作者简介及博士学位论文中英文摘要

论文题目：融合多种谓词信息的语义角色标注方法研究

作者简介：杨海彤，男，1986 年 1 月出生，2011 年 9 月师从于中国科学院自动化研究所宗成庆研究员，于 2016 年 7 月获博士学位。

中 文 摘 要

语义角色标注是一种自然语言处理领域的浅层语义分析技术。它以句子为单位，分析句子中的谓词与其相关成分之间的语义关系，进而获取句子所表达语义的浅层表示。由于语义角色标注可以提供较为简洁、准确、有益的分析结果，因此近年来受到了学术界的普遍重视，并已经成功地应用到信息抽取、自动问答、机器翻译等任务中。

在具体的实现中，语义角色标注以句子中的谓词为核心，分析句子中的其它成分与谓词之间的相互关系，因此谓词在句子的语义表达中处于核心的支配地位，其它成分均为谓词服务。但在现有的大多数研究工作中，谓词的作用仅仅体现在论元分类时作为一种特征，这显然与谓词在谓词－论元结构中的支配地位相悖。因此，本文的研究工作主要围绕如何深入挖掘谓词信息来改善现有的语义角色标注系统展开，本文重点关注了三种谓词信息：谓词先验信息、多谓词信息和双语谓词互补信息。为合理利用这三种信息，本文提出了以下方法：

1. 全局的语义角色标注生成式模型

句子中的谓词与它的语义角色组成了一个统一的整体，相互之间存在着紧密的联系。然而现有的语义角色标注系统却忽视谓词和语义角色之间的联系，每个候选论元的标注过程均独立进行，导致谓词与论元之间的关系也被割裂开来。但是，谓词既有一定的共性，比如对每个谓词来讲核心论元均不重复出现，又有自己独有的特性，比如谓词“销往”总是伴随一个地点论元，这些现象表明了谓词与语义角色之间的紧密联系。本文把这些联系看作是谓词的先验信息，合理地利用谓词的先验信息有利于提升语义角色标注系统的性能。为融入谓词的先验信息，本文用一个新颖的概念来表达谓词和论元之间的联系，并在此基础上提出了一种全局的语义角色标注生成式模型进行求解。实验结果表明，该方法可以有效地处理谓词与论元之间的联系，充分挖掘谓词本身的特性，使得语义角色标注系统的性能有显著的提升。

2. 基于判别式重排序的多谓词语义角色标注方法

现有的语义角色标注系统在分析一个句子时，依次独立地分析每个谓词，即给定句子中的一个谓词，识别出它的论元，然后对它所有的论元完成分类，之后再分析下一个谓词。可以看出该过程每次只关注句子中某个谓词的语义角色标注，却忽略了句中各个谓词之间的相互关系。而且一个句子中包含多个谓词的现象在日常中是普遍存在的。根据统计，在中文命题库中超过 80%的句子包含两个或两个以上的谓词。这些谓词位于同一个句子中，联合起来表达了句子的完整语义，那么它们的语义角色标注结果应具有一定的联系。本文分别调查了多谓词现象对于论元识别和论元分类的影响。具体地，在论元识别阶段，融入了与谓词相关的特征，可以有效地减少论元识别错误；在论元分类阶段，对于多谓词共享论元的分类本文提出了一种判别式重排序的方法，由于该方法充分考虑了多谓词和共享论元的全局信息，因而显著提升了共享论元分类的效果。

3. 基于对偶分解的双语语义角色标注方法

由于机器翻译等跨语言任务需要对双语平行语料进行语义分析，所以本文对双语语义角色标

注问题进行了研究。因为双语平行句对是互为翻译的，所以它们应具有等价的语义。语义的一致性反映在语义角色标注上是双语谓词应具有一致的谓词－论元结构。如果只利用单语语义角色标注系统对平行句对进行分析，会完全忽略双语语义的一致性。而且，双语谓词的谓词－论元结构一旦不一致的话，表明其中一个必然是错误的结果，这启发我们可以利用另一端的信息来改进这种错误情况。因此，本文认为双语谓词的语义角色标注之间存在广泛的互补信息。为了合理利用双语谓词的互补信息，本文提出了基于对偶分解的双语语义角色标注方法。实验表明，该方法可以显著提升双语语义角色标注的效果，并且效率较高。

综上所述，谓词在谓词－论元结构中处于核心地位，而现有的语义角色标注方法仅仅把谓词作为一种分类特征看待，明显与谓词的核心地位不符。因此本论文深入研究了如何利用谓词信息来改善现有的语义角色标注系统。具体地，本文分别研究了谓词先验信息、多谓词信息和双语谓词互补信息等三种谓词信息对于语义角色标注的帮助。实验表明，合理地利用这些谓词信息可以显著地提升语义角色标注的系统性能。

