# 基于词义归纳模型的汉-纳统计机器翻译[1]

周珂，余正涛*，高盛祥，程立，毛存礼
（昆明理工大学 信息工程与自动化学院，昆明 650500）

**摘要：**为了解决汉纳机器翻译中一词多义及词对齐缺失的问题，提出了基于词义归纳翻译模型的汉-纳的统计机器翻译方法。该方法首先对汉语-纳西语词对齐语料进行统计分析，选取词及前后 3 个词上下文作为聚类文本，通过 LDA 的方法对词进行聚类分析，利用聚类出来的词义训练得到词义归纳模型，然后，在汉纳树到串模型的基础上，将词义归纳模型融入到树到串的汉语-纳西的翻译模型中，在解码过程中指导选择出概率最大的译文。最后进行了融合模型前后的对比实验，结果表明融合词义归纳翻译模型的树到串的汉语-纳西的统计机器翻译方法，在解决多义词选择及对齐关系缺少上表现了好的效果。

**关键词：**汉语-纳西语；词义归纳；统计机器翻译；

# Sense-Based Chinese-Naxi Statistical Machine Translation

Ke Zhou, Zhengtao Yu*, Shengxiang Gao, Li Cheng, Cunli Mao
(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500)

**Abstract**: To solve the problem of ambiguity and word alignment missing, a statistical machine translation method based on the sense-based translation model is proposed in this paper. In the method, we firstly make a statistical analysis to the Chinese-Naxi word alignment corpus by clustering and analyzing the text composed of the key word and three words around it via the LDA method. The sense-based model is obtained by training clustered sense. Then, the sense-based model is introduced to the Tree to String Chinese-Naxi translation model. The proposed sense-based translation model enables the decoder to select the most accurate translations. Lastly, we do some comparative experiments about the improved and the former method. The experimental results verifies that the improved method has better effect on ambiguity and word alignment missing.

**Keywords:** Chinese-Naxi; word sense induction; Statistical Machine Translation(SMT);

## 1 引言

纳西文是目前世界上唯一活着的象形文字，汉纳机器翻译能促进汉-纳双语的交流,对纳西象形文字继承和发扬具有非常重要的作用。在机器翻译研究方面,在经历了上世纪90年代初到2000年初的基于词[1-2]和基于短语[3-4]的翻译方法后，基于句法的统计机器翻译成为主流的机器翻译方法，如Yamada提出的串到树模型[5]，刘洋提出的基于对齐模板的树到串翻译模型[6]，熊得意提出的依存树到串翻译模型[7]，Chiang提出的层次短语模型[8]。在汉纳机器翻译方面，张涛提出了一种融合特征约束模型的纳西-汉语双语词语对齐算法[9]，杨秀珍提出了一种基于实体约束的汉纳双语对齐方法[10]，李磊提出一种基于改进的依存树到串的汉语纳西语翻译模板抽取方法[11]，通过源语言词的上下文依存关系获取目标语言中对齐词,实现了汉纳树到串的翻译模型,有好的效果。由于汉语和纳西语言存在很大结构差异，翻译存在很多困难，如汉语句子中“我想长（𡍮）大”中“长”字对应多个意义，在上下文中，“长”字为成长的意思，对于这种一词多义的情况，传统的对齐模板在进行多义词选择时由于词歧义会容易选择错误。另外，由于汉语中一些文字在对应的纳西文

**基金项目：**国家自然科学基金(No.61163022)，科技部科技创新人才基金项目(No.2014HE001)

字没有，存在词对齐缺失问题。如：“小鸟”在纳西语中没有形容词对齐词，只有“鸟（）”字在纳西文中有对齐的词，这样在解码过程中就找不到完全对齐模板。通用的短语或句法翻译模型在处理这些语言现象时，由于仅考虑汉纳双语词语对齐关系，而对于目标语言多义词的翻译选择及未对齐词处理方面效果并不理想，借助于双语语料对齐关系，可以通过对源语言多义词统计建立与目标语言词对齐关系，同样，可以将源语言未对齐的形容词等未对齐词与名词等词语进行合并归纳，建立源语言合并词与目标语言词对齐模型，利用双语对齐语料，训练获得源语言多义词及未对齐词合并词到目标语言词对应关系模型，并利用这个关系在翻译词选择解码时进行约束，在解码过程中指导选择出概率最大的词语，从而在一定程度上解决汉纳翻译过程中的一词多义及未对齐词缺失问题。

