基于字词混合翻译短语的统计机器翻译

万升华　杨沐昀　李生　赵铁军

哈尔滨工业大学教育部—微软语言语音重点实验室,黑龙江 哈尔滨,150001

E-mail: wanshenghua@gmail.com, ymy@mtlab.hit.edu.cn, lisheng@hit.edu.cn, tjzhao@mtlab.hit.edu.cn

摘　要： 汉英翻译中的分词影响着统计机器翻译（SMT）性能。为解决分词工具对 SMT 产生的影响，本文提出了一种基于字词混合翻译短语的方法。它将基于分字和分词的两种方法进行有效融合，缓解了分词带来的不利影响。本文在汉英科技专利文献上进行了实验，表明这种混合策略相比单一的分词的性能最高可以提升 0.56 的 BLEU 分数，并发现使得 SMT 最优化的分词工具的平均分词长度应该在 1.57~1.61 字/词之间。

关键词： 统计机器翻译 汉语分词 平均词长

A Character and Word Phrases Combination Approach to SMT

Wan Shenghua, Yang Muyun, Li Sheng, Zhao Tiejun

MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

E-mail: wanshenghua@gmail.com, ymy@mtlab.hit.edu.cn, lisheng@hit.edu.cn, tjzhao@mtlab.hit.edu.cn

Abstract: To overcome the negative impact on Chinese-English(CE) SMT by Chinese Word Segmentation(CWS) results, this paper proposes a method of combining the word based translation phrases into the character based translation phrases so as to mitigate the adverse effect from CWS. Experiments on CE scientific patent corpus indicate that it produces an improvement of 0.56 BLEU over traditional approach. Moreover, we find that SMT optimized CWS should have an average word length between 1.57 and 1.61 per word.

Keywords: Statistical Machine Translation, Chinese Word Segmentation, Average Word Length

1　引言

在构建高性能的基于短语的 SMT 系统的过程中，SMT 一般需要以词为单位建立双语之间的互译关系。汉语词汇由一个或多个汉字组成，字与字之间没有空白间隔。鉴于词形学上的这些特点，汉语分词是有必要的，并对汉英 SMT 性能产生实质的影响，引起了研究者的重视。

前人的研究结果表明，虽然分词对 SMT 性能影响较大，但分词如何对 SMT 产生影响到目前为止还没有一个清晰地结论。分词对 SMT 性能影响的幅度实际并不如人们所预料的那么大。[1]也提到在最近的国际汉语分词性能评测 SIGHAN 中，运用 CRF 方法的汉语分词系统获得了出色的成绩。然而，与分词性能逊色的采用传统的基于词典的分词方法相比，在 SMT 性能上，采用 CRF 方法的高性能分词工具的 BLEU 得分与传统的基于词典的分词工具几乎一样，例如[2]中基于 CRF 的工具分词性能较后者高 148%，而在 BLEU 得分上，前者仅比后者高 3.7%。[1]中使用多种分词工具进行实验，讨论了分词与 SMT 性能的联系，得出了不同分词标准和分词方法得到的 SMT 性能差异不大的结论，并尝试了运用多分词融合策略优化 SMT 性能，并在同一任务中获得了最高 0.27 到 0.4 的 BLEU 得分提升。

另外，[2]的研究表明，在汉英翻译过程中，虽然语料进行汉语分词处理后，机器翻译的性能相比分字的情况有明显提高。同时，他们也通过实验证实了，针对当前分词标准优化的分词工

具，在衡量汉语分词性能的F值上取得了很高的分数，但是却不能保证在应用到SMT系统上也取得优化的性能。他们发现了分词的一致性和汉语分词粒度是对SMT性能的影响显著的因素，还通过一个分词粒度可调的基于CRF方法的分词工具对分词平均长度按一定步长进行反复试验，获得了0.73的BLEU得分提升。

