日中韓辭典研究所
The CJK Dictionary Institute

LEXICAL RESOURCES FOR DEEP LEARNING

DeepLEX

by Jack Halpern

Revised: December 25, 2020

1. What is Deep Learning

"Deep learning" and "neural networks" are the buzzwords of the day -- they are the most advanced subdomains of artificial intelligence. Based on artificial neural networks, deep learning brings machine learning methods to a whole new level. It plays a major role in advanced technologies as diverse as speech recognition, cybersecurity, and machine translation. The field of natural language processing (NLP), for example, has greatly profited from neural networks, as exemplified by Google Translate's dramatic improvement in translation quality.

2. The benefits of lexical resources

Lexical resources such as large-scale computational lexicons are of great benefit in NLP for such deep learning tasks as dataset generation, the creation of word embeddings, and the implementing of hybrid techniques.

The CJK Dictionary Institute is engaged in the active development of very large-scale lexical resources, referred to as **Lexical Resources for Deep Learning,** or ***DeepLEX Resources*** for short, that can benefit deep learning systems in such domains as . named entity recognition (NER), cybersecurity, neural machine translation (NMT), and speech technology.

2.1 Neural Machine Translation (NMT)

NMT techniques have achieved significant improvements over traditional MT systems, but they do not perform well on relatively low frequency content words, especially named entities such as POIs and personal name variants. The integration of additional information such as discrete probabilistic lexicons into NMT systems has the potential to lead to substantial qualitative increases in accuracy scores (BLEU and NIST). See the following paper for details.

Very Large-scale Lexical Resources to Enhance Chinese and Japanese IR and NLP

2.2 Regularization

Regularization can play a major role in deep learning. Algorithms must perform optimally not only on trained data, but also on unknown input data such as orthographic variants of named entities (erg. 'Ichiroo' vs 'Itirou'). Regularization provides a set of techniques to prevent the overfitting in neural networks. It aims to improve the accuracy of deep learning models when facing completely new input data from the problem domain, such as named entity variants. Large-scale lexicons of named entity variants can be used for compressing vector data and for computing meaningful values for each variant, and has the potential to significantly contribute to higher accuracy.

2.3 Named Entity Recognition (NER)

NER traditionally uses rule-based approaches which work well when supported by domain-specific, exhaustive lexicons. However, since such lexicons are often incomplete, performance is less than optimal. This can be improved with AI-based approaches that automatically discover representations, but in data-rich domains such as romanized personal name variants in Chinese and Arabic, these approaches do not always achieve adequate recall and precision.

To achieve high accuracy, the most practical solution is the integration of comprehensive, hard-coded lexicons covering tens or hundreds of millions of entries, such as those provided by CJKI. In fact, research at the Moscow State University has shown that traditional machine learning techniques (CRF) outperformed neural networks models probably because it is the only technique that uses lexicon features. This demonstrates that lexicon-driven traditional techniques do have a role to play even in the age of neural networks; that is, named entity lexicons have not been supplanted by deep learning techniques.

2.4 Cybersecurity

Large scale entity lexicons can also play a major role in cybersecurity, not only in the recognition of ordinary named entities such as personal names, locations and organization names, but also in recognizing named entity types specific to the cybersecurity domain, which includes names of hackers, hacker groups, software products, viruses, and electronic gadgets (referred to as security entity recognition).

However, cybersecurity entity extraction models suffer from overreliance on machine learning algorithms and tend to ignore the special features of security named entities. Thus cybersecurity can significantly benefit from both traditional CRF-based NER using ordinary entity lexicons, as well as from security entity lexicons fine tuned to the recognition of security specific entities.

2.5 Pretrained models

These considerable potential for resolving problems with establishing word associations. A number of techniques for building these models have emerged in recent years, such as **BERT** (Bidirectional Encoder Representations from Transformers), **ELMo** and **Word2vec**. Building such pretrained models using our

DeepLEX Resources and combining them with other resources such as annotated corpora is a complex task, however it has the potential to lead to satisfactory results, especially for morphologically complex languages like Arabic.

