
Total noun phrases	32,429
General/Ambiguous noun phrases	19,860
Topical/Non-Ambiguous noun phrases	12,569

be drawn from imperfect data. When the costs of mistaken conclusions are high, the minimum alpha needs to be set high as well. In the absence of knowledge of the risks of drawing false conclusions from unreliable data, social scientists commonly rely on coding data with reliabilities $\alpha \geq 0.8$. Upon achieving a Krippendorff's alpha intercoder reliability of 94% accuracy for mapping the topical noun phrases into the topic taxonomy, each of the NPs extracted from our corpus was assigned to two coders for mapping. If there was a disagreement between any pair of coders, then the coders communicated among each other to resolve such disagreements by incorporating the assistance of a third coder as a tie-breaker. Generic topics of large-scale document collections can often be divided into more specific subtopics. Topic hierarchies provide a model for such topic relation structure. The criteria for division between topics and subtopics is that sub-topics occur in the presence of the main topic. Usually human assessment [3] is used to validate the induced topic hierarchies. A sample list of the noun phrase to topic mappings for a pair of topics (i.e. conflict and agreement) and their sub-topics are shown in Figure 3.

4 USE CASE CORPUS AND EXPERIMENTAL EVALUATIONS

This section presents the Ukraine blogs corpus characteristics, our experimental designs and evaluations. As detailed in the framework overview section above, a baseline and two other popular term embedding methods alongside three different clustering algorithms were paired and their performance were evaluated using a set of intrinsic and extrinsic quality measures.

4.1 Ukraine Blogs Corpus

To evaluate our proposed taxonomy induction framework, a moderately sized real-world document corpus compromising 144,448 blog posts from 650 Ukraine related blog domains dated between 2010 and 2020 is utilized. In order to crawl the data, a multi-threaded, scalable and resilient-to-noise crawler described at [49] was used. The crawler was setup and used by initializing it with a seed knowledge containing a list of keywords and blog domains of interest. The keywords that are used in the seed knowledge list can be split into three categories. First category contains sentiments about political and geopolitical arguments related to the Russian hybrid war in the Crimea Peninsula and eastern Ukraine. These keywords include "Russia", "Kremlin", "totalitarian", "ethnic", "tension", "discrimination", etc. The second category of keywords are about key events around the Ukrainian revolution and the Euromaidan movement. These keywords include "Ukrain", "Revolution", "Euromaidan", "Kiev", etc. Third category includes the names of the Ukrainian's political parties as well as the names of their key leaders. The other piece of seed knowledge that is provided to the Ukraine focused crawler is the list of popular blog domain such as "blogpost.com", "wordpress.com" and "livejournal.com". A focused crawling strategy that snowballs from matching keywords and blog domains were then used to crawl and download the blogrolls.

4.2 Evaluation Metrics

We present below four common measures that are widely used for assessing the quality of clustering methods.

4.2.1 Normalized Mutual Information (NMI):. This metric calculates the dependency between two discrete random variables [57]. Based on the knowledge of one random variable it quantifies the uncertainty of the other variable. This metric varies between 0 and 1, where 0 means there is no mutual information between the variables and 1 is total agreement between them. If $p(x, y)$ is the joint probability distribution of two discrete random variables X, Y, then the normalized mutual information can be calculated as shown in (1).

$$NMI(X, Y) = \frac{\sum_{y \in Y} \sum_{x \in X} log\Big(\frac{p(x,y)}{p(x)p(y)}\Big)}{\sqrt{H(X)H(Y)}} \tag{1}$$

where $H(X)$ and $H(Y)$ are marginal entropies calculated as in (2)

$$H(X) = -\sum_{i=1}^{n} p(x_i) log(p(x_i)) \tag{2}$$

4.2.2 Adjusted Rand Index (ARI):. Rand index metric calculates the similarities across the clusters and label sets. As an example if a set of elements as $S = \{a_1, a_2, ..., a_n\}$ is partitioned into two sets of $X = \{x_1, x_2, ..., x_p\}$, $Y = \{y_1, y_2, ..., y_q\}$, the rand index measures the ratio of the agreement between X and Y over the total number of observation pairs. The formula is shown in equation (3).

$$RI(X, Y) = \frac{A}{\binom{n}{2}} \tag{3}$$

where "A" is the total number of agreements and the denominator represents the total number of possible permutations. In case the clustering methods have different number of clusters, the adjusted rand index (ARI) is used to measure the similarity [60] and is explained in (4). It must be noted that the *ARI* calculated value is between 0 and 1 where 0 represents a random labelling and 1 represents identical labels.

$$ARI(X, Y) = \frac{RI(X, Y) - E\{RI(X, Y)\}}{max\{RI(X, Y)\} - E\{RI(X, Y)\}} \tag{4}$$

where "E" is the expected value function.

