# 一种适用于机器翻译的汉语分词方法*

李博渊，奚 宁，黄书剑，张建兵，陈家骏
南京大学 软件新技术国家重点实验室，江苏 南京 210093
南京大学 计算机科学与技术系，江苏 南京 210093
E-mail: {liby, xin, huangsj, zhangjb, chenjj}@nlp.nju.edu.cn

**摘 要**：汉语分词是构建汉语到其他语言机器翻译系统的一项重要工作。基于单语的分词不一定完全适合机器翻译，一个适合于机器翻译所需要的分词方法，应该考虑到机器翻译所具有的双语特点。本文提出了一种单语和双语知识相结合的适应于统计机器翻译系统的分词方法。首先利用对齐可信度的概念从双语平行语料中抽取可信对齐集合，然后根据可信对齐集合对双语语料中的中文部分重新分词；接着将可信对齐分词的结果和单语分词工具的结果相结合，构建出一个新的分词训练语料，并用 CRF 分词模型训练出一个融合了单双语信息的分词工具。本文用该工具对机器翻译所需的训练语料、开发语料和测试语料进行分词并在基于短语的统计机器翻译系统上进行实验。实验结果表明，本文所提的方法提高了系统性能。

**关键词**：中文分词；统计机器翻译；对齐可信度

# Training a MT-motivated Segmenter

Li Bo-yuan, Xi Ning, Huang Shu-jian, Zhang Jian-bing, Chen Jia-jun
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210093
Department of Computer Science and Technology, Nanjing University, Nanjing, 210093
E-mail: {liby, xin, huangsj, zhangjb, chenjj}@nlp.nju.edu.cn

**Abstract:** Chinese word segmentation is the first phase prior to building statistical machine translation (SMT) systems. But the quality of Chinese word segmentation is not equivalent to that of SMT systems. Therefore, it is necessary to get MT-motivated Chinese word segmentation in order to improve the quality of translation. In the paper, we incorporate two kinds of knowledge to train a Chinese word segmenter, the first kind of knowledge comes from the Chinese-character-based bilingual alignment; and the other kind comes from conventional Chinese word segmenters that monolingually motivated. We combine both to train an MT-motivated word segmenter (bilingual segmenter) using Conditional Random Fields. We tokenize the Chinese portion of training, development, and evaluation corpus with bilingual segmenter to build a phrase-based machine translation system. Experiments showed that our approach effectively improve the translation quality by adapting the Chinese corpus to MT-motivated segmentation.

**Keywords:** Chinese word segmentation; statistical machine translation; reliability of word alignment

## 1 引言

获取双语词对齐信息是构建统计机器翻译系统的一项重要工作。在汉英机器翻译系统中，我们首先需要对中文句子做分词处理以适应词对齐工作的要求。从单语的角度而言，目前中文分词方法的研究已经取得了很大进展，并且存在许多成熟的模型和分词工具可供使用。然而，已有研究表明，衡量单语分词质量的 F-score 得分与机器翻译系统的质量之间并无明显关联[1],[2]。因此，如何找到一种更适应统计机器翻译工作的分词方法，已经成为统计机器翻译工作中的一个新的研究方向[1],[2]。

常用的传统分词方法通常由已分好词的单语语料训练得到。这种方法虽然有效利用了单语知识，但却忽略了机器翻译训练语料里包含的双语信息，从而导致分词结果中的汉语词与英文单词不能尽可能多的一一对应，影响了翻译系统的质量。

* 本工作得到国家自然科学基金（61003112）的资助

Ma et al.[3] 和 Paul et al.[4] 试图仅从训练语料的双语对齐信息中寻找更适应机器翻译系统的“分词”方法。这种方法的分词准确率受到词对齐质量的制约，在产生的分词结果中含有大量字符序列不能被识别成词（识别率较低）的现象，进而影响了翻译系统的性能。为了弥补分词识别率上的损失，Ma et al.和 Paul et al.通过在解码器端增大解码空间的方法（牺牲效率）来解决翻译质量问题。

本文介绍了一种融合单语和双语知识的面向汉英机器翻译的分词方法。在双语字对齐的基础上，利用字对齐的可信度得到符合双语对齐知识的词，然后结合单语分词工具对不符合可信度要求的汉语字序列进行修正，得到一种新的分词结果，并用在此结果上进行训练得到最终的分词模型。本文使用常用的基于短语的统计机器翻译系统[5]对本文提出的分词方法进行了测试，实验表明，与传统的分词方法相比，即使采用普通的解码器，使用本文的分词方法也能使统计机器翻译系统的性能得到提升。

