适用于特定领域机器翻译的汉语分词方法*

苏晨[1]，张玉洁[1]，郭振[1]，徐金安[1]

（1.北京交通大学 计算机与信息技术学院，北京 100044）

摘要：在特定领域的汉英机器翻译系统开发过程中，大量新词的出现导致汉语分词精度下降，而特定领域缺少标注语料使得有监督学习技术的性能难以提高。这直接导致抽取的翻译知识中出现很多错误，严重影响翻译质量。为解决这个问题，本文实现了基于生语料的领域自适应分词模型和双语引导的汉语分词，并提出融合多种分词结果的方法，通过构建格状结构(Lattice)并使用动态规划算法得到最佳汉语分词结果。为了验证所提方法，我们在 NTCIR-10 的汉英数据集上进行了评价实验。实验结果表明，本文提出的融合多种分词结果的汉语分词方法在分词精度 F 值和统计机器翻译的 BLEU 值上均取得了提高。

关键词：汉语分词；领域适应；双语引导；Lattice；机器翻译

中图分类号：TP391　　**文献标识码：**A

Chinese word segmentation method for domain-special machine translation

Su Chen[1], Zhang Yu-jie[1], Guo Zhen[1], Xu Jin-an[1]

(1.School of Computer and Information technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: In developing a domain-specific Chinese-English machine translation system, the accuracy of Chinese word segmentation in large-scale training corpus often decreases because of unknown words. The lack of domain-specific annotated corpus makes supervised learning approaches unable to adapt. This problem results in many errors in translation knowledge extraction and therefore seriously affects translation quality. To resolve the domain adaptation problem, we implemented Chinese word segmentation by exploiting n-gram statistical features in raw corpus and bilingually motivated word segmentation information in parallel corpus, respectively. We further propose a lattice-based method to combine multiple results and use dynamic programming algorithm to get the best word segmentation result. For evaluation, we conducted experiments of Chinese word segmentation and Chinese-English machine translation using the data of NTCIR-10 Chinese-English patent task. The experimental results show that the proposed method brought about improvements both in F-measure of the Chinese word segmentation and in BLEU score of the Chinese-English statistical machine translation system.

Key words: Chinese Word segmentation; Domain adaptation; Bilingual motivation; Lattice; Machine Translation

1 引言

在面向特定领域的汉英机器翻译系统开发中，由于汉语语料中的词汇集合和频率分布发生很大变化，大量新词的出现使得已有汉语分词系统的性能下降。关于领域变化对汉语分词系统性能的影响程度，我们在前期工作中进行了评测。在新闻语料上训练的汉语分词系统，在同样领域的测试数据上的召回率和准确率分别为 98.02%和 97.21%，而在科技文献语料上评测时，召回率和准确率分别下降到 86.05%和 81.83%[1]。同时，特定领域中标注语料的缺乏使得有监督的汉语分词方法难以发挥其威力。汉语分词性能的下降表现在两个方面：汉语分词粒度的不合理和错误的分词结果。汉语分词粒度过大或过小都不利于汉英词汇之间的对

* **收稿日期：**　　**定稿日期：**

基金项目：北京交通大学人才基金（KKRC11001532）

作者简介：苏晨（1989—），男，硕士研究生，主要研究方向为机器翻译；张玉洁（1961—），女，教授，主要研究方向为自然语言处理和机器翻译；郭振（1988—），男，硕士研究生，主要研究方向为机器翻译；徐金安(1970—)，男，副教授，研究方向为自然语言处理和机器翻译。

齐处理，影响单词对齐精度。因为翻译知识的获取建立在汉语分词结果和汉英单词对齐的结果上[2-3]，所以明显下降的分词精度会给大规模语料处理带来数量上难以忽视的分词错误，直接导致不正确的翻译知识，从而严重影响翻译质量。

针对这个问题，研究人员在汉语分词的领域自适应方面进行了许多探索。早期的方法是在条件随机场(CRF)统计模型中加入外部词典的特征以实现汉语分词的领域自适应[4]；当缺乏领域词典时，又有研究人员利用大规模生语料中字串的统计特征来提高分词系统的领域自适应能力[1][5]。另一方面，用于机器翻译开发的汉英平行语料中，英语句子的分词信息也可以为对应的汉语句子的分词提供引导信息 [6-7]。

