面向短语统计机器翻译的汉日联合分词研究

吴培昊[1]，徐金安[1]，张玉洁[1]

（1.北京交通大学计算机与信息技术学院，北京 100044）

摘要：未登录词与分词粒度是汉日日汉机器翻译研究的两个主要问题。与英语等西方语言不同，汉语与日语词语间不存在空格，分词为汉日双语处理的重要工作。由于词性标注体系、文法及语义表现上的差异，分词结果的粒度需要进一步调整，以改善统计机器翻译系统的性能。本文提出了面向统计机器翻译的基于汉日汉字对照表及日汉词典信息的汉语与日语的分词粒度调整方法。实验结果表明，该方法能有效地调节源语言和目标语言端的分词粒度，提高统计机器翻译系统的性能。本文通过对比实验结果，分析探讨分词粒度对汉日双语统计系统性能的影响。

关键词： 分词粒度；汉字对照表；汉日机器翻译；

中图分类号：TP391 **文献标识码：**A

Research on Joint Chinese-Japanese Word Segmentation for Phrase-based Statistical Machine Translation

Peihao Wu[1], Jinan Xu[1] ,Yujie Zhang[1]

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

Abstract: Unknown words and word segmentation granularity are two main problems for Chinese-Japanese Machine Translation. Word Segmentation is the first important step for Chinese and Japanese Natural language Processing. As Chinese and Japanese word segmentation is processed with different tagging systems and semantic performance, the granularity of word segmentation results should be readjusted to improve the performance of Statistical Machine Translation (SMT). This paper proposes an approach to adjust the word segmentation granularity for improving the performance of SMT, which combines Hanzi-Kanji comparison table and Japanese-Chinese dictionary. Experimental results express that the proposed method could adjust the granularity between Chinese and Japanese effectively and improve the performance of SMT. At last, this paper analyses the experimental results and discusses the effect of joint Chinese-Japanese word segmentation granularity for Phrase-based SMT.

Key words: segmentation granularity; Kanji-Hanzi comparison table; Chinese-Japanese MT;

1 引言

汉语与日语不同于英语等西方语言，句子不存在空格作为边界的词分隔符[1]。因此，分词为汉语与日语自然语言处理的重要工作。现有的关于汉语与日语分词技术发展较为成熟，存在许多分词工具。然而，由于汉语和日语分词大多根据自身的语言特点指定词性标注体系。词性标注体系的差异导致分词粒度存在差异，同时分词粒度在信息检索、机器翻译等具体应用领域中产生不同的影响效果。另一方面，既有研究成果表明，评价分词性能的 F-score 值与机器翻译系统的性能之间并没有明显关系[2,3,4]。统计机器翻译系统中，一种提高翻译系统性能的方法为通过调整分词粒度，对源语言和目标语言端分词结果进行调整。因此，如何调整汉语和日语分词粒度，以提高统计机器翻译系统的性能，是一个值得探讨的研究课题。

收稿日期： **定稿日期：**

基金项目：科技部国际科技合作计划(K11F100010)；中央高校基本科研业务费专项资(2010JBZ2007)；北京市重点学科共建项目（计算机应用技术）；中国科学院计算技术研究所智能信息处理重点实验室开放课题(IIP2010-4)；北京交通大学人才基金(2011RC034)。

作者简介：吴培昊（1990—），男，硕士研究生，主要研究方向为机器翻译；徐金安(1970—)，男，博士，副教授，研究方向为自然语言处理和机器翻译；张玉洁（1961—），女，博士，教授，主要研究方向为自然语言处理和机器翻译。

导致不同或同种语言分词粒度不同的原因，大致归纳如下：

1. 语系不同导致分词粒度不同。例如，汉语属于孤立语系，日语属于黏着语系，各自形成语义的构成要素存在较大差异。
2. 词性标注体系不同，导致分词粒度不同。
3. 使用目的不同，对分词粒度存在不同要求。
4. 语言文化、语法构成和语义表现等的差异，导致分词粒度不同。
5. 未登录词识别问题导致粒度不同。

