International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 3687-3697

https://doi.org/10.70517/ijhsa463311

Bayesian hierarchical models for English corpora and discourse feature inference

Yani Liu[1,*]
[1] School of Foreign Languages, Liaodong University, Dandong, Liaoning, 118001, China
Corresponding authors: (e-mail: yannie2621@163.com).

Abstract With the continuous development and improvement of textual topic modeling, variational inference, as an effective approximate inference method, is widely used in parameter estimation of topic models. In this paper, combining Bayesian network and hierarchical Delikerian process (HDP) model, an HDP online variational Bayesian inference (Dist-LDA-VB) method is proposed and applied to the task of multidimensional inference of discourse hierarchical features in English corpus. By comparing the corresponding topics derived from the two models, it can be found that they have similar thematic content. In the topic inference of the constructed English corpus, the Dist-LDA-VB model and Markov Topic Models (MTMs) yielded similar topics, which are suitable for corpus discourse hierarchical feature inference. In addition, based on the corpus approach, this paper explores the semantic differences between the English “conclusively infer” class of synonymous adverbs certainly, definitely, necessarily and surely. The results of the study show the spatial distance between the target words and the variables, which is helpful for learners to memorize the correspondence between them, so as to systematize the inter-word differences and reduce the cases of misuse.

Index Terms bayesian networks, hierarchical dirichlet process (HDP), variational inference, English corpus

I. Introduction

In recent years, the rise of corpus linguistics has led to a wide interest in research related to the linguistic output of foreign language learners, especially the study of learners' written discourse [1]. The combination of corpora and computers allows for the analysis and study of features at all levels of language including single features or multiple features [2]. From single linguistic features to discourse studies, the scope of corpus methods applied is gradually widening, and there are systematic differences between language domains, but these differences are far from enough to recognize corpora through individual linguistic features [3]. The Multidimensional Inference Analysis (MDA) method was initially combined with the field of domain variation research, especially in spoken and written English [4]. This research methodology utilizes corpus technology to identify domain variation dimensions by analyzing “co-occurrence” patterns of linguistic features, and to conduct multidimensional comparative analyses of different corpora [5], [6]. This analytical model emphasizes the multidimensionality of the description and believes that no single dimension can fully reveal the features of the domain [7].

At present, domestic and international studies using MDA mainly focus on the study of oral and written language variation and the study of English corpora for specialized purposes. Literature [8] analyzed an English corpus using an improved version of MDA analysis method, aiming to discover the most common patterns of language variation showing different degrees of differences. Literature [9] used MDA to compare the linguistic characteristics of research article abstracts written by authors from the Middle East/North Africa (MENA) region with those written by international authors and found that MENA abstracts tended to be more evaluative and focused on current information, whereas the international abstracts showed more information density and narrative style. Literature [10] sued a corpus-based study which used Hyland's taxonomy and the R program to analyze metadiscourse features in an English corpus and found that interactive metadiscourse features were more prevalent than interactive metadiscourse features. Literature [11] constructed a data-driven model of English language teaching using a multidimensional corpus, proposed a new algorithm to improve the existing artificial intelligence modeling methods, and evaluated it through a simulation study with middle school students. Literature [12] investigated extrapolative reasoning in validity arguments for MELAB speaking assessments using corpus linguistics and MDA analysis, comparing differences in the linguistic features used by students in academic, professional, and conversational speaking domains. Literature [13] describes the construction of the CANELC corpus, a million-word digital corpus of communicative English, and conducts an MDA analysis of the linguistic patterns in the corpus with spoken and

written language from the British National Corpus (BNC), confirming the conjecture that communicative English promotes linguistic feature variation under certain dimensions.

In this paper, the variational inference method is applied to topic modeling to achieve multi-dimensional inference of discourse hierarchical features in an English corpus. Specifically, a variational inference method that combines Bayesian networks with hierarchical Delicacy Process (HDP) models is proposed and further optimized for distributed environments. To test the effectiveness of the proposed method, it is compared with Markov topic modeling, and the method of this paper is applied to inferring internal semantic variance of synonymous adverbs.

II. English corpus construction method

Corpus construction is required before multidimensional inferences can be made about discourse level features in an English corpus. This chapter describes how to create an English corpus and annotate the corpus according to the MAMA (Modeling-Annotation-Modeling-Annotation) cycle in the MATTER loop. The main processes are as follows:

(1) Perform corpus selection

In this paper, by crawling the abstracts of papers from English academic websites, as the corpus of English corpus, after pre-processing, they are put into Word documents for storage with year as the dividing line respectively.

(2) Corpus Creation Based on MAMA Loop

In the MATTER development cycle, the first step is “phenomenon modeling”, i.e., the “M” in the MATTER cycle, which is to build a model for the upcoming task of annotating the English corpus based on the collected data. Firstly, the raw corpus should be classified, and the relevant terms should be organized and described in detail to form a corresponding EXCEL document. Then, the model is represented using XML Document Type Definition (DTD), which can clearly describe the names, elements and attributes of the tasks, and is a popular generalized markup language at present. The corpus is modeled in conjunction with the examination of the English corpus.