本文第二章将详细阐述对齐可信度和单语分词知识相结合的分词方法，第三章介绍本文所进行的实验及实验结果。第四章对文章进行了总结，并对下一步的工作做出了展望。

## 2 基于单语和双语知识相结合的分词模型

双语对齐可以作为分词的重要依据，我们首先以汉语字符为单位得到汉英对齐信息，然后根据汉英对齐中的多对 1 模式来发掘可以成词的潜在字符序列。本文将引入可信对齐[3]的概念，并在 2.2 节利用这一概念对潜在成词字符序列进行筛选合并，得到基于双语知识的分词结果。

### 2.1 可信对齐

给定中英文句对 $C_i^j\{c_1,c_2\cdots\cdots c_j\}$和$E_i^j\{e_1,e_2,\cdots\cdots e_j\}$，用 $A_{C->E}$ 来表示汉语到英语间的字对齐信息。则 $A_{C->E}$ 可以表示为$\{a_1,a_2\cdots\cdots a_i\}$，其中$a_i=<C_i,e_i>, C_i=\{C_{i1},C_{i2}\cdots\cdots C_{im}\}$表示多个中文字符集合 $C_i$ 对齐到一个英文单词 $e_i$。

对任意对齐组合$a=<C_i,e_i>$，用 $COOC(C_i, e_i)$表示 $C_i$ 和 $e_i$ 在平行语料中成对的次数，C(a)表示$<C_i,e_i>$同现时被对齐工具对齐在一起的次数，备选可信对齐的可信度[3]为：

$$AC(a)=\frac{C(a)}{COOC(C_i,e_i)}$$

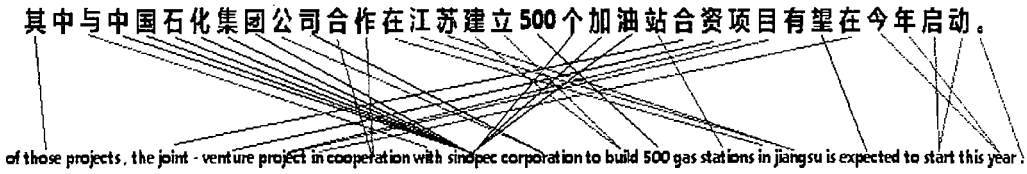
对于某一对齐组合$a=<C_i,e_i>$，只有当 $COOC(C_i, e_i)$和 AC(a)同时满足一定条件时，才称该对齐组合为一个可信对齐组合[1]。

### 2.2 基于可信对齐的汉语分词

若对齐组合$a=<C_i,e_i>$中 $C_i$ 由连续字符组成（即 $C_i$ 中的字符都是相邻字符序列），可以将 $C_i$ 视为潜在的成词序列。在该对齐组合同时满足可信对齐的条件的情况下，将 $C_i$ 合并成一个整体将更有利于简化词对齐训练，提高词对齐的准确率。

如图 2.1 所示，本文在汉英字对齐的基础上，利用可信度概念，将语料中可信对齐组合的汉语部分合并成“词”（图中红色对齐标识部分），非可信对齐部分保持单字不变（图中黑色对齐标示部分），从而得到一组新的“分好词”的训练语料。

[1] 在考虑到语料稀疏性和保证较高准确性的情况下，本文只将$COOC(C_i, e_i)>20$ 并且 AC>0.5 的对齐组合认为是可信的对齐组合[3]。




其中与中国石化集团公司合作在江苏建立500个加油站合资项目有望在今年启动。

图2.1 由可信对齐得到新的分词语料

具体方法和步骤如下：

- 将原始双语训练语料中的汉语部分按字进行切分，利用词对齐工具对训练语料进行基于字的对齐训练。
- 取对齐结果中的所有汉英多对一组合 $a_i =< C_i, e_i >, C_i = \{C_{i1}, C_{i2} \cdots\cdots C_{im}\}$，如果 $C_i$ 由连续相邻字符组成，且 $a_i$ 满足可信度标准的要求，则将 $a_i$ 视为一个可信对齐。
- 将双语训练语料中可信对齐的汉语部分合并，得到“分好词”的新语料。