关键词：语义角色标注，谓词先验信息，多谓词信息，双语谓词，双语语义角色标注，对偶分解

Research on Methods of Mining Multiple Predicate Information for Semantic Role Labeling

YANG Haitong

ABSTRACT

Semantic Role Labeling (SRL) is a kind of shallow semantic parsing in Natural Language Processing. Given a sentence, it analyzes the semantic relations between the predicate and the arguments and the shallow semantic of the sentence can be obtained. Since SRL can provide simple, accurate and useful semantic analysis, it has caught much attention from researchers and has been applied in many NLP tasks such as information retrieval, question answering, and machine translation etc.

Semantic Role Labeling analyzes the semantic relations between the predicate and the arguments in a sentence and the predicate dominates the predicate-argument structure. In other words, the predicate dominates the semantic expression of a sentence and its arguments serve for the predicate. However, in existing methods, the predicate is just treated as a feature in argument classification, which could not match its dominated position of the predicate-argument structure. Therefore, this thesis has conducted researches on how to mine the predicate information to improve semantic role labeling. This thesis focuses on three kinds of predicate information: the predicate prior, the multi-predicate information and bilingual predicates' complementary information. To incorporate the three kinds of information into semantic role labeling, this thesis has proposed the following methods:

1. A global generative model for semantic role labeling

In a sentence, the predicate and its arguments compose of a whole and have close relations. But, existing methods ignore the relations between the predicate and its arguments. In these methods, each argument is classified independently. In fact, there are common points for predicates, like no duplication for core arguments, and characteristics, like a location argument accompanying “销往”. These phenomena show the close relations between the predicate and its arguments and these knowledge can be thought of the predicate prior information. We think the predicate prior is helpful for semantic role labeling. To incorporate the predicate prior information, we construct a novel concept called Predicate-Argument-Coalition (PAC) and based on PAC we have proposed a global generative model for semantic role labeling. The experimental results show that our method can handle the relations between the predicate and its arguments, fully mine the characteristics of each predicate and improve significantly the performance of semantic role labeling.

2. Discriminative reranking-based multi-predicate semantic role labeling. The existing systems of semantic role labeling, for a given predicate, recognize its arguments and perform argument classifications and then analyze the next predicate. In other words, the existing systems perform semantic parsing for each predicate independently. These systems only focus on one predicate's semantic parsing while ignoring other predicates in the same sentence. In the daily life, it is very common that a sentence contains multiple predicates. According to statistics, more than 80% sentences constrains more than two predicates. These predicates are in the same sentence and they convey the whole semantic of the sentence together. Thus, the semantic parsing of these predicates should have relations. In this thesis, we investigate multiple predicates' effects on argument identification and argument classification. In specific, we add some novel predicate-related features in argument identification which are proved to remove many identification errors; in argument classification, we propose a discriminative reranking approach which is proved to improve argument classification significantly with global information.

3. Bilingual semantic role labeling via dual decomposition

Since cross lingual tasks such as machine translation need bilingual semantic analysis, this thesis investigates the task of bilingual semantic role labeling. Bilingual semantic role labeling is a task of making semantic parsing on bilingual parallel corpora. The semantic equivalence of the parallel bi-texts means that they should have the same predicate semantic structure. A conventional way to perform bilingual SRL is using monolingual SRL systems to perform SRL on each side of bi-texts separately. But, this method completely ignore the consistence of bi-texts' semantics. And, if the results for a bi-text are inconsistent, there must be an incorrect one. We could correct the case by utilizing the information of the other end. Therefore, we think there are a lot of complementary information in the two sides of bi-texts. To utilize these information, we proposed a novel method called bilingual semantic role labeling via dual decomposition. Experimental results show that our method yields significant improvements over the state-of-the-art monolingual systems and also is very fast.

In summary, the predicate dominates the predicate-argument structure but in existing methods, the predicate is just treated as a feature in argument classification, which could not match its dominate position. Therefore, this thesis investigates how to utilize the predicate information to improve the system of semantic role labeling. In specific, we have investigated three kinds of predicate information: the predicate prior, multi-predicate information and bilingual predicate complementary information. Experimental results show that these predicate information are very helpful for improving semantic role labeling.

Key words: semantic role labeling, predicate prior information, multi-predicate information, bilingual predicates, bilingual semantic role labeling, dual decomposition