汉蒙多词表达式的抽取及其在机器翻译中的应用

卫林钰[1,2]，李淼[2]，陈雷[2]，杨振新[1,2]，孙凯[1,2]，陈晟[1,2]

（1. 中国科学技术大学 自动化系，合肥，230026;

2. 中国科学院 合肥智能机械研究所，合肥，230031）

摘要：多词表达式的识别与翻译是自然语言处理领域的一项关键技术，对于统计机器翻译也尤为重要，特别是汉语-蒙古语这种形态非对称且语料稀缺的语言对。本文针对汉语和蒙古语的语言特点，总结了汉语-蒙古语的多词表达式模式，提出了一种规则与统计相结合的方法抽取出汉蒙多词表达式，并使用三种融合策略将多词表达式分别融入统计机器翻译的翻译模型和语言模型中。通过实验证明，汉蒙多词表达式有利于提高汉蒙统计机器翻译的性能。

关键词：统计机器翻译；多词表达式抽取；汉语-蒙古语

Improving Chinese-Mongolian Statistical Machine Translation with Multi-word Expressions

Linyu Wei[1,2], Miao Li[1] ,Lei Chen[1] , ZhenXin Wei[1,2],Kai Sun[1,2],Sheng Chen[1,2]

(1.University Of Science And Technology Of China, HeFei, Anhui 230026, China ;

2. Institute of Intelligent Machines, Chinese Academy of Sciences , HeFei, Anhui 230031, China)

Abstract: Identifying and translating Multi-Word Expressions is a key issue for numerous applications of Natural Language Processing including Statistical Machine Translation, especially for translation asymmetric and low-resource language pairs such as Chinese-Mongolian. In this paper, we summarize several Chinese-Mongolian MWE patterns and propose a method to extract bilingual MWEs. Both both statistical and linguistic methods are employed in MWE extraction. In addition, we integrate the MWEs into the Chinese-Mongolian SMT system by three strategies. Experimental results indicate MWEs are effective for improving Chinese-Mongolian SMT performance even in low-resource circumstance.

Key words: Statistical machine translation; Multi-word expression extraction; Chinese-Mongolian

1 引言

目前主流的统计机器翻译方法大都基于大规模语料库，使用大规模的目标语言语料训练语言模型，再使用大规模的双语平行语料训练翻译模型。然而，语料资源匮乏是汉蒙统计机器翻译研究所面临的最大的问题。

如何在语料资源稀缺的情况下提高统计机器翻译系统性能，Haithem 等[1]利用可比语料来扩充平行语料；Wu 等[2]针对特定领域双语平行语料资源匮乏的问题，提出可以通过大规模的领域内单语语料和大规模的领域外的双语平行语料来弥补领域内的双语平行语料的不足。然而，对于汉蒙语言对，这些方法都面临挑战，因为既没有大规模的汉蒙双语可比语料，也没有大规模的蒙古语单语语料。因此，从有限的语料中挖掘更多有效的翻译知识对于提高汉蒙统计机器翻译系统的性能具有重要意义。目前的研究表明多词表达式在自然语言处理的多个领域具有积极作用，包括机器翻译[3][4]。

多词表达式，即原则上无法逐词翻译的具有特殊意义的单词序列，如固定搭配，命

基金项目：中国科学院信息化专项（XXH12504-1-10）；国家自然科学基金（61572462，61502445）

作者简介：卫林钰（1992——），女，硕士研究生，研究方向自然语言处理-机器翻译；李淼（1955——），女，研究员，主要研究方向自然语言处理

名实体等。Sag 等[5]曾简略地将多词表达式定义为跨越词边界（或词空间）的具有特殊意义的翻译。将多词表达式融合到统计机器翻译系统中，已经在部分语言对中证明了其有效性，例如汉英语言对[6]和英法语言对[7]。

本文的目的是通过汉蒙双语多词表达式来提高汉蒙统计机器翻译系统的性能。首先从短语表中抽取出多词表达式，然后再分别采用基于翻译模型的策略和基于语言模型的策略将抽取出的多词表达式融合到机器翻译系统中。

