文章编号：1003-0077(2011)04-0075-07

面向形态丰富语言的多粒度翻译融合

王志洋[1,2]，吕雅娟[1]，刘 群[1]

(1. 中国科学院 计算技术研究所，中国科学院 智能信息处理重点实验室，北京 100190；
2. 中国科学院 研究生院，北京 100049)

摘 要：形态丰富语言由于其复杂的形态变化，会导致大词汇量和数据稀疏问题，这给统计机器翻译带来了巨大挑战。该文通过将这类语言表示为不同的粒度，然后分别进行翻译；由于不同的粒度能表征语言不同层面的特点，通过对不同粒度的翻译结果进行词级系统融合，便可生成更好的译文。维吾尔语、蒙古语到汉语的两组翻译实验表明，这种多粒度系统融合方法改善了翻译效果，BLEU 值比最好的单系统分别提高了+1.41%和+2.03%。

关键词：形态丰富语言；多粒度；系统融合

中图分类号：TP391 **文献标识码**：A

System Combination with Multiple Granularities for Morphologically Rich Language Translation

WANG Zhiyang[1,2], LV Yajuan[1], LIU Qun[1]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Graduate University, Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Morphologically rich language, characterized by complex morphological changes, has huge vocabulary and serious data sparseness issue, which has brought a great challenge to machine translation. In this paper, we first analyze such language and use different granularities to represent and then translate them respectively. As different granularities can catch features of such language in different levels, we integrate the translation hypotheses from different granularities by the system combination approach to generate better results. Experimental results on Uyghur-Chinese and Mongolian-Chinese translation tasks show that system combination with multiple granularities improved the performance of translation, and gained +1.41% and +2.03% compared to the best single system measured by BLEU.

Key words: morphologically rich language; multiple granularities; system combination

1 引言

形态丰富语言的主要特征是高度丰富的形态变化，像曲折(Inflection)、派生(Derivation)、复合(Composition)等。给定一个词根，通过形态变化可以衍生出成百上千种新的形式；例如蒙古语词根UILED，理论上至少可以有 1 710 种变化形式[1]。如果将每一种变化形式都看作单独的词，这会大大增加词汇量，导致语言模型参数估计的不可信，并提高未登录词(Out-Of-Vocabulary, OOV)的比例。对统计机器翻译而言，这种数据稀疏现象会严重影响词语对齐和翻译的质量。这类语言给自然语言处理，尤其是机器翻译带来了巨大的挑战。

我国是一个多民族国家，很多少数民族都有自己的语言文字，并在本民族的各个领域中广泛应用。使用较多的少数民族语言，像维吾尔语、蒙古语、哈萨克语等都属于形态丰富语言。研究这些少数民族

收稿日期：2011-03-22 **定稿日期**：2011-05-23

基金项目：国家自然科学基金重点资助项目(60736014)；国家自然科学基金资助项目(60873167)

作者简介：王志洋(1984—)，男，博士生，主要研究方向为自然语言处理和机器翻译；吕雅娟(1972—)，女，副研究员，主要研究方向为自然语言处理；刘群(1966—)，男，研究员，主要研究方向为自然语言处理和机器翻译。

语言与汉语之间的翻译，对加强民族之间的沟通交流、文化传播、经济发展有重要的意义。而在与中国海陆相邻的二十一个国家中，除中国南部的极少数国家(像越南、缅甸、老挝等)，大部分国家使用的语言都有丰富的形态变化，像俄语、日语、韩语等。通过研究这些语言与汉语之间的翻译，对维护地区稳定、促进交流合作等有重要作用。

