(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105573994 A
(43)申请公布日 2016.05.11

(21)申请号 201610053560.2

(22)申请日 2016.01.26

(71)申请人 沈阳雅译网络技术有限公司
地址 110003 辽宁省沈阳市和平区三好街55号1517室

(72)发明人 肖桐 朱靖波 张春良 高瑜泽

(74)专利代理机构 沈阳优普达知识产权代理事务所(特殊普通合伙) 21234
代理人 张志伟

(51)Int.Cl.
G06F 17/28(2006.01)
G06F 17/27(2006.01)

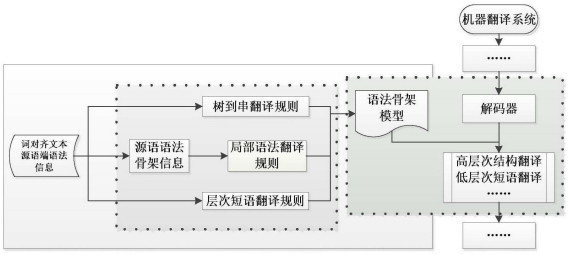
权利要求书2页 说明书12页 附图3页

(54)发明名称

基于句法骨架的统计机器翻译系统

(57)摘要

本发明涉及一种基于句法骨架的统计机器翻译系统,包括以下步骤:1)概率SCFG层次规则抽取方法抽取非句法翻译规则,用于待翻译句子非骨架部分的翻译;2)GHKM规则方法抽取句法翻译规则,用于待翻译句子的骨架部分的翻译;3)非完全句法翻译规则生成:利用句法翻译规则生成非完全句法翻译规则,结合非句法翻译规则和句法翻译规则,实现非句法翻译系统和句法翻译系统两种翻译系统优点的整合;4)模型生成,本发明系统应用句法翻译规则对句法骨架进行翻译以及长距离的调序问,应用非句法翻译系统的规则来处理低层次的词汇翻译和调序。模型易实现,并且效果显著。



CN 105573994 A

1.一种基于句法骨架的统计机器翻译系统,其特征在于包括以下步骤:

1)概率SCFG层次规则抽取方法抽取非句法翻译规则,用于待翻译句子非骨架部分的翻译:

利用抽取层次规则的启发式限制的方法,在经过词对齐但未进行句法分析的平行句对上抽取概率SCFG文法规则,利用层次短语规则即非句法翻译规则处理待翻译句子低层次结构的翻译;

2)GHKM规则方法抽取句法翻译规则,用于待翻译句子的骨架部分的翻译:

利用GHKM规则抽取方法在经过词对齐的平行句对和源语言端的句法分析结果上抽取GHKM规则,利用上述抽取的GHKM规则改写成句法翻译规则。利用句法翻译规则处理高层次骨架结构的生成及翻译;

3)非完全句法翻译规则生成:

利用句法翻译规则生成非完全句法翻译规则,结合非句法翻译规则和句法翻译规则,实现非句法翻译系统和句法翻译系统两种翻译系统优点的整合;

4)模型生成:

根据上述的非完全句法翻译规则,依据不同的翻译任务对句法翻译系统和非句法翻译系统的文法也就是翻译规则集合进行整合,生成非完全句法翻译推导,利用非句法翻译规则处理待翻译句子低层次的词组或短语的翻译,利用句法翻译规则完成待翻译句子的高层次句法骨架结构的翻译任务;利用非完全句法翻译规则指导骨架生成过程和翻译过程;收集非句法翻译规则、句法翻译规则以及非完全句法翻译规则生成一个具有大覆盖度的SCFG文法系统,并通过非完全句法翻译规则完成不同形式文法的结合。

2.按权利要求1所述的基于句法骨架的统计机器翻译系统,其特征在于:利用上述抽取的GHKM规则改写成句法翻译规则即句法翻译规则为:将抽取的GHKM规则,规则形式如下:

源语短语句法标记<以上述句法标记为根节点的源语句法子树片段>→目标语串

其中规则左部的“源语短语句法标记”为通过语言学句法知识所定义短语结构类型标签,即句法非终结符;规则左部的“句法子树片段”为句子句法分析树的片段,是树结构,其叶子节点可以为终结符词语或者非终结符,而这些非终结符必须属于源语句法分析中某一类句法标记;规则右部的“目标语串”为目标语终结符词语和非终结符构成的串,其非终结符标记与源语句法子树片段叶子节点的非终结符一一对应。

通过保持句法子树片段边界的非终结符及舍弃内部的树结构可以将上述GHKM规则改写为句法翻译规则

源语短语句法标记→<源语串,目标语串>

其中“源语串”表示源语终结符词语、非终结符构成和对应的“句法标记”构成的序列,该序列为句法规则所对应GHKM规则中源语句法子树片段的叶子节点序列;“目标语串”为由目标语终结符词语、非终结符和对应的“句法标记”构成的串,其非终结符标记与源语句法子树片段叶子节点的非终结符一一对应。

3.按权利要求1所述的基于句法骨架的统计机器翻译系统,其特征在于:利用非句法翻译规则和句法翻译规则生成非完全句法翻译规则,非完全句法翻译规则形式表述为:

源语短语句法标记→<源语串*,目标语串*>

其中,左部的“源语短语句法标记”为一个非终结符,“源语串*”为源语终结符词语、非终

结符和泛化标记X构成的串，“目标语串*”为目标语终结符词语、非终结符和泛化标记X构成的串，其非终结符标记与源语句法子树片段叶子节点的非终结符一一对应；

非完全句法翻译规则与句法翻译规则的区别在于：非完全句法翻译规则并不要求规则中所有的非终结符必须属于源语句法分析中某一类短语句法标记，而其中的部分非终结符被归约为X，表示该非终结符并不属于任何句法分析类型。

4.按权利要求1所述的基于句法骨架的统计机器翻译系统，其特征在于：实现非句法翻译系统以及句法翻译系统两种翻译系统优点的结合为：

通过源语端的句法翻译规则、非句法翻译规则和非完全句法翻译规则生成的大覆盖度SCFG文法在解码过程中创建句法骨架；

在上述句法骨架结构的生成过程中，捕获对源语言中句法结构中成分间的调序，将待翻译句子高层次的翻译任务分配给句法翻译系统来处理。并且把待翻译句子低层次的翻译任务分配给非句法翻译系统来完成；实现不同翻译系统的优点贡献到各自擅长的翻译任务中。

5.按权利要求1所述的基于句法骨架的统计机器翻译系统，其特征在于：依据不同的翻译任务对非句法翻译系统和句法翻译系统的文法进行整合为：在SCFG系统中，对每一个翻译规则推导进行权重计算，以便更准确的利用各种翻译规则推导，利用下式来计算每个翻译规则推导d的得分：

$$s(d)=\prod_{r_i\in d_s} w(r_i)\times\prod_{r_j\in d_h} w(r_j)\times lm(t)^{\lambda_{lm}}\times\exp(\lambda_{wb}\cdot|t|)$$

其中，s(d)为翻译规则推导d的得分，t为目标语端的字符串，d的得分则定义为多个因子的乘积，包括：

因子1：d中句法骨架(d_s)所包含的所有规则的权重乘积$\prod_{r_i\in d_s} w(r_i)$，其中$r_i$是$d_s$中的第i条规则，w(r*)是规则r*的权重；

因子2：d中非骨架部分(d_h)所包含的所有规则权重的乘积$\prod_{r_j\in d_h} w(r_j)$，其中$r_j$为$d_h$中的第j条规则，w(r*)是规则r*的权重；

因子3：n元语言模型lm(t)的指数加权得分$lm(t)^{\lambda_{lm}}$，λ_{lm}表示n元语言模型的权重；

因子4：词汇奖励$\exp(\lambda_{wb}\cdot|t|)$，其中exp(|t|)表示译文长度的e指数计算结果，当句子越长，这种“奖励”就越大，λ_{wb}是词汇奖励的权重。