本文的方法是基于短语表的，属于统计机器翻译的内部资源，相对于基于可比语言等外部资源的方法而言，本文的方法更加适合于语料资源匮乏语言对的统计机器翻译。此外，据我们的了解，本文第一次将多词表达式应用于汉蒙统计机器翻译，还根据汉语和蒙古语的特点总结了用于抽取多词表达式的汉语-蒙古语的多词表达式模式集。

2 汉蒙统计机器翻译中的多词表达式

2．1 系统框架

本文的汉蒙多词表达式是从已对齐的短语表中抽取出来的，即直接抽取双语多词表达式，然后应用到统计机器翻译中，系统框图如图 1 所示。

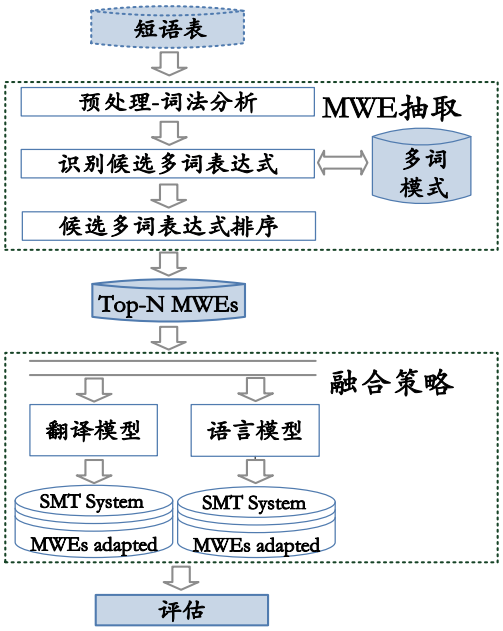

图 1 系统框图

2．2 MWE 的抽取

蒙古语属于形态丰富的黏着语，通过在词干后缀接不同的词缀来实现其语法功能，词与词之已有空格分开。而汉语是词与词之间没有明确的界限，也没有丰富的形态，主要通过独立的虚词和固定的次序来表达语法意义。由于汉蒙的形态差异太大，在语料不充足的情况下，仅仅依赖统计方法抽取的短语表中存在着大量的错误。因此，需要对短语表中的汉语端和蒙古语端进行词法分析，主要涉及汉语端的分词以及词性标注，蒙古语端的词干词缀切分以及词性标注。

汉语分词以及词法分析的相关研究很多，也有很多的开源工具可以使用。蒙古语词法分析的研究相对较少，本文主要参照姜文斌等[8]关于蒙古语词法分析的有向图模型来是实现蒙古语的词干词缀切分以及词性标注。联合切分标注的有向图 G 的概率定义为：

$$P(G) = P(T) \times P(t(T)) \times P(T, t(T))$$

其中，P(T)表示候选树的生成概率，P（t(T)）表示相应的标注树的概率，P(T,t(T))表示平行树结构 T 和 t(T)的映射概率，它定义为平行树中所有节点对的条件概率的乘积。

为了使抽取出的多词表达式符合互译汉蒙多词表达式的词法结构，本文总结了汉语-蒙古语多词表达式匹配模式，如表 1 所示，另外，本文将多词表达式的最大长度设置为 6 个单词。

表 1 汉语-蒙古语多词表达式匹配模式

汉语	蒙古语
Adj-Adi-Noun	Adj-Adj-Noun
Adv-Verb-Verb	Verb-Post-Adv-Verb
Adj-Noun-Verb	Verb-Adj-Noun
Noun-Noun-Noun	Noun-Noun-Noun
Verb-Noun	Noun-Post-Verb
…	…

为了抽取出更加可靠的高质量的双语多词表达式，本文对抽取出的候选多词表达式进行排序，然后挑选出得分相对高的双语多词表达式。排名算法主要涉及三种统计学信息：逐点互信息、对数似然得分、拓展的 C-value。

• 逐点互信息

$$MI(C,M)=\log\frac{n*N(C,M)}{N(C)*N(M)}$$

其中 N(C)和 N(M)分别表示语料中包含汉语多词表达式 C 的包含蒙古语多词表达式 M 的出现句子数，n 表示平行语料的句子数，N(C,M)表示 C、M 共现的句子数。