在对已有研究方法进行改进与扩展的基础上，本文实现了基于生语料的领域自适应和双语引导的分词系统，并提出了将不同的汉语分词结果进行融合的方法，实现了面向特定领域的大规模汉语语料的分词系统。该方法利用格状结构将不同的分词结果进行融合，融合过程中采用半监督学习的方法得到不同分词结果的权重，最后采用动态规划算法获取最优的汉语分词结果。

本文第 2 节介绍基于生语料的领域自适应方法和汉英双语引导的汉语分词方法，然后详细描述融合多种汉语分词结果的算法；第 3 节设计实验评测本文所提方法的性能；第 4 节给出结论和今后的研究课题。

2 多种分词结果的融合方法

特定领域的汉英机器翻译系统开发中，通常需要对大规模的汉英平行语料处理以获取翻译知识，同时作为翻译对象会有大规模的该领域的汉语生语料。我们的目标是充分利用这些资源提高汉语分词的精度，为此我们提出了融合汉语生语料中的 n-gram 统计特征和汉英语料上的分词引导特征的分词方法。它基于以下想法：利用特定领域的汉语生语料的统计特征实现汉语分词向特定领域的自适应，而利用汉英语料上的英语单词边界和双语对齐特征引导汉语分词；为了融合性质不同的特征，分别实现分词系统，再对各自系统的分词结果进行融合。融合系统的总体框架如图 2-1 所示。

