doi:10.3772/j.issn.1002-0470.2009.04.002

基于短语相似度的统计机器翻译模型①

何中军②* ** 刘 群③* 林守勋*

(* 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)
(** 中国科学院研究生院 北京 100049)

摘 要 针对基于短语的统计机器翻译(SMT)模型中由于采用精确匹配策略导致的短语稀疏问题,提出了一种基于短语相似度的统计机器翻译模型。该模型将基于实例的翻译方法引入到统计机器翻译中。翻译时,对于训练语料库中未出现过的短语,通过计算源语言短语之间的相似度,采用模糊匹配策略从短语表中查找相似的实例短语,并根据实例短语为其构造翻译。与精确匹配策略相比,利用相似度进行模糊匹配增加了对短语表的利用程度,缓解了短语稀疏问题。实验表明,该模型能够明显地提高统计机器翻译的质量,效果超过了当前最好的短语系统"摩西(Moses)"。

关键词 相似度,基于短语的统计机器翻译,基于实例的机器翻译

0 引 言

统计机器翻译(statistical machine translation, SMT)是近年来国际上的一个研究热点,且基于短语的统计机器翻译取得了很大进展。基于短语的翻译模型[1,2]以短语作为翻译的基本单位,短语内部隐含了词语的选择和词序的调整,能够较好地翻译习惯用语、固定搭配等。其中,短语可以是任意连续的词串,而不一定具有语言学意义。这种翻译模型在翻译时一般采用精确匹配策略,其缺点是如果一个短语中有一个词匹配不上,就导致整个短语不能使用。例如要翻译短语"于 4 月 1 日访问北京",却不能使用语料库中的短语对 "于 4 月 2 日访问北京, visits Beijing on April 2",因为其中"1 日"和"2 日"不匹配,这是非常可惜的。这样严格的匹配策略一方面使得大量的短语无法被利用,另一方面带来了严重的数据稀疏问题,尤其是对长短语。因为短语长度越长,其精确匹配的可能性就越小。而恰恰是长短语对机器翻译的质量影响很大,因为它内部包含了更丰富的词语上下文信息。对于前面的例子,如果没有长短语进行匹配,我们只能将其拆分为短短语分别翻译,例如"于 4 月 1 日, on April 1"和"访问北京, visits Beijing"。这样还需要对短语之间进行调序才能获得正确的译文"visits Beijing on April 1",这就增加了解码复杂度。这个问题可以通过增强短语的泛化能力得到缓解。一种方法是利用词法(morphology)信息[3,4],例如将动词的各种时态形式归一化为动词原形。这种方法能够起到一定的泛化作用,然而它难以应用于汉语。文献[5]提出了对齐模板方法,将短语中的单词映射到相应的词类中,使得短语具有泛化作用。这种方法需要对单词进行自动聚类,聚类的质量对翻译质量影响很大。文献[6]从双语语料中学习复述短语(paraphrase),未登录短语(unseen phrase)的译文可以根据其复述短语产生。复述短语要求短语是同义的,因此文献[6]中的方法不能用于非同义短语。

本文提出了一种基于短语相似度的统计机器翻译模型,其主要思想是:如果一个短语没有完全匹配,就保留匹配部分的信息,从语料库中查找未匹配部分的翻译,然后组合成一个完整的短语。基于上述思想,我们将基于实例的翻译方法[7,8]引入到统计机器翻译中,在两个层面上进行翻译,即在句子级使用基于短语的统计模型,在短语级使用基于实例的模型,根据短语表中相似的实例短语构造翻译。与文献[5]不同的是,我们的模型无需对单词自动聚类,而是使用词性来指导相似度的计算。与文献[6]相比,我们的模型更具一般性, 对相似度的不同定

① 863 计划(2006AA010108)和国家自然科学基金(60573188,60603095,60736014)资助项目。
② 男,1982 年生,博士生;研究方向:自然语言处理、机器翻译,E-mail:zjhe@ict.ac.cn
③ 通讯作者,E-mail: liuqun@ict.ac.cn
(收稿日期:2008-01-23)

义决定了两个短语在何种层面上是相似的，既可以是词形相似，也可以是结构相似、语义相似等。实验结果显示，这种短语相似度模型能够极大地提高短语表的利用率，缓解短语稀疏问题，提高翻译质量。