基于句法骨架的统计机器翻译系统

技术领域

[0001] 本发明涉及一种统计机器翻译中对源语句法进行建模的技术，具体的来说是一种基于句法骨架的统计机器翻译系统。

背景技术

[0002] 统计机器翻译(Statistical Machine Translation,SMT)中，存在不同的翻译系统，比如基于短语和基于层次短语的非句法翻译系统，树到串以及串到树等句法翻译系统。在不同的翻译系统有着各自的优缺点，比如说，句法翻译系统在处理长距离以及各种成分间复杂的调序问题上有明显的优势，但是当句法翻译统统的翻译规则比较稀疏或者覆盖率比较低时，就会存在系统的鲁棒性问题，可能会导致翻译的效果比较差。并且已经证实了如果对句法系统进行简单的实现，其翻译结果并没有基于短语和基于层次短语等非句法翻译系统取得的效果好。另外，非句法翻译系统在翻译较短的句子片段时，准确率比较高，并且对短片段的层次结构也有比较好的调控能力。可是非句法系统在处理长距离的词序时表现能力比较差。

[0003] 目前，在处理目标语字符串的翻译(比如按照从句法分析数据中获得的树到串的映射关系来替换目标语表面的串)过程中，一种比较流行的方法就是利用源语端句法及句子结构上的信息来指导或者执行解码。这种方式和开始于90年代的基于字或基于词的翻译系统不同，它的源语句法模型是依赖输入的源语端句子的句法解析树生成的。这样做的好处是它可以加强模型处理长距离调动以及各种成分间复杂的调序问题的能力。

[0004] 另外，源语句法的使用在机器翻译中之所以可以有良好的表现是因为它具有呈现句子骨架结构(句法结构)的能力。如果我们用机器翻译系统类比人的翻译行为，这种骨架结构的翻译模式会表现得更为突出：在人为翻译过程中，对于一个给定的源语端输入句子，人们会利用句法上的先验知识首先在意识中对句子产生一个高层次上大致的句子结构或类型，然后根据这个句子结构或类型决定一些句子关键部分的翻译以及顺序，之后再完成词汇的选择及局部的调序工作。既然源语的句子骨架结构可以用源语的句法来表示，那么不免会产生以下问题：是否能够把源语的句法结构信息应用到它在翻译中作用效果最突出的地方?比如说，既能按照源语的骨架结构信息进行翻译，同时又能够利用非句法翻译系统完成良好短语翻译的优势?

[0005] 可是令人失望的是，尽管将句子骨架信息整合至机器翻译中的前景非常令人期待，但能够实现基于句法骨架的统计机器翻译系统尚未见到报道，另外句法系统和非句法系统有着不同的表示形式，在利用时也不尽相同。并且曾经有一些学者尝试利用人工标注的句法骨架数据，效果不好，并且实现过程复杂。

发明内容

[0006] 针对现有技术中句法翻译系统里不能对句子的短片段进行良好的翻译和调序以及规则稀疏而导致的系统鲁棒性问题，并且在非句法翻译系统中模型对长距离的句子成分

不能进行有效的调序问题，人工标注的骨架信息费时费力等问题，本发明要解决的技术问题是提供一种基于句法骨架的统计机器翻译系统，对源语高层次的句法骨架进行建模，并且对低层次的短语进行良好的翻译，同时提出一种句法骨架的新颖表示方式，以便机器翻译系统使用。

[0007] 为解决上述技术问题，本发明采用的技术方案是：

[0008] 本发明一种基于句法骨架的统计机器翻译系统，包括以下步骤：

[0009] 1)概率SCFG层次规则抽取方法抽取非句法翻译规则，用于待翻译句子非骨架部分的翻译：

[0010] 利用抽取层次规则的启发式限制的方法，在经过词对齐但未进行句法分析的平行句对上抽取概率SCFG文法规则，利用层次短语规则即非句法翻译规则处理待翻译句子低层次结构的翻译；

[0011] 2)GHKM规则方法抽取句法翻译规则，用于待翻译句子的骨架部分的翻译：

[0012] 利用GHKM规则抽取方法在经过词对齐的平行句对和源语言端的句法分析结果上抽取GHKM规则，利用上述抽取的GHKM规则改写成句法翻译规则。利用句法翻译规则处理高层次骨架结构的生成及翻译；

[0013] 3)非完全句法翻译规则生成：

[0014] 利用句法翻译规则生成非完全句法翻译规则，结合非句法翻译规则和句法翻译规则，实现非句法翻译系统和句法翻译系统两种翻译系统优点的整合；

[0015] 4)模型生成：

[0016] 根据上述的非完全句法翻译规则，依据不同的翻译任务对句法翻译系统和非句法翻译系统的文法也就是翻译规则集合进行整合，生成非完全句法翻译推导，利用非句法翻译规则处理待翻译句子低层次的词组或短语的翻译，利用句法翻译规则完成待翻译句子的高层次句法骨架结构的翻译任务；利用非完全句法翻译规则指导骨架生成过程和翻译过程；收集非句法翻译规则、句法翻译规则以及非完全句法翻译规则生成一个具有大覆盖度的SCFG文法系统，并通过非完全句法翻译规则完成不同形式文法的结合。

[0017] 利用上述抽取的GHKM规则改写成句法翻译规则即句法翻译规则为：将抽取的GHKM规则，规则形式如下：

[0018] 源语短语句法标记<以上述句法标记为根节点的源语句法子树片段>→目标语串

[0019] 其中规则左部的“源语短语句法标记”为通过语言学句法知识所定义短语结构类型标签，即句法非终结符；规则左部的“句法子树片段”为句子句法分析树的片段，是树结构，其叶子节点可以为终结符词语或者非终结符，而这些非终结符必须属于源语句法分析中某一类句法标记；规则右部的“目标语串”为目标语终结符词语和非终结符构成的串，其非终结符标记与源语句法子树片段叶子节点的非终结符一一对应。

[0020] 通过保持句法子树片段边界的非终结符及舍弃内部的树结构可以将上述GHKM规则改写为句法翻译规则

[0021] 源语短语句法标记→<源语串，目标语串>

[0022] 其中“源语串”表示源语终结符词语、非终结符构成和对应的“句法标记”构成的序列，该序列为句法规则所对应GHKM规则中源语句法子树片段的叶子节点序列；“目标语串”为由目标语终结符词语、非终结符和对应的“句法标记”构成的串，其非终结符标记与源语

句法子树片段叶子节点的非终结符一一对应。

[0023] 利用非句法翻译规则和句法翻译规则生成非完全句法翻译规则，非完全句法翻译规则形式表述为：

[0024] 源语短语句法标记→<源语串*，目标语串*>

[0025] 其中，左部的“源语短语句法标记”为一个非终结符，“源语串*”为源语终结符词语、非终结符和泛化标记X构成的串，“目标语串*”为目标语终结符词语、非终结符和泛化标记X构成的串，其非终结符标记与源语句法子树片段叶子节点的非终结符一一对应；

[0026] 非完全句法翻译规则与句法翻译规则的区别在于：非完全句法翻译规则并不要求规则中所有的非终结符必须属于源语句法分析中某一类短语句法标记，而其中的部分非终结符被归约为X，表示该非终结符并不属于任何句法分析类型。

[0027] 实现非句法翻译系统以及句法翻译系统两种翻译系统优点的结合为：

[0028] 通过源语端的句法翻译规则、非句法翻译规则和非完全句法翻译规则生成的大覆盖度SCFG文法在解码过程中创建句法骨架；

[0029] 在上述句法骨架结构的生成过程中，捕获对源语言中句法结构中成分间的调序，将待翻译句子高层次的翻译任务分配给句法翻译系统来处理。并且把待翻译句子低层次的翻译任务分配给非句法翻译系统来完成；实现不同翻译系统的优点贡献到各自擅长的翻译任务中。

[0030] 依据不同的翻译任务对非句法翻译系统和句法翻译系统的文法进行整合为：在SCFG系统中，对每一个翻译规则推导进行权重计算，以便更准确的利用各种翻译规则推导，利用下式来计算每个翻译规则推导d的得分：