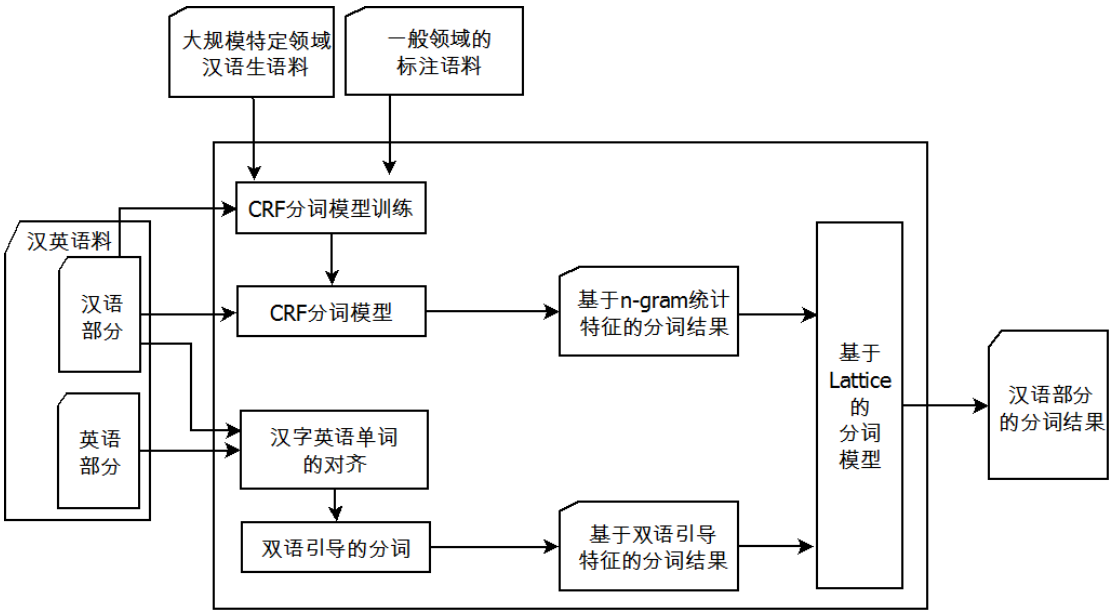

图 2-1 多种分词结果融合的汉语分词方法的整体框架

为了利用特定领域的汉语生语料的统计特征，我们实现了基于 n-gram 统计特征的汉语

分词系统；而为了利用汉英语料的分词引导特征，我们实现了基于单词对齐的分词，分别在下面两小节介绍，并在第三小节详细描述融合多种分词结果的方法。

2．1 基于 n-gram 统计特征的汉语分词

本文实现了利用生语料的统计特征的汉语分词系统[5]，具体步骤如下所述。

在利用 UPENN 的汉语标注数据的基础上，增加了特定领域的汉语大规模生语料中的统计特征，并利用 CRF 工具进行学习。汉语标注数据的特征抽取使用文献[8]提到的模板，采用五字滑动窗口提取特征，即最远使用前后各两个字作为当前字标注的依据。生语料的统计特征包括两种 n-gram 统计量：n-gram 频度值和 n-gram AV(Accessor Variety)值[9]，n-gram 频度值为 n 元字串在语料中出现的次数，n-gram AV 值为 n 元字串在语料中出现的上下文环境数。然后使用开源工具 CRF++[1]进行模型训练得到分词模型，具体流程如图 2-2。

对于每一个汉语句子，CRF 分词模型可以输出 n-best 分词结果以及相应的概率得分，以往的分词工作通常只采用 1-best 的分词结果。但是通过对 n-best 以内分词结果的观察，我们发现 1-best 结果中的错误切分部分，有可能在排名靠后的结果中获得正确切分，如图 2-3 中的例子所示。在图 2-3 中，1-best 结果中将“甘氨酸”部分切分为“甘”和“氨酸”，与标准分词结果不同，而 3-best 分词结果中将“甘氨酸”切成一个单词。鉴于这一观察结果，我们将充分利用 CRF 分词模型的 n-best 以内结果，期望从中选出正确的切分部分。本文取 10-best 以内结果，相应的概率得分表示为 $Conf_{CRF1}$, $Conf_{CRF2}$, …, $Conf_{CRF10}$，并把它们分别作为对应分词结果中单词的置信度。