## 2 词义归纳

首先定义多义词词类，对于任意一个多义词，可能对应不同意思，要根据上下文信息对这些对一词进行确定，比如句子“我想长大后成为一名科学家”中的“长”词类有不同,分别表示时间和成长等词类，可以根据词类的上下文信息自动推导出这个词类的词义，同样，对句子“这有一只小鸟”中的“小”字对齐语料缺失需要与“鸟”字合并翻译问题也可以归结到词义归纳的问题。因此，词义归纳也可以看成是一个主题模型问题，一个词类的词义数可以看成是潜在的所有的主题，我们按照主题模型这个思想从大规模的训练语料库中推导词义。定义伪文本是由给定词的上下文 N 个词组成的。对于汉语句子“我想长大后成为一名科学家”，我们假设 N=3，如汉语词“长”的伪文本，具体见表 1。以此类推，我们可以抽取出词类的很多伪文本，这些伪文本的集合可以组成对应这个词类的语料库。对于这个语料库，我们根据主题模型的方法推导出这个语料库的所有的主题。

**表 1 中文句子中抽取“长”词类的伪文本**

| 句子 | 我想长大 |
|---|---|
| “长”的伪文本 | 我 想\|\|大 后 成为 |

图 1（a）给出了一次词类的 W 的基于 LDA 的词义归纳，$w_{j,i}$表示给定词类 W 的第 j 个文本的第 i 个词。$s_{j,i}$ 是$w_{j,i}$指定的词的词义，第 j 个伪文本的主题分布$\theta_j$被看作是给定的词类 W 的词义。LDA 在词义归纳上面的使用如下步骤所示：

(1) 对于每一个伪文本 $D_j$，使用 Dirichlet 分布 $Dir(\alpha)$ 生成它的词义分布 $\theta_j$。

(2) 对于每一个在伪文本 $D_j$ 中的词条 $w_{j,i}$，首先使用词义聚类生成服从 $\mathrm{Multinomial}(\theta_j)$ 分布的词义 $s_{j,i}$，然后产生一个词 $w_{j,i}$，使这个词服从 $\varphi_{s_{j,i}}$ 分布,而 $\varphi_{s_{j,i}}$ 是 $w_{j,i}$ 使用 $Dir(\beta)$ 产生的词义 $s_{j,i}$ 分布。

由于LDA模型需要手工的指定主题(词义）的个数，而我们想让 LDA 模型从训练数据中自动的确定主题的个数即每一个词类的词义。因此，我们借助无参的 LDA，即 HDP。HDP 在词义归纳上的产生过程如图 1（b）所示：

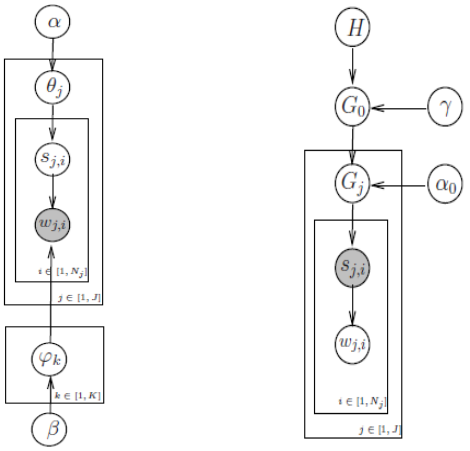
(1)给定一个调节参数 $\gamma$ 和一个基本的分布 $H$，使用狄利克雷过程 $DP(\gamma,H)$ 生成一个基本分布 $G_0$。

(2)对于每一个伪文本$D_j$，产生一个$G_j$使之服从$DP(\alpha_0, G_0)$分布。

(3)对于伪文本$D_j$中的每一个词条$w_{j,i}$，产生一个词义聚类$s_{j,i} \sqcup G_j$以及$w_{j,i} \sqcup \varphi_{s_{j,i}}$

其中，$G_0$对于词义聚类的一个全局分布，下面被$G_j$使用。$G_j$是每一个文本对应于这些词义聚类的词义分布，在对应过程中可以指定文本词义聚类的权重。$\gamma$，$\alpha_0$是 HDP 的核心参数，用来调节全局分布$G_0$和指定文本分布$G_j$的词义的变化。

基于 LDA 的词义归纳方法符合分布式假设，即在相同的句子或者文章中出现的词具有相似的含义。因此，我们在 LDA 的词义归纳的基础上建立基于词义归纳模型，并把词义归纳模型融合到机器翻译中。




(a)　　(b)