本文主要研究形态丰富语言到汉语的翻译。由于这类语言形态变化复杂，而且双语资源相对匮乏。为了充分利用有限的双语语料，缓解数据稀疏问题，本文将这类语言表示为不同的粒度，并分别进行翻译，然后利用系统融合技术将不同粒度的翻译结果进行融合，以提高机器翻译的性能。具体来说，对同样一份双语语料，我们将源语言(形态丰富语言)用不同的粒度(词、词干、词素等)表示，并使用同一个翻译系统分别翻译；然后将不同粒度的翻译结果进行词级系统融合。维吾尔语、蒙古语到汉语的两组翻译实验表明，这种多粒度融合方法改善了翻译效果，BLEU值[2]比最好的单系统分别提高了+1.41%和+2.03%。

2 相关工作

在机器翻译任务中，当源语言为形态丰富语言时，一般有以下几种处理思路。一种是选择合适的粒度，尝试通过不同的词干词缀组合来改善翻译效果。Lee[3]先对阿拉伯语进行词法分析，然后通过合并或删除某些词缀，来平衡阿拉伯语和英语之间的词级语义；类似的工作还有文献[4]等。另外一种方式是预调序，让源句子的语序更接近目标句子，最有代表性的是Collins等人的工作[5]，类似的预调序方式还有文献[6-8]等。这类方法往往需要借助句法分析技术，这对很多语言，尤其是形态丰富语言往往是不可得的。还有一种思路是尽量利用形态句法信息。Koehn等[9]提出了基于要素(Factor)的模型，这能够更好地融合形态和句法信息；但若使用要素过多，会影响调参效果和翻译速度。Dyer等[10]将源句子词法分析的结果表示为词图(Lattice)形式，使输入更具容错性，在阿拉伯语到英语的翻译任务上取得了一定的效果。

在本文中，我们将源语言切分表示为不同的粒度，分别抽取翻译模型进行翻译；然后将不同粒度的翻译结果进行系统融合。相比Koehn等人[9]的方法，不同粒度的翻译模型都是单独调参的，这样即使引入更多的粒度，也不会影响调参效果；跟基于词图的方法比，我们的方法简单而直接。

与本文工作类似的是Gispert等人的工作[11]，他们通过使用不同的切分工具对源语言切分，然后使用最小贝叶斯风险(Minimum Bayes Risk，MBR)[12]的方法对翻译结果进行融合。而本文使用同一个词法分析工具，获得源语言句子的不同粒度表示，像词、词干、词缀等。此外，文献[11]的融合方式是句子级的，更像是一种重排序(Re-rank)技术；而本文使用的是词级系统融合，这往往能产生更好的融合效果[13]。

3 多粒度翻译

在上一节中提到，翻译中一个常见思路是选择合适的粒度来表示形态丰富的语言端，然后再进行翻译。但合适的粒度往往与双语语料的规模以及翻译语言对本身有关。在本文中，我们使用不同的粒度进行翻译，然后再将翻译结果进行词级融合。因为不同的粒度表征了语言不同层面的特征；直觉上，不同粒度的翻译结果融合应该可以生成更好的结果。例如，词(Word)粒度的翻译规则更精确，但丰富形态变化导致的数据稀疏，会使规则覆盖面有限；词干(Stem)能表征词的大部分语义，使用词干粒度能够大大缓解词稀疏的问题，但会引发某些歧义；而词素(Morpheme)粒度，融入了更多的句法信息，可以生成更符合句法的结果，但词素粒度过小，给词语对齐和翻译调序都带来了负担。

由于不同的粒度表示各有其优缺点，我们将其分别翻译，然后将翻译结果融合，尽量利用各种粒度的优点，以改善翻译质量。图1是一个维语句子经过词法分析后，不同粒度表示的结果。

词 ：چىقىلىدۇ قاراپ لۇنۇپبۇ گۇرۇپپىلارغا

词干：چىق قاراپ بۇلۇپ گۇرۇپپا

词素：ۇ+ىد+ال+چىق قاراپ ۇن+بۇلۇپ غا+لار+گۇرۇپپا

词干+词缀：الىدۇ+چىق قاراپ ۇن+بۇلۇپ لارغا+گۇرۇپپا

图1 维语句子的不同粒度表示(对应的中文是“分组审议”)