1 短语相似度模型

在短语相似度模型中，一个源语言句子 F_1^J 的翻译过程如下：

(1) 对源语言进行短语划分，$F_1^J = \tilde{f}_1 \cdots \tilde{f}_K$；

(2) 对每个短语 $\tilde{f}_k (k = 1, \cdots, K)$，按照精确匹配策略从短语表中查找对应的双语短语，如果找到，则转到(4)；

(3) 对于训练语料库中没有出现过的短语 $\tilde{f}_{k'} (k' = 1, \cdots, K)$，按照模糊匹配策略从短语表中查找相似的实例短语 $(\tilde{f}, \tilde{e}, a)$，为之构造新的短语对 $(\tilde{f}_{k'}, \tilde{e}_{k'}, a)$；

(4) 按照翻译模型，将目标短语进行组合，产生最终译文。

本文重点讨论短语匹配时所用的模糊匹配策略，即第(3)步。我们采用基于实例的方法建立短语相似度模型，为此，需要解决以下三个问题：一是如何衡量两个短语是相似的，相似程度多大？二是如何根据实例短语构造新的短语？三是如何计算新短语的概率？下面，我们分别进行详细介绍。

1.1 短语相似度的定义

两个短语之间的相似度可以用多种方式来衡量，例如根据短语中单词的词形、词义，或者短语的结构信息等。针对统计机器翻译而言，在训练过程中，我们能从语料库中学习到大量的短语。考虑到时空开销，需要设计简单高效的短语相似度计算方法。为此，我们利用单词词形计算相似度，并使用词性信息加以约束。

一种直观的认识是：两个短语包含的相同单词越多，则这两个短语越相似。因此，可以用 Dice 系数来计算两个词串 s_1 和 s_2 的相似度：

$$Dice(s_1, s_2) = 2 \times \frac{| s_1 \cap s_2 |}{| s_1 | + | s_2 |} \quad (1)$$

不过，公式(1)没有考虑短语中单词的位置，例如“枪手 被 警察 击毙”和“警察 被 枪手 击毙”，这两个短语虽然内部的单词相同，即 $Dice(s_1, s_2) = 1.0$，但是却不能精确匹配。因此，我们对公式(1)进行了改进，计算时比较相同位置上的词语是否一样。对于两个短语 $\tilde{f}_1^I = c_1, c_2 \cdots c_I$，$\tilde{f}'^J_1 = w_1, w_2 \cdots w_J$，其相似度计算如下：

$$SIM(\tilde{f}_1^I, \tilde{f}'^J_1) = \frac{\sum_{j=1}^{J} \delta_{c_j w_j}}{J} \quad (2)$$

其中

$$\delta_{ij} = \begin{cases} 1, & if\ i = j \\ 0, & if\ i \neq j \end{cases} \quad (3)$$

这样 $SIM(\tilde{f}_1^I, \tilde{f}'^J_1) = 1.0$ 就意味着这两个短语是精确匹配。

另外在计算相似度时，我们还使用了词性信息进行约束，要求相似的两个短语必须具有相同的词性序列。词性序列在一定程度上反映出短语的结构信息，例如“在/p 政治/n 领域/n 的/udel 合作/vn, cooperation in the political field”和“在/p 经济/n 方面/n 的/udel 交流/vn, communication in the economic area”。按照公式(2)，尽管它们的相似度比较低，但是相同的词性序列使得它们具有相同的结构，都是一个介词短语(PP)修饰一个名词短语(NP)。汉语中 PP + NP 的短语翻译到英语一般都是将介词短语后置，即 NP + PP。因此，词性序列可以帮助我们找到结构相似的短语，弥补了公式(2)的不足。

1.2 新短语的构造

当为一个短语 f_1^I 找到相似的实例短语对 $(f'^I_1, e'^{I'}_1, a)$ 后，可以通过以下步骤构造新的短语对：

(1) 比较：比较 f_1^I 和 f'^I_1 每个位置上的单词，得到未匹配单词的位置集合 $P = \{j \mid f_j \neq f'_j, j = 1, \cdots, J\}$；

(2) 替换：用 f_j 替换 f'_j，并根据对齐矩阵 a，从 $e'^{I'}_1$ 中删除 f'_j 对应的单词 e'_{a_j}，其中 $j \in P$；