然而，上述研究并未真正弄清分词对SMT带来的影响，其不利因素也没有消除。针对这一问题，本文提出了字词混合短语测策略，即将基于字的和基于词的翻译短语表进行现行融合，以弥补分词错误对翻译模型带来的干扰。实验结果表明，相比分词的SMT性能，这种策略在同等条件下可提高BLEU得分0.56，相对性能改进达1.3%。

本文内容安排如下：我们将在第二部分介绍汉语分词的基本概念并验证不同分词系统对SMT的性能影响；在第三部分中，我们将阐述本文的研究思路——基于字词混合翻译策略；在第四部分，我们将给出实验和结果分析，第五部分总结全文。

2 汉语分词及其对统计机器翻译性能的影响

汉语文本中词与词之间没有明确的分隔标记。汉语分词就是将连续的汉字串分割成一个一个词，因为词是最小的能够独立活动的有意义的语言成分。

现有的分词方法可分为三大类：基于词典的分词方法、基于理解的方法和基于统计的方法。常用的分词方法是基于词典的使用字符串匹配方法和基于统计的机器学习模型分词方法。现在流行的基于机器学习模型的分词方法又有隐马尔可夫模型（HMM）和条件随机场模型（CRF）。

由于对词的定义没有一个共识，当前的各个研究机构的分词工具采用的分词标准不尽相同，而且国际上也没有一个统一的分词标准。目前知名的分词标准有北京大学分词标准（PKU）、宾州树库分词标准（Penn Chinese Treebank）、哈工大树库分词标准（HIT Chinese Treebank）等。

国际上定期举行分词评测会议SIGHAN，对分词系统给予不同任务测试。每年最好的系统都不同，并且各任务的最好系统也不一样。这反映了分词系统对不同任务的敏感。在2008年SIGHAN评测中，针对CITYU语料，在F值上，最好比最差的系统高了7.5%。而针对语料CTB，得分最好比最差的系统竟然高了9.2%。分词系统对语料和任务的敏感性可见一斑，保持稳定成绩很难。

实际上，左右分词性能的是分词方法和分词标准，而且不同的分词方法和分词标准对SMT性能的影响也不一样。下面是本文将使用的分词工具的简介：

ICTCLAS[5]是中科院计算所的基于层叠隐马尔可夫模型（Hierarchical Hidden Markov Model）的分词工具，使用北大分词标准(PKU)；CEMT2K[6,7]是哈工大机器智能与翻译实验室的基于隐马尔可夫模型（Hidden Markov Model）的分词工具，使用哈工大树库标准(HIT)；SCS（Stanford Word Segmenter）[8]是美国斯坦福大学人工智能实验室自然语言处理组的基于随机条件场（Conditional Random Field）模型的工具，使用北大标准(PKU)或宾州树库标准(CTB)。

利用上述分词工具对同一汉英科技语料（详见第四部分）中的汉语进行分词处理，其结果的差异如表1所示。注意，这给出的是SMT中经典的Moses系统对分词后的汉语句子过滤后的统计数字，由于分词结果不同，实际上翻译系统可接收的训练语料已经有了明显的变化。

名称	句子数	字数	分词数	平均词长
ICTCLAS	284872	8603927	5259957	1.6357
CEMT2K	284946	8603232	5249361	1.6389

SCS-CTB	285493	8635751	5185284	1.6654
SCS-PKU	256214	8685200	5118859	1.6967

表 1 分词性能

表 1 是上述四组实验的分词过程后的词数和平均词长。这些变化对最终的翻译性能也产生了是指影响，其翻译结果的 BLEU 和 NIST 评分如表 2 所示。

名称	分词标准	BLEU-4(%)	NIST-5
ICTCLAS	北大标准	44.77	9.8438
CEMT2K	哈工大标准	46.28	10.0592
SCS-CTB	宾州树库标准	45.24	9.8838
SCS-PKU	北大标准	44.59	9.8392

表 2 SMT 翻译测试

可见使用不同的分词工具导致了 BLEU 值的不同，其中表现最好的 CEMT2K 系统比最差的 SCS-PKU 相对要高 3.8%，而且对于 NIST 值，它们的表现也不稳定。表 2 的平均词长的统计表明，平均分词长度也因分词工具的不同而变化。参考文献[2]对平均分词长度对 SMT 性能的影响做了细致的分析，并指出平均分词长度与 SMT 性能有一定的联系，但却没有给出明确的关联方式。