由于异种语言间的词汇、语法和语义层面上大多是非同构的，很难达到词与词之间的一一对应关系[5]。既有的单语分词结果在使用于机器翻译时，需要同时考虑源语言与目标语言的词法特点，对双语分词粒度进行整合，以期改善统计机器翻译系统的性能。分词粒度对汉日双语统计机器翻译的影响，还有待深入的研究。

目前，面向统计机器翻译的汉语分词粒度研究的主流方法是依据另一端语言分词信息，对汉语分词粒度进行调整。在汉英统计机器翻译领域，Wang 等[6,7]的实验表明，细粒度分词结果，能提升统计机器翻译系统的性能。Ma 等[8]提出基于训练语料的自适应方法，采用可信对齐构建字格(word lattice) 对汉语端进行粒度调整，以提升分词的领域适应能力。奚宁等[5]描述一种基于可信对齐与单语分词相融合的策略对汉语分词进行调优。Bai 等[9]依据汉英词典对齐信息抽取汉语粒度切分的规则模板，使用模板进行汉语分词粒度调整。Wang 等[6]采用一种半自动(semi-automatic)的学习方法，对汉语分词进行短单元(short-unit)的调整。Dyer[10]和 Zhang[3]等人基于多策略汉语分词对汉英统计机器翻译解码过程进行优化。

由于日语端不存在空格作为词分隔符，无法确定上述在中英有效的方法在汉日机器翻译中是否有效。汉语与日语语言中均使用汉字，因此在汉日机器翻译中可使用汉字对照表作为特征信息进行粒度调整。Chu 等[11]使用汉日汉字对应信息，通过日语端分词结果对汉语端分词结果进行调优，该方法没有对汉语的分词粒度进行考察，也没有同时调整汉日双语的分词粒度。

为系统地考察通过改善分词粒度提高汉日双语统计机器翻译系统性能的可行性，本文使用简体汉字与日语汉字对照表以及日汉词典相结合，提出一种提高统计机器翻译系统性能的汉日双语分词粒度调整策略。实验结果表明，提出的方法能有效调节汉日双语分词粒度，提升机器翻译系统的性能。

本文第一节讨论阐述汉日汉字对照表的构建以及词典的处理方法；第二节阐述使用汉日汉字对照表及词典对分词粒度调整的策略，并分析汉日双语分词粒度之间的差异；第三节介绍本文的实验方法、实验结果和分析；最后，对本文进行总结并展望未来工作。

2 汉字对照表构建及词典处理

2.1 汉日汉字对照表构建

汉字在汉语与日语中均被广泛使用[12]，日语汉字来源于古汉语，因此日语汉字与汉语汉字（包含简体汉字与繁体汉字）在很多情况下是相同的。然而，如表 1 所示，日语汉字与汉语汉字的对应关系十分复杂。Goh 等[13]使用日汉字典，通过直接匹配的方法，将日语汉字转化为汉语汉字；Chu 等[12]使用开源资源构建日语汉字、繁体汉字、简体汉字对照表。

表 1 汉字不同表现

日语汉字	愛	国	書	氷
繁体汉字	愛	國	書	冰
简体汉字	爱	国	书	冰

汉日双语翻译系统中，汉语端通常只包含简体汉字，因此本文构建日语汉字与简体汉字的对照表。图 1 为本文提出的日语汉字与简体汉字对照表构建的流程图。该流程中，本文共使用三类字典信息：