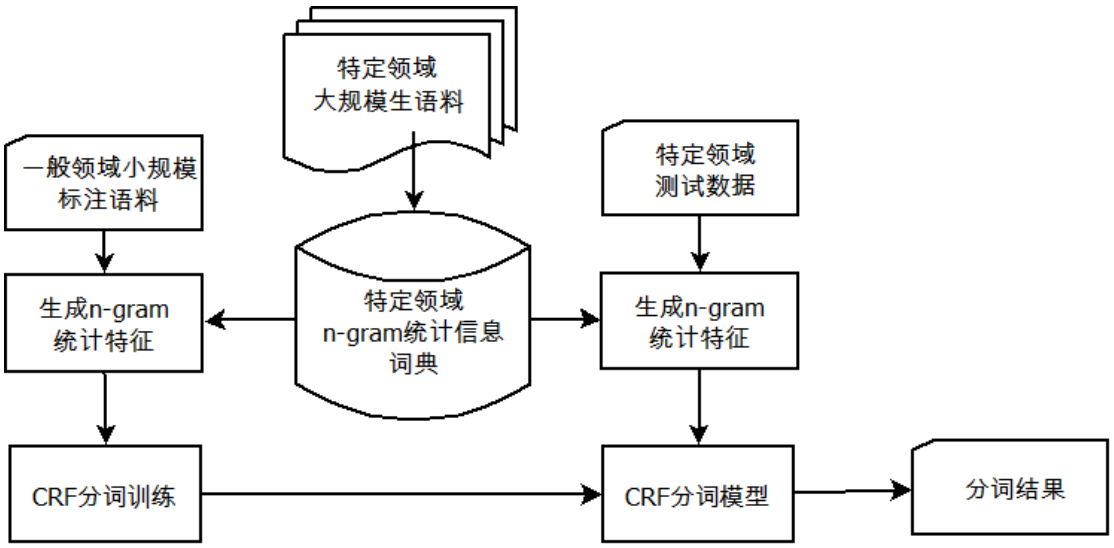

图 2-2 特定领域上汉语分词模型的自适应框架

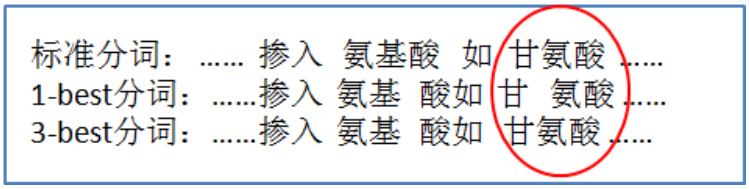

图 2-3 特定领域汉语分词结果比较：标准分词、1-best 分词和 3-best 分词结果

2．2 双语引导的汉语分词

在汉英句子平行语料中，英语部分有明确的单词界线，利用汉字字串与英文单词之间的

[1] https://code.google.com/p/crfpp/

对齐关系可以引导汉语分词。图 2-4 所示的例子是汉字字串“……癸二酸衍生单体……”与英文部分“……sebacic acid-derived monomer……”的对齐结果，这一对齐结果指示出，该字串可以被切分成 “癸二”、“酸衍生”和“单体”。在本节中，我们描述基于对齐结果置信度的分词方法。

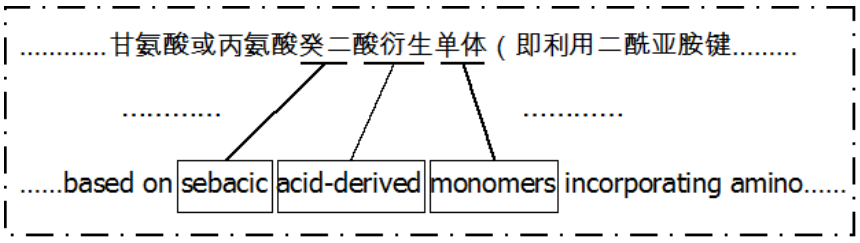

图 2-4 双语引导的汉语分词结果

给定汉英平行句对 $C_1^J = c_1 c_2 \ldots c_J$ 和 $E_1^I = e_1 e_2 \ldots e_I$，其中 $c_j(1 \le j \le J)$表示汉语句子中第 j 个汉字，$e_i(1 \le i \le I)$表示相应的英文句子中第 i 个单词。首先使用 GIZA++[2]对齐工具对汉英语料进行对齐，其中汉语以汉字为对齐单位，英语以单词为对齐单位。对齐处理进行双向训练(汉语到英文和英文到汉语)，并使用 grow-diag-final[3]启发式算法合并双向对齐结果。a_i=<e_i, C>表示英文单词 e_i 与汉字集合 C 的对齐，当集合 C 中的汉字在汉语句子中是连续的时候，该对齐结果可以用来引导这一字串为一个汉语单词。

定义 $Count(e_i, C)$表示 e_i 和 C 在平行语料中共现的次数，$Count\ (a_i)$表示对齐 a_i=<e_i, C>出现的次数，而 $Conf(a_i)$表示当 e_i 和 C 在平行语料中共现时的对齐置信度，由公式（2-1）计算得到。

$$Conf(a_i) = \frac{Count(a_i)}{Count(e_i, C)} \tag{2-1}$$

双语引导的汉语分词步骤如下：

1）将汉语句子切分为单个汉字，使用对齐工具(GIZA++)进行汉-英双向对齐，得到对齐结果并合并；

2）根据公式 2-1 计算所有对齐的置信度 $Conf\ (a_i)$；

3）对于每一个对齐结果 a_i=<e_i, C>，如果 C 在汉语句子中是连续的字串，则将 C 切为一个单词，并将 $Conf\ (a_i)$作为该切分结果的置信度。

2．3 多种分词结果的融合方法

在前面的两小节中我们分别介绍了基于生语料的领域自适应的汉语分词和双语引导的汉语分词的实现方法，本小节将介绍融合多种分词结果的方法。

为了综合利用汉语生语料中的 n-gram 统计特征和双语语料中的分词引导特征，我们对基于这些特征的多种分词结果进行整合，借助线性模型获取其中的正确信息以得到最佳的汉语分词结果。

我们取 CRF 分词模型的 10-best 以内结果作为 10 种基于大规模生语料 n-gram 统计特征的汉语分词结果，将每个结果的概率得分 $Conf_{CRF}$ 作为结果中每个单词的置信度；再取双语

[2] https://code.google.com/p/giza-pp/
[3] http://www.statmt.org/moses/?n=FactoredTraining.AlignWords