图 1 (a)用 LDA 做 WSI 的图模型表示 (b) 用 HDP 做 WSI 的图模型表示

## 3 词义标注

词义标注是使用词义归纳方法去自动的预测词的词义去标注源语言词，对于每一个词类逐个建立词义归纳模型，并用数据训练这些建立的模型。具体的过程如下所述。

1）数据预处理：移除训练集、开发集和测试集中源语言端的停用词。

2）词义标注：从预处理的数据中，对每一个源语言端的词类抽取出所有可能的伪文本，然后把抽取出来的一个词类的所有的伪文本作为一个语料库来归纳出这个词类的所有词义，我们以最大的概率输出的词义标注对应的词。

## 4 词义归纳模型

词义归纳模型是用来评估在给定上下文信息描述的词义信息的情况下一个源语言词$c$被翻译为目标语言$\tilde{e}$的概率，我们规定目标短语$\tilde{e}$的最大长度为三个单词、最短为空。模型的基本组成为基于最大熵的分类器，我们使用这个最大熵分类器去预测词义的概率$p(\tilde{e}|C(c))$，最大熵分类器的具体描述如等式（1）。

$$p(\tilde{e}|C(c)) = \frac{exp(\sum_i \theta_i h_i(\tilde{e}, C(c)))}{\sum_{\tilde{e}'} \theta_i h_i(\tilde{e}', C(c))} \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

其中，$h_i$是二值特征，$\theta_i$为这些二值特征的权重，$C(c)$为词$c$的上下文信息。我们定义了两组二值特征：1）词汇特征；2）词义特征。这些特征可以被描述成等式（2）。

$$h(\tilde{e}|C(c)) = \begin{cases} 1, & if\ \tilde{e} = \square\ and\ C(c) \cdot \mu = \upsilon \\ 0, & else \end{cases} \quad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

其中，□是一个可能的目标翻译的占位符（最长为 3，最短为 null），$\mu$为源语言$c$上下文特征（词汇特征或者词义特征），符号$\upsilon$表示特征$\mu$的值。

我们以词$c$为中心词以$\pm k$个词为上下文窗口抽取词汇和词义特征。例如，词

汇特征被定义为词$c$的前$k$个词和后$k$个词以及词$c$本身：$\{c_{-k},...,c_{-1},c,c_1,...,c_k\}$。词义特征被定义为上下文这些词被预测到的词义：$\{s_{c_{-k}},...,s_{c_{-1}},s_c,s_{c_1},...,s_{c_k}\}$。虽然这两个特征都是利用的上下文窗口，但是这两个特征并不冲突，因为，第一，基于 HDP 的词义归纳的词义特征在处理数据上和词汇特征相比采用不同的方法。第二，词义特征包含语义分布信息。

给定一个源语言句子$\{c_i\}_1^I$，提出的词义归纳模型$M_s$可以表示为公式（3）：

$$M_s=\prod_{c_i\in W}\left(\tilde{e}_i\,|\,C(c_i)\right) \qquad \text{.....（3）}$$

其中，$W$是用来建立最大熵分类器的词的集合。

## 5 融合词义归纳模型的树到串翻译模型

### 5.1 树到串翻译模型

树到串翻译模型，是用概率化的规则来描述源语言的结构树到目标语言的串之间的转换关系。它在源语言端使用句法分析器生成句法树，在目标语言端找到对应的串。该方法的基本思想：首先用句法分析器获得源语言的短语结构树，然后利用树到串对齐模板(TAT)将源语言的树映射到目标语言上，因此解码过程就是一个树到串的转换过程。其中树到串的对齐模板是一个三元组<CT,VS,A>,汉语短语句法树 CT、纳西语串 VS、两者之间的对齐关系。其训练过程是在源语言句法树和目标语言串满足对齐一致性时，使用抽取算法后序遍历源语言句法树，抽取出对齐模板。其解码过程是采用自底向上的柱搜索算法，后序遍历输入的源语言句法树，对每个节点所对应的源语言短语搜索推导，当处理完根节点后，就可以翻译整个句子。此模型的一大优点能够自动获取树到串对齐模板，从而捕获语言学驱动的局部（词）重排序和全局（短语、子句）重排序。

### 5.2 训练

首先是词义归纳模型的训练，该模型的训练主要是就是评估特征权重$\theta$，对于每一个源语言的词类都建立一个分类器，当经过基于 HDP 的词义归纳处理并做过词义标注后，我们就可以使用这个数据去训练词义归纳模型。训练需要上下文元素$C(c)\cdot\mu=\upsilon$以及目标翻译$\tilde{e}$。