4　词级系统融合

在自然语言处理中，几个功能相似的系统执行同一任务时，可能有多个输出结果，系统融合就是将这些结果进行融合，以改善最终结果。具体到机器翻译任务，每个翻译模型都有相应的优点，生成的译文也有所差别；系统融合可以将多个系统的译文融合起来，充分利用各种系统的优点，输出更好的译文。

在机器翻译中有多种系统融合方式，根据操作目标语言句子粒度的不同，可以分为三类：

a) 句子级系统融合：对同一个源语言句子，利用 MBR 解码[12]或重打分的方法比较多个系统的翻译结果，将最优结果输出。句子级系统融合方法不会产生新的翻译假设(Hypothesis)，它只是在已有的翻译假设中挑选出“最好”的一个，本质上属于一种重排序技术。

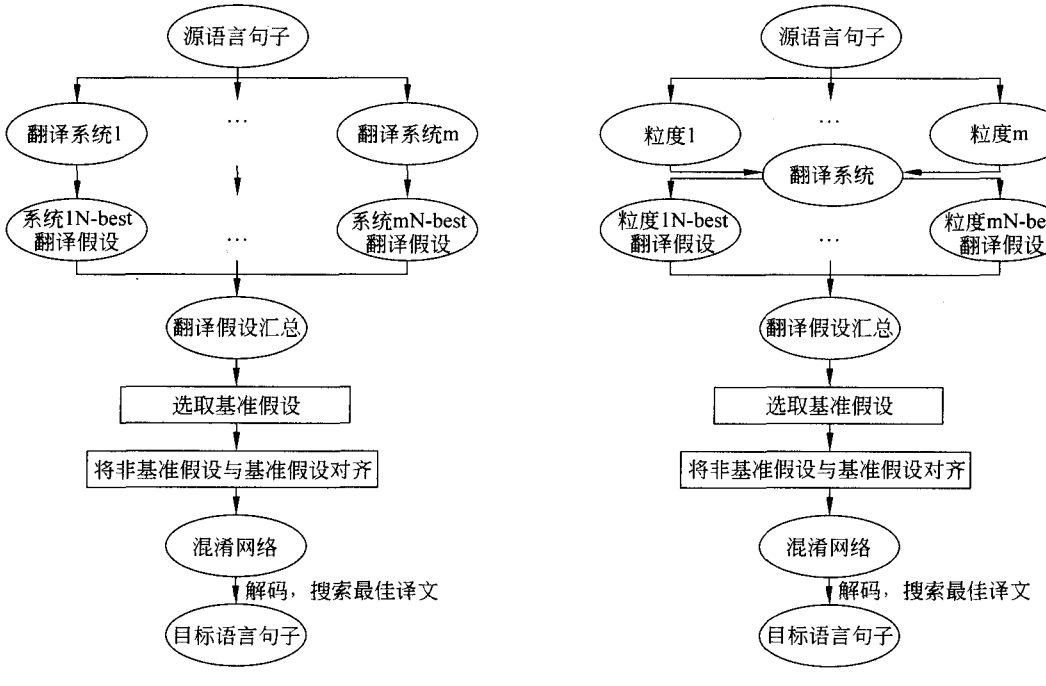
b) 短语级系统融合：根据多个系统输出的结果，重新抽取或生成与测试集相关的短语表，再利用新的短语表对测试集重新解码。

c) 词级系统融合：首先将不同系统的输出的翻译结果利用词对齐方法构建混淆网络 (Confusion Network)，再选取一定的特征在混淆网络上进行解码。

在实际融合性能上，Macherey 等[13]对这三种融合方法进行了经验性的比较。实验结果显示，相关度较小的翻译系统之间进行融合，在性能上词级系统融合最好，句子级最差。本文采用的融合方法是词级系统融合。