3 基于混合策略的统计机器翻译

当前解决分词对 SMT 性能影响的思路主要有两个：一是将多分词结果进行融合以改善翻译性能[2]，但它并不能彻底消除分词带来的错误。一种思路是调整分词模型的参数以优化翻译性能[1]，但目前这一参数与 SMT 的内在联系并不显著，只能部分缓解分词错误对 SMT 性能的影响。

本文的思路是采取字词混合的策略进行翻译，即将汉语分字训练出的短语表与汉语分词的短语表进行合并，使新的短语表的粒度介于分字与分词之间，兼有两者的特性。在第四部分的实验中，我们将研究上述系统在分字向分词短语表融合和相反两个方向的 SMT 性能，并且我们将在本文就两个极端情况进行实验，即分别以分词短语表、分字短语表为目标，融合进另一个短语表。

虽然分字含有的语言学信息比分词要少，进行词对齐之后，但是它翻译效果仍然值得一提。不过分字带来更低的粒度增大了数据稀疏性，缩小了短语表，相比分词，它的 SMT 性能稍逊一筹，可它便于调整整个语料的分词粒度。为了尽可能地利用分词和分字的各自优点，我们采用了折中的办法去探索对 SMT 来说最佳的分词粒度：

(1)将汉语语料进行分词和分字处理，并进行训练抽取分字短语表 Tc 和 Tw

(2)将上述两个短语表的相同短语按如下公式进行融合，不相同的短语对齐条目直接收录：

$$H(e|f)=\sum_i h_i(e|f)g\alpha_i \tag{1}$$

其中，α_i 是对每个短语表的加权值，为了归一化，必须满足 $\sum_i \alpha_i = 1$。$h_i(e|f)$ 是每个短语对齐条目中的第 i 个特征，包括：短语的翻译模型 $p(e|f)$，反向翻译概率 $p(f|e)$，关于词对齐的词法概率 $lex(e|f,a)$，反向词法概率 $lex(f|e,a)$ 等。

本来确定参数α_i一般也需要进行最小错误率训练进行调整参数，由于这样的训练非常费时，我们仅对公式(1)的极端情况进行实验，但实验的结果仍然具有指导意义。

4 实验及结果分析

4.1 实验语料

本实验训练语料来自中信所的英汉科技语料，包含约 30 万双语语料。训练和测试语料都来自全国第四届机器翻译研讨会(CWMT08)的英汉科技翻译评测的科技语料训练集和测试集。测试语料包含 1008 组句子，其中每组句子含有 1 句英语和相应的 4 句汉语译文。而汉英翻译的测试语料也包含 1008 组句子，其中，每组句子含有 1 句汉语和相应的 1 句英语译文，详见表 3。

训练集含有的句对数	303704
训练集含有的汉字数	21451600
训练集含有的英语单词数	9907073
调参集含有的句子数	915
测试集含有的句子数	1008
英汉翻译测试集含有的英语单词数	24227
汉英翻译测试集含有的汉字数	36626

表 3 实验语料统计情况

4.2 实验设置及结果

在本实验中，分字只是简单将中文句子按字分开，分词长度恒为 1。ICTCLAS 和 CEMT2K 分别使用 PKU 和 HIT 标准，而 SCS 可以选择两种分词标准，第一种是 PKU，另一种是 CTB。

对于短语表，本次实验采用了三种处理方式，第一种是不改变短语表，直接进行测试，另外两种是通过融合分字和分词的短语表，重新进行调参，再进行测试。处理方式是将分字方式和分词方式训练出来的短语表进行融合，对于中英短语相同而特征不同的表项，Plan A 取分字短语表中的特征作为新的特征，称为 WtoC 融合，相当于分字权重为 1，分词权重为 0；Plan B 则反之，取分词短语表中的特征作为新的特征，称为 CtoW 融合。这是两个极端情况。