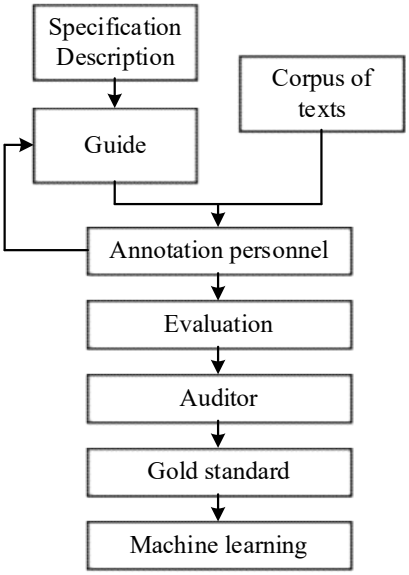


Figure 1: Annotation process

(3) Labeling and Review

After creating a corpus and a model, the actual annotation process, i.e., the “A” in the MATTER cycle, begins. The raw corpus is annotated according to the annotation model guidelines.In this paper, the English corpus is annotated using the Natural Language Processing Framework (GATE).GATE is an infrastructure for the development and deployment of software components for processing human languages.It is used for components for a variety of language processing tasks, such as parsers, morphology, tagging, information retrieval tools, information extraction components for a variety of languages, and many other components. GATE supports documents in a variety of formats, including XML, RTF, email, HTML, SGML, and plain text.

Firstly, the word document of the simple equation problem type of elementary school mathematics is imported into GATE, and then it is manually annotated according to the specification created before, and the final output document in XML format means that the annotation of the raw corpus is completed. However, problems may occur when entering information during the annotation process, and wrong labels may be filled in by chance if one is too

fatigued or not concentrated enough during the annotation. Therefore, after the annotation is complete, the labeled data is used to calculate annotation agreement (IAA) scores, and if these scores are low, the model is modified and then relabeled. If the scores are more favorable, an audit can be performed on the data to produce a gold standard corpus, which can then be used to train and test machine learning algorithms. This phase is called the MAMA cycle. The annotation process is shown in Figure 1, where the key part is to audit the results of obtaining annotations from the annotators and using it to generate an English corpus suitable for training machine learning.

III. Multidimensional inferential modeling of corpus discourse hierarchical features

In this chapter, by combining the Hierarchical Delicacy Process Model (HDP) with Bayesian network inference methods, the Distributed Approach to Online Variational Bayesian Inference for HDP (Dist-LDA-VB) is proposed for the problems faced by HDP online variational Bayesian inference in distributed environments to realize multidimensional inference of discourse hierarchical features of the English corpus.

III. A. Bayesian networks

Bayesian networks (BN) [14], also known as belief networks, are one of the extension methods based on Bayes' theorem. Compared to other models, Bayesian networks can achieve timely updating of the original beliefs under the observation of new evidence while successfully modeling conditional dependencies between variables, and have now become one of the most effective models in the field of uncertainty analysis and reasoning.

Bayesian networks consist of two parts: graphical structure and network parameters. The graphical structure is a directed topological graph describing the relationship between the variables of the research object, including two parts: nodes and directed edges, where nodes represent the relevant variables existing in the system, which are divided into child nodes and parent nodes. The directed edges represent the conditional dependencies between nodes and represent this relationship through directed connections from parent to child. The network parameters are quantitative reflections of the attributes of the nodes and the dependencies between them, and are the basis for probabilistic inference using the network, including the initial a priori probability and the conditional probability between the nodes, which are based on Bayesian theory for their principles and inference capabilities.

III. A. 1) Bayesian network correlation modeling

(1) Plain Bayesian Network Modeling

The plain Bayesian network (NBN) [15] is one of the most classical Bayesian network models, which is simple and efficient, especially suitable for processing high-dimensional data and text data. Compared with other network models, the typical feature of NBN is that it is based on the conditional independence assumption, i.e., for the sample data set $D=\{d_1,d_2,\cdots,d_n\}$, the corresponding set of feature attribute variables is $X=\{x_1,x_2,\cdots,x_d\}$, the set of class variables is $C=\{c_1,c_2,\cdots,c_m\}$, and D can be categorized into c_m categories, where all the attribute variables are conditionally independent of the class variables.The construction of an NBN is usually based on the definition of the graphical structure of the network, and then using relevant algorithms such as maximum likelihood estimation, to train the dataset to estimate the network parameters, thus outputting the full model for use in specific tasks. Since the traditional NBN suffers from an exponential explosion of its conditional probability table (CPT) when the attribute variables increase, and the huge amount of computation will seriously increase the processing time of computers, some scholars have proposed an improved NBN model based on the traditional model. The improved NBN switches the direction of the oriented edges in the graph structure, which greatly improves the computational speed while realizing the same effect of the traditional model.