4.1　词级系统融合

图 2(a)是传统的词级系统融合的流程图。首先收集各系统的翻译假设，然后按照 MBR 方法为每个系统选取一个基准假设，按照一定的对齐方法将每个非基准假设和基准假设对齐以构建混淆网络。最后在构建好的混淆网络上搜索最优路径，将最优路径上的词拼接起来便得到最终译文。



(a) 传统的系统融合过程　　(b) 本文采用的方法

图 2　词级系统融合流程图

词语对齐在构建混淆网络的时候尤为重要，它影响最终译文的生成。在基于混淆网络的系统融合模型中，应用最广泛的是 Rosti 等[14]和 He 等[15]提出的方法。两者的主要区别在于对齐方法的不同，前者采用类似编辑距离的 TER(Translation Edit Rate)作为标准进行对齐，后者采用基于间接隐马尔科夫模型(Indirect HMM，IHMM)的方法来进行对齐。由于 IHMM 的方法不仅考虑了两个目标词之间的字面相似度，还考虑了它们之间的语义相似度，进而取得了更好的对齐效果。本文采用了基于

IHMM 的对齐方法。

IHMM 方法将基准假设中的词看成是隐马模型的状态，翻译假设中的词看作是隐马模型的观察序列，基准假设和翻译假设之间的词对齐关系当作隐变量，于是可以使用一阶隐马模型来估计翻译假设相对于基准假设的条件概率：

$$p(e_1^J \mid e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j \mid a_{j-1}, I) p(e'_j \mid e_{a_j})]$$

其中，e_1^I 为基准假设，e_1^J 为翻译假设，a_1^J 表示对齐序列，a_j 表示词 e'_j 对齐到的词在基准假设中的位置。

发射概率 $p(e'_j \mid e_{a_j})$ 可以根据词之间的相似度进行建模，一般看作是语义相似度 $p_{sem}(e'_j \mid e_{a_j})$ 和字面相似度 $p_{sur}(e'_j \mid e_{a_j})$ 的线性插值：

$$p(e'_j \mid e_{a_j}) = \alpha \cdot p_{sem}(e'_j \mid e_{a_j}) + (1-\alpha) p_{sur}(e'_j \mid e_{a_j})$$

其中 α 为插值因子。语义相似度 $p_{sem}(e'_j \mid e_{a_j})$ 计算时，将源语言词作为隐变量，通过双语词典便可进行计算。字面相似度 $p_{sur}(e'_j \mid e_{a_j})$ 则是考虑两个词语的最大公共字串。在单语对齐中，语义相似度可以处理同义词问题，而字面相似度则可以很好的处理同根词等使用 GIZA++工具[16]对齐时很难处理的问题。

转移概率 $p(a_j \mid a_{j-1}, I)$ 对词序重排进行建模，它取决于对齐的词之间的跳转距离，一般将其分成几类，并赋予一定的经验值。由于是单语对齐，对同序的对齐给予奖励，而给非同序的对齐一定的惩罚。

最终的对齐结果可以通过 Viterbi 算法搜索得到：

$$\hat{a}_1^I = \text{argmax}_{a_1^I} \prod_{j=1}^{J} [p(a_j \mid a_{j-1}, I) p(e'_j \mid e_{a_j})]$$

在生成最终的译文时，我们使用了一些特征来进行打分，它们通过对数线性模型组合在一起。

$$E^* = \text{argmax}_E (\lambda_{AL} p_{AL} + \lambda_{LM} p_{LM} + \lambda_{NULL} N_{NULL}(E) + \lambda_{WORD} N_{WORD}(E))$$

其中 p_{AL} 为词的置信度，p_{LM} 为语言模型得分，$N_{NULL}(E)$ 为插入空词的惩罚，$N_{WORD}(E)$ 为长度惩罚。λ 为对应的权重。

