Revealing the Research Deviation of AI Research Between China and the U.S.*

Han Sun[1,†], Guo Chen[1,†],

[1] *Nanjing University of Science and Technology, Nanjing ,210994, Jiangsu, China*

Abstract

China and the U.S. are recognized as leading forces in artificial intelligence (AI) research. Understanding the research differences between these two nations is crucial for grasping the global AI landscape. This paper moves beyond traditional methods reliant on frequency statistics and topic analysis. By analyzing both co-occurrence and vector semantic fields, we delineate the research focuses and content preference on specific domain entities in AI between China and the U.S. This framework enables a thorough examination of the distribution of research efforts within each zone, providing valuable insights into the distinctive research profiles and potential collaboration pathways in AI between these two technological giants..

Keywords

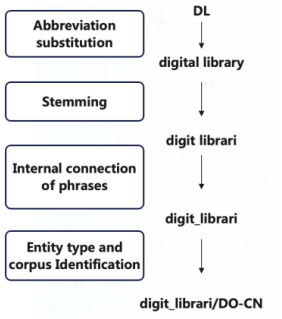
Research difference, Semantic deviation, Domain entity, Semantic field, Artificial Intelligence

1. Introduction

Artificial Intelligence (AI) has become a critical catalyst for economic and cultural progress [1]. Within the sphere of AI, China and the United States are acknowledged leaders in the global arena [2]. However, in the comparative analysis of AI development, the aforementioned studies do not extensively address the semantic nuances in the disparities of research between China and the U.S. within this domain. Thus, this study adopts a theoretical framework grounded in semantic deviation and semantic fields to undertake a comparative analysis of AI research disparities between China and the U.S.

2. Overview of the papers

Based on the aforementioned approach, we propose the implementation process illustrated as follows:



Joint Workshop of the 2th Innovation Measurement for Scientific Communication (IMSC) in the Era of Big Data (IMSC2024), Dec 20th, 2024, Hong Kong, China and Online

[1]* Corresponding author.

[†] These authors contributed equally.

1719801337@qq.com (H. Sun); delphi1987@qq.com (G. Chen);

0009-0002-0121-8055 (H. Sun); 0000-0003-2873-1051 (G. Chen);

Figure 1: Process of entity identification.

Data processing and construction of word vectors: This study takes 404,168 journal articles in the field of AI from the WOS core collection from January 1996 to May 2023 as the data source. After data processing and identifying problem and method entities from titles and abstracts, each entity is represented by a 100-dimensional vector using the Word2Vec model.

Two-dimensional matrix analysis: To calculate the quantitative values of each domain entity, we focus on research scale and semantic deviation: the research scale for a given entity in a particular country can be represented by the document frequency of the entity in the two corpora, and the semantic deviation between the two corpora can be quantified by weighted vector distance.

Analysis based on semantic field: Semantic field analysis consists of two distinct parts: co-occurrence semantic field and semantic distance semantic field. Identifying the top 10 words that have the smallest vector distance and highest co-occurrence with with the selected words in the corpus provides insights into the research scale and content preference.

3. Result analysis

3.1. Overall analysis

The resulting distribution is detailed as follows.

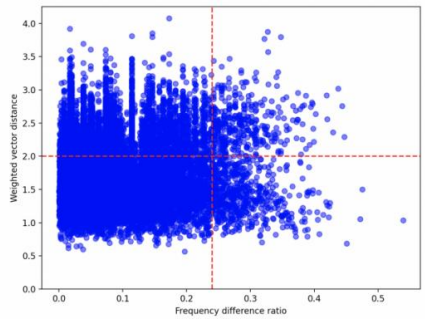


Figure 2: Distribution of all entities.

From the data, we can see that there are 4020 entities with large semantic differences, accounting for 37.55% of the total, and the rest are those with small semantic differences. More than one-third of the topics have large differences in content preferences, which is a considerable proportion. Subsequent analysis will delve deeper into the differences and the reasons behind them from a smaller perspective.

3.2. Macro level analysis - top 100 entities

To further explore the differences in research focuses between China and the U.S., we construct two co-occurrence networks of top 100 high-frequency entities in each corpus, followed by topic clustering and visualization as shown in Figure 3 and Figure 4. In the figure, the color of the nodes indicates the cluster category to which the entities belong, the shape of the nodes represents the types of semantic differences associated with the entities.