• 对数似然得分

$$LL(C,M)=\log\sum_{i=1}^{4}4*[X_i*\log\frac{n*X_i}{N(C)*N(M)}]$$

其中 X_1,X_2,X_3,X_4 分别表示语料中 C、M 共现的句子数，仅 C 出现的句子数，仅 M 出现的句子数，C 和 M 均未出现的句子数。

• 拓展的 C-value

$$Cvalue(m)=(|m|-1)*(N(m)-\frac{T(m)}{C(m)})$$

其中 m 表示一个单语的多词表达式，|m|表示 m 中的单词数，N(m)表示语料中 m 出现的次数，T(m)表示将 m 作为子串的所有多词表达式的出现次数，C(m)表示将 m 作为子串的多词表达式的个数。

通常使用 Cvalue(m)来度量单语多词表达式 m 的合法性，本文对其进行拓展来计算双语多词表达式(C，M)的合法性。首先分别计算汉语和蒙古语的 C-value；然后进行排名，r(C)和 r(M)分别表示汉语多词表达式 C 和蒙古语多词表达式 M 的排名；最后将拓展的 C-value 定义为

$$Cvalue(C,M)=\frac{r(M)+(M)}{2}$$

2.3 融合策略

• 使用多词表达式更新短语表

解码器通过查询短语表来获悉如何将输入的源语言翻译成目标语言。然而，由于短语表抽取算法[9]依赖自动词对齐的结果，且对空对齐的点采用双向扩展的方式抽取短语对，导致原始的短语表中存在许多没有意义甚至错误的短语对。为了缓解这个问题，本文对短语表中多词表达式的翻译概率和正反向词汇化概率进行修改。对于短语表中我们抽取出的多词表达式，和 Ren 等[6]的研究一样，我们将这些概率简化为 1。修改过的短语表将会更加合理，因为这些抽取出的多词表达式无论是词汇意义还是对齐的准确率都有了很大的提升。本文将该方法标记为“Update-PT”。

• 添加一个多词表达式特征

Ren 等[6]和 Carpuat 等[3]通过实验证明，在机器翻译系统中添加一个多词表达式特征可以提高系统性能。在他们的研究中，为翻译模型添加一个“0”或者“1”特征来表示该短语对是否为多词表达式。本文认为该特征过于武断，我们对该特征进行了改进，采用 R(MI)、R(LL)和 R(Cvalue)的均值作为新特征，其中 R(MI)、R(LL)和 R(Cvalue)分别表示多词表达式通过 MI、LL 和 C-value 得分的排名。本文将该方法标记为“Add-FEA”。

• 双语言模型

语言模型是统计机器翻译系统的重要组成部分，通常大规模语料训练出来的语言模型效果更好。由于蒙古语语料资源匮乏，原始系统中的语言模型十分简陋。本文使用抽取的多词表达式来扩充原始语料并重新训练语言模型，再使用双语言模型即原始语言模型和新的语言模型进行解码。统计的语言模型 $P(\omega_1,\omega_2,...,\omega_n)$为单词序列分配概率，我们的方法可以提高多词表达式的出现频率，且多词表达式也包含语义和词法知识，因此改进后的语言模型性能更优。本文将该方法标记为“Double-LM”。

3 实验

3．1 实验数据和工具

为了验证本文方法的有效性，我们通过 CWMT2009 提供的汉蒙平行语料进行实验验证，实验数据如表 2 所示。

表 2 汉-蒙平行语料数据

	汉语	蒙古语
训练集	67288	
开发集	400	400*4
测试集	600	600*4

实验使用 Stanford Word Segmenter 进行分词，汉语部分的词性标注使用 Stanford Log-linear Part-Of-Speech Tagger，蒙古语的

词性标注使用 Mglex 词法分析工具[8]。

所有的实验都在开源平台 Mose[10]上进行，使用 GIZA++工具包[9]进行词对齐，使用带 Kneser-Ney 平滑[11]的改进的 SRILM 工具包训练三元语言模型。基线系统使用基于短语的机器翻译系统，并用 BLEU[12]对所有系统翻译结果进行打分，且每个测试语句对应 4 个参考译文。