[0031] $$s(d)=\prod_{r_i\in d_s} w(r_i)\times\prod_{r_j\in d_h} w(r_j)\times lm(t)^{\lambda_{lm}}\times \exp(\lambda_{wb}\cdot|t|)$$

[0032] 其中，s(d)为翻译规则推导d的得分，t为目标语端的字符串，d的得分则定义为多个因子的乘积，包括：

[0033] 因子1：d中句法骨架(ds)所包含的所有规则的权重乘积$\prod_{r_i\in d_s} w(r_i)$，其中$r_i$是$d_s$中的第i条规则，w(r*)是规则r*的权重；

[0034] 因子2：d中非骨架部分(d_h)所包含的所有规则权重的乘积$\prod_{r_j\in d_h} w(r_j)$，其中$r_j$为$d_h$中的第j条规则，w(r*)是规则r*的权重；

[0035] 因子3：n元语言模型lm(t)的指数加权得分$lm(t)^{\lambda_{lm}}$，λ_{lm}表示n元语言模型的权重；

[0036] 因子4：词汇奖励$\exp(\lambda_{wb}\cdot|t|)$，其中exp(|t|)表示译文长度的e指数计算结果，当句子越长，这种“奖励”就越大，λ_{wb}是词汇奖励的权重。

[0037] 本发明具有以下有益效果及优点：

[0038] 1.本发明系统使用了自己定义的特殊的句法结构信息(句法骨架或简称为骨架)进行翻译的方法，可以对源语高层次的句法骨架进行建模，以便机器翻译系统使用，它在一个框架中良好的结合了两个优点：1)应用句法翻译规则对句法骨架进行翻译以及长距离的

调序问题;2)应用非句法翻译系统的规则来处理低层次的词汇翻译和调序。

[0039] 2.本发明的模型非常灵活,可以通过一个单独简洁的句法解码范式来涵盖非句法、非完全句法或全句法翻译规则的推导,可以实现句法翻译规则和非句法翻译规则之间双向逐渐过度,使翻译系统可以在句法翻译系统和非句法翻译系统之间有选择性的使用翻译系统。因此,非句法翻译系统和句法翻译系统可以被视为利用此方法得到的两种特例,模型易实现,并且效果显著。

[0040] 3.本发明系统也适用于一般基于同步上下无关文法(SCFGs)框架的翻译系统,可以在一个支持SCFG文法解码器的翻译系统中简便的实现,并证实对系统的翻译进行加速。

[0041] 4.本发明定义了一种新颖的骨架结构表示方式,是首例对句法骨架信息进行自动获取,它可以在句法翻译规则、非完全句法翻译规则和非句法翻译规则的指导下,实现骨架信息的自动获取,避免了标注骨架信息浪费的大量人工劳动。

[0042] 5.本发明与传统的句法翻译系统不同,在翻译系统的解码过程中,该发明实现了先对句法结构框架的翻译,并对调序进行控制,然后在良好的句法骨架下实现局部片段的非句法性翻译,这在目前的翻译系统中是第一次使用该种方法。

附图说明

[0043] 图1为本发明系统的模型框架图;

[0044] 图2为本发明系统中抽取非句法翻译规则和句法翻译规则的样例图;

[0045] 图3为本发明从一样本句法中产生句法骨架的过程图;

[0046] 图4为本发明中基于树的系统解码一条句法翻译规则的过程图;

[0047] 图5为骨架深度对翻译质量的影响图示;

[0048] 图6为不同系统产生翻译结果的对比图。

具体实施方式

[0049] 下面结合说明书附图对本发明作进一步阐述。

[0050] 如图1所示,本发明一种基于句法骨架的统计机器翻译系统包括以下步骤:

[0051] 1)概率SCFG层次规则抽取方法抽取非句法翻译规则,用于待翻译句子非骨架部分的翻译:

[0052] 利用抽取层次规则的启发式限制的方法,在经过词对齐但未进行句法分析的平行句对上抽取概率SCFG文法规则,利用非句法翻译规则即非句法翻译规则处理待翻译句子低层次结构的翻译;

[0053] 2)GHKM规则方法抽取句法翻译规则,用于待翻译句子的骨架部分的翻译:

[0054] 利用GHKM规则抽取方法在经过词对齐的平行句对和源语言端的句法分析结果上抽取GHKM层次式规则,利用上述抽取的GHKM规则改写成句法翻译规则即句法翻译规则,处理待翻译句子的高层次结构,也就是句子句法结构的句法性翻译;

[0055] 3)非完全句法翻译规则的生成:

[0056] 利用句法翻译规则生成非完全句法翻译规则,并结合非句法翻译规则的使用,实现非句法翻译系统以及句法翻译系统两种翻译系统优点的结合;

[0057] 4)模型生成:

[0058] 根据上述的非完全句法翻译规则，依据不同的翻译任务对非句法翻译系统和句法翻译系统的文法(翻译规则集合)进行整合，生成非完全句法翻译推导，通过非完全句法规则，对不同翻译任务进行识别，利用非句法翻译规则处理文本低层次(词组或短语)的翻译，利用句法翻译规则和非完全句法翻译规则完成文本高层次(句法结构)的翻译任务；收集非句法翻译规则、句法翻译规则以及非完全句法翻译规则生成一个具有大覆盖度的SCFG文法系统。

[0059] 步骤1)中，概率SCFG层次规则抽取：本发明利用经过词对齐，但未进行句法解析的平行句对上，利用抽取层次短语规则的启发限制方法抽取概率SCFG文法规则，利用层次短语规则即非句法翻译规则处理待翻译句子低层次结构的翻译；

[0060] 步骤2)中，在源语言句法树信息指导下，利用词对齐平行句对数据抽取GHKM规则，并改写成SCFG式的句法翻译规则，即将抽取的GHKM规则，从如下的规则形式：

[0061] 源语句法短语标记(源语串属性源语串句法结构源语串)→目标语译文

[0062] 通过保持句法树片段的边界的非终结符并舍弃内部的树结构改写为句法翻译规则的形式：

[0063] 源语短语句法标记→<源语串，目标语串>

[0064] 其中“源语串”表示源语终结符词语、非终结符构成和对应的“句法标记”构成的序列，该序列为句法规则所对应GHKM规则中源语句法子树片段的叶子节点序列；“目标语串”为由目标语终结符词语、非终结符和对应的“句法标记”构成的串，其非终结符标记与源语句法子树片段叶子节点的非终结符一一对应。

[0065] 步骤3)中，利用源语言端句法信息，获取句法骨架信息，通过对句法翻译规则和非句法翻译规则的调控和改编，得到非完全句法翻译规则，非完全句法翻译规则的形式为：

[0066] 源语短语句法标记→<源语串*，目标语串*>

[0067] 其中，左部的“源语短语句法标记”为一个非终结符，“源语串*”为源语词(终结符)、非终结符和泛化标记X构成的序列，；“目标语串*”为由目标语词(终结符)、非终结符和泛化标记X构成的串，其终结符标记与源语句法子树片段叶子节点的非终结符一一对应；

[0068] 非完全句法翻译规则与句法翻译规则的区别在于：非完全句法翻译规则并不要求规则中所有的非终结符必须属于源语句法分析中某一类短语句法标记，而其中的部分非终结符被归约为X，表示该非终结符并不属于任何句法分析类型。

[0069] 对于每一条句法翻译的规则，可以把它的形式进行改写，得到非完全句法翻译规则，具体方式是通过把规则右部的一个或两个非终结符泛化成X，并且保持左部不变，可以转化为非完全句法翻译规则。

[0070] 在句法翻译规则、非句法翻译规则、非完全句法翻译规则收集完全之后，利用所有规则生成一个较大的SCFG文法系统，通过非完全句法翻译规则实现待翻译句子解码过程中推导的指导，并且产生对应的句法结构，在不同句子层次中利用不同翻译方式的优点。处理低层次(比如短语)的翻译时可以利用非句法翻译系统的优点，处理高层次(比如句法结构)的翻译任务时能够利用句法翻译系统的优点。

