Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 2
CII 2023

A Computational Linguistic Approach to English Lexicography

Fan Yang

Kashi University, Kashi, 844000, China

Abstract. Focusing on computational linguistic approaches to English linguistics, this research explores how computational methods can be applied to dissect, understand and utilise the English language. We first looked at text analysis and processing, delving into natural language processing techniques such as text categorisation, sentiment analysis and machine translation, and their application to social media and automated text processing. In the area of lexicography and semantics, we explored how techniques such as distributed word vectors, semantic role labelling and sentiment analysis can deepen our understanding of vocabulary and semantics. We highlight the importance of these techniques in natural language processing tasks such as sentiment analysis and information retrieval. In addition, we focus on cross-language comparative and multilingual research, emphasising how big data and cross-language comparative research can reveal similarities and differences between languages and their implications for global linguistics. Finally, we explore corpus linguistics and big data analytics, highlighting the richness of linguistic data and tools they provide for linguistic research. Overall, this study highlights the importance of computational linguistic approaches to English linguistics and how they have transformed the way linguistics is studied and language technology has evolved. Future research trends will continue to drive the further development of computational linguistics methods, leading to a closer integration of linguistics with big data analytics and computational methods, creating more opportunities for the future of the field of linguistics.

Keywords: Computational Linguistics; Natural Language Processing; Text Analysis; Language Generation; Language Understanding.

1. Introduction

English lexicography has always been an important research area in the field of linguistics and natural language processing. With the rapid development of computer science and big data technology, the application of computational linguistics methods in English lexicography research has become more and more widespread, providing new tools and perspectives for us to better understand and analyse the properties and semantic relationships of English vocabulary [1]. This study aims to explore computational linguistic approaches to English lexicography and the importance and potential impact of these approaches in linguistics and practical applications.

As a global language, English has a rich and diverse lexical system, including synonyms, polysemous words, and implicit relations of word meanings. The study of English vocabulary not only contributes to the development of linguistic theories, but also has far-reaching impacts on application areas such as natural language processing, information retrieval, machine translation, and sentiment analysis [2]. However, traditional lexicographic research methods are often limited by data size and complexity, making it difficult to comprehensively understand and utilise the various features of English vocabulary [3]. The main purpose of this study is to explore the application of computational linguistics methods in English lexicographic research and to examine how these methods can provide new tools and perspectives to address various issues in lexicography.

The application of computational linguistics methods in English lexicography is valuable for advancing linguistic research, improving the performance of natural language processing applications, and solving practical problems [4]. They help improve technologies such as search engines, sentiment analysis tools, machine translation systems, intelligent assistants, etc., and also provide more data and analysis methods for linguistic research.

Through this study, we hope to gain a deeper understanding of how computational linguistic methods can enrich and expand the study of English lexicography, as well as provide important insights into the field for researchers, practitioners, and policymakers. We will also explore future research directions and perspectives of the field to generate more research interest and exploration.

2. Text Representation and Lexical Feature Extraction

In the study of computational linguistic approaches to English lexicography, text representation and lexical feature extraction are key steps for transforming textual data into computer-processable forms for further analysis and research.

Bag-of-words modelling is a classical text representation method, the basic principle of which is to regard text as an unordered collection of words, ignoring the order and grammatical structure of words [5]. Key points include the basic principles and concepts of the bag-of-words model. How to transform text into bag-of-words representation and decompose text into independent lexical items. Word Frequency (TF) calculation and Document Frequency (DF) calculation to determine the importance of words. Advantages of bag-of-words models, such as simplicity and efficiency, and their limitations, such as the inability to capture relationships between words [6]. Applications of bag-of-words models in text classification and clustering, and the corresponding algorithms and techniques.

TF-IDF is a text representation based on word frequency and inverse document frequency, which is used to measure the importance of a word in a collection of text. Key points include the calculation method of TF-IDF weights, including how to calculate word frequency and inverse document frequency [7]. The concept and role of inverse document frequency, which is used to measure the importance of a word in a whole collection of text. How to represent text using TF-IDF weights to capture the lexical features of each text.The application of TF-IDF in information retrieval to improve the relevance of retrieval results. Methods for improving and extending TF-IDF in different domains to suit the needs of specific tasks.