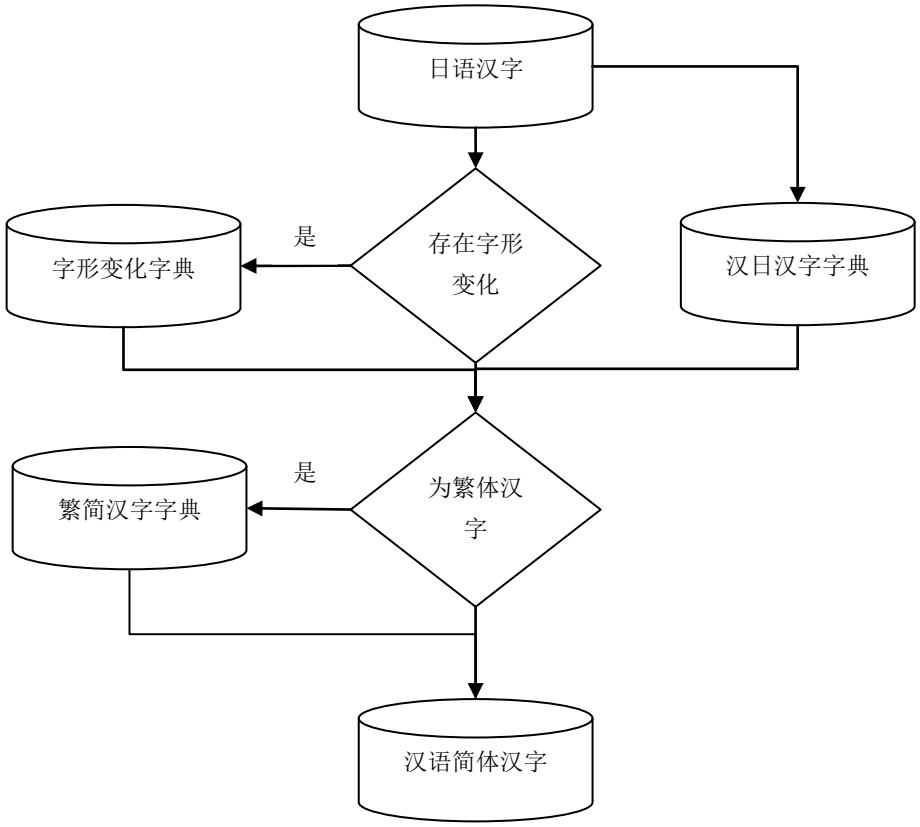

图 1 日语汉字转化为简体汉语汉字流程

1. 字形变化字典。一个汉字在汉日字典中可能存在多种不同字形，因此在构建字典时，可以枚举每种字形情况进行对应关系抽取。Unihan Database[1]为 Unicode Consortium 的中日韩三语的知识数据库。该数据库中包含每个汉字的变型（variants）特征信息，该特征信息记录了日语汉字与汉语汉字之间的关系。本文采用 variants 对日语汉字进行字形变化，若两个汉字之间通过 variants 存在联系，则说明两个汉字可以相互转化。

2. 汉日汉字字典。本文使用 Kanconvit[2] 中的汉日汉字转化表作为汉日汉字字典，该字典共包含了 1,159 个词表变型(variants)不同的汉字对信息。

3. 繁简汉字字典。如表 2 所示，繁简汉字之间并非简单的一一对应关系。本文使用 Chinese Encoding Converter[3] 中的繁简汉字转化表作为繁简汉字字典。该表含有 6,740 对繁简单词转化信息。

本文通过上述方法与资源构建简体汉字与日语汉字转化表。

表 2 繁简汉字对应表

繁体汉字	萹，稿	變	並，併	佈
简体汉字	扁	变	并	布

在 EDR 词典中，存在如表 3 所示情况，相同语义的词并没有对应关系。因此，本文使用两个步骤对词典进行整合：1）使用汉日汉字对照表将日语的汉字转化为汉语汉字；2）若任意两行词典信息中存在相同词，则认为两行词典中的所有词均为同义词，并将两行的数据合并。

通过上述两个步骤，获得最终的日汉对照词典。

[1] http://unicode.org/charts/unihan.html
[2] http://kanconvit.ta2o.net
[3] http://www.mandarintools.com/zhcode.html

表 3 词典中同义词信息样例

日语端	汉语端
華語	汉语
漢語	汉语　中文　中国话
支那語	汉语
中国語	中国语　汉语
唐言	中国话　汉语
唐詞	汉语

3 汉日双语分词粒度调整

3.1 双语粒度差异抽取

Bai 等[9]表明调整分词粒度使得双语词素间达到一一对齐关系，能优化对齐结果，从而提升机器翻译精度。本文使用汉日汉字对照表以及日汉词典对双语平行语料进行分词粒度处理，抽取出双语分词粒度不同的单词对，以进行下一步工作。