Although the strong assumptions of the plain Bayesian network do not always hold in practical problems, the method is logically simple and has good robustness in the face of different types of data.

(2) Tree-Augmented Simple Bayesian Network Modeling

The Tree Augmented Simple Bayesian (TAN) network [16] improves on the NBN, which overcomes the assumption of conditional independence followed by the NBN by allowing each attribute variable to have up to one attribute variable as a parent node in addition to the class variables, and considers the interrelationships among the attribute variables. Typically TAN networks, like NBN, include a set of class variables C and a set of attribute variables X, and each attribute variable x_i consists of n states, denoted by the set $S_X=\{x_{i1},x_{i2},\cdots,x_{in}\}$. Unlike NBN, the graph structure of TAN networks usually also relies on data learning to determine the optimal network structure by calculating the conditional mutual information between attribute variables and iterating until the output is optimal.

Compared with NBN, TAN network maintains the computational complexity and robustness, and also takes into account the existence of interconnections between variables, which makes the network model closer to the real situation and the analysis results more accurate.

III. A. 2) Bayesian network inference

The core purpose of Bayesian network construction is to update the original beliefs with the help of its powerful inference ability to obtain the a posteriori probability distribution of the uncertain event in question, so as to provide support for relevant decisions. The ability of the Bayesian network to update the node prior probabilities given new observational evidence E is based on Bayes' theorem, and the calculation of the node posterior probabilities is specifically shown in equation (1):

$$P(X \mid E) = \frac{P(X,E)}{P(E)} = \frac{P(X,E)}{\sum_{X} P(X,E)} \tag{1}$$

where $P(X \mid E)$ is the posterior probability and X is a set of variables $(x_1, x_2, \cdots, x_n)$ in BN.

According to the purpose and direction of reasoning, Bayesian network reasoning can usually be categorized into three types: forward reasoning, reverse diagnosis, and mixed reasoning.

(1) Forward reasoning

Forward reasoning, i.e., from cause to effect, also known as causal inference, this inference mode inputs the observation information of the parent node as evidence into the model, and updates the occurrence probability of the child node after inference, which is usually used for estimating the probability distribution of the untyped coming event.

(2) Reverse Diagnosis

Reverse diagnosis, i.e., pushing the cause from the effect, also known as diagnostic reasoning, the reasoning is the opposite of forward reasoning, is based on the observed evidence of the child node inference of the parent node posterior probability distribution of the inference mode. Reverse diagnosis is usually used to determine the cause of a known outcome and to identify the key factors that have a significant impact on the child node based on the probability change of the parent node's factors.

(3) Hybrid reasoning

Hybrid reasoning synthesizes the above two types of reasoning, usually using observations of child nodes and some of the parent nodes as input evidence to reason about the posterior probability distributions of other parent nodes. Hybrid reasoning can analyze the correlations between all parent nodes that have an effect on a child node.

III. B. Distributed Optimization and Implementation of Hierarchical Dirichlet Process Models

LDA [17] better solves the problem of text topic clustering for fixed corpus, and its inability to solve the problem of topic number variation in the face of text stream topic clustering. Hierarchical Dirichlet Process (HDP) [18] re-models the text and solves the topic evolution problem in streaming computing scenarios well. In this section, the HDP model is further optimized for distributed environment.

III. B. 1) HDP online variational Bayesian inference

In order to support the incremental update of the model, this paper improves the HDP Variational Bayesian algorithm (LDA-VB) by updating the corpus layer variational parameters when all the documents have been trained, and adjusting it to updating the corpus layer variational parameters when the training of each document or each small batch of documents has been completed to form the HDP Online Variational Bayesian Inference algorithm.

Like LDA online Bayesian inference, HDP online variational Bayesian inference algorithm samples the window training method, only a small portion of documents in the corpus are trained in each window, and the parameters are updated when the window training is completed. In order to effectively fuse the local model obtained from window training with the global model, the parameter values of the local model obtained from window training was amplified. According to the principle of document similarity, their effects on model updating should be equivalent, which leads to the following formula for updating the corpus layer variational parameters for HDP online variational Bayesian inference:

$$\begin{aligned} \Delta u_k &= -u_k + 1 + D\sum_{t=1}^{T} \varphi_{jtk} \\ \Delta v_k &= -v_k + \gamma + D\sum_{t=1}^{T}\sum_{l=k+1}^{K} \varphi_{jtl} \\ \Delta \lambda_{kw} &= -\lambda_{kw} + \eta + D\sum_{t=1}^{T} \varphi_{jtk} \left(\sum_{n} \zeta_{jnt} I[w_{jn} = w] \right) \end{aligned} \tag{2}$$

$$
\begin{aligned}
u_k &= u_k + \rho_{t_o} \Delta u_k \\
v_k &= v_k + \rho_{t_o} \Delta v_k \\
\lambda_{kw} &= \lambda_{kw} + \rho_{t_o} \Delta \lambda_{kw}
\end{aligned} \tag{3}
$$

where ρ_{t_o} denotes the learning rate.