引导的汉语分词结果作为第 11 种分词结果，将对齐的置信度 *Conf* (*a*)作为相应的单词的置信度。为了融合这 11 个分词结果，我们设计了如公式(2-2)所示的线性模型。

$$F_{i,j} = \lambda_1 \cdot Conf_{CRF1} \cdot seg_1(i,j) + \cdots + \lambda_{10} \cdot Conf_{CRF10} \cdot seg_{10}(i,j) + \lambda_{11} \cdot Conf_{i,j} \quad (2\text{-}2)$$

$$w_{i,j} = \frac{F_{i,j}}{\sum_j F_{i,j}} \quad (2\text{-}3)$$

公式 2-2 中，$F_{i,j}$表示汉语句子中第 i 个节点（如图 2-5 中所示的标有 0 到 10 的圆圈，表示汉字的边界）到第 j 个节点之间的汉字C_i^j构成一个单词的支持度；$Conf_{CRFl}(1\le l\le 10)$是 CRF 分词模型的 10-best 以内结果对应的概率得分；$seg_l(i, j)$是一个二值函数，当 CRF 分词模型的第 l 种分词结果中C_i^j为一个单词时，$seg_l(i, j)$值为 1，否则为 0；$Conf_{i,j}$是双语引导的汉语分词结果中C_i^j切成一个单词的置信度；λ_l $(1\le l\le 11)$表示 11 种分词结果的特征权重。

按照公式(2-3)对 $F_{i,j}$进行归一化后得到支持度 $w_{i,j}$。我们使用格状结构(Lattice)表示 $w_{i,j}$，如图 2-5 所示。节点表示句子中汉字之间的边界，并标有序号。节点 i 和 j 之间的边<i, j>表示节点 i 和 j 之间的汉字构成一个单词，边的上面标有该单词的支持度 $w_{i,j}$。Lattice 的解码是一个动态规划的过程，寻找一个支持度乘积最大的分词结果。

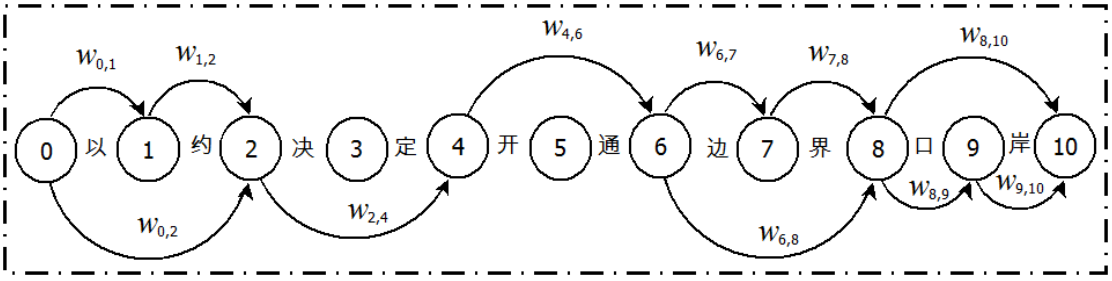

图 2-5 多种汉语分词结果的 Lattice

本文采用基于网格的线性优化算法[10]训练参数 λ_l $(1\le l\le 11)$。首先在 11 维参数空间中初始化一个点；然后迭代优化参数，每步迭代在固定其他维度参数条件下，优化一个维度的参数使得相应的分词结果 F 值最高；当分词结果的 F 值收敛到了某种期望的程度，结束迭代。为了避免训练参数局部最优，我们选择多个不同的初始点进行参数训练。

3 实验设计及结果

为了验证本文提出的对多种分词结构的融合方法，我们模拟机器翻译系统开发的实际场景。为此，我们在 NTCIR-10[4]的科技文献专利领域汉英翻译任务的数据上设计了实验，通过对其中大规模汉语语料的分词处理，从汉语分词的精度和机器翻译质量两个方面进行了评价。