抽取过程主要包含以下两个方面：

1. 抽取字表信息相同的词对：若某一端单词通过汉日汉字对照表进行汉字转化，得到的结果与另一端的连续单词序列完全相同，则称该词对的字表信息相同。例如，汉语端单词“中国人”通过对照表可转化为“中國人”，同时日语端分词结果存在单词序列“中國 人”，则“中国人”与“中國 人”的字表信息相同。通过汉日汉字对照表，抽取所有字表信息相同但双语端分词粒度不同的词对。

2. 抽取字典信息相同的词对：如果字表信息不同，则依据字典信息，抽取单语端为词，另一语言端为词序列，并且存在于词典中的词对。例如，日语端“刻削な（残忍 的）”，通过词典信息可以查询到该单词汉语端应为“残忍的”，在汉语分词结果中，“残忍的”被切分为“残忍”和“的”两个单词。通过词典抽取字典信息相同，汉日双语分词粒度不同的词对。

本文使用 CWMT2011 汉日新闻语料进行测试，通过本节所述方法进行词对抽取。本文对字表信息或字典信息相同，但分词粒度不同的词语进行归纳总结，主要存在以下几类汉日切分中的不同。

以下两小节将从 1）汉语细粒度分析；2）日语细粒度分析两个方面进行双语分词粒度差异分析。

3.2 汉语细粒度分析

汉语端单词被切分为细粒度的原因主要如下几类：

1. 汉语中出现的日语专有名词无法正确切分。主要包括日语中特有的命名实体，即人名、地名、组织名等。例如“山田”为日本人名，而汉语分词时无法识别，切分为“山 田”造成错误。

2. 汉语结构助词。汉语中结构助词“的”、“地”、“得”用法较为复杂，例如“恐れながら（冒昧 地）”、“うれしい（高兴 的）”、“思わず（不由 得）”等。中文端将结构助词单独成词，日语端由于语法及语义的原因，汉语结构助词信息往往包含于日语单词中，从而造成汉日切分粒度不同。

3. 日语缩略语。日语存在大量的汉语缩略语，例如日语端单词“急变（突然变化）”，汉语端将该词切分成多个单词“突然”和“变化”，造成汉日分词粒度不同。

4. 汉语“不”字问题。“不”在在汉语中常表示对后续词的否定，被独立切分成词。而日语语法中一般使用词尾变化表示否定意义，例如“つまらない（不 值钱）”中，使用后缀“ない”表示否定，而汉语端切分为“不”和“值钱”两个单词，类似的还有“めちゃくちゃ（不 合理）”、“不仲（不 和睦）”、“不作法（不 礼貌）”等，日语端均为一个单词，而汉语端为多个单词，造成切分粒度不同。

5. 日语熟语。日语存在的固有熟语，例如“おはよう（您 早）”、“乗り物（交通 工具）”、“乗り合い（公共 马车）”等，在汉语端均切分为多个词语。

6. 日语动词后缀问题。类似于“不”字，日语均使用后缀变化进行动词的时态等的变化，因此“乗れる（能 乘坐）”、“吐き出せる（能 吐出）”等词的汉语端粒度均无法与日语端一致。需要注意的是，日语中不同词所使用的否定意义的词缀不同，不易将日语端词缀进行切分。

3.3 日语细粒度分析

日语端单词被切分为细粒度的原因主要如下几类：

1. 数词、时间词。汉语分词将数词和相关的后续词合并，日语端则分开处理。例如“16日”，“1．95 V”等均进行了分割。

2. 汉语专有名词。主要包括汉语中的专有名词例如人名“丁美媛”、“一年生”、“中央军事委员会”等日语中均无法进行正确切分。

3. 汉语熟语。汉语中存在一些固定用语的情况，如“一海知义（一海知義）”、“一瞬间（一瞬间）”等，由于固定用语并不存在于日语分词词典中，日语分词中无法与汉语端粒度相对应。