然后是树到串翻译模型的训练，我们首先使用句法分析器得到汉语句法树，再按照先序遍历抽取树到串的模板。如图 2 所示：

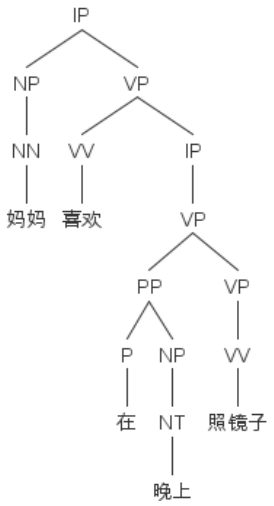



图 2 源语言句法树和词对齐

抽取的模板如表 2 所示：

**表 2 抽取的树到串模板**

| 编号 | 树 | 串 | 对齐 |
|---|---|---|---|
| 1 | （NP 妈妈） |  | 1:1 |
| 2 | （VV 喜欢） |  | 1:1 |
| 3 | （PP 在晚上） |  | 1:1 |
| 4 | （VP 照镜子） |  | 1:1 |
| 5 | （IP（NP 妈妈）（VP（VV 喜欢）（PP 在晚上）（VP 照镜子））） |  | 1:1<br>2:2<br>3:4<br>4:3 |

### 5.3 解码

解决，我们设计如图 3，图中详细描述了融合了词义归纳模型的统计机器翻译的整个过程。

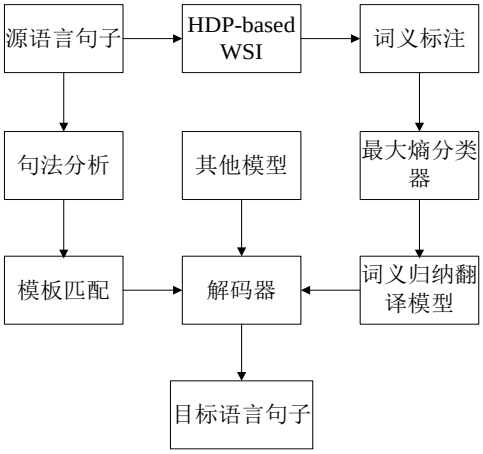



图 3 融合词义归纳模型的汉纳翻译系统流程

在翻译一个源语言句子之前，首先对源语言句子进行预处理，使用基于 HDP 的词义归纳模型去预测一个词类的词义，接着使用已经训练好对应这个词类的最大熵分类器去指导解码，当一个新的源语言词类$c$被翻译时，我们通过词对齐找到它的目标翻译$\tilde{e}$，然后根据最大熵分类器去评估等式（1）的翻译概率$p(\tilde{e}|C(c))$。通过这种方法，我们就可以很容易的得到基于词义的翻译模型对应的得分。例如：汉语句子“我想长大”的翻译。当在对句子“我想长大”时，先进行句法分析生成源语言的句法树。当源语言树上有歧义词时，如搜索（VV 长）时，就会有不同的对应翻译 (zhang) 和 (chang)，然后词义归纳模型可以根据“长”的上下文词义找到“长”正确的翻译 (zhang)。

## 6 实验和分析

### 6.1 实验数据及工具

为了进行汉纳机器翻译方法的评测，收集整理了 53000 词对齐平行句对，从中选取 42000 对汉语-纳西双语句子作为训练集，并选取 3000 对词对齐的汉语-纳西双语句子作为开发集，随机选取 2000 对汉语-纳西双语句子作为测试集，具体实验语料见表 3。

**表 3 训练集和测试集语料信息**

| 数据集 | | 中文 | 纳西文 |
|---|---|---|---|
| 训练集 | 句数 | 42000 | |
| | 词数 | 341254 | 21986 |
| 开发集 | 句对 | 3000 | |
| | 词数 | 25624 | 15203 |
| 测试集 | 句数 | 2000 | |
| | 词数 | 17021 | 11478 |

为了验证融合词义归纳模型的汉语-纳西语的统计机器翻译模型效果，设计了一个对比验证实验，以刘洋提出的短语树到串的模型为基准系统，使用中科院的开源工具 ICTCLAS 对中文进行分词，并使用 ctbparser 中文短语句法分析器解析源语言端的短语句法树，使用 SRILM toolkit 工具训练汉语 3 元语言模型，使用 C++HDP toolkit(Wang and Blei,2012)[12] 训练获得基于 HDP 的词义归纳模型，抽取词的上下文 a±3 窗口作为伪文本，使用 MaxEnt tool 去训练我们的最大熵分类器，采用 BLEU 评测标准，使用最小错误率算法 MERT (Och,2003)去训练对数线性模型并获取各个模型的模型参数。