3．1 实验数据

NTCIR-10 的汉英翻译任务提供了 1,000,000 句对的训练集、2,000 句对的开发集和 2,000

[4] http://research.nii.ac.jp/ntcir/ntcir-10/

句对的测试集。我们从 1,000,000 句对的训练集中随机抽出 300 句对，作为人工标注集，记为 AS；其余的句对作为机器翻译系统的训练数据，记为 TS。本文的目标就是提高 TS 中汉语语料的分词精度，改进机器翻译质量。依照宾州中文树库的分词标注标准[11]，我们对 AS 中的 300 句汉语句子进行了分词标注。随后对标注的 300 句随机地一分为二，记为 AS1 和 AS2，分别用于线性模型的参数训练和汉语分词精度的评测。

作为一般领域的标注数据，我们使用宾州中文树库 CTB 5.0 的数据训练 CRF 分词模型，训练数据文件包括 1-270 篇、400-931 篇和 1001-1151 篇。

3．2 汉语分词及其评价

首先建立领域自适应的 CRF 分词模型，使用宾州中文树库的数据作为标注数据；使用 TS 中汉语语料作为领域生语料提取 n-gram 统计特征。然后使用获得的 CRF 分词模型对 TS 中的汉语语料进行分词处理，并取 10-best 以内的分词结果。

接着进行双语引导的分词处理，使用汉英平行语料 TS 进行单词对齐，根据对齐信息得到 TS 中汉语语料上的双语引导的分词结果。

然后构建线性模型，使用 AS1 数据训练得到线性模型的参数。

最后利用线性模型融合 CRF 的 10-best 以内的分词结果和双语引导的分词结果，最终获得训练语料 TS 中汉语语料的分词结果。

为了评测分词结果的精度，我们对 AS2 的 150 句进行了同样的分词处理，评测结果显示在表 3.1 中。从表 3.1 中可以看出，融合多种分词结果的分词方法的召回率和准确率相对于 CRF 的 1-best 结果均高出 1 个百分点，F 值提升了 1.257%。

表 3.1 特定领域上汉语分词的评测结果

汉语分词方法	准确率	召回率	F 值
双语引导的汉语分词	73.1312%	61.4480%	66.7825%
CRF 1-best	90.2439%	90.7710%	90.5067%
多种分词结果融合的方法	91.6650%	91.8614%	91.7631%

通过分析这些分词结果，我们发现多种分词结果融合的方法有效地利用了多种分词结果中的正确分词的信息对 CRF1-best 中错误的分词进行了修正。表 3.2 中给出了这样的例子，下面对第一行的例子进行说明。CRF1-best 将汉语句子中的“甘氨酸”分为两个单词“甘”和“氨酸”，而 3-best、7-best、9-best 将“甘氨酸”分为一个单词；在平行语料中，对应的英文句子中有单词”glycine”，其汉语译语是“甘氨酸”，指示了汉语句子中应该把“甘氨酸”分成一个单词。融合算法有效地利用了这些信息，得到了正确的分词结果。因此我们认为多种分词结果融合的算法修正了 CRF1-best 中错误的分词结果，说明本文提出的融合方法在处理特殊领域的汉语分词任务时，具有较好的领域适应能力。

表 3.2 CRF1-best 中的错误分词被融合方法修正的例子

CRF1-best 的结果	融合方法的结果
甘　氨酸	甘氨酸
聚合　物	聚合物
碳原子	碳　原子
碘复合物	碘　复合物
抗　微生物	抗微生物

进一步查看训练语料 TS 中汉语语料的分词结果，总共得到了 37,109,126 个汉语单词，

在如此大规模的数据上，分词精度的微小提高具有数量上的实际意义。我们期待分词精度的提升能够改进机器翻译系统的质量。

3．3 机器翻译系统构建及评测

接下来，我们使用开源统计机器翻译工具 Moses[5]，在 NTCIR-10 的汉英数据 TS 上搭建基于短语的统计机器翻译系统，TS 中的汉语语料使用 3.1 节中获得的分词结果。然后使用 2,000 句对的开发集进行最小错误率训练[12]。最后使用 2,000 句对的测试集进行 BLEU[13] 评测，评测结果列于表 3.3。

为了与其他分词系统进行比较，我们也采用了现有公开的汉语分词工具 Stanford 汉语分词工具[6]和 NLPIR 汉语分词工具（ICTCLAS 2013 版）[7]，分别进行汉语分词处理，并搭建统计机器翻译系统。评测结果也列在表 3.3 中。