For each batch of documents learned in the window, the update formula for the corpus layer variant parameters is shown in Eq. (4):

$$
\begin{aligned}
\Delta u_k &= -u_k + 1 + \frac{D}{S} \sum_{j \in J_{t_o}} \sum_{t=1}^{T} \varphi_{jtk} \\
\Delta v_k &= -v_k + \gamma + \frac{D}{S} \sum_{j \in J_{t_o}} \sum_{t=1}^{T} \sum_{l=k+1}^{K} \varphi_{jtl} \\
\Delta \lambda_{kw} &= -\lambda_{kw} + \eta + \frac{D}{S} \sum_{j \in J_{t_o}} \sum_{t=1}^{T} \varphi_{jtk} \left(\sum_{n} \zeta_{jnt} I[w_{jn} = w] \right)
\end{aligned} \tag{4}
$$

where J_{t_o} denotes the indexed set of documents in the t_o window, and $I[w_{jn} = w]$ is the indicator function, whose value is 1 when the n th word in the document j is the w th word in the vocabulary, and 0 otherwise.

III. B. 2) Data parallelism

Data parallelism is the main way machine learning systems are used to deal with large-scale problems, and it has two typical characteristics:

(1) Sample data is distributed on a cluster. Distributed file systems are usually utilized to manage the sample data, for example, Hadoop File System (HDFS) and Distributed Resilient Dataset (RDD) can transparently implement the distributed storage function of the sample data for the users, and provide good fault tolerance and scalability.

(2) Migrate computation towards data. In order to achieve data locality and reduce the overhead of remote access to data during training, distributed systems usually assign training tasks to be executed at the node that owns the data.

In the data parallel system, the sample data is dispersed in each node of the cluster, and each node is responsible for training its own part of the sample data, and the node obtains the parameters needed for training from the model side through the network, and updates the results to the model through the network after the training is completed. Data parallelism uses the multi-core characteristics of the cluster to divide the training task of a large number of sample data into multiple sub-tasks, which are given to different CPU cores to compute respectively.Spark is a typical data parallel distributed system, whose RDD divides the data into multiple partitions, and each partition will form an independent task during data computation, so as to realize the computation to data migration. In this paper, Spark system as the basis of HDP distributed will be used.

III. B. 3) Model Parallelism

Model parallelism refers to a machine learning system that distributes the maintenance of model parameters in order to achieve large-scale parameter storage and access with high performance. The system structure of data parallelism and model parallelism is shown in Fig. 2, in which both sample data and parameter data are distributed among cluster nodes, and the two types of data interact with each other through the network. Data parallelism and model parallelism can be applied to machine learning scenarios with large-scale sample sets and a large number of model parameters, which is the mainstream structure of today's large-scale distributed machine learning systems.

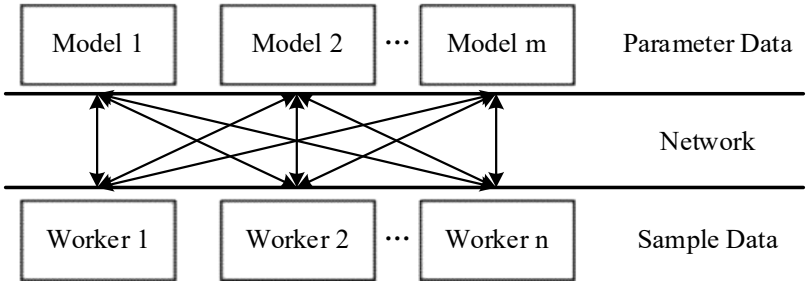


Figure 2: Data parallelism and model parallelism

III. B. 4) Variational parameter update strategy

In order to realize HDP online variational Bayesian inference applied in the parameter server framework, the parameter update strategy for each training partition need to be formulated. Now consider the multiple partitions concurrent training scenario, assuming that the document training in the window is divided into P partitions, and each partition contains the number of document articles of S_p, which will be satisfied:

$$\sum_{p=1}^{P} S_p = S \tag{5}$$

where s is the number of document pieces in the window.

To ensure the equivalence of the parameter updating effect of window multi-partition training and stand-alone training, it can be obtained:

$$\Delta u_k = \sum_{p=1}^{P} \Delta u_{pk}; \quad \Delta v_k = \sum_{p=1}^{P} \Delta v_{pk}; \quad \Delta \lambda_{kw} = \sum_{p=1}^{P} \Delta \lambda_{pkw} \tag{6}$$

where $\Delta\lambda_{pkw}, \Delta u_{pk}, \Delta v_{pk}$ denote the update amount of the corpus layer variant parameters after the training of the pth partition is completed, respectively.

