中文剖析系統

本院覽號

05T-1060526

公告日期

智財權狀態

know-how

摘要

句子的結構是語義分析及了解的必要訊息。要電腦具有智慧型的語言處理能力，例如機器翻譯、語言分析、語言了解、資訊抽取，電腦系統都必須先能分析句子結構。因此，中文句子自動剖析的工作成了語言理解不可或缺的技術。基本上句子自動剖析利用語法規律和斷詞後的文本做比對，找出可能的短語結構，由於存在歧義的短語結構，因此，如何利用結構出現的機率及檢測結構中詞與詞之間搭配的合理性成為解決結構歧義的方法。 本系統採用機率式無語境規律的模型(Probabilistic Context-free Grammar)為基本剖析架構並加入結構中詞彙搭配關係機率解決結構歧義。在結構決定之後，本系統可選擇是否對結構進行語義角色的指派。分詞與詞類標記採用本實驗室發展的中文斷詞與詞類標記系統。

技術優勢

本系統實作研究的二個主要部份為句法抽取與結構剖析。 1.句法抽取：研究如何從Sinica Treebank中抽取句法規則，並尋找出有效的語法普遍化及精確化方法，得到覆蓋率高且精確的句法規則，以加強中文剖析器的剖析效能。統計相關的規則機率、中心語機率值及中心語與搭配語機率，作為剖析器歧義結構挑選與機率統計的依據。 2.結構剖析：研究如何從無數的剖析歧義結構中有效率的找到最佳結構。除了利用規則機率外並考慮詞彙搭配的合理性作為歧義結構評估的方法。

應用範圍

完成中文句子析剖析系統，包含斷詞／斷詞標記／中文剖析／角色指派。

創作人

陳克健、馬偉雲

中央研究院
ACADEMIA SINICA