Distributed word vectors are a method of representing words by mapping them to a continuous vector space, allowing similar words to be closer together in the vector space. Key points include the fundamentals and concepts of distributed word vectors. an introduction to Word2Vec models, including continuous bag-of-words (CBOW) and skip-gram (Skip-gram) models [8]. How to train word vector models, including context windows and negative sampling methods. Spatial representation of word vectors and similarity computation for finding similar words [9]. Applications of word vectors to natural language processing tasks, such as lexical classification, sentiment analysis, and grammatical analysis. Other word vector models, such as FastText and GloVe, and their comparison and applications [10].

Deep learning methods that have demonstrated excellence in text feature extraction, including applications of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Key points include the application of Convolutional Neural Networks (CNN) to text feature extraction, such as the design of text convolutional layers and the use of convolutional kernels [11]. Textual representations of Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks (LSTM), including how to deal with sequential and textual data. Generation of word embeddings and sentence embeddings with methods such as Word2Vec and Doc2Vec [12]. Text classification and sentiment analysis using deep learning methods to mine sentiment information in text. Deep learning based text generation and summarisation methods such as text generation models and applications of Generative Adversarial Networks (GANs).

3. Lexical Semantic Analysis

Lexical semantic analysis is an important area in the study of computational linguistic approaches to English lexicography, aiming at understanding and extracting semantic information from words.

Applications of semantic role labelling in information extraction, question and answer systems and machine translation [13]. Challenges, such as polysemy and ambiguity, and improved methods. Lexical semantic relation extraction aims to automatically extract and represent semantic relations between words, such as superordinate-inferior relations and synonym relations. Key points include the definition and classification of lexical semantic relations, such as superclass and subclass relations, synonym relations, and antonym relations [14]. How to automatically extract and represent these relations using computational linguistics methods such as semantic networks and knowledge graph construction. Applications of semantic relation extraction in knowledge graph construction and information retrieval. Criteria and datasets for evaluating the performance of lexical semantic relation extraction systems.

Lexical disambiguation is the task of solving polysemous word problems where the goal is to determine the exact semantics of a word in a given context. Key points include the concepts and challenges of word sense disambiguation, including the problem of polysemous words and the importance of context [15]. Context-based word sense disambiguation methods, including supervised and unsupervised learning methods. Applications of word sense disambiguation to natural language processing tasks such as machine translation and information retrieval. Criteria and datasets for evaluating the performance of word sense disambiguation systems.

Lexical Sentiment Analysis is an important task aimed at identifying sentiment information in text, such as positive, negative and neutral sentiment. Key points include the fundamentals of lexical sentiment analysis, including sentiment vocabulary and sentiment polarity [16]. Sentiment analysis tasks such as sentiment classification and sentiment intensity analysis. How to use computational linguistics methods to automatically identify sentiment information in text. Applications of sentiment analysis to social media monitoring, product review analysis and public opinion analysis.

4. Computational Analysis of Vocabulary Learning Habits

Understanding and analysing the vocabulary learning habits of individuals and groups is of great importance in the study of computational linguistic approaches to English lexicography. Individual vocabulary learning histories are analysed in terms of vocabulary accumulation and frequency of use. Examine how computational methods can be used to track the vocabulary learning history of individual learners, such as those using learning apps or electronic dictionaries [17]. Exploration of the impact of vocabulary learning experiences on vocabulary comprehension and production to reveal learners' trajectories and strategies [18]. Vocabulary learning methods are analysed, including traditional methods such as word cards, root word and contextual methods, as well as modern methods such as online applications and learning resources. Computational analyses to study the effectiveness and efficiency of different learning methods so as to provide learners with more effective learning strategies.

Computational analyses of vocabulary learning outcomes, including vocabulary quantity, vocabulary quality and vocabulary depth. Assessment of vocabulary learning outcomes using computational tools such as vocabulary tests and language proficiency tests to quantify learners' level of vocabulary knowledge and skills [19]. Investigate the impact of vocabulary learning outcomes on language application skills, such as reading, listening and oral communication.

Analyse individual differences, e.g. differences in vocabulary learning between learners, including learning speed, learning strategies and learning outcomes. Conduct group analyses, including group dynamics and trend analyses of vocabulary learning, to reveal common features and differences among different groups of learners. Use data mining techniques to identify and understand common features and differences between different groups of learners to provide a basis for personalised learning.