Based on the document similarity assumption, according to Eqs. (4) and (6), the update amount of the corpus layer variant parameters after the training of the pth partition is completed is:

$$\begin{aligned}
\Delta u_k &= -\frac{S_p}{S} u_k + \frac{S_p}{S} + \frac{D}{S} \sum_{j \in J_{to}} \sum_{t=1}^{T} \varphi_{jtk} \\
\Delta v_k &= -\frac{S_p}{S} v_k + \frac{S_p}{S} \gamma + \frac{D}{S} \sum_{j \in J_{to}} \sum_{t=1}^{T} \sum_{l=k+1}^{K} \varphi_{jtl} \\
\Delta \lambda_{pkw} &= -\frac{S_p}{S} \lambda_{kw} + \frac{S_p}{S} \eta + \frac{D}{S} \sum_{j \in J_{to}} n_{tsw} \phi_{tswk}
\end{aligned} \tag{7}$$

where η, γ is used to describe the parameter update step, for which the amount of change in the p-partitioned corpus layer variant parameter is corrected:

$$\begin{aligned}
\Delta u_{pk} &= -\frac{S_p}{S} u_k + 1 + \frac{D}{S} \sum_{j \in J_{to}} \sum_{t=1}^{T} \varphi_{jtk} \\
\Delta v_{pk} &= -\frac{S_p}{S} v_k + \gamma + \frac{D}{S} \sum_{j \in J_{to}} \sum_{t=1}^{T} \sum_{l=k+1}^{K} \varphi_{jtl} \\
\Delta \lambda_{pkw} &= -\frac{S_p}{S} \lambda_{kw} + \eta + \frac{D}{S} \sum_{j \in J_{to}} n_{tsw} \phi_{tswk}
\end{aligned} \tag{8}$$

where J_{t_o} denotes the indexed set of documents in the t_o window.

The p partitioning needs to be parameterized for the corpus layer variant uploaded by the model as:

$$\begin{aligned}
\Delta u'_{pk} &= \rho_{t_o} \Delta u_{pk} \\
\Delta v'_{pk} &= \rho_{t_o} \Delta v_{pk} \\
\Delta \lambda'_{pkw} &= \rho_{t_o} \Delta \lambda_{pkw}
\end{aligned} \tag{9}$$

where denotes the learning rate of the t_o window.

Up to this point, this paper derives a partitioned update strategy for the HDP variational Bayesian algorithm in a distributed environment. According to this update strategy, each document training node can concurrently update the model parameters, thus realizing the distributed HDP online variational inference method for English corpus.

IV. Multidimensional Inference of Discourse Hierarchical Features Based on English Corpus

IV. A. Algorithm Validation and Model Analysis

In this study, the dataset of English dissertation abstracts was divided by year, and a total of 10 English corpora were obtained. The details of the English corpus in different years are shown in Table 1.

It can be seen that the difference in the number of documents in the English corpus in different years is not obvious, all of them are around 1400, and the number of words in the corpus is slightly different, but the difference is not big. However, with the increase of the year, the average number of words per document in the corpus shows an upward trend.

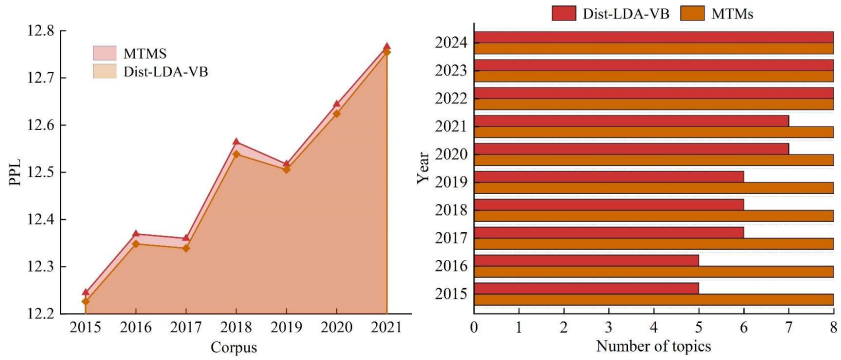
Table 1: English corpora of different years

Corpus	Number of documents	Number of words	The average number of words
In 2015	1462	28643	19.59
In 2016	1485	29048	19.56
In 2017	1432	28115	19.63
In 2018	1487	30124	20.26
In 2019	1476	31135	21.09
In 2020	1445	30947	21.42
In 2021	1412	30854	21.85
In 2022	1427	32478	22.76
In 2023	1399	33831	24.18
In 2024	1415	35946	25.40
Total	14440	311121	21.55

In order to verify the effectiveness of the proposed HDP online variational Bayesian inference distributed method (Dist-LDA-VB) in inferring discourse hierarchical features of the English corpus, it is compared with Markov topic models (MTMs) for the corpus topic inference experiments and the perplexity degree (PPL) is chosen as an evaluation index, where the smaller the perplexity degree also implies that the model selects the number of topics better.

When the number of topics in the English corpus changes, the perplexity degree changes accordingly. The changes of perplexity and the number of topics in the corpus under the two models are shown in Figure 3, where (a) denotes the PPL metric and (b) denotes the number of topics.