4.2 多粒度系统融合

图 2(b)是多粒度系统融合的流程图。相比传统的利用多个翻译系统来进行融合，本文只使用一个翻译系统；并将源语言通过多种不同粒度来表示，而不是单一的表示。由于不同的粒度可以刻画形态丰富语言不同层面的特征，使用同一个翻译系统来翻译不同的粒度，得到不同粒度的翻译结果，再进行词级系统融合，直觉上可以生成更好的翻译结果。

由于对源语言有不同粒度的表示，在翻译假设对齐过程中，计算语义相似度时需要考虑不同的粒度，并使用相应粒度的双语词典来计算单词间的语义相似程度。

5 实验

形态丰富语言众多，这里我们仅以维吾尔语和蒙古语为例。通过维吾尔语、蒙古语到汉语的翻译实验，来验证我们的方法。

将源语言表示为多种粒度，需要通过词法分析工具来完成。我们按照姜文斌等[17]的有向图思想实现了维语词法分析工具，重现了蒙古语词法分析工具。这里我们使用了词、词干、词素三种粒度来进行融合。

基于短语的 Moses① 系统作为基线翻译系统，翻译质量使用基于词的 BLEU-4 来衡量。在利用 Moses 进行翻译时，语言模型是根据对应训练集的中文部分，利用工具 SRILM[18] 训练的五元模型；系统融合时，语言模型是使用约 41M 的 LDC 中文语料②训练的五元模型。

5.1 维吾尔语到汉语翻译

我们收集了面向新闻领域和政府文献的约 120K 维—汉平行句对，通过去重，过滤掉单词数超过 100 的句对，最终得到的有效句对数目为 117 419 句对。然后随机各抽取出 1 000 句作为开发集和测试集，剩余部分作为训练集。这里，开发集和测试集均为单参考译文。

在训练集上的统计信息如表 1 所示。经过词法分析后，数据稀疏现象得到较大缓解，词干和词素粒度都大大减少了词汇量。

当源语言使用不同粒度表示时，翻译结果如表 2所示。显然使用词干和词素粒度都在一定程度上改善了翻译效果。

然后我们将不同粒度的翻译结果的 100-best，进行词级系统融合。由于源语言采用不同粒度表示，

① http://www.statmt.org/moses/，著名的开源工具。

② 包括 LDC2002E18，LDC2003E07，LDC2003E14，Hansards portion of LDC2004T07，LDC2004T08 和 LDC2005T06 的中文部分。

表 1　维—汉训练语料统计信息

语言	粒度	# of token	# of type	Avg length
维吾尔语	word	2 690 477	83 921	23.310 5
	stem	2 690 477	22 936	23.310 5
	morpheme	4 168 165	23 285	36.113 3
汉语	word	2 500 273	37 030	21.662 6

表 2　不同粒度表示的翻译结果

翻译粒度	Dev	Test
Word	0.415 003	0.396 2
Stem	0.426 706	0.402 4 +0.62%
Morpheme	0.420 646	0.398 9 +0.27%

因此在融合时需要选择一个粒度作为基准对齐，开发、测试集源端以及双语词典都使用同样一种粒度表示。表 3 是实验结果，可以看出，不论使用哪种粒度作为基准，系统融合的结果都有稳定提高，BLEU 值都提高了 1 个点以上。

表 3　维—汉翻译系统融合结果（和单系统最好结果比较）

基准粒度	Dev	Test
Word	0.437 1	0.413 3 +1.09%
Stem	0.441 8	0.416 2 +1.38%
Morpheme	0.439 6	0.416 4 +1.40%

实际上，在 Moses 生成的 N-best 结果中，有一些是重复的。表 4 是 N-best 列表去重前后的总数目的变化，有 50%以上的翻译假设都是重复的。

表 4　N-best 结果去重前后总数量对比

	Word		Stem		Morpheme	
	Dev	Test	Dev	Test	Dev	Test
去重前	100 000	100 000	100 000	100 000	100 000	100 000
去重后	42 904	44 403	43 276	45 335	34 087	36 527

