摘要

檢索或理解文件，首要工作是先對文件中的實詞（content word）進行標記，而具名實體（named entity）是實詞中最難識別的一類。近年來，藉由雙語語料（bilingual corpus）抽取資訊信息之技術研發正快速地推動自然語言相關研究領域之進展，因此如何從雙語語料中自動擷取互為對應之具名實體進而運用其所蘊含之訊息逐漸成為目前熱門之研究議題。本論文之目的在提出一個新的作法，有效地對雙語平行語料庫（parallel corpora）進行具名實體對應（named entity alignment），所採用之方法及論文貢獻簡述如下：

1. 對於中英文具名實體翻譯（translation）之處理，我們提出一個統計式片語翻譯模型（statistical phrase translation model），並將此模型機率函數表示成兩個獨立機率函數：詞彙翻譯機率函數（lexical translation probability function）及位置對應機率函數（position alignment probability function）。此作法之優點是統計式片語翻譯模型的參數可藉由給定之片語詞組資料自動訓練而得，且片語資料事先可不需經人工切割對應；除此之外，透過片語翻譯模型可有效降低具名實體翻譯候選詞組個數的產生。

2. 對於中英文音譯（transliteration）之處理，我們提出一個統計式音譯模型（statistical transliteration model），依據中英文發音結構特性，可有效地描述

音譯模型機率為音譯單元（transliteration unit）對應機率及音譯單元長度對應機率之組合關係。相對於前人作法，本法之優點是我們既不需英文音譯名詞之實際發音訊息，也不需人工給定音譯單元對應分數，且模型的參數只需藉由給定之資料自動訓練而得，如此，也使得我們的作法將來轉移至其他不同語言時，更加可行。

3. 同時，我們也引入其他知識訊息，可進一步的提高具名實體對應之精確率。透過中文人名辨識模型（Chinese person name model），可有效找出英文人名與中文人名之對應關係;藉由字串縮寫比對（abbreviation handling）模組，可協助找出翻譯時對應之中文具名實體簡稱;而英文簡稱擴展（acronym expansion）模組，則可藉由還原英文簡稱之原始名稱進而找出對應之中文具名實體。

4. 藉由大量實驗測試評估，在音譯名詞對應實驗上，我們分別對朗文字典例句、中英文科學人雜誌以及光華雜誌等雙語語料進行測試，實驗結果，詞精確率分別為 94.2%、94.0%及 93.0%。在具名實體對應實驗上，我們分別對光華雜誌雙語語料及香港新聞雙語語料進行測試，實驗結果，詞精確率分別為 91.13%及 80.18%；同時，我們也與 IBM Model 4 進行比較，無論在哪一個測試語料，實驗結果皆顯示我的作法優於 IBM Model 4。

Abstract

Named entities make up a bulk of documents. Extracting named entities is crucial to various applications of natural language processing. Although efforts to identify named entities within monolingual documents are numerous, aligning named entities in bilingual documents has not been investigated extensively due to the complexity of the task. In this dissertation, we introduce statistical phrase translation and transliteration models to align bilingual named entities in parallel corpora. In our approach, we model the process of translating an English named entity phrase into a Chinese equivalent using lexical translation/transliteration probabilities for word translation and alignment probabilities for word reordering. The method involves automatically learning phrase position alignment and acquiring word translation from a bilingual phrase dictionary and parallel corpora, and automatically discovering transliteration transformations from a training set of name-transliteration pairs. Unlike previous approaches, the proposed transliteration model does not involve the use of either a bilingual pronunciation dictionary for converting source words into phonetic symbols or manually assigned phonetic similarity scores between source and target words. The method for aligning bilingual named entities also involves

language-specific knowledge functions, including abbreviation handling, Chinese person name recognition, and acronym expansion. At run time, the proposed models are applied to each source named entity in a pair of bilingual sentences to generate and evaluate the target named entity candidates, and the source and target named entities are aligned based on the computed probabilities. Experimental results demonstrate that the proposed approach, which integrates statistical models with extra knowledge sources, is highly feasible and offers significant improvement in performance. The proposed methodology is applicable to a wide range natural language processing, such as machine translation, cross-language information retrieval, and bilingual lexicon acquisition. Finally, we conclude the proposed approach with an emphasis on the main contributions of aligning bilingual named entities and some directions on future work.