3．2 实验结果与讨论

在我们的实验中，仅通过模式匹配可以抽取 22525 对汉蒙多词表达式，分别使用 3 种不同的策略融合到机器翻译系统中，结果如表 3 所示。

表 3 使用 3 种融合策略的实验结果对比

	TM-based		LM-based
基线	Update-PT	Add-FEA	Double-LM
18.53	18.95	18.99	18.68

由表 3 可以看出，添加新特征的方法效果最好，但是仅提升了 0.46 个 BLEU 值，效果不明显。因为仅仅通过模式匹配抽取的双语多词表达式中仍然存在很多噪声，因此，需要对其进行过滤，即逐点互信息、对数似然得分、拓展的 C-value 三个统计度量值对多词表达式进行评分并排名，然后抽取排名靠前的 10000 个（实验证明 top10000 的效果最佳，限于文章篇幅，不列出所有的实验结果）多词表达式使用 Add-FEA 策略融入机器翻译系统，结果如表 4 所示。

表 4 使用不同方法抽取出的多词表达式融合到机器翻译系统中的效果

System	BLEU（%）
S1：基线系统	18.53
S2：S1+MWE 模式匹配	18.99
S3：S2+MI	19.14
S4：S2+LL	19.34
S5：S2+Cvalue	19.29
S6：S2+MI+LL+C-value	**19.72**

由表 4 可知，S3、S4、S5、S6 的效果均优于 S2，即统计信息与模式匹配相结合的方法比单纯依靠模式匹配的方法效果要好，其中 S6 的效果最好，比基线系统高出 1.19 个 BLEU 值，即三种统计方法组合后效果更好。此外，比较表 4 与表 3 可知，并不是融合到翻译系统中多词表达式的数量越多越好，是多词表达式的质量而不是数量决定了实验的效果。

3．3 对比实验

为了验证本文方法的有效性，还设置了多词表达式与随机抽取的短语对的对比实验。

为了验证表 3 和表 4 实验效果的提升是由多词表达式引起的，我们从短语表中随机选取 22525 个短语对，然后与多词表达式一样使用三种策略融合到机器翻译系统中，对比实验结果如表 5 所示。

表 5　多词表达式与随机短语对对比结果

融合策略	+多词表达式	+随机短语对
Baseline	18.53	
TM-based		
Update-PT	19.01(+0.48)	18.33(-0.20)
Add-FEA	19.34(+0.81)	18.45(-0.18)
LM-based		
Double-LM	18.91（+0.38）	18.37（-0.16）

表 5 的对比实验结果表明了多词表达式在机器翻译系统中的积极作用，融合相同数量的随机抽取的随机短语对的实验效果甚至比基线更差。

4 相关工作

本文方法的核心技术包括双语多词表达式的抽取和融合到机器翻译系统的策略。

对于双语多词表达式的抽取工作，Ren 等[6]首先采用基于对数似然比层次衰减算法从单语语料中抽取出单语多词表达式，再从短语表中获取该单语多词表达式的翻译，最终得到双语的多词表达式。Bouamor 等[13]通过形态句法分析抽取单语多词表达式，然后采用向量空间模型完成多词表达式的对齐以得到最终的多词表达式。他们的方法均包含两个步骤：首先，分别从单语语料中抽取出单语的多词表达式；然后再使用各种对齐工具对抽取到的多词表达式对齐。本文直接从已对齐的短语表中抽取双语多词表达式，因为先抽取再利用短语表对齐的必要条

件是抽取到的短语必须在短语表中。因此，本文直接从短语表中抽取多词表达式，既充分利用了统计机器翻译系统中重要的内部资源短语表，又无需再从语料中抽取单语多词表达式。

另外，对于抽取算法，主要包含基于规则的方法（如 Bouamor 等[13]的形态句法分析法）和基于统计的方法（如 Ren 等[6]的基于对数似然比层次衰减算法）以及混合的方法（如 Boulaknadel 等[14]）。由于汉蒙语料资源有限，本文采用混合方法，即基于规则的多词表达式模式匹配以及三种统计值相结合的方法。

对于融合策略，传统的方法包括：静态融合和动态融合。但是 Semmar 等[7]的实验证明静态方法效果很差。动态方法主要包括：重新训练翻译模型，添加新的短语表以及添加特征三种方法。重新训练语言模型就是将抽取出来的双语多词表达式作为平行语料添加到原始的平行语料中，然后重新训练翻译模型代替原始的翻译模型；添加新的短语表的方法是由于 Moses 支持多个短语表，因此可以用抽取到的双语多词表达式构造一个新的短语表，并将短语表中的正反向翻译概率以及正反向词汇化概率简化为 1；添加新特征的方法，就是遍历短语表，与抽取到的双语多词表达式进行匹配，并为短语表中的所有短语对添加一个“0/1”特征，用来表征该短语是否属于多词表达式。这些方法都是以不同的方式改变翻译模型，本文还提出一种基于语言模型的方法，即通过多词表达式对语言模型进行改善以提高机器翻译系统的性能。