在图2.1中，例句中含有中文字符33个，使用可信度分词法可确定词10个，被确定词中包含字符17个，占总字符比例为51.5%。可以看出，基于对其可信对齐的分词结果中，词语的识别率较低。

本文将本步骤生成的语料叫做“可信对齐分词语料”，将本节所述的分词方法称为“可信对齐分词”。

## 2.3 可信对齐和单语分词相结合

使用可信对齐分词方法得到的分词语料往往具有较低的分词识别率。为了提高分词结果中的识别率，本文将可信对齐分词的结果和单语知识分词的结果进行融合。用基于单语知识的分词结果对可信对齐分词无法判断的汉字序列进行修正。

**待分词句子：**
其中与中国石化集团公司合作在江苏建立500个加油站合资项目有望在今年启动。
**可信对齐分词结果：**
其 中 与 中 国 石 化 集 团 公司 合作 在 江苏 建立 500 个 加 油 站 合 资 项目 有 望 在 今年 启动 。
**单语分词工具分词结果：**
其中 与 中国 石化 集团公司 合作 在 江苏 建立 500 个 加油站 合资 项目 有望 在 今年 启动 。
**合并的结果和标注：**
其中 与 中国 石化 集团 公司 合作 在 江苏 建立 500 个 加油 站 合资 项目 有望 在 今年 启动 。

图2.2 “双语知识分词法”分词示例

如图2.2所示，对于一个待分词句子，分别利用可信对齐分词方法和单语知识分词方法对齐进行分词，再将两种分词方法的结果进行合并。合并时以可信对齐分词的结果为主，单语知识分词方法的结果为辅。如“公司”在可信对齐分词结果中是一个确定的词，因此“集团公司”的最终的分词结果应为“集团 公司”。

为表述简便，本文后续部分将可信对齐分词与单语知识分词相结合的方法称为“双语知识分词”法。

## 2.4 训练分词模型

在统计机器翻译系统中，汉语分词的一致性对机器翻译系统的性能有着重要影响。因此在对系统的训练语料和测试语料进行分词时需要使用统一的方法。然而，由于缺乏相应的双语语料，测试语料的分词工作无法使用双语知识相结合的分词方法，因此，需要寻找一种新的分词方法来对

训练和测试语料进行分词。并且新方法应满足如下条件：

- 新方法中应包含双语知识信息；
- 新方法能对中文单语语料分词。

本文通过使用条件随机场分词模型[6],[7],[8]来解决上述问题。

- 通过本章 2.3 节描述的方法，从双语训练语料中得到“分好词”的中文语料；
- 将得到的“分词”语料作为条件随机场模型的训练语料。得到基于条件随机场模型的分词器。

由于使用了可信对齐分词和单语知识分词相结合的结果作为训练语料，故在训练得到的模型中一定包含双语知识信息。而条件随机场分词模型本身可以对仅含中文的单语语料进行分词。因此，通过上述方法得到的分词器满足本文所提的条件，本文以此分词器得到的分词结果作为本文的最终分词结果。

# 3 实验及分析

## 3.1 数据和实验环境

本文以 1998 年 1 至 6 月份已分好词的人民日报作为单语知识分词的训练语料，从 LDC2003E14 语料中选取了 19 万句中英平行句对作为统计机器翻译系统的训练语料，使用 NIST'06 测试集作为系统的开发语料，NIST'08 测试集作为测试语料。最后，使用 SRILM 对 Gigaword 中的 Xinhua 部进行分训练，得到了一个 5 元语法模型作为机器翻译系统的语言模型。

在单语分词方面，本文用条件随机场模型结合人民日报的语料训练出一个分词工具 CRF-Based。汉语单字与英文单词的对齐结果则由开源的词对齐工具 GIZA++训练得到。在计算对齐可信度之后，本文尝试将可信度分词的结果与不同单语分词工具得到的结果相结合（CRF-Based、ICTCLAS[1]、Stanford Chinese Segmenter[2]），并将最终结果用于统计机器翻译系统的训练。

在机器翻译系统方面，本文采用了一个类似 Moses 的基于短语的统计机器翻译系统，并采用最小错误率训练方法（minimum error rate training，MERT）[9]进行参数训练。最后用系统翻译结果的 BLEU 得分[10]对系统性能做出评价。

## 3.2 单语知识分词

本文使用 CRF++[3]作为条件随机场模型的训练工具。采用四字位标注法[11]（见表 3.1 所示）和基于子串的序列化标注方法进行分词[12]，其中子串部分为基于规则识别出的英文单词和表示数字的词。分词模板如表 3.2 所示。