[0071] 实现非句法翻译系统以及句法翻译系统两种翻译系统优点的结合为：

[0072] 通过生成的大覆盖度的SCFG文法系统，利用非完全句法翻译规则，实现从句法翻译系统到非句法翻译系统的逐渐过渡，在推导过程中创建句法骨架；

[0073] 利用上述非完全句法翻译规则和句法翻译规则捕获对待翻译句子中不同组成成分间的调序，并且把低层次的翻译任务分配给非句法翻译规则来处理；将高层次骨架部分的翻译任务分配给句法翻译规则和非完全句法规则来处理。

[0074] 步骤4)中，根据上述的非完全句法翻译规则，依据不同的翻译任务对非句法翻译系统和句法翻译系统的文法进行调控，生成三种类型规则组成的大覆盖度的SCFG文法系统，在SCFG系统中，不但可以对句子骨架成分进行良好的调序，并且实现了句子的句法骨架的生成，其中会对每一个翻译规则推导进行权重计算，以便更准确的利用各种翻译规则推导，利用下式来计算每个翻译规则推导d的得分：

[0075] $$s(d)=\prod_{r_i\in d_s} w(r_i)\times\prod_{r_j\in d_h} w(r_j)\times lm(t)^{\lambda_{lm}}\times \exp(\lambda_{wb}\cdot|t|)$$

[0076] 其中，s(d)为翻译规则推导d的得分，t为目标语端的字符串，d的得分则定义为多个因子的乘积，包括：

[0077] 因子1:d中句法骨架(d_s)所包含的所有规则的权重乘积$\prod_{r_i\in d_s} w(r_i)$，其中$r_i$是$d_s$中的第i条规则，w(r*)是规则r*的权重；

[0078] 因子2:d中非骨架部分(d_h)所包含的所有规则权重的乘积$\prod_{r_j\in d_h} w(r_j)$，其中$r_j$为$d_h$中的第j条规则，w(r*)是规则r*的权重；

[0079] 因子3:n元语言模型lm(t)的指数加权得分$lm(t)^{\lambda_{lm}}$，λ_{lm}表示n元语言模型的权重；

[0080] 因子4:词汇奖励$\exp(\lambda_{wb}\cdot|t|)$，其中exp(|t|)少表示译文长度的e指数计算结果，句子越长，“奖励”越大，λ_{wb}是词汇奖励的权重。

[0081] 解码应用：

[0082] 本模型在解码中应用时，通过利用生成的大覆盖度的SCFG同步上下文无关文法对源语端待翻译的句子进行句法解码，在分析的过程中利用非完全句法规则和句法规则对待翻译句子按照句法骨架的结构进行分析，在分析的过程中，产生句子的句法骨架，并利用生成的大覆盖度SCFG同步上下无关文法中规则的目标语推导部分产生目标语端的翻译。每个片段如果有非完全句法翻译规则对应的推导，则可以获得局部片段的结构信息，如果没有找到对应的非完全句法翻译规则，模型会在推导空间(包含句法翻译规则、非句法翻译规则、非完全句法翻译规则)寻找最佳翻译推导。

[0083] 在本发明中，基于句法骨架的机器翻译模型框架大体可以分为三部分：规则获取、模型生成、模型应用等。模型框架如图1所示。

[0084] 首先采用上面所述方法在双语对齐数据和源语句法树信息上，利用不同的方式抽取不同类型的翻译规则，然后依据源语句法特征，改写部分句法翻译规则，生成恰当的非完全句法翻译结构推导，连接各种不同类型的推导规则。最后在解码中利用骨架模型根据不同层次翻译任务寻找合适的推导方式。

[0085] 一.翻译规则获得：

[0086] 本发明中，不同的规则采用不同的方法抽取：

[0087] 1)非句法翻译规则抽取：

[0088] 由于本发明是基于SCFG文法之上实现的，对于SCFG文法规则，可以使用下面形式进行表达：

[0089] LHS→〈α,β,～〉

[0090] 其中LHS是一个非终结符，α和β分别是源语端和目标语端由终结符和非终结符组成的词序列，～则表示α和β中非终结符的一一对应关系。

[0091] 对于非句法翻译规则，利用抽取层次规则的启发式限制的方法，在经过词对齐但未进行句法分析的平行句对上抽取概率SCFG文法规则，对于获得的概率SCFG文法，一个给定的翻译句子可以通过寻找最可能、概率最大的规则来推导进行解码。图2给出了一个抽取非句法翻译规则的例子，其中非终结符只被标记为X。如果有一些这样的SCFG规则组成的序列集合可以完整的覆盖并推导出源语句子，则认为它是此源语句子的一个SCFG推导文法。比如图中的规则h_5、h_1和h_3能够产生出一个句子对的推导。

[0092] 2)句法翻译规则获取：

[0093] 非句法翻译规则的形式和正规的句法(句法)翻译规则的形式基本上是一样的，只是非句法翻译规则不是按照(源语端或目标语端的)句法的约束生成的。如果利用任意一边(源语端或目标语端)的句法信息进行约束，我们可以得到符合句法信息的推导规则，也就是句法翻译规则，为此，我们可以利用下述方式获取句法翻译规则。

[0094] GHKM规则抽取：

[0095] 为了生成句法形式的句法规则，本发明利用主流的方法----利用源语端的句法树信息作为约束和指导，在有词对齐信息的双语句对上抽取GHKM规则。

[0096] 在抽取GHKM的方法中，本发明在从源语言句法树到目标语串之上建模，一条GHKM规则是由源语片段s_r，目标语片段t_r和它们片段(源语片段和目标语片段)中非终结符的对应关系组成的，例如下式是一条GHKM规则：

[0097] VP(VV(提高)x_1:NN)→increase x_1

[0098] GHKM规则改写：

[0099] 本实施例将上述规则形式改写成SCFG规则形式，具体操作是保持最前端的非终结符注释不变，抛弃非终结符内部的树结构信息，比如：

[0100] VP→〈提高NN_1,increase NN_1〉

[0101] 其中，VP为动词短语，VV为动词词性，NN为名词词性，x_1为非终结符，NN_1为词性是名词的一个变量。

[0102] 发明中参考SCFG规则对GHKM规则进行转化，由于所有的非终结符都被源语言端的句法标签标记，所以应用生成句法翻译规则时都被会受到正确句法的约束。

[0103] 图2中给出了从一个源语树和目标语串对中抽取句法翻译规则的过程，本实施例忽略了原始GHKM规则的多层次树结构，但保留了规则前端的节点，所以这样的改写操作可以让系统对新句子翻译结果有一个比较好的产生能力。

[0104] 此外，利用句法翻译规则进行解码可以看做是SCFG句法分析过程。一种比较流行的方法就是串解析(或基于串的解码)，这种方法可以在一个表格解码器中对输入句子进行解码(比如，CYK解码器)。并且在测试集有源语端解析信息情况下，我们可以利用树解析(或基于树的解码)的方法对解析树进行解码。这种情况下，由于所有的推导都必须遵循输入的

句法解析树，源语端句法信息可以被看为用来施加硬约束，以增加准确性。

[0105] 3)非完全句法翻译规则(本发明定义)获取

[0106] 非句法翻译系统和句法翻译系统都有各自的优缺点，比如，非句法翻译模型在词汇选择和调序方面有着遵循词汇化规则的突出能力，但是在处理复杂的成分运动时会有很多约束。句法翻译类的模型能够通过语言学上的句法注释来描述成分的层次性的运动，并且它在高层次的句法调序上也有出色的表现。可是这两种模型都存在稀疏和有限覆盖度问题。

[0107] 在理想情况下，可以把两种模型的优势应用到其作用程度最大的地方：1)句法翻译模型能够处理高层次的句法骨架的生成和句法成分之间的调序；2)非句法翻译规则能够处理低层次的词汇翻译和调序。为了达到这个目的，本发明提出了一种能够在一个模型里结合两种优点的方法。在翻译中重新利用非句法翻译和句法翻译的句法，并且开发了一种新型的规则---非完全句法翻译规则，用来把句法翻译规则和非句法翻译规则过渡性的连接起来。