In MTMs, the number of topics is 8 for all corpora. However, in Dist-LDA-VB, the optimal number of topics for the two corpora in 2015 and 2016 is 5, the optimal number of topics for 2017~2019 is 6, and the optimal number of topics for 2017 and 2018 is 7. For the 10 corpora in the whole dataset, only three corpora have the same number of optimal topics as the number of topics in the dataset, and the remaining seven have changed accordingly. The difference in the optimal number of topics between different corpora can be explained by the increase in the average number of words in the documents. Combining Fig. 3(a) and Fig. 3(b) shows that in Dist-LDA-VB, when the optimal number of topics of a corpus decreases, the perplexity of the corpus decreases as well.



(a) PPL (b) Number of topics

Figure 3: The changes in confusion and the number of topics in the corpus

In order to better study the effect of the change in the number of topics on the corpus, this study takes 2015 as an example to start the analysis. the top 3 topics of the 2015 corpus under the MTMs model and the Dist-LDA-VB model, and the top 10 words in terms of contribution rate are listed respectively as shown in Table 2.

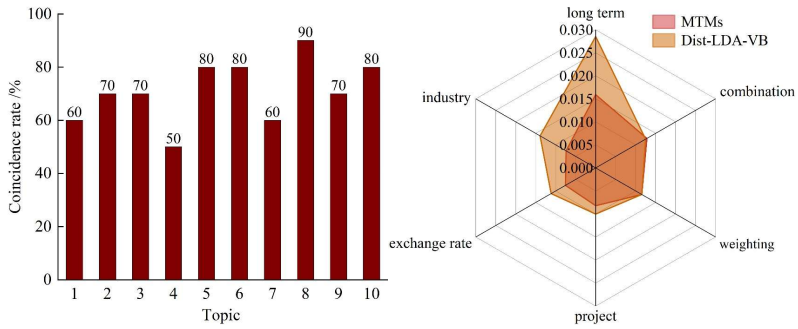
By comparing the corresponding topics derived from the two models, it can be found that they have similar subject content. Specifically, topics 1 to 3 are related to economy, rural education, and production performance, respectively. These results indicate that the two different topic models yielded similar topics, verifying the effectiveness of this paper's model in making inferences about corpus discourse level features.

Table 2: The top ten words in terms of topic contribution rate

Topic type					
Economic topic		Educational topic		Production topic	
MTMs	Dist-LDA-VB	MTMs	Dist-LDA-VB	MTMs	Dist-LDA-VB
Long term	Industry	Input	Countryside	Performance	Performance
Combination	Long term	Education	Survey	Control	Control
Weighting	Ability	Decision making	Ability	Logistics	Evaluate
Project	Industry	Process	Process	Effect	Energy
Exchange rate	Combination	Give	Decision making	Process	Effect
Industry	Weighting	Ability	Transfer	Survey	Value
Renminbi	Exchange rate	Condition	Education	Decision making	Process
Currency	Project	Countryside	Evaluate	Value	Algorithm
Equalization	Cluster	Fund	Fund	Evaluate	Logistics
Scale	Outputs	Labor force	Labor force	Mass	Mass

Then, the top 10 topics in the 2015 corpus were studied, and the results of the word research are shown in Figure 4, in which (a) is the topic overlap rate of the two models, and (b) is the contribution rate of the six words "long-term""combination" "weight""project""exchange rate" and "industry" in the MTMs and Dist-LDA-VB in topic 1 "economy".

Figure 4(a) shows that in the 2015 corpus, the top ten words in the 10 topics ranked by the two models have a high overlap rate of more than 50%, with the overlap rate of the eighth topic reaching 90%. While observing Figure 4(b), although these words are not necessarily ranked in the top six in Dist-LDA-VB, it can be found that, except for the word “combination”, the contribution rate of the other five words is not lower than that in MTMs in Dist-LDA-VB. In addition, the total contribution of these six words is higher in Dist-LDA-VB. This indicates that the topics in Dist-LDA-VB are more representative, more ordered, and the words are more concentrated, and fewer words can be used to represent the topic, reflecting the superiority of Dist-LDA-VB in inferring discourse-level features of the corpus.



(a) Word coincidence rate (b) Topic 1 Contribution Rate

Figure 4: Word research in the topic

IV. B. Corpus-based exploration of semantic differences within synonymous adverbs

In this section, the proposed Dist-LDA-VB method is applied to explore the internal semantic differences of the synonymous adverbs "certainly""definitely" "necessarily" and "surely" in the English corpus.

IV. B. 1) English corpus processing

Before the experimental investigation, the corpus in the English corpus needs to be processed, which mainly includes 3 steps of corpus extraction, corpus labeling and corpus statistics. First, all target word index lines are downloaded from the constructed English corpus and non-adverbial index lines and duplicate index lines are exhaustively deleted manually, and then the purified examples are classified and stored according to the types of domains, and finally 1,200 examples of each of the four target words are randomly selected as the final annotated and analyzed examples in this study by adopting the stratified sampling method. Then, the final examples are analyzed one by one, and the target word features are manually labeled. The selection of variables in this study is mainly based on the syntactic function and syntactic structure of English intonational adverbs. Finally, the annotation results were frequency summarized and converted into relative frequencies, and some of the relative co-occurrence frequencies are shown in Table 3.