(3) 翻译：从短语表中找到单词 f_j ($j \in P$) 的翻译 e，并将其放到对应的英语位置 a_j。

通过以上三步，我们构造了一个新的短语对 (f_1^I, e_1^I, a)。如图 1 所示，我们利用实例短语“全市 出口 总值 的 半数, half of the entire city 's export volume”为“全省 出口 总值 的 25.5%”构造了翻译“25.5% of the entire province 's export volume”。可以看出，新短语对的构造充分利用了实例短语的信息：保留了已匹配的词语的译文，并且利用对齐信息实现了单词的调序。经过简单的比较和替换，就可以为未登录短语构造较高质量的译文，从而缓解了数据稀疏问题。另外需要注意的一点是，由于一个短语可能对应多个相似的实例短语，而一个单词也可能对应多个翻译，因此，对于一个短语 f_1^I，我们可以构造多个译文。

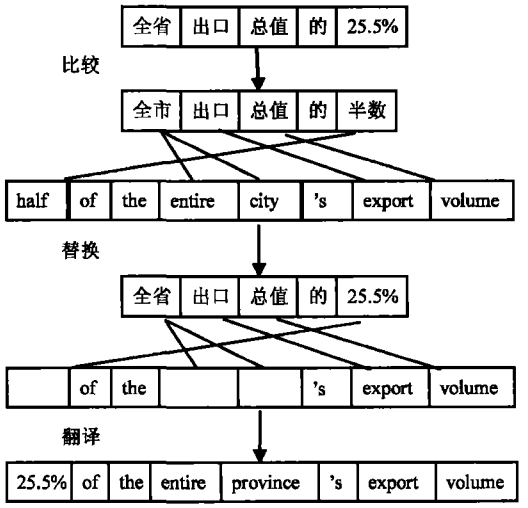

图 1　由相似的实例短语构造新的短语对

1.3　新短语的概率计算

当新短语对构造出来以后,为了使其能够参与句子翻译,需要为它们赋予翻译概率。同其它基于短语的模型[2]类似,我们使用 4 个概率来刻画一个汉语短语和一个英语短语互为翻译的好坏程度:

- 短语翻译概率 $p(\tilde{f}\mid\tilde{e})$ 和 $p(\tilde{e}\mid\tilde{f})$;
- 词汇化翻译概率 $p_w(\tilde{f}\mid\tilde{e})$ 和 $p_w(\tilde{e}\mid\tilde{f})$,它根据 IBM 模型 1[9]由短语内部单词的翻译概率计算得到:

$$p_w(f_1^J\mid e_1^I,a)=\prod_{j=1}^{J}\frac{1}{|\{i\mid(j,i)\in a\}|}\sum_{\forall(j,i)\in a}p(f_j\mid e_i) \tag{4}$$

假设 $(\tilde{f},\tilde{e},a)$ 是由相似实例短语 $(\tilde{f}',\tilde{e}',a)$ 构造的,我们将 $(\tilde{f}',\tilde{e}',a)$ 的短语翻译概率赋予 $(\tilde{f},\tilde{e},a)$,即 $p(\tilde{f}\mid\tilde{e})=p(\tilde{f}'\mid\tilde{e}')$,$p(\tilde{e}\mid\tilde{f})=p(\tilde{e}'\mid\tilde{f}')$。

对于词汇化翻译概率,可以根据替换的单词重新计算得到。假设 $S'\{(f',e')\}$ 是短语 $(\tilde{f}',\tilde{e}')$ 中被替换的单词对集合,它被 $S\{(f,e)\}$ 替换以构造新的短语对 $(\tilde{f},\tilde{e})$,则词汇化概率可以按照下式重新计算:

$$p_w(\tilde{f}\mid\tilde{e},a)=\frac{p_w(\tilde{f}'\mid\tilde{e}',a)\times\prod_{(f,e)\in S\{(f,e)\}}p_w(f\mid e)}{\prod_{(f',e')\in S'\{(f',e')\}}p_w(f'\mid e')} \tag{5}$$

这样,新构造的短语也具有了 4 个概率,可以同原本就存在于短语表中的短语一起参与句子翻译。

2　实验及分析

基于短语相似度模型,我们研制了一个汉英翻译系统"Mencius",它使用对数线性(log-linear)模型[10]融合了以下 8 个特征:短语翻译概率 $p(\tilde{f}\mid\tilde{e})$ 和 $p(\tilde{e}\mid\tilde{f})$,词汇化翻译概率 $p_w(\tilde{f}\mid\tilde{e})$ 和 $p_w(\tilde{e}\mid\tilde{f})$,译文长度惩罚,语言模型,词汇化调序模型[11],短语相似度。为了测试系统性能,我们以当前最好的短语系统"Moses"[11]作为基线系统,它在选择短语时采用精确匹配策略。