[0108] 如果一条规则的左部(LHS)是一个源语端的句法标签，并且右部(RHS)至少有一个非终结符带X标志。下面是一条非完全句法翻译规则：

[0109] VP→<提高X_1，increase X_1>

[0110] NT→<提高X_2，increase X_2>

[0111] 其中左部代表了一个动词短语(VP)，右部和标准的非句法翻译规则一样，包含了非终结符X。可以在一个部分非句法翻译推导中应用这种规则，并且产生一个以VP为根节点的推导规则。然后句法翻译的规则可以像平时在句法机器翻译系统中一样代替这个VP推导，从而实现了从句法翻译系统到非句法翻译系统的过度。

[0112] 二、骨架模型生成

[0113] 由于非完全句法翻译规则能够把非句法翻译规则和句法翻译规则连接起来，所以可以利用这两种所有的规则来组建非完全句法翻译推导规则，构成可以生成句法骨架的文法系统，也就是骨架模型的基础。图3给出了一个从非句法翻译规则，句法翻译规则以及非完全句法翻译规则构建的推导。在此推导中，非句法翻译规则(h_3，h_6和h_8)被应用到低层次的翻译。通过在X的部分推导上应用句法规则(非完全句法翻译规则p_3和句法翻译规则r_1和r_4)，建立一个符合句子句法骨架的推导。

[0114] 本发明中这种句法结构是通过源语端的句法规则((图3右上角)创建的，它被称作句法骨架。它大体上是一种具有高层次句法并且叶子节点上有终结符或非终结符的树片段。通过使用这种骨架结构，可以很容易地捕获到“对NP VP”中成分间的调序，并且把低层次的翻译(“回答”和“表示满意”)分配给非句法翻译规则来处理。

[0115] 为了获取非完全句法翻译规则，使用一种简单却直接的方法。对于每一条句法翻译的规则，通过把右部(RHS)的一个或两个非终结符归约成X，并且保持左部(LHS)不变，可以把它转化成非完全句法翻译规则。比如说基于树的系统解码一条句法翻译规则的过程图(图4)中的r_5(VP→<对NP_1 VP_2，VP_2 with NP_1>)，可以得到三条非完全句法翻译规则：

[0116] VP→<对X_1 X_2，X_2 with X_1)

[0117] VP→<对X_1 VP_2，VP_2 with X_1>

[0118] VP→<对NP_1 X_2，X_2 with NP_1>

[0119] 一旦所有的规则包括，非句法翻译规则，句法翻译规则和非完全句法翻译规则准备完毕，就利用它们建立一个更大的SCFG推导文法并把它应用到解码器中。利用权重化的对数线性方法来计算规则的权重。和标准化的基于SCFG模型一样，对于LHS→<α，β，～>有以下几个特征：

[0120] 1.翻译概率P(α|β)和P(β|α)使用相关的频次进行估算，这两个概率分别是正向翻译概率和反向翻译概率。

[0121] 2.词汇的权重Plex(α|β)和Plex(β|α)使用启发式学习的方法进行估算。

[0122] 3.对于非句法翻译规则，句法翻译规则和非完全句法翻译化规则的规则奖励(exp(1))是分别不同的。

[0123] 4.定义了指示胶水规则，词汇规则和非词汇化规则的指示器，能够允许模型学习选择特定的规则。

[0124] 5.源语端非完全句法翻译规则中非终结符X的数目(exp(#))，它控制着模型冒犯句法的相容度。

[0125] 本发明在模型中定义了推导权重(得分)。定义d是上述句法的推导。为了把句法性的规则(也就是句法翻译规则和非完全句法翻译规则)和非句法翻译规则区别开，定义d是一个元组<d_s,d_h>，其中d_s是骨架结构的局部推导，d_h是用来组建d剩余部分推导的规则集合。例如，在图2中，d_s＝{r_4,r_1,p_3}，另外d_h＝{h_6,h_8,h_3}。

[0126] 定义t是目标语端编码的字符串，然后d的得分就可以定义为拥有n-gram语言模型lm(t)的连乘积和词汇奖励exp(|t|)的规则权重的结果。

[0127] $$s(d)=\prod_{r_i\in d_s} w(r_i)\times\prod_{r_j\in d_h} w(r_j)\times lm(t)^{\lambda_{lm}}\times \exp(\lambda_{wb}\cdot|t|)$$

[0128] 其中w(r*)是规则r*的权重，λ_{lm}和λ_{wb}分别是语言模型和词汇奖励的特征权重。

[0129] 另外，对于本发明模型，框架非常灵活，它特别的包括了句法翻译和非句法翻译模型。比如，d如果只是由非句法翻译规则组成(也就是说$d_h=\varphi$)，那么它就是一条非句法翻译推导。同样，如果一个推导d只是由句法翻译规则组成(也就是说$d_s=\varphi$)，那么它就是一条句法翻译式推导。本发明阐明的就是如何用非完全句法翻译规则引导到非句法翻译和句法翻译式的推导空间。解码器可以根据模型得分从扩大的推导规则中选择最好的推导。

[0130] 三、模型在解码中应用：

[0131] 本发明的模型在使用时可以看做是一个串解析的问题，因为它使用源语言端的句法规则对源语端的文本串进行解析，使用目标语端的规则推导信息来生成目标语的翻译结果。所以，翻译结果可以被当作是由规则推导产生并且具有最高得分的目标语串。此发明中，系统是利于基于CYK解码器上实现的，并且解码器已经利用了beam search和cube pruning技术，能够使用经过同步二值化方法获得的二值化规则。

[0132] 由于会有大量的非完全句法翻译规则被引入，导致解码速度非常的慢。为了提速解码系统，进一步使用几种剪枝方法对搜索空间进行剪枝，减小搜索空间。首先，丢弃那些作用范围大于3的词法或非完全句法翻译规则。之所以清除这些规则是因为它们是降低解码速度的一个主要原因，而且它们对最后的翻译结果并没有很大的帮助。另外，舍弃那些右部(RHS)只有非终结符X的非词法规则和非完全句法翻译规则。在大多数的情况下，这种类

型的规则并不能起到句法上的限制引导作用。例如说，规则VP→<X_1X_2，X_2X_1>存在的太普遍，如果在两个没有任何词法或句法迹象的连续块中引入一个VP句子成分，是非常不理智的，因为这样做并不能起到什么效果。

[0133] 除了对规则进行剪枝外，还可以用一个参数w_s来控制句法骨架的深度。如果赋给w_s一个很小的值，那么系统会被强制的使用一个更小的句法骨架(和更少的句法规则)。在极端情况下，如果参数w_s＝0，系统则会回退成一个典型的非句法翻译系统；同样地，如果参数值w_s＝+∞，系统可以考虑任意深度的句法骨架。所以我们可以在测试集上对参数w_s调优来寻找一个平衡点。

[0134] 为了加速系统，我们还应用一些树解析的技术。除了源语句子，我们还把源语的句法解析树加进解码器。首先我们利用非句法翻译系统中普遍使用的非句法翻译规则对源语句子进行解析，但是我们处理源语中与句法树成分对应的片段时，并没有对应用规则的距离进行限制。然后，我们在句法分析树上利用句法翻译规则。如果源语端的一条句法翻译规则可以匹配到一个输入树片段，则：1)这条规则会被转化成非完全句法翻译规则(见第3部分)；2)句法翻译以及对应的非完全句法翻译规则会被添加到规则列表里，这些列表和与源语句法树片段对应的CYK网格单元链接。图4给出了一个解码器中树匹配的例子。之后，剩余的解码步骤(比如说构建翻译超图，语言模型交叉)会正常处理。这种方法可以有效地匹配解码需求的(非完全)句法翻译规则，而且，不需要对规则进行二值化处理。由于源语句法树所给出的是硬约束，作为一个权衡处理，我们会引入一些对句法敏感的推导。