表 3.1 四字位标注集的定义

| 标记 | 单字词与对字词的字位标注举例 |
|---|---|
| B、I、F、S | S，BF，BIF，BIIF，... |

在四字位标注法中，用 B 表示词首，I 表示词中，F 表示词尾，S 表示单字词

表 3.2 CRF 分词工具采用的模板

| 特征模板 |
|---|
| $C_n$, n = –2, –1, 0, 1, 2 |
| $C_nC_{n+1}$, n = –2, –1, 0, 1 |
| $C_{-1}C_1$ |
| $F(C_0)$ |
| $F(C_{-1})F(C_0)F(C_1)$ |

[1] http://www.ictclas.org/

[2] http://nlp.sttanford.edu/software/segmenter.shtml

[3] http://crfpp.sourceforge.net/

表 3.2 中 $C_n$ 表示单个字符作为特征，n = 0 表示当前字符，n = –1 表示当前字符的前一个字符，n = 1 表示当前字符的后一个字符。$C_nC_{n+1}$ 表示相邻的两个字符组合作为特征。$C_{-1}C_1$ 表示当前字符的前后字符组合作为特征。$F(C_n)$的值表示该字符是否是汉字、标点或者子串。

本文首先使用人民日报 1998 年 1 至 5 月份的已分词语料作为模型的训练语料，训练得到一个可以用于分词工作的 CRF 模型 CRF-Based。使用人民日报 6 月份的语料对 CRF-Based 进行分词测试，正确率为 97%，召回率为 97.1%。使用公式(3-1)进行计算并取 β = 1，得到 F-score 为 0.97。

$$F = \frac{(\beta^2 + 1) \times 准确率 \times 召回率}{\beta^2 \times 准确率 + 召回率} \tag{3-1}$$

表 3.3 CRF-Based 性能表

| 正确词数 | 分得词数 | 准确率 | 召回率 | F 值 |
|---|---|---|---|---|
| 1232237 | 1234194 | 0.970 | 0.971 | 0.970 |

## 3.3 机器翻译实验

本文首先使用 CRF-Based 分词模型对机器翻译所用的训练和测试等语料进行分词，可以得到用于统计机器翻译系统的完整语料。经测试，该系统的 BLEU 得分为 21.82。

用 3.2 节所描述的 CRF 模板对 LDC2003E14 双语平行语料的可信对齐分词结果进行训练，得到一个基于可信对齐的 CRF 分词模型 AC-Based。使用 AC-Based 对机器翻译所需语料进行重新分词，最终得到的 BLEU 得分为 21.05。由于可信对齐分词中含有大量的未识别字符，因此 AC-Based 分词结果中存在大量单字词，影响了机器翻译的质量。

将 LDC2003E14 双语平行语料的可信对齐分词结果与单语分词模型 CRF-Based 得到的分词结果相结合，并使用 3.2 节所描述的 CRF 模板对合并结果进行训练可以得到一个结合单双语知识的分词模型 AC+CRF。使用 AC+CRF 对机器翻译的各项语料进行分词，系统的 BLEU 得分为 22.46。可见，使用本文所提的分词方法，可以使统计机器翻译系统的性能得到一定的提升。表 3.4 展示了分别使用三种方法的统计机器翻译系统的 BLEU 得分。

表 3.4 使用三种分词方法的 SMT 系统的 BLEU 得分

| | DEV | TEST |
|---|---|---|
| CRF-Based | 27.26 | 21.82 |
| AC-Based | 25.05 | 21.05 |
| AC+CRF | 26.71 | 22.46 |

其中 DEV 表示使用 NIST06 语料作为系统开发集时的最终 BLEU 得分，TEST 表示该系统在 NIST08 语料上进行测试的 BLEU 得分。

为了探讨 AC+CRF 使机器翻译性能提高的原因，我们对 CRF-Based 与 AC+CRF 的分词结果进行了对比，表 3.5 展示了二者的一些分词差异。

表 3.5 CRF-Based 与 AC_CRF 分词示例

| CRF-Based 汉语分词结果 | AC+CRF 汉语分词结果 | English |
|---|---|---|
| 核反应炉 | 核/ 反应炉 | Nuclear reactor |
| 北韩 | 北/ 韩 | North Korea |
| 原子能 | 原子/ 能 | Atomic energy |
| 世贸/ 组织 | 世贸组织 | WTO |
| 更/ 多 | 更多 | More |
| 特别/ 是 | 特别是 | Especially |