将去重后的 N-best 结果进行系统融合，结果如表 5 所示。

表 5　N-best 去重后维—汉翻译结果

基准粒度	Dev	Test
Word	0.442 6	0.416 0 +1.36%
Stem	0.441 7	0.413 1 +1.07%
Morpheme	0.440 8	0.416 5 +1.41%

总的来说，N-best 去重后再融合，词和词素粒度为基准粒度时，BLEU 值略有提高；但当词干粒度作为基准系统时，反而有所下降，不如去重前的效果。

5.2　蒙古语到汉语翻译

蒙汉翻译实验使用的是 CWMT09[①] 的蒙汉语料的口语部分，共有 34 135 句对。各随机抽出 500 句进行开发测试，剩下的 33 135 句对作为训练集。

表 6 是不同粒度的翻译结果，使用词干和词素粒度都改善了翻译质量，提高了 2 个点以上。

表 6　蒙—汉不同粒度的翻译结果

翻译粒度	Dev	Test
Word	0.156 734	0.137 3
Stem	0.183 705	0.161 0 +2.37%
Morpheme	0.179 559	0.161 5 +2.42%

表 7 是取 100-best 进行系统融合的结果。和维—汉翻译的结果类似，这种多粒度系统融合的方式，都能带来翻译质量的稳定提高；跟最好的单系统结果相比，这里 BLEU 值也都有 1 个点左右的提高。尤其是词素粒度作为基准时，提高了 1.69 个点。

表 7　蒙—汉翻译系统融合结果（和单系统最好结果比较）

翻译粒度	Dev	Test
Word	0.192 1	0.173 7 +1.22%
Stem	0.191 6	0.171 4 +0.95%
Morpheme	0.201 6	0.178 4 +1.69%

表 8 是对翻译结果的 N-best 去重后的结果。去重后，融合结果相比去重前的结果，都有一定的提高。其中当使用词干粒度作为基准粒度时，BLEU 值比去重前提高了约 1 个点，共计提高了 2.03 个点。

表 8　N-best 去重后蒙—汉系统融合结果

基准粒度	Dev	Test
Word	0.196 3	0.175 4 +1.39%
Stem	0.201 3	0.181 8 +2.03%
Morpheme	0.202 4	0.179 2 +1.77%

总体来说，在蒙—汉翻译任务上，N-best 去重后再

① http://www.icip.org.cn/cwmt2009/index.html

进行融合，结果更稳健。

6 结论

当待翻译的源语言为形态丰富语言时，本文将其切分为不同的粒度，分别使用翻译引擎进行翻译，并将不同粒度的翻译结果通过词级系统融合技术进行融合优化，从而改善翻译质量。通过将不同粒度的结果进行词级融合，可以优势互补，生成更好的译文。在维汉和蒙汉机器翻译实验上，本方法都取得了不错的效果。

本方法直接而有效，在下一步工作中，可以在其他形态丰富语言上进行尝试。此外，本文只是利用了三种粒度来进行融合，可考虑融入更多的粒度并在可获得的更大规模的平行语料库上进行实验，来进一步改善翻译质量。

7 致谢

感谢内蒙古大学和新疆大学提供的语料，感谢新疆大学的麦热哈巴·艾力老师在维吾尔语知识层面的帮助。

参考文献

[1] 那顺乌日图，刘群，巴达玛敖德斯尔. 面向机器翻译的蒙古语生成[C]//全国第六届计算语言学联合学术会议论文集，清华大学出版社，2001.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation[C]//Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, 2002:311-318.

[3] Young-Suk Lee. Morphological Analysis for Statistical Machine Translation [C]//Proceedings of HLT-NAACL 2004, 2004:57-60.

[4] Sonja Nießen and Hermann Ney. Statistical Machine Translation with Scarce Resources using Morpho-syntactic Information [J]. Computational Linguistics, 2004, 30: 181-204.