4. 词类后缀。在汉语中“市”、“县”、“部”、“街”、“人”、“化”等词语后缀均与相关词汇合并为一个词，而日语中将此类词单独成词。

3.4 汉日双语分词粒度调整

3.2 与 3.3 节的分析表明，由于汉日分词工具分词结果的不同，汉日双语分词粒度差异严重，双语粒度并没有达到一一对应的效果。

本文使用 3.1 节所述方法，通过汉日汉字转化词典、日汉对照词典，从分词后的语料中，抽取分词粒度不相同，但字表信息相同，或字典信息相同的词对。

本文处理中，只考虑抽取的词对中，存在一端为单词的情况。对于字表与字典信息相同的词对，我们采取不同的处理方式。

若该词对字典信息相同，则将词对的任意端都合并成一个单词处理。

若该词对字表信息相同，由于可以正确获取到每个单词的对应信息，因此，我们可以使用如下两种方法处理。

1. 词对中一端单词依据另一语言端的词序列，切分成与另一语言端序列一致的单词序列。

2. 将分词结果为词序列的一端，合并成一个单词进行处理。

例如，中文端单词“中国人”，为一个单词，而日语端为词序列“中國 人”。既可以考虑使用方法 1，将中文端“中国人”切分为词序列“中国 人”；也可以考虑参照方法 2，将日语端词序列“中國 人”合并为单词“中國人”。

下一章对本节提出方法进行实验测试，研究分词粒度变化对汉日双语统计机器翻译系统性能的变化。

4 实验及结果分析

4.1 实验数据及工具

本文使用 CWMT2011 汉日新闻语料，使用经过处理后的 282,476 句对作为实验训练集，498 句对作为开发集，948 句对作为测试集。使用 NLPIR2013[4]作为汉语分词工具，选用 mecab[5]作为日文分词工具。本文所有实验均采用 moses[6]进行翻译模型的训练以及解码工作，使用 GIZA++[7]作为对齐工具，Srilm[8]构建语言模型。汉日语言模型均使用 5-gram 模型；moses 中

[4] http://ictclas.nlpir.org
[5] https://code.google.com/p/mecab/
[6] http://www.statmt.org/moses
[7] http://code.google.com/p/giza-pp/

使用 grow-diag-final-and 优化对齐结果。实验结果均使用 BLEU 及 NIST 作为测评标准。

4.2 双语粒度融合实验

Wang 等[6,7]提出细粒度的分词结果能提升统计机器翻译系统的性能。本文为验证当双语分词粒度不同时，双语粒度融合与统计机器的影响，使用 3.4 节所述方法对双语粒度不同的词对进行抽取，对训练语料进行如下处理，得到不同的分词结果：

1. 使用分词工具进行分词的基线结果(baseline)
2. 双语分词粒度不同的词对中，汉语端词序列合并为单词（cn-mix）
3. 双语分词粒度不同的词对中，日语端词序列合并为单词（ja-mix）
4. 双语分词粒度不同的词对中，双语端词序列合并为单词（bi-mix）
5. 双语分词粒度不同的词对中，汉语端单词根据日语端词序列粒度，进行切分，形成词序列(cn-split)
6. 双语分词粒度不同的词对中，日语端单词根据汉语端词序列粒度，进行切分，形成词序列(ja-split)
7. 将方法 5 与方法 6 的结果进行融合，得到双语粒度均进行细切分的结果(bi-split)

对于上述 7 种分词粒度不同的分词结果，我们在汉日与日汉两个方向，分别进行一组基于短语的统计机器翻译性能测试。

经统计，在 282,476 句对的训练语料中，仅存在 23,274 句对需要进行分词粒度调整，存在粒度调整的语料占全部语料的比例较小。因此本文抽取出存在粒度调整的 23,274 句对，并且从剩余句对中随机抽取 80,000 句对与其混合，提高粒度调整语料占所有语料的比例，再次在汉日与日汉两个方向进行一组实验。上述四组实验的结果如表 4 所示。

4.3 实验结果分析

通过 3.2 节的实验结果我们可以得到如下结论：

1）通过对双语分词粒度进行调整，能提升汉日双语间统计机器翻译系统的性能。

2）并非所有的粒度调整都能提升统计机器翻译系统的性能。

本文提出一种衡量双语语料平行句对间的粒度差异的方法，其表达式如公式 1 所示：

$$\mathrm{dis(Corpus)} = \frac{\sum_{i=1}^{N} abs(len(J_i) - len(C_i))}{N} \quad (1)$$