2.6 Large-scale frequency dictionaries

Frequency dictionaries such as those in the CJKI's DeepLEX Resources can be used for constraining much of the complexity arising from orthographic variants, such as multiple versions of the same name (e.g. over 150 variants of Mohamed) or the multiple surface forms commonly used in Japanese.

3. DeepLEX Resources

CJKI's **DeepLEX Resources** include tens of millions of CJK named entities specifically designed to support NLP applications such as NER and speech technology. These consists of tens of millions of CJK and Arabic named entities, some of which are described below.

1. Chinese Personal Name Variants (CNV)

   This is a multilingual database of Chinese personal names in Mandarin and four Chinese dialects, with over 1.6 million Chinese seed names and more than ten million entries covering all the major and popular romanization systems.

2. Japanese Orthographical Database (JOD)

   This comprehensive database of Japanese orthographic variants enhances the accuracy of information retrieval and machine translation by disambiguating variants of identical related meanings, such as *neko* 'cat' written 猫, ねこ or ネコ and *kakiarawasu* 'write out, publish' written 書き著す, 書 著す, 書き著わす or 書著わす.

3. Japanese Personal Name Variants (JNV)

   This database contains approximately 3.5 million Japanese personal names and variants, including approximately 550,000 seed names, and covers all the major romanization systems, their variants, and hybrids.

4. Arabic Full-Form Lexicon (ArabLEX)
   This resource, covering over 600 million entries, provided exhaustive treatment of all inflected, declined and conjugated forms in Arabic, including part-of-speech codes, detailed grammatical attributes and romanized forms. This is the ultimate resource for Arabic NLP, and is ideally suited for NER, speech technology, and AI tasks like answer generation.

5. Database of Arabic Names (DAN)
   Our comprehensive database of nearly 6.5 million romanized Arab personal names and their variants, which also includes the corresponding names in Arabic in both vocalized and unvocalized Arabic.

6. Japanese-Multilingual Place Names and POIs (JMP)
   This is a large-scale database of Japanese place names and POIs (stations, schools, airports, etc.) in CJK, European, and other Asian languages. The data covers approximately 3.1 million entries spread over 14 languages and covers numerous POI types.

These resources are being leveraged by some of the world's largest IT companies in NLP and AI applications such as speech technology, morphological analysis, and machine translation, and in AI for applications such as natural language generation and speech technology. These resources can also support embedded MT, TTS and ASR technology by integrating them into the architecture of AI chips.

About Jack Halpern

Jack Halpern (春遍雀來), CEO of The CJK Dictionary Institute, is a lexicographer by profession. For sixteen years was engaged in the compilation of the New Japanese-English Character Dictionary, and as a research fellow at Showa Women's University (Tokyo), he was editor-in-chief of several kanji dictionaries for learners, which have become standard reference works.

Jack Halpern, who has lived in Japan over 40 years, was born in Germany and has lived in six countries including France, Brazil, Japan and the United States. An avid polyglot who specializes in Japanese and Chinese lexicography, he has studied 18 languages (speaks ten fluently) and has devoted several decades to the study of linguistics and lexicography.

The CJK Dictionary Institute

The CJK Dictionary Institute, Inc. (CJKI) specializes in CJK and Arabic computational lexicography. The institute creates and maintains CJK (Chinese, Japanese and Korean) and Arabic lexical databases currently covering approximately 50 million entries. Located in Saitama, Japan, CJKI is headed by Jack Halpern, editor-in-chief of the world-renowned New Japanese-English Character Dictionary and of various other CJK dictionaries.

CJKI plays a leading role in helping the IT industry penetrate the lucrative East Asian market by providing software developers with high quality dictionary data. This includes comprehensive databases of general vocabulary, proper nouns and technical terms for CJK languages, including Chinese dialects such as Cantonese and Hakka. CJKI also maintains databases and romanization systems of Arabic proper nouns, a large-scale Spanish-English dictionary, and various multilingual databases of proper nouns and geographic data.

CJKI has become one of the world's prime sources for CJK lexical resources. It is contributing to CJK and Arabic information processing technology by providing high-quality lexical resources and professional consulting services to some of the world's leading software developers and IT companies, including Fujitsu, Sharp, Sony, IBM, Google, Microsoft, Yahoo, Amazon and Baidu.