英汉翻译实验经过预处理数据、建立语言模型、训练翻译模型、调整参数、测试并评估之后，我们将各组融合分字与分词的短语表，再进行英汉翻译实验，得到结果，详见表 4 和图 1。

序号	实验名称	BLEU-4(%)	同比提升
1	Char	40.07	——
2	Word_ICT	44.77	+ 0.0 %
3	WtoC_ICT	41.63	- 7.0 %
4	CtoW_ICT	42.37	- 5.4 %
5	Word_HIT	46.28	+ 0.0 %
6	WtoC_HIT	46.62	+ 0.7 %
7	CtoW_HIT	46.68	+ 0.9 %
8	Word_SCS-CTB	45.24	+ 0.0 %

9	WtoC_SCS-CTB	45.82	+ 1.3 %
10	CtoW_SCS-CTB	44.95	- 0.6 %
11	Word_SCS-PKU	44.59	+ 0.0 %
12	WtoC_SCS-PKU	45.15	+ 1.3 %
13	CtoW_SCS-PKU	45.05	+ 1.0 %

表 4 英汉翻译实验结果

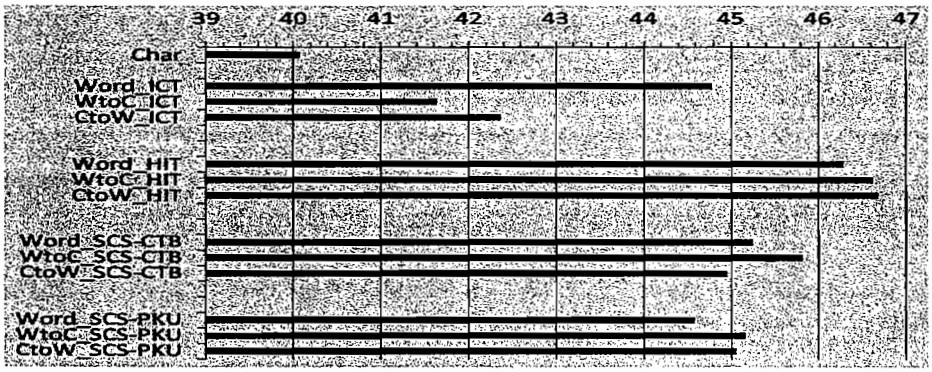

图 1 英汉翻译实验结果

我们采用类似的方式进行汉英翻译实验，它们只在参考译文数量上不同，实验结果见图 2。

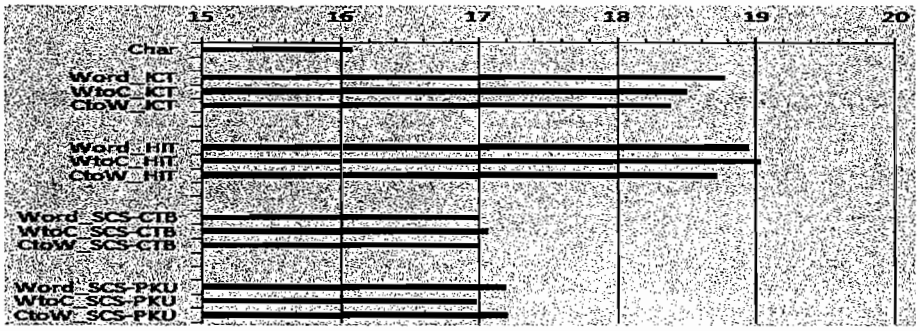

图 2 汉英翻译实验结果

4.3 分析与讨论

可见，本次实验反映出汉语分词系统与统计机器翻译系统的翻译效果不稳定性有一定关系：某些分词系统并不能保证在英汉双向翻译中保持稳定的表现，除 CEMT2K 外，其他 3 种分词过程（ICTCLAS，SCS-PKU，SCS-CTB）都不能保证在双向翻译评测取得一个稳定的排位。