Table 3: Relative co-occurrence frequency (Partial)

Target word	certainly	definitely	necessarily	surely
Modal verb	0.685	0.348	0.328	0.351
Action verb	0.061	0.189	0.145	0.153
Verb of judgment	0.079	0.201	0.093	0.148
Existential verb	0.035	0.092	0.131	0.106
Psychological verb	0.045	0.034	0.004	0.031
Relative verb	0.009	0.020	0.052	0.022
Directional verb	0.014	0.007	0.003	0.009
Causative verb	0.006	0.010	0.130	0.009
Dummy verb	0	0.002	0	0.002
…	…	…	…	…

IV. B. 2) Target Phrase Differentiation Analysis

Hierarchical cluster analysis was first used to explore the intergroup variability of the four target words, and then correspondence analysis was used to examine the internal variability of the target words in detail.

(1) Inter-group differences of target words

Based on the clustering results, "definitely" and "surely" form the first cluster, "necessarily" reclusters with the results of the first cluster to form the second cluster, and "certainly" reclusters with the results of the second cluster. Therefore, "definitely" and "surely" have the strongest semantic similarity, while "certainly" has the weakest semantic similarity with "definitely" and "surely". The two p-values of the "first cluster" are 99 and 91 respectively, both greater than 92 and 76 of the "second cluster". It can be known from the clustering analysis algorithm that the larger the p-values of the two types are, the easier they are to cluster. Therefore, "definitely" and "surely" are the most likely to cluster, and the p-value of "definitely" and "surely" clustering is as high as 99%, indicating that they have the strongest semantic similarity and tend to be used the same. Cluster analysis and p-values reveal that "definitely" and "surely" have the most similar semantics, while "certainly" has the greatest difference from the other three.

(2) Differences in the internal usage of the target word

The corresponding analysis results explain 90.4% of the data variation, indicating that this analysis is relatively stable. It presents all the variables and converts the co-occurrence frequency between the target words and each variable into a visual two-point distance. The distance is inversely proportional to the correlation between the data points, that is, the farther the distance between the target words and between the target words and the variables, the weaker the correlation. The letter "definitely" and "surely" in the same quadrant are the closest in distance and have the strongest correlation. The result that "certainly" has the greatest distance from "definitely" and "surely" and the difference between them is consistent with the result of cluster analysis, and the two confirm each other. The variables significantly correlated with "certainly" mainly include "willingness V""tendency V""living subject""imperative sentence""negative target word", etc. The variables coexisting with "necessarily" mainly include "relation V""cause V""no living subject""negative", etc. The variables associated with "definitely" mainly include "judgment V""negative structure" "sentence beginning", etc. The variables associated with "surely" mainly include "action V""judgment V""adjective", etc. Therefore, the four target words all have internal differences at the levels of "willingness V""tendency V""relation V""cause V""judgment V""adjective""subject type""syntactic position and sentence type" "semantic rhyme"and "negative structure".

(3) Semantic rhyme

At the semantic rhyme level, "necessarily" differs most significantly from the other three, with only "necessarily" being the closest to the negative semantic rhyme. Nearly half of the index lines in the example show the combination of "necessarily" with words that convey negative semantic rhymes, expressing the speaker's definite speculation that something being caused will have a negative effect, particularly highlighting the speaker's strong dissatisfaction. Therefore, the frequent co-occurrence of "necessarily" and “Order V” provides a basis for the negative semantic rhyme it reveals, and this is the second reason for the semantic difference between "necessarily" and the other three. However, this usage feature of "necessarily" has rarely been mentioned in previous studies. In conclusion, the four target words show subtle differences in different usage patterns as shown in Table 4. Among them, "+" indicates strength and "-" indicates non-existence.

Table 4: The main differences in the usage of target words

Variable	certainly	definitely	necessarily	surely
Degree of certainty	+	++	+++	++++
Subjectivity	+++	++	+	-
Living subject	+++	++	++	+
No Living subject	+	++	++	+++
Generally refers to the subject	+	-	-	-
Co-occurrence frequency with “want”	++	+	+	+
Co-occurring frequency with “will”	+	++	++	++
Causative verb	+	+	+	++
Firmness of will	+	-	-	-
Imperative tone	+	-	-	-
Attitude marking function	+	-	-	-
Speech act function	+	+	+	-
Discourse marker	-	+	-	-
Negative semantic rhyme	-	-	-	+
Sentence beginning	++	++	+	-
Private use	+	+	-	-
Negative target word	+	-	-	-

V. Conclusion

This paper proposes a distributed variational inference method that combines Bayesian networks with hierarchical Delicacy Process (HDP) models to realize multidimensional inference of discourse hierarchical features based on an English corpus.