5. Cross-Language Comparisons of Vocabulary Learning

Conducting cross-linguistic comparisons is an important area of research in the study of computational linguistic approaches to English lexicography, which aims to understand patterns and differences in vocabulary learning across languages. Compare lexical features such as vocabulary size, polysemy and morphological structure between different languages. Examine lexical similarities and dissimilarities between different languages. Examine differences in specific lexical categories in different languages, such as colour words, emotion words and verb categories. Compare vocabulary learning strategies of learners of different languages, including memorisation, contextual learning and morphological learning [20]. To explore the most effective vocabulary learning strategies chosen by learners in different language contexts. Examine the evolution of vocabulary learning strategies at different stages of language learning.

To compare the vocabulary learning outcomes, including vocabulary size, vocabulary depth and vocabulary quality, of learners of different languages. To examine the performance of different language learners on different vocabulary learning tasks, such as translation, phonological learning and text comprehension. To explore the effects of language exposure on cross-language vocabulary learning outcomes. Examine vocabulary learning tools and technologies used by different language learners, such as electronic dictionaries, online learning platforms and mobile applications. Compare the utility and effectiveness of different language learning tools to understand which tools are more advantageous in different language contexts. Explore the cross-cultural adaptation and personalisation features of cross-language learning tools.

6. Corpus Linguistics and Big Data Analytics

Corpus linguistics and big data analytics is a key area in the study of computational linguistic methods in English linguistics, aiming to utilise large-scale linguistic data to reveal linguistic patterns and develop linguistic theories. Basic concepts of corpus linguistics, including what a corpus is and how to build one. How to design and maintain large-scale linguistic datasets to support linguistic research. Applications of corpus linguistics to different linguistic subfields such as phonetics, syntax, semantics and lexicography. Fundamentals of big data analytics, including data collection, cleansing and storage. Data mining and machine learning methods for extracting useful information from large-scale linguistic data. Natural language processing techniques such as text classification, entity recognition and sentiment analysis for analysing textual data. How big data analytics methods can be applied to study linguistic phenomena and patterns. Big data applications of speech recognition and speech synthesis, such as voice assistants and voice search. Text data analysis, including the construction and utilisation of large-scale text corpora. Big Data applications in natural language processing tasks, such as machine translation, question and answer systems and information retrieval. Applications of big data analytics to the study of language change and evolution, such as the development and evolution of language models. How big data analytics can be used to test and validate linguistic theories, such as syntactic and semantic theories. The correlation between linguistic theories and actual language use, and the new laws of linguistic phenomena revealed through big data mining. The challenges and impetus of big data analytics to linguistic theories, such as the development of structuralism and generative grammar.

7. Conclusion

The study of computational linguistic methods in English linguistics provides a wealth of knowledge and tools for a deeper understanding of the structure, semantics and usage of language. Through the application of large-scale data analysis, machine learning, deep learning and natural language processing techniques, researchers are able to achieve unprecedented advances in the field of linguistics. Computational linguistics methods have become indispensable tools for studying linguistic phenomena and solving practical problems. They have brought new perspectives and approaches to the field of linguistics. Text analysis and processing techniques have been widely used

for natural language processing tasks such as text classification, sentiment analysis and machine translation. These techniques have had a profound impact on social media, search engines and automated text processing. Natural language generation and understanding systems continue to improve, enabling computers to better interact with humans in language. This has important implications for intelligent assistants, automated question and answer systems, and machine translation. Research in lexicography and semantics has benefited from techniques such as distributed word vectors, semantic role labelling and sentiment analysis. This contributes to a deeper understanding of the semantics and context of words. Big data and cross-linguistic comparative studies have fuelled comparative research across languages, revealing global commonalities and differences in linguistics. Corpus linguistics and big data analytics provide rich linguistic data and tools for linguistic research. They have had a profound impact on linguistic theory and practical applications. Future research trends will continue to drive the development of computational linguistics approaches, including more advanced natural language processing techniques, larger-scale corpus and data analyses, the deepening of cross-linguistic and multilingual research, and the convergence of linguistic theories with big data. Computational linguistics will continue to play a key role in the fields of language technology, social media, natural language processing, machine learning, and language education.