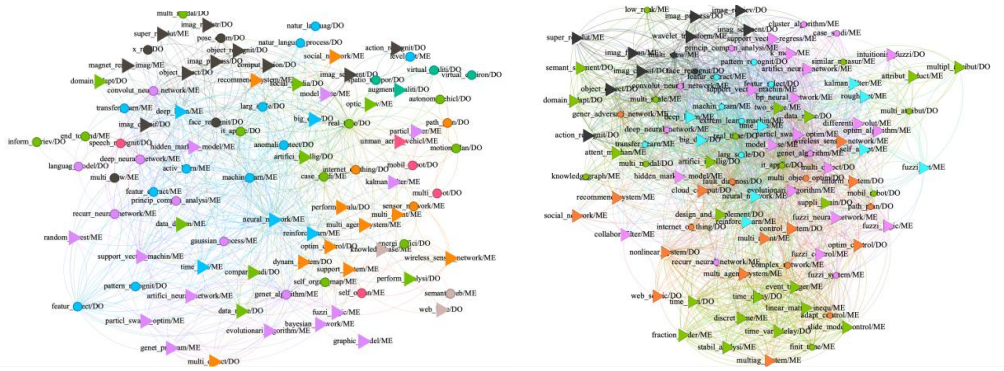


Figure 3: The U.S. author and Chinese author corpus.

As shown in the figures, the results align closely with the overall findings shown in Figure 2, indicating that for nearly 80% of popular research topics in AI, China and the United States have similar levels of semantic deviation (content preference).

3.3. Micro level analysis - case study

To further explore the details and underlying causes of semantic deviation at a micro level, we taking the term "facial recognition" as an example.

Table 1
Comparison of semantic neighbors of "Face Recognition"

	Top 10 semantic neighbors
Chinese corpus	ear_recognit/DO;facial_express_recognit/DO;singl_sampl_face_recognit/DO;robust_face_recognit/DO;face_verif/DO;palmprint_recognit/DO;gender_classif/DO;face_hallucin/DO;heterogen_face_recognit/DO;micro_express_recognit/DO
The U.S. corpus	face_verif/DO;3d_face_recognit/DO;pose_invari_face_recognit/DO;gait_recognit/DO;facial_express_recognit/DO;face_identif/DO;face_imag/DO;face_detect/DO;hierarch_bayesian_network/ME;swir_band/DO
Overlap rate	0.2

Table 1 shows that the differences lie in the areas of research emphasis. From the semantic neighbors, we see that China's research on facial recognition leans more towards specific individual traits. In contrast, U.S. research on facial recognition is more ocused on feature fdetection and image differentiation. Overall, China's research is conducted at a finer level of granularity.

These differences are closely related to privacy concerns. In China, historical practices have fostered greater acceptance of facial recognition technology, leading to detailed research. Conversely, U.S. citizens prioritize privacy protection [3]. Consequently, U.S. research focuses more on technologies that are less connected to personal identity.

This case study demonstrates that the semantic field constructed by semantic neighbors uncovers more detailed information.

The last paper in this section is by Qiu and Li, "Research on Paper Semantic Novelty Measurement Based on Large Language Model", they proposed a semantic novelty measurement model for scientific papers using a large language model to generate question and method words. Enhanced by LoRA and prompts, the model achieved high precision and recall, proving effective and robust, with optimal cost- effectiveness at 3,000 training samples..

4. Conslusion

The innovative approach we proposed that integrates a word embedding model with semantic field analysis vectors to investigate differences in semantics and research applications across various entities. This novel method surpasses traditional co-occurrence-based semantic field studies. Nonetheless, the study acknowledges certain methodological limitations, which is inherent to the word2vec model used in the analysis. Future research should focus on identifying more precise methods for representing entity semantics.

Acknowledgements

This study is supported by the Social Science Foundation of Jiangsu Province (No. 24TQB001).

Declaration on Generative AI

(by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Benko, A., &Lányi, C. S. : History of artificial intelligence. In Encyclopedia of Information Science and Technology, Second Edition (pp. 1759-1762). IGI Global(2009).doi:10.4018/978-1-60566-026-4.

[2] Saveliev, A., & Zhurenkov, D. : Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China. Kybernetes, 50(3), 656-675(2021). doi:10.1108/k-01-2020-0060..

[3] Zhang, S., Feng, Y., & Sadeh, N. Facial recognition: Understanding privacy concerns and attitudes across increasingly diverse deployment scenarios. In Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021) (pp. 243-262).doi:10.1037/t25710-000.