DOI:10.6129/CJP.201909_61(3).0004
中華心理學刊 民 108，61 卷，3 期，231-246
Chinese Journal of Psychology 2019, Vol.61, No.3, 209-224

DeepLEX: Toward a Knowledge-yielding Approach and Resource for AI

Shu-Kai Hsieh and Yu-Hsiang Tseng
Graduate Institute of Linguistics, National Taiwan University

Deep learning and neural network has gained substantial progress in recent years. After the introduction of word embeddings, a form of distributional vector semantics, computers could better simulate the lexical semantic relationships between words. However, the hierarchical nature of human language and concepts are still difficult to modeled by current approach. In computational linguistics, researchers developed lexical resources from different theoretical perspectives. These language resources attempt to bridge the gap between syntagmatic relationships, which computers can readily modeled from data, and paradigmatic knowledge, that are not readily grasped by computers. These knowledge are essential for the capability to reason in an unfamiliar context with only few data, and are also vital to develop empathy of human emotions. The commonality of these capabilities involves the high context variance, in which individual, social and cultural context intertwined, render a great challenge for computers to learn in a data-hungry way. Current study considers, as one would argue in computational functional linguistics, lexicon as an explicit knowledge base of human language. It is human annotation aided by automatic extraction the essential building block of strong artificial intelligence. Moreover, the knowledge stored in lexicon not only contains the pairing between forms and meanings, it should also address the fluidity of formulae and the dynamics between form-meaning pairings. The goal of current study is thus to integrate and develop a novel lexicon model called DeepLex that includes multilevel lexical properties, such as linguistic, psychological and pedagogical. A web-based tool is also developed to help users to freely determine and annotate formulae in Chinese. Further applications of DeepLex is also discussed.

Keywords: *AI, computational lexicon, dialogue system, semantic representation*

Abstract

This paper proposes a dynamically integrated lexical model called DeepLEX. With its modularized architecture, the aim of DeepLEX is to provide a fine-grained yet scaled and multidimensional lexical resource that empowers language scientists to pursue a wide array of previously unanswerable research questions. It can also be used to foster AI applications.

When seeking to understand natural language tasks, the main focus should not only be on the "representation" of the message, but also on how the exchange of information, understanding, and corresponding speech acts are realized in human-computer communication. In this regard, the real battlefield of AI-language processing tasks may still be the Turing test in conversations. This can be confirmed by the tendency of the conversational agent to be highly valued by the academic community and the industry. However, existing dialogue systems, whether using template engineering or the neural network seq2seq machine learning mode, mostly presuppose the "real" existence of vocabulary (or word boundaries), and its static correspondence of meaning. Nonetheless, as functional linguistics reveals, "form and meaning pairs" are conventionalized in nature, as in the actual language used in spontaneous dialogue. Following the Wray's (2005) observation, a huge amount of our everyday language is formulaic and seems to be stored in (semi-fixed) chunks, in addition to being prefabricated. Therefore, in the proposed lexicon, we further integrate the "formulaic sequences," which may better cover vocabulary knowledge warehousing.

As for the meaning part, lexical data at different levels are flexibly modularized, such as in the syntax-semantics module, emotion module, discourse and pragmatic module, diachronic module, etc. The architecture does not serve as a theoretical assumption, but rather as a way to draw together various perspectives so that researchers from different fields can initiate new cooperation based on it. In addition to multidimensional data, the architecture also provides various quantitative measures (via application programing interfaces) for exploring the data in a reproducible manner. We also propose the "fluid annotation flow" framework, which is convenient for annotators wanting to select different granularity units and markup the varied meanings under different contexts, and automatically update the lexicon.

To demonstrate the possibilities of how the proposed DeepLEX can serve as an integrated lexical knowledge resource for language studies, an exploratory analysis and a pilot experiment using an AI dialogue system based on DeepLEX were conducted. The exploratory analysis demonstrated that in addition to frequency, "age" and sentimental polarity also co-predicted the reaction time in the lexical decision task; and by incorporating DeepLEX resources into the sequence to sequence neural network, the dialogue system thus trained could become more emotionally aware.

In summary, our approach expands on previous efforts and calls for an open collaboration in which lexical knowledge is semantically founded, symbolically operationalized, and empirically gleaned. This lexical resource highlights the importance of the nature of lexicons. The findings from this research provide insights and a basis for the further development of knowledge-yielding transparent AI systems.