表 3.3 基于不同汉语分词方法搭建的统计机器翻译系统的评价结果

汉语分词方法	BLEU
CRF 1-best	30.53%
CRF10-best 以内结果与双语引导的分词结果的融合	31.15%
Stanford 汉语分词工具	30.98%
NLPIR 汉语分词工具	30.56%

在统计机器翻译系统的评测实验中，我们以 CRF 的 1-best 分词结果搭建的翻译系统作为 Baseline，它的 BLEU 值为 30.53%。当采用本文提出的多种分词结果融合的方法时，BLEU 值相对于 Baseline 系统提升了 0.62%。

3.2 节的实验结果已经显示融合方法的分词性能优于 CRF1-best 的性能，这是 BLEU 值提升的一个直接原因；另一方面，融合方法中引入了双语语料的分词引导特征，相对于 CRF1-best 的分词增加了双语单词对齐的信息，这有利于随后的单词对齐处理，提高了单词对齐结果的精度以及短语模型的精度，最终改善了翻译系统性能。以上实验和分析说明了本文提出的分词方法不仅在分词精度上有提高，而且直接带来了特定领域上统计机器翻译系统的性能提升。作为对比评测的另外两种分词方法， Stanford 汉语分词工具和 NLPIR 汉语分词工具的 BLEU 值分别为 30.98%和 30.56%。这两种分词方法都略逊于本文提出的汉语分词方法。

4 总结

本文实现了基于生语料的 n-gram 特征统计的汉语分词和双语引导的汉语分词，提出了一种融合多种汉语分词结果的方法，为特定领域汉英机器翻译开发中大规模汉语语料的分词问题提供了一种有效的解决方案。通过在 NTCIR-10 的科技领域汉英机器翻译开发数据集上的评测实验，显示了该方法在汉语分词精度 F 值和汉英统计机器翻译的质量 BLEU 值上都得到了提高。本文的方法具有很好的拓展性，可以融合更多基于不同特征的分词结果。在今后的工作中，我们考虑可以利用对齐中的单词翻译概率和对齐概率分布等信息提高双语引导的分词方法的性能，进一步提高机器翻译的质量。

参考文献

[1] Guo Z, Zhang Y, Su C, et al. Exploration of N-gram Features for the Domain Adaptation of

[5] http://www.statmt.org/moses/
[6] http://nlp.stanford.edu/software/segmenter.shtml
[7] http://ictclas.nlpir.org/

Chinese Word Segmentation[M]//Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2012: 121-131.

[2] Och F J, Ney H. The alignment template approach to statistical machine translation[J]. Computational linguistics, 2004, 30(4): 417-449.

[3] Chiang D. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 263-270.

[4] 张梅山, 邓知龙, 车万翔, 等. 统计与词典相结合的领域自适应中文分词[J]. 中国计算语言学研究前沿进展 (2009-2011), 2011.

[5] Wang Y, Kazama J, Tsuruoka Y, et al. Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data[C]//Proceedings of 5th International Joint Conference on Natural Language Processing. 2011: 309-317.

[6] Ma Y, Way A. Bilingually motivated domain-adapted word segmentation for statistical machine translation[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 549-557.

[7] 奚宁, 李博渊, 黄书剑, 等. 一种适用于机器翻译的汉语分词方法[J]. 中文信息学报, 2012, 26(3): 54-58.

[8] Jin Kiat Low, Hwee Tou Ng and Wenyuan Guo. A Maximum Entropy Approach to Chinese Word Segmentation. In Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing (SIGHAN05), 2005:161–164.

[9] Haodi Feng, Kang Chen, Xiaotie Deng and Weimin Zheng. Accessor variety criteria for Chinese word extraction[J]. Computational Linguistics,2004,30(1):75—93

[10] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. Numerical Recipes in C++. Cambridge University Press, Cambridge, UK.

[11] Xia F. The segmentation guidelines for the Penn Chinese Treebank (3.0)[J]. 2000.

[12] Och F J. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 160-167.

[13] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.