doi:10.1631/FITEE.1500067

题目：基于中文维基的大规模命名实体识别语料自动生成方法

目的：命名实体识别作为自然语言处理领域一项重要的基础性工作，当前主流方法是基于有监督的机器学习方法。该类方法依赖于特定语种和领域的标注语料，而语料的标注过程需耗费大量的人力、物力。本文提出一种基于中文维基的大规模命名实体识别（NER）语料自动生成方法。利用该方法能自动抽取并标记中文维基中的句子，从而为中文 NER 任务提供有效的语料支持。

创新点：本文根据中文维基的特点设计出四类启发式规则，并结合有监督的命名实体分类器，实现中文维基条目的命名实体类型的准确、全面识别；为避免缺失的维基链接引发的标注缺失，本文利用出链接的边界信息发现维基文档中的隐式指称项，并利用实体链接技术识别歧义指称项的实体类型；本文提出一种基于核心条目扩展的标注语料选择方法，实现测试数据的领域自适应。

方法：本文方法的整体流程如原文图 2 所示。该方法主要包括显式指称项的实体分类、隐式指称项的类型识别和标注语料选择三个主要步骤。在显式指称项的实体分类中，为实现准确、全面的实体类型识别，采用基于启发式规则与有监督实体分类器相结合的方法；在隐式指称项的类型识别中，提出一种新方法发现维基文档中的隐式指称项并识别歧义指称项的实体类型；在标注语料选择中，提出一种基于核心条目扩展的方法，实现测试数据的领域自适应。

结论：根据实验结果，采用本文方法能自动生成大规模的中文 NER 语料。此外，将生成语料与标准语料结合时，训练获得的 NER 模型性能更优。

关键词：NER 语料；中文维基；实体分类；领域自适应；语料选择