会议公告信息抽取系统

付宝宇 李侯运 吕克林
河南财经学院计算中心 河南科技情报研究所（郑州，450002）

Conference Communque Info. Extraction System

Fu Bao-yu, Li Hou-run an Lu ke-len
Computation Center of Henan College of Finance and Economic
Henan Science and tech. info. Institute

本文给出一个用于会议公告的信息抽取系统，为了问题的简化，该系统把问题领域限制在会议公告的范围。虽然基于知识的方法对于受限领域的信息抽取是很有效的，但是构造数量巨大的语言模板仍然相当困难，本系统通过句法分析模块的输出和脚本数据库，自动生成语言模板。系统可以基本满足应用要求。

信息抽取(Information Extraction) 也称为信息文摘，信息摘要(Information Abstraction) 系统，是可以由计算机代替人工，对各种有价值的图书、资料、公文、报告、文献进行自动摘要，产生一个精炼的浓缩的输出，从中摘出人们感兴趣的内容的系统。这种系统对于资料的保存、检索、浏览都十分有用，可广泛用于情报、信息、图书馆等部门，具有巨大的意义。

基于知识的语言处理技术已成功地应用于文本的信息抽取，我们的系统广泛使用了语言模板(Linguistic pattern)，此系统采用 J.Kim 的方法自动获取语言模板，使用一个用作训练的语料库，输入一批样本会议公告文本，用来构造语言模板，获取模板的方法是使用句法分析模块，根据句法分析模块的输出，或者重新反馈给句法分析模块，或者由此构造出一个新的初级模板，通过多级的语义约束优化，最终形成一个规范的模板加入知识库中。
把会议公告的训练文本输入到系统中，该系统的句法分析模块（本模块是系统的重点）可以产生三种输出：

1. 正确的信息抽取输出 ——说明已有对应的语言模板
2. 没有任何信息输出 ——说明没有对应的语言模板，系统生成对应的语言模板，加入知识库。本模块是系统的核心，实现比较复杂。
3. 不正确的信息抽取输出 ——说明语言模板不正确，则修改知识库

通过充分多的训练文本，就能生成初步实用的语言模板知识库，基于该知识库，系统可以不断丰富、修正、完善，最终接近达到实用程度。对于常见的会议公告，系统可以抽取出以下信息：会议名称、会议地点、开始时间、结束时间、召开部门、主持人、主要参加人、参加人、一般参加人、主要议题、主要观点、主要决议等，基本能够满足一般文档信息抽取要求。