This paper constructs an English corpus from 2015 and compares the top 10 words in terms of contribution rate for the first 3 topics of the corpus in the MTMs model and the Dist-LDA-VB model introduced in this work. The results show that the corresponding topics derived from the two models have similar thematic content, and the 1st to 3rd topics are all related to economy, rural education, and production performance, respectively, which verifies the validity of this paper's model in making inferences of discourse-level features of the corpus. Meanwhile, in the 2015 corpus, the overlap rate of the top ten words in the 10 topics of both models is above 50%, and the highest rate reaches 90%. And in Topic 1, the total contribution rate of the six words listed is higher in Dist-LDA-VB, which indicates that the topics in Dist-LDA-VB are more representative and organized, reflecting the superiority of Dist-LDA-VB in corpus discourse hierarchy feature inference.

The Dist-LDA-VB model was used to explore the semantic differences between the synonymous adverbs "certainly""definitely""necessarily" and "surely" of "conclusive inference", and the conclusions were as follows: the subjects of "certainly" are mostly living nouns, which often co-occur with the verb "to", and the expresser's willingness and commitment to let a certain act be carried out are used in imperative sentences, which have strong subjective emotions. The syntactic position of “definitely” is more flexible; when it is located at the beginning of a sentence, it serves as a discourse marker and has the function of articulating a certain subjective attitude and emotion on the part of the speaker. “Surely” cannot be used exclusively. The subject of “necessarily” is mostly an inanimate noun, which focuses on expressing the strong objective necessity of the proposition, showing the negative semantic rhyme. This paper adopts the correspondence analysis method to crystallize the inter-word differences, which will help learners to grasp and reduce the misuse cases.

References

[1] Leech, G. (2014). The state of the art in corpus linguistics. English corpus linguistics, 8-29.

[2] Monaghan, P., & Rowland, C. F. (2017). Combining language corpora with experimental and computational approaches for language acquisition research. Language Learning, 67(S1), 14-39.

[3] Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. Literary and linguistic computing, 25(4), 447-464.

[4] Goulart, L., & Staples, S. (2023). Multidimensional analysis. In Conducting genre-based research in applied linguistics (pp. 127-148). Routledge.

[5] McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. Annual Review of Applied Linguistics, 39, 74-92.

[6] Dash, N. S., Arulmozi, S., Dash, N. S., & Arulmozi, S. (2018). Features of a Corpus. History, features, and typology of language corpora, 17-34.

[7] Cobb, T. (2018). From corpus to CALL: The use of technology in teaching and learning formulaic language. In Understanding formulaic language (pp. 192-210). Routledge.

[8] Clarke, I. (2022). A Multi-dimensional analysis of English tweets. Language and Literature, 31(2), 124-149.

[9] Alamri, B. (2023). A multidimensional comparative analysis of MENA and international english research article abstracts in applied linguistics. SAGE Open, 13(1), 21582440221145669.

[10] Farahani, M. V. (2019). Metadiscourse in academic English texts: A corpus-based probe into British academic written English corpus. Kalbų Studijos, (34), 56-73.

[11] Chen, D. (2022). Constructing a Data‐Driven Model of English Language Teaching with a Multidimensional Corpus. Mathematical Problems in Engineering, 2022(1), 2715408.

[12] LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. Language Testing, 34(4), 451-475.

[13] Knight, D., Adolphs, S., & Carter, R. (2014). CANELC: Constructing an e-language corpus. Corpora, 9(1), 29-56.

[14] Mohsen Jafari Songhori, Claudia Fecarotti & Geert Jan van Houtum. (2025). Simulation supported Bayesian network approach for performance assessment of complex infrastructure systems. Reliability Engineering and System Safety,261,111045-111045.

[15] Hongwei Huang, Chen Wu, Mingliang Zhou, Jiayao Chen, Tianze Han & Le Zhang. (2024). Rock mass quality prediction on tunnel faces with incomplete multi-source dataset via tree-augmented naive Bayesian network. International Journal of Mining Science and Technology,34(3),323-337.

[16] Ruz Gonzalo A.,Araya Díaz Pamela & Henríquez Pablo A. (2022). Facial biotype classification for orthodontic treatment planning using an alternative learning algorithm for tree augmented Naive Bayes. BMC Medical Informatics and Decision Making. 22(1),316-316.

[17] Fei Li,Huishang Li,Xin Dai,Hongjie Ren & Huaiyang Li.(2025).Does Online Public Opinion Regarding Swine Epidemic Diseases Influence Fluctuations in Pork Prices?—An Analysis Based on TVP-VAR and LDA Models.Agriculture,15(7),730-730.

[18] Walaa Gamaleldin, Osama Attayyib, Linda Mohaisen, Nadir Omer & Ruixing Ming. (2025). Developing a hybrid model based on Convolutional Neural Network (CNN) and Linear Discriminant Analysis (LDA) for investigating anti-selection risk in insurance. Journal of Radiation Research and Applied Sciences,18(2),101368-101368.