实验也表明采用字词混合策略之后的 SMT 性能均有不同幅度的提升，其中提升最明显的一组有 0.56 的 BLEU 分值增长，但实验结果也反映有一些组在采用混合策略之后，SMT 性能不升反降。

考虑到对短语表的融合相当于改变了汉语分词粒度，不同工具使用的分词标准也造成了汉语分词粒度的不同，本实验还对这 4 个分词过程短语表的粒度进行了测量。这个实验的对象是短语表中英语单词对应一个中文词的数目（称为 1 比 1 中文词数）和汉语平均词长。实验结果见表 5。

编号	实验组名	平行短语数	1 比 1 中文词数	平均词长
1	ICT	8877585	117808	1.5722588
2	HIT	8809606	120471	1.5792556

3	SCS_CTB	8947795	75963	1.6052368
4	SCS_PKU	8917568	73501	1.6257051

表 5 四个分词过程的短语表的粒度-汉语平均词长

对比表 4 和图 1 等，可见对统计机器翻译产生最佳影响的 CEMT2K 的平均分词长度处于中间，效果比较差的 ICTCLAS 和 SCS_PKU 的汉语平均分词长度处于两个极端。

与[4]不同，我们对汉英双向进行了 SMT 实验，并发现就 BLEU 评分来说，英汉翻译中，分词向分字短语表进行融合的方式是最佳的，这一点在汉语环境中验证了他们在日语环境中的研究结果。然而，我们在汉英翻译中发现，在英汉翻译中的最佳融合方式与汉英翻译中的完全相反，不能一致。由此看来，有必要开发一个面向优化 SMT 性能的专用分词工具，它能保持性能的稳定并使性能最优化。对于面向统计机器翻译并且使其保持性能稳定的汉语分词工具的平均分词长度和产生的短语表中 1 比 1 中文词数应该避免极端，，而选择走折中路线，才可能取得比较理想的翻译效果，具体的范围应该是：平均分词长度应该保持在 1.57-1.61 字/词之间。

5 结束语

本文对三个通用汉语分词工具的性能进行了研究，提出了英汉互译环境下，面向提升 SMT 性能的专用汉语分词工具的技术指标要求。首先，我们验证了对于 SMT 性能来说，汉语分词作为预处理比只是简单的分字要好很多；其次，我们发现，使用这些通用汉语分字工具时，SMT 的性能不能保持稳定；然后，我们在英汉互译的环境下，对面向优化 SMT 性能的专用分词工具的技术指标进行了研究，采用字词混合策略进行短语表融合，发现这样的方法可以提升 SMT 性能的 BLEU 得分多达 0.56，并得出的结论是平均分词长度要取折中值，而不是追求单调极端值，并给出了日后研发针对统计机器翻译进行优化的专用分词工具的具体技术指标的范围，平均词长应该保持在 1.57-1.61 字/词之间，为接下来的工作做了指引。

参考文献

[1] Ruiqiang Z, Keiji Y and Eiichiro S. Chinese Word Segmentation and Statistical Machine Translation. ACM TSLP, 2008.

[2] PiChuan C, Michel G and Christopher DM. Optimizing Chinese Word Segmentation for Machine Translation Performance. Proc. of the 3rd Workshop on Statistical Machine Translation of ACL, 2008.

[3] Philipp K, et al. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of ACL, 2007.

[4] Junguo Z, et al. A Phrase Combination Approach to Patent SMT. Proceedings of the 11th JCIS, 2008.

[5]Huaping Z, et al. HHMM-based Chinese Lexical Analyzer ICTCLAS. Proc. of 2nd SIGHAN Workshop on CLP 2003.

[6]赵铁军，吕雅娟，于浩，杨沐昀，刘芳. 提高汉语自动分词精度的多步处理策略, 中文信息学报, 2001.

[7]吕雅娟，赵铁军，杨沐昀，于浩，李生. 基于分解与动态规划策略的汉语未登录词识别, 中文信息学报, 2001.

[8]Huihsin T, et al. A Conditional Random Field Word Segmenter. Proc. of the 4th SIGHAN Workshop on Chinese Language, 2005.