References

[1] Skoufaki, S., & Petrić, B. (2021). Exploring polysemy in the Academic Vocabulary List: A lexicographic approach. Journal of English for Academic Purposes, 54, 101038.

[2] Frankenberg-Garcia, A., Rees, G. P., & Lew, R. (2021). Slipping through the Cracks in e-Lexicography. International Journal of Lexicography, 34(2), 206-234.

[3] Dollinger, S. (2020). English Lexicography: A global perspective. The Handbook of English Linguistics, 525-546.

[4] Zhumasheva, K. B., Sapargaliyeva, M. E., Sarkulova, D. S., Kuzhentayeva, R. M., & Utarova, A. G. (2022). Representation of gender metaphor in lexicography as a reflection of culture. International Journal of Society, Culture & Language, 10(3), 151-162.

[5] Hargraves, O. (2021). Lexicography in the Post-Dictionary World. Dictionaries: Journal of the Dictionary Society of North America, 42(2), 119-129.

[6] Maziarz, M., Grabowski, Ł., Piotrowski, T., Rudnicka, E., & Piasecki, M. (2023). Lexicalisation of Polish and English word combinations: an empirical study. Poznan Studies in Contemporary Linguistics, 59(2), 381-406.

[7] Costa, R., Salgado, A. D. C., & Almeida, B. (2021). SKOS as a key element for linking lexicography to digital humanities. Information and Knowledge Organisation in Digital Humanities, 178-204.

[8] D'Souza, J., Hoppe, A., Brack, A., Jaradeh, M. Y., Auer, S., & Ewerth, R. (2020). The STEM-ECR dataset: grounding scientific entity references in STEM scholarly content to authoritative encyclopedic and lexicographic sources. arXiv preprint arXiv:2003.01006.

[9] Khumalo, L., & Nkomo, D. (2022). The intellectualization of African languages through terminology and lexicography: methodological reflections with special reference to lexicographic products of the University of KwaZulu-Natal. Lexikos, 32(2), 133-157.

[10] Bellandi, A. (2023). Building linked lexicography applications with LexO-server. Digital Scholarship in the Humanities, fqac095.

[11] Grimm, N. (2022). Documentary Approaches to Lexicography. In Current Issues in Descriptive Linguistics and Digital Humanities: A Festschrift in Honor of Professor Eno-Abasi Essien Urua (pp. 551-567). Singapore: Springer Nature Singapore.

[12] Pankratz, E., Arppe, A., & Lachler, J. (2022). Low hanging fruit and the Boasian trilogy in digital lexicography of morphologically rich languages: Lessons from a survey of Indigenous language resources in Canada. Nordlyd, 46(1), 193-204.

[13] Salgado, A. M. D. C. F. (2022). Terminological Methods in Lexicography: Conceptualising, Organising, and Encoding Terms in General Language Dictionaries.

[14] Costa, R., Salgado, A., Ramos, M., Almeida, B., Silva, R., Carvalho, S., ... & Romary, L. (2023). A crossroad between lexicography and terminology work: Knowledge organization and domain labelling. Digital Scholarship in the Humanities, 38(Supplement_1), i17-i29.

[15] Liu, S. A Comprehensive Review of Lexicography, Corpus Linguistics and Machine Translation. International Journal of Education and Teaching Research, 45.

[16] Tiberius, C., & Heylen, K. (2022). ELEXIS Pathfinder to Computational Lexicography for Developers and Computational Linguists. DARIAH-Campus.

[17] Chaga, A. (2023). Den 'Den' skoj: A Lexicographic Portrait of a Russian Microsyntactic Unit. In Literature, Language and Computing: Russian Contribution (pp. 21-30). Singapore: Springer Nature Singapore.

[18] Rojas Díaz, J. L. (2023). 'Arm's length' phraseology? Building bridges from general language to specialized language phraseology–a study based on a specialized dictionary of International Commerce and Economics in Spanish and English. Terminology, 29(1), 1-44.

[19] Munson, M. (2021). Lexicography, the Louw–Nida Lexicon, and Computational Co-Occurrence Analysis. Hutchings, Tim; Clivaz, Claire (Hg.): Digital Humanities and Christianity. An Introduction, Berlin, 169-192.

[20] Arias-Badia, B., & Torner, S. (2023). Bridging the gap between website accessibility and lexicography: information access in online dictionaries. Universal Access in the Information Society, 1-16.