我们设计了两组实验:一个是小规模实验,训练语料包含 3 万句对,其中有 840K 汉语词和 950K 英语词;另一个是大规模实验,训练语料包含 250 万句对,其中有 68M 汉语词和 74M 英语词。小规模的数据稀疏问题严重,能够明显地考察短语相似度模型的作用。另外,我们希望相似度模型在大规模数据上也能够提升翻译系统的性能。

我们使用 GIZA++[12]训练汉英和英汉两个方向的词语对齐,并采用"grow-diag-final"方法[2]对词语对齐进行优化。在进行短语抽取时,限制源语言端的短语长度最大为 7。使用 SRI 语言模型训练工具[13]训练了一个 4 元的语言模型,训练语料是 LDC 发布的 Gigaword 新华社部分,包含 190M 单词。使用 2002 年 NIST 评测的测试集作为开发集,2005 年 NIST 评测的测试集作为测试集。评测指标采用 BLEU-4。

2.1　小规模语料的实验

在小规模语料的实验中,基线系统"Moses"在 NIST2005 上的 BLEU-4 值是 0.2444,而我们的系统"Mencius"是 0.2531,提高了 0.87 个百分点。

在训练中,我们从 3 万句对中抽取了约 1M 短语对。"Moses"采用精确匹配方法选择短语,可用于翻译开发集和测试集的短语对只有 0.2M,仅占 20%。这说明,采用精确匹配的方法对短语表的利用很不充分,浪费了大部分短语。另外,我们对这 0.2M 的短语进行了统计,它们的长度分布如表 1。从中可以看出长度大于 3 的短语仅占总数的 2.31%,这说明长短语的数据稀疏严重。

采用短语相似度模型进行模糊匹配,可用短语对增加到 0.85M,占总数的 85%,说明相似度模型能够提高短语的利用率,从而缓解数据稀疏问题。

表 1 小规模语料针对 NIST2002 年和 2005 年测试集的短语分布表

短语长度	1	2	3	4	5	6	7
所占比例	48.75%	40.11%	8.83%	1.74%	0.46%	0.09%	0.02%

另外,我们分析了两个系统在测试集上的运行结果,并统计了它们所用短语分布情况,见表 2。在"Moses"中,共使用 24618 个短语,其中长度小于等于 3 的短语占了 99.6%,长度大于 3 的短语占了 0.4%。说明在解码过程中,"Moses"所采用的精确匹配策略难以使用长短语。在"Mencius"中,共使用短语 21195 个,其中长度小于等于 3 的短语占了 97.7%,长度大于 3 的短语占了 2.3%,在所有这 21195 个短语中,有 14.3%是由相似的实例短语构造出来的(即模糊匹配)。这说明短语相似度模型能够缓解短语的数据稀疏问题。

表 2 小规模实验所用短语的分布情况

系统名称	匹配方式	源语言短语长度							短语总数	NIST05 BLEU-4
		1	2	3	4	5	6	7		
Moses	精确匹配	19485	4416	615	87	12	2	1	24618	0.2444
Mencius	精确匹配	14750	2977	387	48	10	1	0	21195	0.2531
	模糊匹配	0	1196	1398	306	93	17	12		

2.2 大规模语料的实验

在大规模实验中,"Moses"在测试集上的 BLEU-4 值是 0.3045,而"Mencius"是 0.3096,提高了 0.51 个百分点。

加大训练语料库的规模,固然可以一定程度上缓解数据稀疏问题,但是对于长度较长的短语,如果按照精确匹配策略,仍然难以匹配。而且,大部分的短语仍然难以被利用。我们从 250 万句对抽取了大约 110M 短语对,可用于翻译开发集和测试集的短语对只有 7.4M,仅占 6.7%。而使用短语相似度模型,则可用短语所占比例增加到 52%。由此可见,相似度模型极大地提高了短语的利用率。另一方面,模糊匹配策略能够匹配较长的短语,获得更为流畅的译文。例如,在测试集中有这样的句子:

但是 经济 产出 的 长期 趋势 将 显示 经济 成长 趋缓

"Moses" 将其翻译为:

长期 ‖ 经济 产出 ‖ 但是 ‖ 的 ‖ 趋势 ‖ 将 ‖ 显示 ‖ 趋 ‖ 缓 ‖ 经济 成长

long term ‖ economic output ‖ , but ‖ the ‖ trend ‖ will ‖ show ‖ a ‖ slow down ‖ economic growth