[0135] 四、实验

[0136] 本发明在英汉(en-zh)以及汉英(zh-en)翻译上实验他们的方法。

[0137] 1)baseline系统实验设置

[0138] 本发明使用从NIST12 OpenMT里挑选的274万汉英双语句对。在利用GIZA++工具让双语文本产生双向词对齐之后，本发明使用grow-diag-final-and的方法获得对称化的词对齐文件。对于句法分析，本发明首先使用伯克利parser对两边数据分别进行处理，然后利用流行的最左推导方法对句法分析树进行二值化，以便测试集上更好的产生式。基于句法(或句法翻译)规则从整个训练数据集中抽取，并且规则中最多只能有五个非终结符。而对于非句法翻译系统，层次性规则(非句法翻译)是从94万的句子子集上抽取，并且每条规则中的非终结符不超过两个，而短语规则则是从整个训练集上抽取。这里所有的规则都是使用开源工具包NiuTrans获取的。

[0139] 本发明训练了两个5元语言模型：一个是在英语Gigaword数据中的新华部分和双语数据的英语部分上训练的，这个模型使用在汉英的翻译系统中；另一个是在汉语Gigaword数据的新华部分和双语数据的汉语部分上训练的，这个模型被应用到英汉翻译系统中。所有的语言模型都使用修正过的Hneser-Ney平滑方法进行平滑。

[0140] 对于汉英翻译系统，本发明分别在新闻领域和网上数据对系统进行评价。本发明的调优集(新闻领域：1198个句子、web数据：1308个句子)是引用NIST机器翻译04-06的评测数据和GALE数据。测试集(新闻领域：1779个句子，web：1768个句子)则包含NIST08、12机器评测和08-progress中所有的新闻领域和网络数据的评测数据。对于英汉翻译系统，本发明的调优集(995个句子)和测试集(1859个句子)分别是SSMT07和NIST MT08汉英翻译记录的评价数据。所有源语端句法分析树都使用和处理训练数据一样的方法进行处理。

[0141] 2)基于句法骨架的机器翻译系统实验

[0142] 本发明按照模型在解码中应用部分提及到的方法来实现他们的CYK解码器。默认设置下，实验中使用到了串解析，初始情况下，把参数w_s设置到$+\infty$。所有的特征权重都使用MERT的方法进行调优。由于MERT有得到局部最优结果的可能，所以我们对每个实验进行了5次操作，并且每次都赋予不同的初始特征值。在评价部分，我们分别使用未修改的BLEU4和未修改的BLEU5对汉英以及英汉翻译系统进行评价。

[0143] 3)基于句法骨架的机器翻译系统实验结果

[0144] 表1是实验结果，其中基于句法骨架的系统被简写成SYNSKEL。首先可以看到SYNSKEL系统在3个测试集上都有明显的提高。使用CTB式的句法分析树获得了一个平均在0.6以上的BLEU值得提高，经过二叉化的句法树可以得到平均在0.9以上的BLEU值改善。并且利用解析树的方法可以很好地在正常使用非句法翻译规则时应用(部分)句法规则，获得了很好的结果。它获得了和串解析方法相当的BLEU值。然而，在一个二叉化的森林中放入更多的树对结果并没有什么提高效果。这些有趣的结果表明，在一个已经很大的推导空间里，很难通过考虑更多经过二叉化的可选句法结构来引入一些新颖的推导。

[0145]

类型		汉-英(新闻) 调优集 (1198)	汉-英(新闻) 测试集 (1779)	汉-英(网络) 调优集 (1308)	汉-英(网络) 测试集 (1768)	英-汉 调优集 (995)	英-汉 测试集 (1859)	平均 提升性能 测试集
非句法翻译系统		35.70	31.76	27.29	22.61	33.12	30.59	0
CTB树库	句法翻译系统	34.41	30.76	25.59	21.69	32.29	30.20	-0.77
	非句法+ 源语语法特征	36.02	31.99	27.35	22.76	33.34	30.90	+0.23
	非句法+ 源语语法SAMT型规则	36.09	32.09	27.47	22.86	33.46	30.99*	+0.32
	SYNSKEL	36.34*	32.43*	28.00*	23.15	33.70	31.50*	+0.71
	SYNSKEL (树解析方式)	36.44*	32.35	27.95	23.11	33.49	31.32*	+0.61
二叉化树	句法翻译系统	34.82	31.21	25.83	21.88	33.03	30.69	-0.39
	非句法+ 源语语法特征	36.09	31.98	27.42	22.87	33.40	30.97	+0.29
	非句法+ 源语语法SAMT型规则	36.14*	32.18*	27.46	22.90	33.40	30.92	+0.35
	SYNSKEL	36.70*	32.75*	28.09*	23.29*	33.82*	31.77*	+0.95
	SYNSKEL (树解析方式)	36.67*	32.64*	27.92	23.39*	33.95*	31.66*	+0.91
	SYNSKEL (森林解析方式)	36.77*	32.56*	28.05*	23.45*	33.95*	31.79*	+0.94

[0146] 表1不同系统下的实验结果

[0147] 另外，在不同骨架最大深度(也就是参数w_s)下研究了系统的结果。图5说明了太大的骨架并不一定能够获得更好的结果，其中BLEU为评价翻译质量的指标。控制参数$w_s \leq 5$时使用骨架系统能够获得令人满意的提高，和全部使用骨架系统时相比，减少了将近27％的解码时间。

[0148] 不同各类规则的使用率如表2所示，可见本发明定义的非完全句法翻译类型的规则使用率最高，并且取得了良好的翻译效果。

[0149]

推导类型	汉-英(新闻)	汉-英(网络)	英-汉
句法.	1.4%	0.6%	14.4%
非句法.	19.7%	26.6%	11.6%
非完全句法.	78.9%	72.8%	74.0%

[0150] 表2调优集上不同推导的使用率

[0151] 4)实验结果分析

[0152] 本发明在试验之后研究了一下系统调用不同类型推导的频率。表2展示了在三个不同的任务中，系统选择非完全句法翻译推导以及非句法翻译推导时的倾向。英汉翻译任务中表现出对句法和非完全句法翻译推导的重度使用，紧跟其后的是汉英翻译的新闻领域以及网络数据翻译任务。这个结果在一定程度上反映了分析质量在不同语言和领域数据是有差异的。

[0153] 1.翻译质量提升：

[0154] 试验结果表明，本发明基于句法骨架的系统被简写成SYNSKEL。首先可以看到SYNSKEL系统在3个测试集上都有明显的提高。使用CTB式的句法树获得了一个平均在0.6以上的BLEU值得提高，经过二叉化的句法树可以得到平均在0.9以上的BLEU值改善。并且利用解析树的方法可以很好地在正常使用非句法翻译规则时，应用句法翻译规则和非完全句法翻译规则，获得了很好的结果。它获得了和串解析方法相当的BLEU值。

[0155] 2.良好的调序控制：

[0156] 根据本发明中给出的对比实验数据可知，图6调优集中排列最前的5个翻译结果，用来对比同一个调优集上不同的翻译结果。另外，非句法翻译系统输出的翻译结果凌乱，并且调序也是错的，句法翻译翻译系统对本来句法解析就困难的“对$_{3}$……困难$_{20}$”结构的翻译效果也比较差。对比之下，SYNSKEL系统使用了跨度源语单词“还$_{2}$……担忧$_{22}$”的所有上述规则，并且由于是在非句法翻译系统中，所以系统对被局部非句法翻译推导覆盖的“对$_{3}$……困难$_{20}$”结构的翻译结果也比较好。

[0157] 3.句法结构更好识别：

[0158] 图6的底部展示了一个由这条规则推导产生的真实翻译例子，可以看到SYNSKEL系统中这个规则覆盖了源语单词“对$_{3}$……年薪$_{20}$”，并且成功的识别出“…的…”调序结构。注意，虽然非句法翻译系统中也有这样的规则$X\rightarrow\langle X_1$的X_2, X_2 of $X_1\rangle$能够翻译“…的…”(de)的结构。但是当跨度变大时，比如翻译一个表较长的词序列“该$_{8}$……美元$_{15}$”，非句法翻译系统就会丧失这种调序的能力。