其中，Corpus为双语语料，C_i与J_i分别为源语言与目标语言的第i个句子，len(C_i)与len(J_i)分别为C_i与J_i的句子分词后的词总数，N为双语语料的句对总数。

表 4 不同分词粒度与数据规模下汉日统计机器翻译性能

	语料规模		baseline	cn-mix	ja-mix	bi-mix	cn-split	ja-split	bi-split
汉日方向	282,476	BLEU	0.1643	0.1636	**0.1675**	0.1674	0.1520	0.1572	0.1563
		NIST	4.4543	4.4277	**4.5012**	4.4569	4.3575	4.4118	4.4063
	103,274	BLEU	0.1431	0.1444	0.1438	0.1579	0.1532	**0.1614**	0.1541
		NIST	4.1938	4.1702	4.1938	4.3460	4.2855	**4.4105**	4.3038
日汉方向	282,476	BLEU	0.1383	0.1351	0.1361	**0.1385**	0.1380	0.1360	0.1332
		NIST	4.2593	4.1969	4.2375	**4.2584**	4.2696	4.1745	4.1558
	103,274	BLEU	0.1269	0.1257	0.1269	**0.1280**	0.1258	0.1226	0.1209
		NIST	3.9700	3.9393	3.9741	**3.9527**	3.9655	3.8947	3.8028

本文定义，根据指定双语语料Corpus中所有句子计算出的dis(Corpus)，为该语料中双语的绝对粒度差值。同时，双语语料A与双语语料 B 间分词粒度若存在差异，令A中存在分词粒度差异的

[8] http://www.speech.sri.com/projects/srilm/

句对集为A′，B中存在分词粒度差异的句对集为B′，定义dis(A′)为A、B语料对间A的相对粒度差值，dis(B′)为A、B语料对间的B相对粒度差值，根据公式 2 比较语料A与语料B相对粒度差值之间的差异。

$$\mathrm{diff}(A,B) = \mathrm{dis}(A') - \mathrm{dis}(B') \quad (2)$$

根据上述定义，本文以 baseline 为基准，同组的其余实验均与 baseline 进行比较，根据公式 2，计算diff(baseline, T)，其中，T 为同组其余实验中的任意一组。本文比较diff(baseline, T)与统计机器翻译中 BLEU 值之间的关系，得到如下结果，图 2 为 3.2 节大规模训练集的试验中，汉日(左图)与日汉(右图)的翻译性能 BLEU 值与diff(baseline, T)之间的关系结果图。

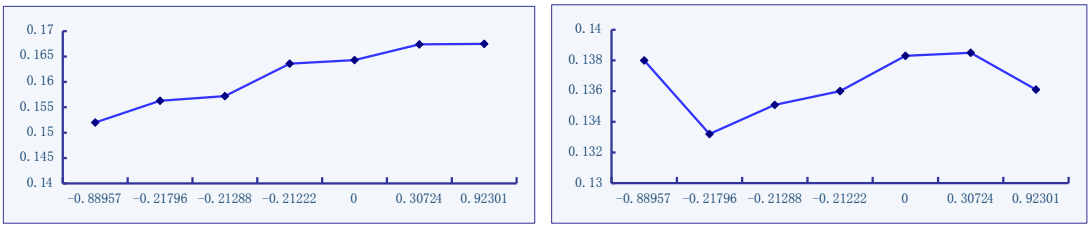

图 2 汉日双语机器翻译中相对粒度差值差异与 BLEU 影响

其中横轴为diff(baseline, T)，竖轴为T 的机器翻译性能评价指标 BLEU 值。

图 2 左图表明，在本文实验条件下，汉日统计机器翻译的性能与相对粒度差值diff(baseline, T)之间存在正相关关系，即相对于 baseline 而言，训练语料分词的相对粒度越小，统计机器翻译系统的性能越好。

图 2 右图中除去一个特殊点外其余结果表明，日汉统计机器翻译中也存在与汉日统计机器翻译结果相同的性质。

由于绝对粒度与相对粒度差值呈正相关关系，依据图 2 结果，本文推测，双语语料的绝对粒度差值dis(Corpus)在一定范围内时，粒度差值与统计机器翻译的性能呈正相关关系。

