日中韓辭典研究所
The CJK Dictionary Institute

用于深度学习的词库资源

DeepLEX

作者：Jack Halpern（春遍雀來）
修订日期：2020 年 12 月 25 日

1. 什么是深度学习

“深度学习”及“神经网络”都是当下的热门词汇，因为它们体现了人工智能最为先进的分支领域。基于人工神经网络，深度学习将机器学习方式推升至全新高度，在语音识别、网络安全及机器翻译等各类先进技术中发挥了重要作用。例如，神经网络显著推动了自然语言处理（NLP）领域的发展，谷歌翻译在翻译质量方面的巨大提升就是极好示例。

2. 词库资源的益处

大规模的计算机词库等词库资源有助于 NLP 中的数据集生成、词嵌入创建及混合技术实施等深度学习任务。

日中韩辞典研究所正积极致力于开发大规模词库资源，即**面向深度学习的词库资源**，简称“***DeepLEX 资源***”。深度学习系统在命名实体识别（NER）、网络安全、神经机器翻译（NMT）和语音技术等方面均可以从 DeepLEX 资源中受益。

2.1 神经机器翻译(NMT)

与传统的 MT 系统相比，NMT 技术已取得长足的进步，但 NMT 在出现频率相对较低的内容词汇上表现不佳，特别是 POI 及人物姓名变体等命名实体。将离散概率词库等额外信息整合进 NMT 系统中可大幅定性提升准确率得分（BLEU 和 NIST）。详见下页。

强化中文及日文 IR 和 NLP 的非常大规模词库资源

2.2 正则化

正则化在深度学习中能发挥重要作用。算法不仅需要在训练过的数据上发挥最优性能，在未知输入数据上（如命名实体的正字法变体（erg. 'Ichiroo' vs 'Itirou'））也需要发挥最优性能。正则化是一套防止神经网络中过度拟合的技术。其旨在提升深度学习模型在面向来自问题领域（如命名实体变体）中的全新输入数据时的准确性。大规模的命名实体变体的词库可用于压缩矢量数据、计算每一变体的有意义取值，并可极大提高准确性。

日中韩辞典研究所
352-0001 日本埼玉县新座市东北 2 丁目 34-14 室
电话：048-473-3508 传真：048-486-5032
电子邮件：jack@cjki.org 网站： www.cjk.org

2.3 命名实体识别（NER）

NER 通常采用基于规则的方法，该方法在特定领域详尽词库的支持下效果良好。然而，由于这类词库通常不够完整，因此效果并不理想。可通过基于人工智能的自动发现表征方式来改善使用效果，但在数据量很大的领域，如中文及阿拉伯语的罗马化人名变体，此方法并不总能达到足够的召回率及精确度。

为实现高精确度，最实用的解决方案是整合涵盖数千万或数亿条词条的综合硬编码词库，如 CJKI 提供的词库。事实上，莫斯科国立大学研究表明，可能由于传统的机器学习技术（CRF）是唯一使用词库特征的技术，其效果优于神经网络模型。这表明，即使在神经网络时代，词库驱动的传统技术也仍然有一席之地；也就是说，命名实体词库尚未被深度学习技术所取代。

2.4 网络安全

大规模实体词库在网络安全领域也发挥了重要作用，它不仅可以识别一般的命名实体，如人名、地点和组织名称，还可以识别网络安全领域特有的命名实体类型，如黑客、黑客组织、软件产品、病毒及电子小工具的名称（简称安全实体识别）等。

然而，网络安全实体提取模型存在过度依赖机器学习算法的问题，并且往往忽略了安全命名实体的特殊性。因此，网络安全既可以从使用普通实体词库、基于 CRF 的传统 NER 中受益，也能从根据安全特定实体识别进行微调的安全实体词库中受益。

2.5 预训练模型

通过建立词语关联，能够促进问题的解决。近年来涌现了一些建立此类模型的技术，如 **BERT**（基于 Transformer 的双向编码器表征）、**ELMo** 和 **Word2vec**。使用我们的 DeepLEX 资源建立此类预训练模型，并将其与注释语料库等其他资源结合起来是一项复杂任务，但潜在效果令人满意，特别是对于像阿拉伯语这样形态复杂的语言。

2.6 大规模频率词典

频率词典，如 CJKI 的 DeepLEX 资源中的频率词典，可降低由正字法变体导致的诸多复杂问题，如同一名字的多个版本（如 Mohamed 的 150 多个变体）或日语中常用的多种表面形态。