由表 3.5 可见，在对组合型词语进行切分时，AC+CRF 分词方法可以使分词结果中的汉语词和其对应英文翻译中的英文单词之间尽可能多的一一对应。因此能够在 CRF-Based 分词的基础上降低词对齐工作的难度，提高机器翻译系统的性能。

为了进一步验证本文方法的可行性，本文将可信对齐分词的结果分别与单语的分词工具的结果（ICTCLAS、Stanford Chinese Segmenter）相结合，训练统一的分词模型后用于统计机器翻译系统。如表 3.6 所示，性能上，结合单、双语知识进行分词的机器翻译系统均优于使用单语分词工具的机器翻译系统。

表 3.6 可信对齐分词结合不同分词工具后的 SMT 系统 BLEU 得分

| | DEV | TEST |
|---|---|---|
| AC-Based | 25.05 | 21.05 |
| ICTCLAS | 26.93 | 22.03 |
| AC+ICT | 26.79 | 22.37 |
| Stanford_PKU | 27.61 | 22.13 |
| AC+PKU | 27.35 | 22.79 |
| Stanford_CTB | 27.33 | 22.00 |
| AC+CTB | 27.09 | 22.30 |

AC+ICT 为可信对齐结合 ICTCLAS 的结果，AC+PKU 和 AC+CTB 分别为可信对齐结合 Stanford Chinese Segmenter 中 PKU 模型和 CTB 模型的结果。

## 4 总结

本文从汉英机器翻译中的中文分词工作入手，旨在寻找一种更适应于机器翻译系统的分词方法。本文通过对单语知识分词和双语对齐知识分词的优劣分析，提出了一种新的结合双语知识的分词方法。与传统分词方法相比，本文提出的分词方法可以使汉语词与英文单词间的对应关系更加明确，具有更好的机器翻译性能。

由于训练模板的限制，我们提出的分词模型并不能达到 100%的准确率。其中包含的分词错误仍然会影响机器翻译的性能。另一方面，基于 Lattice[13]的解码已经得到广泛讨论，基于 Lattice 的解码方法可以摆脱机器翻译解码时对某一种分词结果的依赖。因此，在未来的工作中，我们将尝试将本文的分词方法和基于 Lattice 的解码相结合，以克服分词错误对系统性能带来的影响。

## 参 考 文 献


[1] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing Chinese word segmentationfor machine translation performance. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 224-232, 2008.

[2] Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. Improved statistical machine translation by multiple Chinese word segmentation. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 216-223, 2008.

[3] Yanjun Ma and Andy Way. Bilingually Motivated Domain-Adapted Word Segmentation for Statistical Machine Translation. In Proceedings. of the 12th EACL, pages 549-557, 2009.

[4] Michael Paul, Andrew Finch and Eiichiro Sumita. Integration of Multiple Bilingually-Learned Segmentation Schemes into Statistical Machine Translation. In Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 400-408, 2010.

[5] Philipp Koehn, Franz Josef Och and Daniel Marcu. Statistical Phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 923-940, 2003.

[6] John D. Lafferty, Andrew McCallum and Fernando C. N. Pereira. Conditional Random Field: Probabilistic models for segmenting and labeling sequence data. In Proceedings 18th International Conference on Machine Learning, pages. 282-289, 2001.

[7] Fuchun Peng, Fangfang Feng and Andrew McCallum. Chinese segmentation and new word detection using Conditional Random Fields. In Proceedings of the 20th international conference on Computational Linguistics, pages 562-568, 2004.

[8] Jun-Sheng Zhou, Xin-Yu Dai, Rui-Yu Ni and Jia-Jun Chen. A hybrid approach to Chinese word segmentation around CRFs. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pages 196-199, 2005.

[9] Franz Och. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational, 2003.

[10] Kishore Papineni, Salim Roukos, ToddWard, andWei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311-318, 2002.

[11] Nianwen Xue and Libin Shen. Chinese word segmentation as LMR tagging. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pages 176-179, 2003.

[12] 赵海, 揭春雨. 基于有效子串标注的中文分词. 中文信息学报, 21(5): 8-13, 2007.

[13] Christopher Dyers, Smaranda Muresan, and Philip Resnik. Generalizing word lattice translation. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pages 1012-1020, 2008.