4.2.3 Adjusted Mutual Information (AMI):. Similar to the ARI metric, [60] proposed adjusted mutual information (AMI), which is given in (5). This metric varies between 0, which shows completely different partitions, and 1 that indicates identical partitions. [47] shows that AMI is more practical for the cases when labels are unbalanced, and the number of clusters are small and ARI is preferred

Table 3: Performance evaluation on the Ukraine blog data

Term Embeddings	Clustering	Purity(%)	NMI[a]	AMI[b]	ARI[c]
GloVe	Agglomerative	70.52	0.480	0.433	0.445
GloVe	**K-means**	**73.69**	**0.695**	**0.657**	**0.689**
GloVe	K-medoids	67.86	0.453	0.484	0.453
Word2Vec	Agglomerative	55.32	0.360	0.361	0.334
Word2Vec	K-means	65.74	0.354	0.330	0.357
Word2Vec	K-medoids	55.92	0.289	0.313	0.315
TF-IDF	Agglomerative	38.01	0.194	0.126	0.105
TF-IDF	K-means	40.21	0.182	0.155	0.089
TF-IDF	K-medoids	37.89	0.114	0.106	0.068
LDA[d]	——	38.44	0.278	0.215	0.197
Guided LDA	——	42.63	0.323	0.302	0.231
HLDA[e]	——	44.89	0.370	0.391	0.441
HPAM[f]	——	43.65	0.426	0.385	0.285

[a] Normalized Mutual information
[b] Adjusted Mutual Information
[c] Adjusted Rand Index
[d] Latent Dirichlet Allocation
[e] Hierarchical Latent Dirichlet Allocation
[f] Hierarchical Pachinko Allocation Model

when the the labels have large and similarly sized volumes.

$$AMI(X,Y) = \frac{MI(X,Y) - E\{MI(X,Y)\}}{max\{RI(X,Y)\} - E\{MI(X,Y)\}} \tag{5}$$

4.2.4 Purity: Purity metric measures the accuracy of a cluster assignment. In order to compute this metric, each cluster is assigned to the most frequent label in the cluster, and then the accuracy of this assignment is calculated by counting the number of correctly assigned observations. Finally for normalization, the calculated value is divided by total number of observation N. The calculation of the purity metric is shown in (6).

$$\text{purity}(X,Y) = \frac{1}{N}\sum_{k}\max_{j}|x_k \cap y_j| \tag{6}$$

where $X = \{x_1, x_2, \ldots, x_K\}$ is the set of labels and $Y = \{y_1, y_2, \ldots, y_J\}$ is the set of clusters. The calculated value of the purity is between 0 and 1, where 0 represents a poor clustering and 1 represents a perfect clustering.

4.3 Quantitative Results

Overall, the evaluation of a taxonomy induction is a challenging task since there is no ground-truth taxonomy for a particular domain and a particular corpus. Secondly, words can have multiple senses or semantic meanings. To tackle these difficulties, in this work first we focused on inducing a gold-standard taxonomy alongside its codebook. Next a group of ten communication scientists were trained on the codebook to map all unambiguous/topic specific terms-to-topics in the glod-standard taxonomy. This allowed us to employ all popular intrinsic and extrinsic measures that require ground-truth, in our experimental evaluations.

By applying the aforementioned term embedding and clustering methods on the input data the purity, NMI, AMI, and ARI metrics have been calculated. As shown in Table 3, each of the term embeddings method were paired with each of the three clustering algorithms and their joint performance were ranked in order to find the optimal combinations. The highest purity metric overall is 73.69% for the GloVe term embedding and K-means clustering. This combination resulted in $NMI = 0.695$, $AMI = 0.657$, and $ARI = 0.689$, which are also higher than the measures reported by other combinations. GloVe gets a holistic view of the entire corpus' term co-occurrence patterns while training its global vector representations as compared to Word2Vec which gleans its training performance from the local term contexts only. Choosing GloVe embeddings with K-means clustering results in top performance for the topic taxonomy induction and term-to-topic mapping tasks for the Ukraine blog dataset by all measures. It must be noted that extrinsic measures were calculated against the gold-standard topic taxonomy that was inducted by the team of experts.