5 总结

本文首先从短语表中抽取出多词表达式，然后分别采用三种不同的策略融入到机器翻译系统中，实验结果表明多词表达式有效的提高了汉蒙机器翻译系统的性能。此外，本文的方法与相关工作的主要区别在于：

- 本文通过多词表达式提高汉蒙统计机器翻译系统的性能，就我们的了解，这应该是首次将多词表达式应用于蒙古语，此外，本文总结的第一个汉蒙多词表达式模式集。
- 对于语料资源匮乏的语言对，无法获取大规模的平行语料和可比语料，本文的方法仅仅依靠统计机器翻译中的短语表，不需要大规模的外部语料资源。且本文方法还结合了汉语和蒙古语的语言特点，在小规模语料的情况下，准确率更高。
- 关于融合策略，本文提出的基于语言模型的融合策略更加适合语料资源匮乏的语言对，因为小规模语料训练的语言模型通常都比较差，数据稀疏严重，本文的方法涉及词干还原，在一定程度上可以缓解数据稀疏。对于大规模的语料，本文基于翻译模型的融合方法同样具有借鉴意义，因为本文是直接更新短语表，而不是重新构造新的短语表，在一定程度上避免了数据冗余。

然而，本文的融合策略还不够精细，特别是基于语言模型的融合方法，在未来的工作中，仍然需要更加深入的研究，将多词表达式更好的融入到机器翻译。

参考文献

[1] Haithem A., Lo ̈ıc B., and Holger S. Multimodal Comparable Corpora as Resources for Extracting Parallel Data: Parallel Phrases Extraction.[C] Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October , 2013:14-18

[2] Wu, H., Wang, H., Zong, C. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora.[C] Proceedings of the 22nd International Conference on Computational Linguistics,COLING. 2008: 993–1000.

[3] Carpuat, M., Diab, M.T.: Task-based evaluation of multiword expressions: a pilot study in statistical machine translation.[C] Proceedings of Human Language Tech-nologies: Conference of the North American of the Association of Computational Linguistics. 2010: 242–245

[4] Arcan, M., Turchi, M., Tonelli, S., Buitelaar, P.: Enhancing statistical machine translation with bilingual terminology in a cat environment.[J] Proceedings of AMTA 2014, vol. 1: MT Researchers,Vancouver, BC ,2014.

[5] Sag, I.A., Baldwin, T., Bond, F., Copestake, A.A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP.[C] Proceedings Computational Linguistics and Intelligent Text Processing, Third International Conference. 2002:1–15.

[6] Ren, Z., u, Y.L., Cao, J., Liu, Q., Huang, Y.: Improving statistical machine translation using

domain bilingual multiword expressions.[C] Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications. 2009: 47–57.
[7] Bouamor, D., Semmar, N., Zweigenbeaum, P.: Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective.[C] Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III) COLING. 2012:95–108
[8] 姜文斌,吴金星,乌日力嘎, 那顺乌日图, 刘群. 蒙古语有向图形态分析器的判别式词干词缀切分[J] 中文信息学报，2011，第 4 期.
[9] Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. computational linguistics.[J] Computational Linguistics. vol. 38, 2003:9–51.
[10] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N.,Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A.,Herbst, E.: Moses: Open source toolkit for statistical machine translation.[C] ACL.2007: 177–180.
[11] Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling.[C] Proceedings of 34th Annual Meeting of the Association for Computational Linguistics, ACL. 1996:310–318.
[12] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. [C] ACL. 2002:311–318.
[13] Bouamor, D., Semmar, N., Zweigenbaum, P.: Identifying bilingual multi-word expressions for statistical machine translation. [C] Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC). 2012:674–679
[14] S. Boulaknadel, B. Daille, and A. Driss. A multiterm extraction program for arabic language.[C] Proceedings of LREC, Marrakech, Morocco. 2008.