[5] Michael Collins, Philipp Koehn, and Ivona Ivona Ku?erová. Clause restructuring for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005:531-540.

[6] Marine Carpuat, Yuval Marton, and Nizar Habash. Improving Arabic-to-English Statistical Machine Translation by Reordering Post-verbal Subjects for Alignment[C]//Proceedings of the ACL 2010 Conference Short Papers, 2010: 178-183.

[7] Dmitriy Genzel. Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation [C]//Proceedings of the 23rd International Conference on Computational Linguistics, 2010:376-384.

[8] Peng Xu, Jaeho Kang, Michael Ringgaard, Franz Josef Och. Using a Dependency Parser to Improve SMT for Subject-Object-Verb Languages [C]//Proceedings of 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009:245-253.

[9] Philipp Koehn and Hieu Hoang. Factored Translation Models[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007: 868-876.

[10] C. Dyer, S. Muresan, and P. Resnik. Generalizing Word Lattice Translation[C]//Proceedings of ACL-08: HLT, 2008: 1012-1020.

[11] Adrià de Gispert, Sami Virpioja, Mikko Kurimo, William J. Byrne. Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions[C]//Proceedings of 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009: 73-76.

[12] S. Kumar, W. Byrne. Minimum Bayes Risk Decoding for Statistical Machine Translation [C]//Proceedings of HLT-NAACL 2004, 2004:169-176.

[13] W. Macherey, F. J. Och. An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007:986-995.

[14] Antti-Veikko I. Rosti, Spyros Matsoukas, and Richard Schwartz. Improved Word-level System Combination for Machine Translation[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007:312-319.

[15] Xiaodong He, Mei Yang, Jangfeng Gao, Patrick Nguyen, and Robert Moore. Indirect-HMM-based Hypothesis Alignment for Computing Outputs from Machine Translation Systems[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008:98-107.

[16] Frans J. Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models[J]. Computational Linguistics, 2003, 29:19-51.

[17] 姜文斌，吴金星，长青，等. 蒙古语词法分析的生成式统计建模[C]//少数民族青年自然语言处理技术研究与进展，2010 年.

[18] Andreas Stolcke. SRILM-an Extensible Language Modeling Toolkit[C]//Proceedings of International Conference on Spoken Language Processing，2002：901-904.

（上接第 70 页）

[J]. 汉语语言与计算学报，2003，13(2)：197-214.

[2] 中文元数据标准研究项目组. 国外元数据标准比较研究报告—中文文献元数据标准系列报告之一[R]. http://www.idl.pku.edu.cn/pdf/metadata1.pdf. 2000.

[3] 冯志伟. 标准通用置标语言 SGML 及其在自然语言处理中的应用[J]. 当代语言学(试刊). 1998，(4)：1-11.

[4] 鲁·伯纳，麦克·苏宝麦昆，马德伟著，谢筱琳，黄韦宁译. TEI 使用指南—运用 TEI 处理中文文献[OL]. http://ablogtags. info/2011/tei-chinloc-2ndprinted-gj-ba/.

[5] David Mertz 博士. TEI—文本编码规范[OL]. [2003 年 10 月 01 日]. http://www.ibm.com/ developerworks/cn/xml/x-matters/part30/.

[6] 扎西加，顿珠次仁. 自然语言处理用藏语格助词的语法信息研究[J]. 中文信息学报，2010，24(5)：41-45.

[7] Roma：制作 TEI 的文件模型档[OL]. http://www.tei-c.org/Roma/.

[8] 圣才学习网. 图书馆资源描述标准[OL]. [2010-10-19 11：49]. http://www.100bjcb.com/HP/20101019/OTD 246998.shtml.

[9] 吴守用，古丽拉·阿东别克. 哈萨克文语料库 XML 格式标注规范初探[C]//中国少数民族语言文字信息处理研究与发展. 民族出版社，2010.