To evaluate STIF's performance more extensively, we also show the experimental results for four other commonly used topic modeling algorithms including Latent Dirichlet Allocation (LDA)[6], Guided LDA[20], Hierarchical Latent Dirichlet Allocation (HLDA)[5] and Hierarchical Pachinko Allocation Model (HPAM)[51]. As shown in Table 3, none of these models could obtain the accuracies that can be obtained thru the utilization of any of the modern term embeddings, paired with a clustering algorithm.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we present the Semi-supervised Taxonomy Induction Framework (STIF) and evaluate its performance using a moderately sized real-word blog corpus comprising 145,000 posts from 650 Ukraine-related blog domains related to the ongoing Ukrainian crises. The Ukrainian crisis is the collective name for the 2013–14 Euromaidan protests associated with an emergent social movement supporting integration of Ukraine into the European Union, the 2013–14 Ukrainian revolution and the ensuing pro-Russian unrest. The contributions of this paper include a text processing pipeline that initially splits extracted noun phrases into general/ambiguous

vs. topical/non-ambiguous ones in order to facilitate taxonomy induction and term-to-topic mappings at high accuracies while reducing the overall human coding and mapping efforts significantly (i.e. by up to 74%) due to the high coherence, contrast, and purity of its intermediate topic clusters.

A team of communication scientists utilized STIF to induce a gold-standard topic taxonomy alongside its codebook. We experimented with various combinations of term embedding and clustering methods, evaluated their performance towards inducing a hierarchical topic taxonomy and associated term-to-topic mappings. Our experimental evaluations show that GloVe term embeddings paired with K-means provide the optimal results in terms of purity, NMI, AMI, and ARI quality measures. Additionally, four commonly used topic models have been assessed and it is shown that STIF yields better performance by all measures. Our future work includes folding GloVe term embeddings into other contextual document representations, such as BERT [13] or RoBERTa [30] to see if meaningful and interpretable document groups can be detected.

REFERENCES

[1] Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*. 85–94.

[2] Haodong Bai, Frank Z Xing, Erik Cambria, and Win-Bin Huang. 2019. Business taxonomy construction using concept-level hierarchical clustering. *arXiv preprint arXiv:1906.09694* (2019).

[3] AV Belyy, MS Seleznova, AK Sholokhov, and KV Vorontsov. 2018. Quality evaluation and improvement for hierarchical topic modeling. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*. 1–15.

[4] Abhijeet SANDEEP Bhardwaj, Akash Deep, Dharmaraj Veeramani, and Shiyu Zhou. 2021. A Custom Word Embedding Model for Clustering of Maintenance Records. *IEEE Transactions on Industrial Informatics* (2021).

[5] David M Blei, Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, et al. 2003. Hierarchical topic models and the nested Chinese restaurant process.. In *NIPS*, Vol. 16.

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

[7] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.

[8] B Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and Bruce Croft. 2021. An Intent Taxonomy for Questions Asked in Web Search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 85–94.

[9] Roberto Cerchione and Emilio Esposito. 2017. Using knowledge management systems: A taxonomy of SME strategies. *International Journal of Information Management* 37, 1 (2017), 1551–1562.

[10] Yoon Ho Cho and Jae Kyeong Kim. 2004. Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert systems with Applications* 26, 2 (2004), 233–246.

[11] Renato Cordeiro De Amorim and Christian Hennig. 2015. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information sciences* 324 (2015), 126–145.

[12] Knut De Swert. 2012. Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. *Center for Politics and Communication* 15 (2012).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. Supervised distributional hypernym discovery via domain adaptation. In *Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1-5; Austin, TX. Red Hook (NY): ACL; 2016. p. 424-35.* ACL (Association for Computational Linguistics).

[15] Amit Gupta, Rémi Lebret, Hamza Harkous, and Karl Aberer. 2017. Taxonomy induction using hypernym subsequences. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1329–1338.

[16] Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*.

[17] Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods* 53, 1 (2021), 232–246.