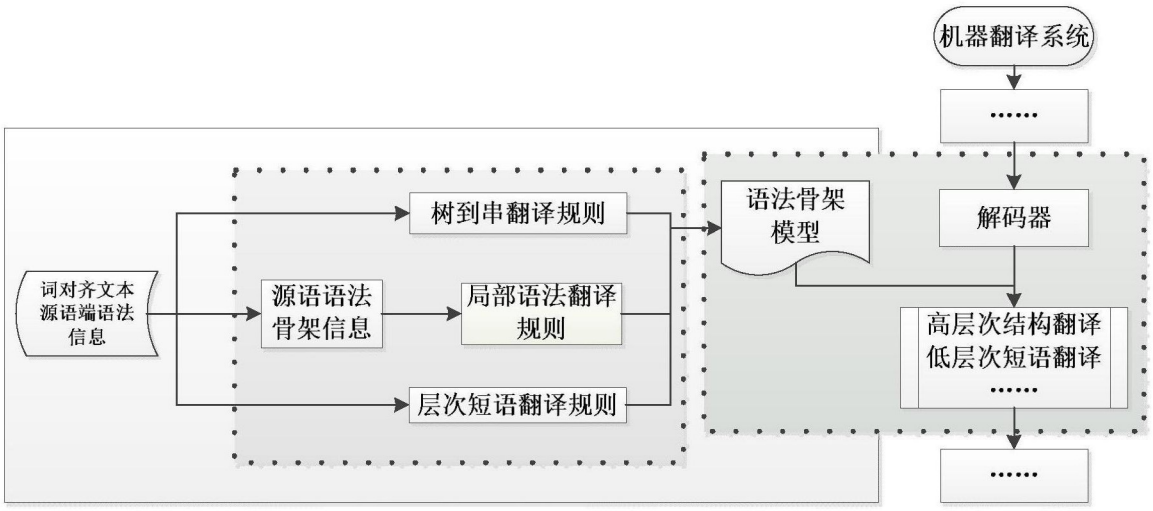
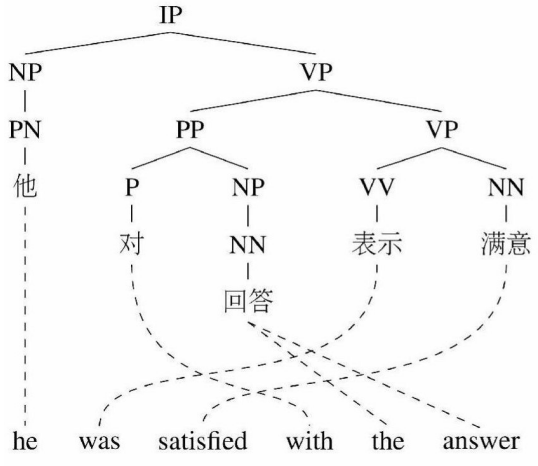


图1



非句法翻译规则

h_1	X → ⟨他, he⟩
h_2	X → ⟨对, with⟩
h_3	X → ⟨回答, the answer⟩
h_4	X → ⟨表示 满意, was satisfied⟩
h_5	X → ⟨X_1 对 X_2 表示 满意,
...	X_1 was satisfied with X_2⟩

从GHKM转化的句法翻译规则

r_1	NP → ⟨ 他, he ⟩
r_2	NP → ⟨ 回答, the answer ⟩
r_3	VP → ⟨ 表示 满意, was satisfied ⟩
r_4	IP → ⟨ NP_1 VP_2, NP_1 VP_2 ⟩
r_5	VP → ⟨ 对 NP_1 VP_2, VP_2 with NP_1 ⟩
...	

图2

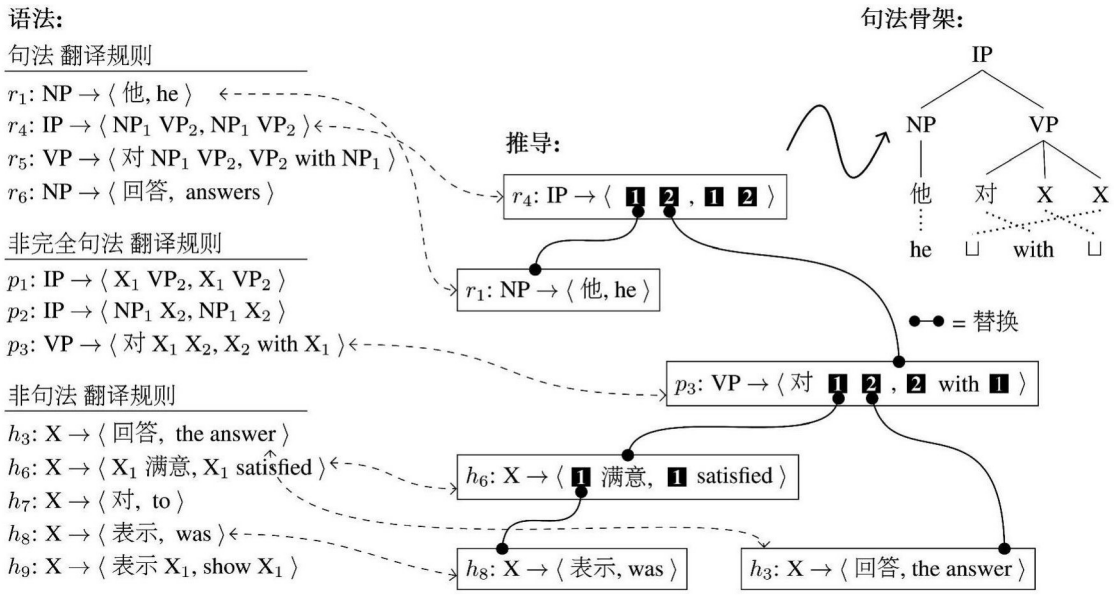

语法:
句法 翻译规则
r_1: NP → ⟨ 他, he ⟩
r_4: IP → ⟨ NP$_1$ VP$_2$, NP$_1$ VP$_2$ ⟩
r_5: VP → ⟨ 对 NP$_1$ VP$_2$, VP$_2$ with NP$_1$ ⟩
r_6: NP → ⟨ 回答, answers ⟩
非完全句法 翻译规则
p_1: IP → ⟨ X$_1$ VP$_2$, X$_1$ VP$_2$ ⟩
p_2: IP → ⟨ NP$_1$ X$_2$, NP$_1$ X$_2$ ⟩
p_3: VP → ⟨ 对 X$_1$ X$_2$, X$_2$ with X$_1$ ⟩
非句法 翻译规则
h_3: X → ⟨ 回答, the answer ⟩
h_6: X → ⟨ X$_1$ 满意, X$_1$ satisfied ⟩
h_7: X → ⟨ 对, to ⟩
h_8: X → ⟨ 表示, was ⟩
h_9: X → ⟨ 表示 X$_1$, show X$_1$ ⟩
推导:
r_4: IP → ⟨ 1 2 , 1 2 ⟩
r_1: NP → ⟨ 他, he ⟩
p_3: VP → ⟨ 对 1 2 , 2 with 1 ⟩
h_6: X → ⟨ 1 满意, 1 satisfied ⟩
h_8: X → ⟨ 表示, was ⟩
h_3: X → ⟨ 回答, the answer ⟩
= 替换
句法骨架:
IP
NP VP
他 对 X X
he ␣ with ␣