本文实验结果中，细粒度的分词结果对汉日机器翻译有着一定程度上的帮助。因此，本文在原有 4.2 节实验的基础上，对两个方向四种情况下的训练语料均添加一个最细分词粒度（基于字）的翻译系统，以测试最细粒度的分词结果对日汉双语统计机器机器翻译的影响。实验结果如表 5 所示。

表 5 最细分词粒度下汉日统计机器翻译性能

	汉日方向		日汉方向	
语料规模	282,476	103,274	282,476	103,274
BLEU	0.1415	0.1298	0.1168	0.1128
NIST	4.1723	4.0831	3.8016	3.5718

表 5 结果表明，最细粒度分词结果的统计机器翻译系统性能，相对于其他系统有明显降低。为探究降低的主要原因，本文对测试集的翻译结果进行了比较和分析。

日汉方向上，虽然翻译结果中局部结果较好，然而上下文之间连续性较差，存在提升空间。此外，翻译结果只在局部进行了调序，相对于其他系统而言，不存在长距离调序。

汉日方向上，除存在与日汉方向上相同的问题外，还发现翻译结果的平均长度较长，出现许多冗余信息，主要包括“で”、“す”等日语中的格助词等，在译文中不断重复出现降低了译文质量。

因此，分词粒度应当在一定粒度范围内时，才能有效的提升统计机器翻译系统的性能。

5 总结与展望

本文通过使用现有开源资源构建汉日汉字对照表，并使用构建的汉字对照表对 EDR 词典进行优化。通过根据上述方法构建的资源，对汉日双语语料的不同分词粒度进行数据分析，

在一定程度上解析了汉日分词粒度不同现象产生的原因。

本文提出了使用汉日汉字对照表及词典信息对双语分词粒度进行调整的方法。实验结果表明，本文提出的方法能有效地调节双语分词粒度，提升汉日双语间统计机器翻译系统的性能。本文根据实验结果，对汉日双语统计机器翻译性能与双语句对粒度上的差异进行了分析与预测。

今后工作中，作者将继续扩大词典规模、补充汉日汉字对照表，进一步验证本文提出的方法的有效性，对汉日双语统计机器翻译性能与双语句对间词数量上的差异进行更加深入的分析与研究，并且尝试在层次短语模型中测试本方法的实用性与可扩展性。同时，对于在基础分词中分词粒度不同的词对，将根据词对的特征信息如词性等，对词对进行置换操作，从而提升统计机器翻译的系统性能。

参考文献

[1] Chu C, Nakazawa T, Kawahara D, et al. Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation[C]//EAMT 2012, Proceedings of the 16th Annual Conference of the European Association for Machine Translation. Trento. 2012: 35-42.

[2] Chang P C, Galley M, Manning C D. Optimizing Chinese word segmentation for machine translation performance[C]//Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008: 224-232.

[3] Zhang R, Yasuda K, Sumita E. Improved statistical machine translation by multiple Chinese word segmentation[C]//Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008: 216-223.

[4] Xu J, Zens R, Ney H. Do we need Chinese word segmentation for statistical machine translation[C]//Proceedings of the Third SIGHAN Workshop on Chinese Language Learning. 2004: 122-128.

[5] 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.

[6] Wang Y, Uchimoto K, Kazama J, et al. Adapting Chinese word segmentation for machine translation based on short units[C]//LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation. La Valetta, Malta. 2010: 1758-1764.

[7] 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.

[8] 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.

[9] Bai M H, Chen K J, Chang J S. Improving word alignment by adjusting Chinese word segmentation[C]//Proceedings of the Third International Joint Conference on Natural Language Processing. 2008: 249-256.

[10] Dyer C, Muresan S, Resnik P. Generalizing word lattice translation[R]. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2008/

[11] Chu C, Nakazawa T, Kurohashi S. Japanese-chinese phrase alignment using common chinese characters information[C]//Proceedings of MT Summit. 2011, 13: 475-482.

[12] Chu C, Nakazawa T, Kurohashi S. Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese[C]//Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12). 2012.

[13] Goh C L, Asahara M, Matsumoto Y. Chinese word segmentation by classification of characters[J]. Computational Linguistics and Chinese Language Processing, 2005, 10(3): 381-396.