[18] Jin Huang, Zhaochun Ren, Wayne Xin Zhao, Gaole He, Ji-Rong Wen, and Daxiang Dong. 2019. Taxonomy-aware multi-hop reasoning networks for sequential recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 573–581.

[19] Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1928–1936.

[20] 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*. 204–213.

[21] Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M Kaplan, Timothy P Hanratty, and Jiawei Han. 2017. Metapad: Meta pattern discovery from massive text corpora. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 877–886.

[22] Kalpana D Joshi and PS Nalwade. 2013. Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing* 2, 7 (2013), 219–223.

[23] Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. Txtract: Taxonomy-aware knowledge extraction for thousands of product categories. *arXiv preprint arXiv:2004.13852* (2020).

[24] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1001–1012.

[25] Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing*. 1110–1118.

[26] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. 2001. On Semi-Automated Web Taxonomy Construction.. In *WebDB*. Citeseer, 91–96.

[27] Jiaqing Liang, Yi Zhang, Yanghua Xiao, Haixun Wang, Wei Wang, and Pinpin Zhu. 2017. On the transitivity of hypernym-hyponym relations in data-driven lexical taxonomies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[28] Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88, 3 (2000), 265.

[29] Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1433–1441.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[31] Steven Loria. 2018. textblob Documentation. *Release 0.15* 2 (2018).

[32] Anh Tuan Luu, Jung-jae Kim, and See Kiong Ng. 2014. Taxonomy construction using syntactic contextual evidence. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 810–819.

[33] Anh Tuan Luu, Yi Tay, Siu Cheung Hui, and See Kiong Ng. 2016. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 403–413.

[34] 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*. 3111–3119.

[35] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[36] David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*. 633–640.

[37] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. 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*. 1003–1011.

[38] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 1135–1145.

[39] Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. 1318–1327.

[40] Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[42] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. explanation of glove lib. https://nlp.stanford.edu/projects/glove/.

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[44] William M. Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. *J. Amer. Statist. Assoc.* 66, 336 (1971), 846–850.

[45] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[46] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* (2004).

[47] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. 2015. Adjusting for chance clustering comparison measures. *arXiv preprint arXiv:1512.01286* (2015).

[48] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[49] Anal Kanti Roy. 2019. *Automation of Crawling Blogosphere Based on Pattern Recognition*. Ph.D. Dissertation. University of Arkansas at Little Rock.

[50] Pedro V Sander, Xianfeng Gu, Steven J Gortler, Hugues Hoppe, and John Snyder. 2000. Silhouette clipping. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 327–334.

[51] Deepak Sharma, Bijendra Kumar, and Satish Chand. 2017. A survey on journey of topic modeling techniques from SVD to deep learning. *International Journal of Modern Education and Computer Science* 9, 7 (2017), 50.

[52] Rob Shearer and Ian Horrocks. 2009. Exploiting partial information in taxonomy construction. In *International Semantic Web Conference*. Springer, 569–584.

[53] Jiaming Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, and Jiawei Han. 2017. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 288–304.

[54] Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. 2018. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2180–2189.

[55] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in neural information processing systems*. 1297–1304.

[56] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A taxonomy of queries for e-commerce search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1245–1248.

[57] Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* 3, Dec (2002), 583–617.

[58] Robert L Thorndike. 1953. Who belongs in the family? *Psychometrika* 18, 4 (1953), 267–276.

[59] Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics* 39, 3 (2013), 665–707.

[60] Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research* 11 (2010), 2837–2854.

[61] Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*. 1063–1064.

[62] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. 2013. A phrase mining framework for recursive construction of a topical hierarchy. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 437–445.

[63] Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1190–1203.

[64] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.

[65] Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *COLING 2004: Proceedings of the 20th international conference on Computational Linguistics*. 1015–1021.

[66] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 481–492.

[67] Hui Yang and Jamie Callan. 2009. A metric-based framework for automatic taxonomy induction. 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*. 271–279.

[68] Li Yang and Yejun Wu. 2019. Creating a Taxonomy of Earthquake Disaster Response and Recovery for Online Earthquake Information Management. *KO KNOWLEDGE ORGANIZATION* 46, 2 (2019), 77–89.

[69] Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. 2017. Efficiently answering technical questions—a knowledge graph approach. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[70] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2701–2709.

[71] Yuchen Zhang, Amr Ahmed, Vanja Josifovski, and Alexander Smola. 2014. Taxonomy discovery for personalized recommendation. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 243–252.