作者简介：

博士，中国科学院计算技术研究所副研究员，中国科学院大学岗位教授，硕士研究生导师。任中国中文信息学会青年工委委员，中国计算机学会中文信息技术专委会委员。研究方向为自然语言处理和机器翻译，在 ACL、IJCAI 和 CL 等机器学习和自然语言处理顶级国际会议和刊物上发表论文数十篇，培养及协助培养博士与硕士研究生十余名。当前研究兴趣为多语言处理、机器翻译与智能问答。

依存文法的跨语言同构化

1.引言

由于语言之间固有的词句法差异以及语言学家分析标准的差异，互译语句的语言分析结果之间不一定具有较好的一致性。对于许多跨语言应用如多语言检索和机器翻译而言，语言分析结果中特定语言单位或实体之间的关系，被期望在多种语言之间保持一致，即具有跨语言同构性。跨语言应用最乐观的情景是，存在充分同构化的多语言分析结果，用以进行跨语言应用系统的知识表示、特征提取和分类决策。因此，语言分析结果的跨语言同构性程度，是提升跨语言应用性能需要考虑的重要因素。我们旨在针对依存分析提出一种自动跨语言同构算法，提升依存文法在跨语言上层应用中的性能。本文仅对问题和方法进行简要阐述，详细的技术方案和实验分析请参考相应论文[1]。

2.依存文法同构化的难点与可行性

语言分析的跨语言同构化并不容易，难点在于既要提升语言分析的跨语言同构化程度，又要保证单语语言分析的内部一致性程度。换句话说，跨语言同构化要在不损失句法知识的前提下提升跨语言同构性。针对成分句法分析，研究者尝试过借助转换规则进行结构变换以提升跨语言同构性。相比于成分文法，依存文法是词汇化的且没有显式的层级结构。这种简洁的词汇化结构直接描述了词语之间的语义或句法关联，在用于一些跨语言应用如机器翻译时具有独到的优势。提升依存结构的跨语言同构性具有同样重要的意义，然而依存结构的这种特性却使得借助转换规则进行结构变换的策略难以适用。

现代依存分析模型通常涉及两个基本操作，决策和搜索。基于图的依存分析模型需要对词语之间的依存关系作出决策，并在全联通图中搜索最优生成树结构[2]；基于状态转移的模型需要依据当前状态决策下一步可能的移进归约操作，然后搜索最优的移进归约操作序列[3]。依存分析被因子化为基本判别决策，词语之间的依存关系或特定状态的下一步操作。我们设想，依存分析的跨语言同构化也可以因子化为基本判别决策的跨语言同构化。本工作针对依存分析提出一种自动跨语言同构算法。该算法在依存分析的判别决策层面，同步地对源语言和目标语言进行跨语言同构化调整，而非借助转换规则进行依存结构变换。在此基础上，通过迭代跨语言协同学习，在保证所得依存文法内部一致性的同时，逐步提升依存文法的跨语言同构性程度。

3.依存文法自动跨语言同构化方法

结构化分析任务可以因子化为基本判别决策，相应地，结构化分析结果的风格调整也可以因子化为基本判别决策的调整。对于依存分析的跨语言同构化而言，我们在依存分析的判别决策层面进行跨语言同构化。对于最大生成树模型，依存分析的基本决策为判别词语之间的依存关系。相应地，依存文法的跨语言同构化因子化为词语依存关系判别决策的同构化。本工作中，我们探究在判别决策层面进行跨语言同构化的方法，避免采用转换规则进行依存结构的调整带来的困难；同时采用迭代协同学习策略进行渐进式的跨语言同构化，保证所得的依存文法具有内部一致性。

依存结构的跨语言同构性程度，可以设计自动化的计算指标加以度量。依存句法分析以句子为单位，因此我们采用双语句对作为跨语言同构性程度度量的最小对象。同构性程度的计算需要依据词汇在两种语言之间的对齐信息，因此我们需要为双语句对获取词汇对应信息，这可以通过手工标注或自动化词语对齐算法如 GIZA++获得。对于给定的双语句对以及相应的依存结构，该评估指标借助词汇之间的对齐信息，计算这两个依存结构中依存边的一致性程度。对于双语篇章的跨语言同构性程度，我们简单定义为双语篇章内所有双语句对的同构性程度的平均值。

最大生成树模型的基本判别决策是依存分类器对候选依存边的分数评估。当前语言基本判别决策的跨语言同构化，可以定义为当前语言和参考语言的依存分类器给出的评估分数的某种融合运算。对于当前语言句子的两个词语，它们在当前语言依存分类器下的评估分数，可由依存分析模型的结构评分公式计算得到。对于这两个词语在参考语言依存分类器下的评估分数，可借助当前语言语句在参考语言中的对应语句（即对应译文）计算得到。综合这两个评估分数重新计算句中可能的依存边，运行解码搜索算法即可得到跨语言同构化调整之后的依存结构。在此同构化调整算法基础上，迭代协同学习同时对两种语言进行同构化调整，并以迭代训练过程逐步提升跨语言同构性程度和自身的内部一致性。

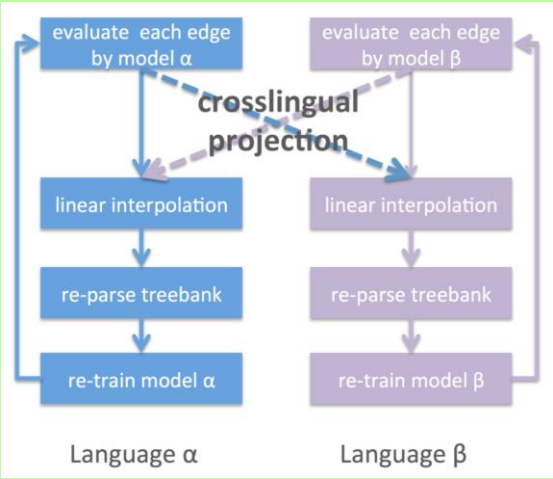

图 1. 基于迭代协同学习的跨语言文法同构化流程图

4.同构化依存文法的评估与应用

我们需要检验跨语言同构化操作的实际效果。跨语言同构化所得的依存文法不能是平凡的，比如扁平的依存结构，这种结果尽管可能具有很高的同构性程度，但却没有应用价值。对于跨语言同构化算法所得的依存文法，我们需要测量它的

知识含量，以及用于最终应用场景的效果。

在汉语和英语的依存文法任务上，跨语言同构化使得文法的跨语言同构化程度获得显著提升。同构化之后的文法经过标注风格迁移学习仍能取得和原始文法差不多的性能，意味着同构化操作并未明显损失句法知识。我们以基于依存句法的机器翻译为例，验证同构化依存文法在跨语言实际应用场景中的性能。实验显示，同构化文法能够带来显著的翻译性能提升。

5.参考文献

[1] Wenbin Jiang, Wen Zhang, Jinan Xu, and Rangjia Cai. 2016. Automatic Cross-Lingual Similarization of Dependency Grammars for Tree-based Machine Translation. In *Proceedings of EMNLP*, pages 501-510.

[2] Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of ACL*, pages 91–98.

[3] Joakim Nivre, Johan Hall, Jens Nilsson, Gulsen Eryigit, and Svetoslav Marinov. 2006. Labeled pseudoprojec- tive dependency parsing with support vector machines. In *Proceedings of CoNLL*, pages 221–225.