3. DeepLEX 资源

CJKI 的 **DeepLEX 资源**包含数千万条日中韩命名实体，专门用于支持 NLP 应用，如 NER 及语音技术。它包括数千万条日中韩和阿拉伯语的命名实体，下文将描述部分内容。

日中韩辞典研究所
352-0001 日本埼玉县新座市东北 2 丁目 34-14 室
电话：048-473-3508 传真：048-486-5032
电子邮件：jack@cjki.org 网站： www.cjk.org

1. 中国人名变体 (CNV)
   这是包含使用普通话及四种中国方言的多语种中国人名数据库，其中有 160 多万个中国人名词源，超过 1000 万个条目，涵盖所有主要及流行的罗马化系统。

2. 日语异体词数据库 (JOD)
   这是全面的日语异体词数据库，通过消除相同的相关含义变体，提高了信息检索及机器翻译的准确性，如neko“猫”写成猫、ねこ或ネコ，kakiarawasu“写出、发布”写成書き著す、書著す、書き著わす或書著わす。

3. 日本人名变体 (JNV)
   该数据库包含约 350 万个日本人名及变体，包括约 55 万个词源名称，涵盖所有主要的罗马化系统、变体及混合体。

4. 阿拉伯语全变体数据库 (ArabLEX)
   该资源涵盖了 6 亿多词条，详尽地处理了阿拉伯语中所有的转折词、降音词和同源词形式，包括词性代码、详细的语法属性及罗马化形式。这是阿拉伯语 NLP 的终极资源，非常适合于 NER、语音技术和答案生成等人工智能任务。

5. 阿拉伯语人名数据库 (DAN)
   我们的综合数据库包括近 650 万个罗马化的阿拉伯人名及其变体，其中还包括相应带母音及不带母音的阿拉伯语姓名。

6. 日语多语地点数据库 (JMP)
   这是一个大规模的日本地名及POI（车站、学校、机场等）数据库，采用日中韩、欧洲和其他亚洲语言。其数据涵盖了约 310 万个条目，覆盖于 14 种语言，并涵盖了众多POI类型。

全球最大的一些IT公司均在使用此类资源进行NLP及人工智能应用，如语音技术、形态分析和机器翻译，也将此类资源运用于在人工智能领域，进行自然语言生成及语音技术应用等。此类资源还可以通过整合至人工智能芯片架构，以支持嵌入式MT、TTS和ASR技术。

日中韩辞典研究所
352-0001 日本埼玉县新座市东北 2 丁目 34-14 室
电话：048-473-3508 传真：048-486-5032
电子邮件：jack@cjki.org 网站： www.cjk.org

日中韓辭典研究所
The CJK Dictionary Institute

关于 Jack Halpern （春遍雀來）

日中韩辞典研究所所长春遍雀來是职业词汇学家。16 年来，他始终致力于《新日英汉字词典》的编纂工作。他曾任昭和女子大学（东京）的研究员，主编的若干供学习者使用的日文汉字词典已成为标准的参考用书。

春遍雀來在日本已生活了 40 多年，他出生于德国，曾在法国、巴西、日本和美国等共计六个国家生活过。他是一名饱含热情的多语言学家，专注于研究日语及汉语词汇学。他熟悉 18 种语言（可流利使用 10 种语言），几十年来一直热衷于语言学和词汇学的研究。

日中韩辞典研究所

日中韩辞典研究所（CJKI)专门从事日中韩及阿拉伯文计算机词汇学的研究。CJKI 创建并维护日中韩和阿拉伯文词库，目前涵盖约 5000 万个条目。CJKI 位于日本埼玉县，负责人为世界知名的《新日英汉字词典》及其他日中韩词典的主编春遍雀來。

CJKI 通过向软件开发商提供高质量的字典数据，在推动 IT 行业进入东亚市场方面发挥了主导作用。CJKI 的数据库覆盖日中韩语言中的一般词汇、专有名词、技术术语的综合数据库，包括广东话和客家话等中文方言。CJKI 还维护阿拉伯语专有名词数据库及罗马化系统、一部大规模西班牙语—英语词典以及含有各种专有名词和地名数据的多语种数据库。

CJKI 已经成为世界上主要的日中韩词库资源来源之一。通过向包括富士通、夏普、索尼、IBM、谷歌、微软、雅虎、亚马逊和百度等世界领先的软件开发商和 IT 公司提供高质量的词库资源和专业咨询服务，CJKI 为日中韩和阿拉伯语信息处理技术做出了卓越贡献。

日中韩辞典研究所
352-0001 日本埼玉县新座市东北 2 丁目 34-14 室
电话：048-473-3508 传真：048-486-5032
电子邮件：jack@cjki.org 网站： www.cjk.org