图3

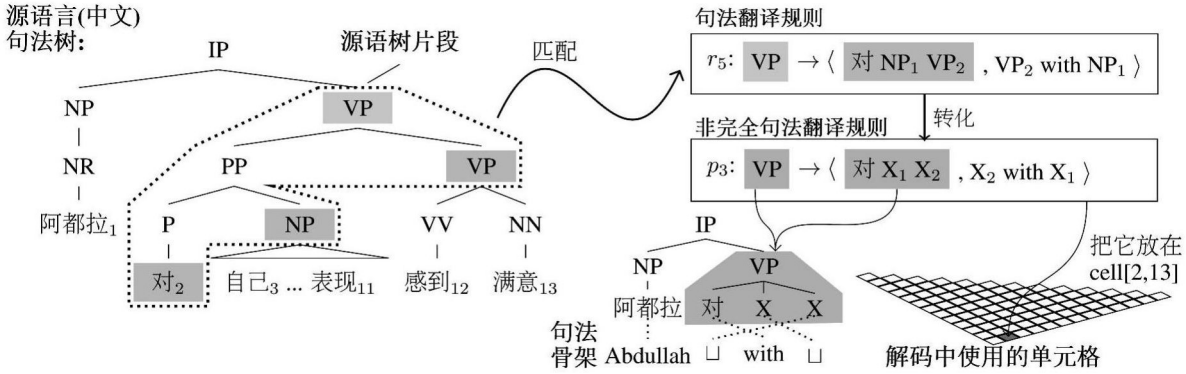

源语言(中文)
句法树:
IP
NP
NR
阿都拉$_1$
VP
PP
VP
P
NP
VV
NN
对$_2$
自己$_3$ ... 表现$_{11}$
感到$_{12}$
满意$_{13}$
源语树片段
匹配
句法翻译规则
r_5: VP → ⟨ 对 NP$_1$ VP$_2$, VP$_2$ with NP$_1$ ⟩
转化
非完全句法翻译规则
p_3: VP → ⟨ 对 X$_1$ X$_2$, X$_2$ with X$_1$ ⟩
IP
NP
VP
阿都拉
对 X X
句法
骨架 Abdullah ␣ with ␣
把它放在
cell[2,13]
解码中使用的单元格

图4

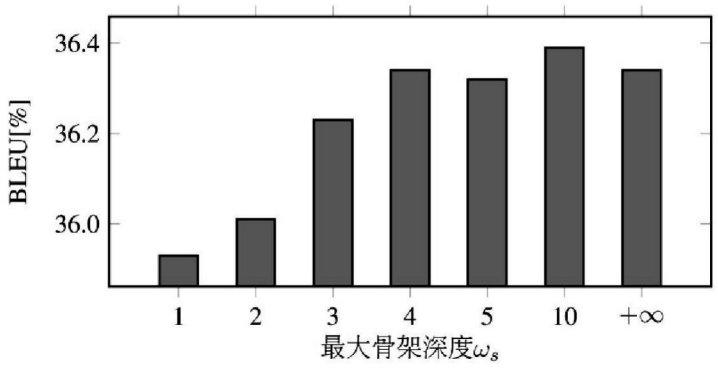

BLEU[%]
36.4
36.2
36.0
1
2
3
4
5
10
+∞
最大骨架深度ω_s

图5

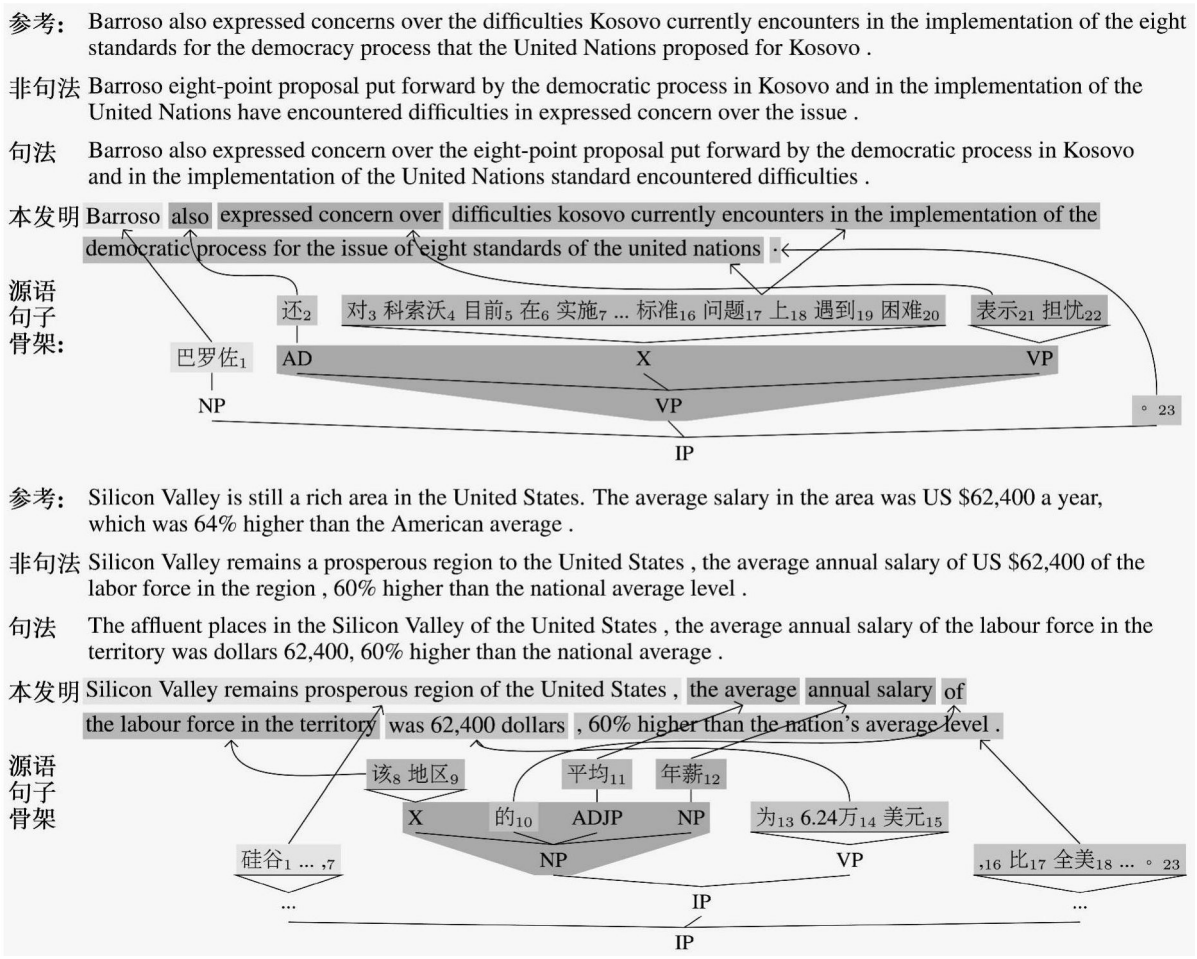

参考: Barroso also expressed concerns over the difficulties Kosovo currently encounters in the implementation of the eight standards for the democracy process that the United Nations proposed for Kosovo .
非句法 Barroso eight-point proposal put forward by the democratic process in Kosovo and in the implementation of the United Nations have encountered difficulties in expressed concern over the issue .
句法 Barroso also expressed concern over the eight-point proposal put forward by the democratic process in Kosovo and in the implementation of the United Nations standard encountered difficulties .
本发明 Barroso also expressed concern over difficulties kosovo currently encounters in the implementation of the democratic process for the issue of eight standards of the united nations .
源语句子骨架:
还2
对3 科索沃4 目前5 在6 实施7 ... 标准16 问题17 上18 遇到19 困难20
表示21 担忧22
巴罗佐1
AD
X
VP
NP
VP
。23
IP
参考: Silicon Valley is still a rich area in the United States. The average salary in the area was US $62,400 a year, which was 64% higher than the American average .
非句法 Silicon Valley remains a prosperous region to the United States , the average annual salary of US $62,400 of the labor force in the region , 60% higher than the national average level .
句法 The affluent places in the Silicon Valley of the United States , the average annual salary of the labour force in the territory was dollars 62,400, 60% higher than the national average .
本发明 Silicon Valley remains prosperous region of the United States , the average annual salary of the labour force in the territory was 62,400 dollars , 60% higher than the nation's average level .
源语句子骨架
该8 地区9
平均11
年薪12
X
的10
ADJP
NP
为13 6.24万14 美元15
硅谷1 ... ,7
NP
VP
,16 比17 全美18 ... 。23
...
IP
...
IP

图6