而"Mencius"将其翻译为

但是 ‖ 经济 产出 的 长期 趋势 ‖ 将 ‖ 显示 ‖ 趋 ‖ 缓 ‖ 经济 成长

but ‖ the long-term trend of economic output ‖ will ‖ show ‖ a ‖ slow down ‖ economic growth

其中,"‖"用来表示短语划分。可以看出,由于测试语料中没有出现"经济 产出 的 长期 趋势"这样的短语,"Moses"只好将其拆分为 4 个短短语"经济 产出"、"的"、"长期"、"趋势",然后使用调序模型进行调序。遗憾的是,不正确地调序导致了错误的译文。而使用相似度模型进行模糊匹配,我们在语料库中可以找到实例短语"经济 发展 的 必然 趋势, the inevitable trend of economic development"。由此,可以为"经济 产出 的 长期 趋势"构造翻译,得到正确的译文"the long-term trend of economic output"。可见,通过模糊匹配策略,相似度模型能够为长短语构造翻译,减轻了短语调序模型的负担,提高了译文质量。

3 结 论

本文在统计机器翻译框架下,引入了基于实例的翻译思想,提出了基于短语相似度的统计机器翻译模型。该模型采用模糊匹配策略,通过计算源语言短语的相似度,能够根据相似的实例短语,为未登录短语构造新的短语对。这一方面增强了短语的泛化能力,能够充分利用从语料库中抽取的双语短语;另一方面,缓解了数据稀疏问题,能够使用更多的长短语,获得更流畅的译文。实验表明,相比于现在性能最好的短语系统"Moses",该短语相似度模型能够明显地提高译文的质量。

本文给出了相似度模型的基本思想:使用模糊匹配策略进行短语匹配。其中关键的一点是相似度

的定义和计算。目前,我们将相似度定义在词形上,并使用词性加以限制。这种方法简单易行,取得了较好的效果。在将来的工作中,还可以考虑引入更丰富的信息,例如使用同义词词林计算词义的相似度。另外,对齐质量对于相似度模型影响比较大,这是因为在短语构造过程中,使用了词语对齐信息,在以后的工作中,可以采用更好的词语对齐工具。

参考文献

[1] Marcu D, Wong W. A phrase based joint probability model for statistical machine translation. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, USA, 2002. 133-139

[2] Koehn P, Och F J, Marcu D. Statistical phrase-based translation. In: Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting, Edmonton, Canada, 2003. 127-133

[3] Niessen S, Ney H. Statistical machine translation with scarce resources using morpho-syntactic Information. *Computational Linguistics*, 2004, 30(2): 181-204

[4] Goldwater S, McClosky D. Improving statistical MT through morphological analysis. In: Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, 2005. 676-683

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

[6] Callison-Burch C, Koehn P, Osborne M. Improved statistical machine translation using paraphrases. In: Proceedings of North American chapter of the Association for Computational Linguistics Annual Meeting, USA, 2006. 17-24

[7] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle. In: Proceedings of the international NATO symposium on Artificial and human intelligence, 1984. 173-180

[8] Sumita E, Iida H. Experiments and prospects of example-based machine translation. In: Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics, USA, 1991. 185-192

[9] Brown P F, Della Pietra S A, Della Pietra V J, et al. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 1993,19(2): 263-311

[10] Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics, USA, 2001. 295-302

[11] Koehn P, Hoang H, Birch A et al. Moses: open source toolkit for statistical machine translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, Czech, 2007. 177-180

[12] Och F J, Ney H. Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics, 2000. 440-447

[13] Stolcke A. Srilm-an extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing, 2002. 901-904

A phrase similarity-based model for statistical machine translation

He Zhongjun[*,**], Liu Qun[*], Lin Shouxun[*]

([*] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**] Graduate University of Chinese Academy of Sciences, Beijing 100049)

Abstract

In consideration of the phrase sparseness problem caused by the exact matching strategy in phrase-based statistical machine translation (SMT) models, the paper presents a phrase similarity-based SMT model. The model introduces the example-based method into SMT. During decoding, when facing source phrases which do not appear in the training corpus, the model firstly computes the similarity between source phrases and finds similar examples from the phrase table by fuzzy matching. Then the model produces translations for these source phrases according to the examples. Compared to the exact matching strategy, fuzzy matching can increase the utilization rate of the phrase table, and to some extent, solves the problem of phrase sparseness. The experiments show that the phrase similarity-based model outperforms the state-of-the-art phrase-based SMT system "Moses" and achieves significant improvements.

Key words: similarity, phrase-based statistical machine translation, example-based machine translation