### 6.2 实验结果与分析

训练语料库中和测试语料库中的基于 HDP 的词义归纳信息见表 4，其中，训练语料库中和测试语料库中分别有 3156 和 563 个经过预处理的词类，对于这些词类，分别抽取 75018 和 2620 个伪文本。其中，在训练语料库和测试语料库中的每一个词类分别对应 25 和 8 个伪文本。使用基于 HDP 的词义归纳方法从训练语料和测试语料的伪文本中学习得到 9143 和 1025 个词义。

**表4 训练集和测试集的词义归纳**

| | 训练集 | 测试集 |
|---|---|---|
| 词类 | 3156 | 563 |

续表

| | 训练集 | 测试集 |
|---|---|---|
| 总共的伪文本 | 75018 | 2620 |
| 平均的伪文本 | 25 | 8 |
| 总共词义 | 9143 | 1025 |
| 平均词义 | 3 | 2 |

以上的数据训练完词义归纳模型后，把词义归纳模型融合到短语树到串翻译模型中进行训练、解码，对比实验结果见表5。

**表 5 融合词义归纳模型的汉纳翻译系统实验结果**

| 翻译系统 | 实验集 | 开发集 | | 测试集 | |
|---|---|---|---|---|---|
| | 评测指标 | BLUE（%） | 准确度(%) | BLUE（%） | 准确度（%） |
| 短语树到串模型 | | 25.87 | 56.84 | 28.59 | 60.48 |
| 融入词义翻译模型的短语树到串模型 | | 27.45 | 68.01 | 30.13 | 71.41 |

表中给出的实验结果显示，在相同的语言模型情况下，在开发集实验中，融合词义归纳模型的翻译系统比基准系统，BLUE值提高了1.58%，准确度提高了11.17%。在测试集实验中，BLUE值提高了1.54%，准确度提高了10.93%。实验结果表明，融合了词义归纳的翻译模型的翻译系统可以提高翻译效果。

## 7 总结

本文提出了融合词义归纳模型的树到串汉纳翻译模型，能够提高翻译效果，并在一定程度上能解决多义词选择及对齐缺失问题，实验也证明了方法的有效性。进一步研究工作重要集中在基于篇章结构语义信息融合的汉纳机器翻译方法的研究。

## 参考文献


[1] Peter F Brown, et al. A Statistical Approach to Machine Translation[J].Computational Linguistics, 1990,16(2):79-85.

[2] Peter F Brown, et al. Analysis, Statistical Transfer, and Synthesis in Machine Translation[C]// Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, 1992:83-100.

[3] Och F J. Statistical Machine Translation: From Single-Word Models to Alignment Templates[D]. PhD dissertation: 2002.

[4] Och F J. Minimum error rate training in statistical machine translation[C]// Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. 2003:160-167.

[5] Kenji Y and Kevin K.A Decoder for Syntax-Based Statistical MT[C]// Proceedings.of Association for Computational Linguistics(ACL02),2002:303–310.

[6] Liu Y, Liu Q, Lin S X. Tree-to-String Alignment Template for Statistical Machine Translation[C]// Proceedings.of Association for Computational Linguistics, 2006:345-353.

[7] Xiong D Y, Liu Q, Lin S X. A Dependency Treelet String Correspondence Model for Statistical Machine Translation[C]// Proceedings of the Second Workshop on Statistical Machine Tranlaion.2007:40-47.

[8] David Chiang.A Hierarchical Phrase-Based Model for Statistical Machine Translation[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.2005:263-270.

[9] 张涛，余正涛，郭剑毅，等. 融合特征约束模型的纳西-汉语双语词语对齐算法[J].西安交通大学学报,2011,45(10)48-53.

[10] Yang X Z, Yu Z T, Guo J Y, et al. Naxi-Chinese Bilingual Word Alignment Method Based on Entity Constraint[C]// Proceedings of 14th Workshop on Chinese Lexical Semantics, 2013: 378-386.

[11] Li L, Yu Z T, Mao C L, et al. The Extracting Method of Chinese-Naxi Translation Template Based on Improved Dependency Tree-To-String[J]. Lecture Notes in Computer Science, 2014,8229(3):350-358.

[12] C.Wang and D. M. Blei. 2012. A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process[J]. Arxiv preprint arXiv:1201.1657. 2012.