·开发研究与工程应用· 文章编号：1000-3428(2019)04-0288-08 文献标志码：A 中图分类号：TP391

基于释义信息的维汉机器翻译系统融合研究

王亚娟[1,2,3,4]，李 晓[1,3]，杨雅婷[1,3]，米成刚[1,3]

（1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011；2. 中国科学院大学，北京 100049；
3. 新疆民族语音语言信息处理实验室，乌鲁木齐 830011；4. 新疆警察学院 信息安全工程系，乌鲁木齐 830011）

摘 要： 针对维汉机器翻译中单个翻译模型翻译效果差且多个翻译模型间翻译差异较大的问题，提出一种基于释义信息的系统融合方法。通过提取汉语端释义信息对汉语翻译假设进行词对齐，利用词对齐信息构建并解码混淆网络，从而得到维汉机器翻译系统融合结果。实验结果表明，与单个翻译系统 HPSTW 相比，该方法能够有效提高翻译质量。

关键词： 维汉机器翻译；释义信息；系统融合；翻译假设词对齐；释义表过滤

中文引用格式： 王亚娟，李晓，杨雅婷，等. 基于释义信息的维汉机器翻译系统融合研究[J]. 计算机工程，2019，45(4)：288-295，301.

英文引用格式： WANG Yajuan，LI Xiao，YANG Yating，et al. Research of Uyghur-Chinese machine translation system combination based on paraphrase information[J]. Computer Engineering，2019，45(4)：288-295，301.

Research of Uyghur-Chinese Machine Translation System Combination Based on Paraphrase Information

WANG Yajuan[1,2,3,4]，LI Xiao[1,3]，YANG Yating[1,3]，MI Chenggang[1,3]

（1. Xinjiang Technical Institute of Physics and Chemistry，Chinese Academy of Sciences，Urumqi 830011，China；
2. University of Chinese Academy of Sciences，Beijing 100049，China；
3. Xinjiang Laboratory of Minority Speech and Language Information Processing，Urumqi 830011，China；
4. Department of Information Security Engineering，Xinjiang Police College，Urumqi 830011，China）

【Abstract】 Aiming at the problem that the translation effect of single translation model in Uyghur-Chinese machine translation is poor and the translation difference between multiple translation models is large，a system fusion method based on paraphrase information is proposed. By extracting the Chinese interpretation information，the Chinese translation hypothesis is word aligned，and the word alignment information is used to construct and decode the confusion network，and the fusion result of the Uyghur-Chinese machine translation system is obtained. Experimental results show that compared with the single translation system HPSTW，this method can effectively improve the translation quality.

【Key words】 Uyghur-Chinese machine translation；paraphrase information；system combination；translation hypothesis word alignment；paraphrase table filtration

DOI： 10.19678/j.issn.1000-3428.0050313

0 概述

语言是人类沟通的桥梁，机器翻译是建立这一桥梁的重要工具，研究维汉机器翻译是国家“一带一路”发展战略的重要基础。维汉机器翻译中的源语言和目标语言差异导致翻译难度较大[1]。维汉机器翻译模型主要包括基于短语的翻译模型[2]和基于层次短语的翻译模型[3]。基于短语的翻译模型将短语作为最小翻译单元，其原理简单且功能强大，但是无法处理长距离调序[4]。基于层次短语的翻译模型由于借用形式化语法的结构使得翻译过程变得层次化，能够处理复杂的远距离重排序，但是由于数据稀疏性和噪声影响，在解码时搜索空间较大，可能导致搜索错误。在维汉机器翻译中，2 个模型均达不到理想的翻译效果[5]。因此，将多个翻译系统进行融合具有重要意义。

基金项目： 国家自然科学基金(U1703133)；中科院西部之光人才培养引进计划(2017-XBQNXZ-A-005)；中国科学院青年创新促进会项目(2017472)；新疆维吾尔自治区重大科技专项(2016A03007-3)；新疆维吾尔自治区高层次人才引进工程(Y839031201)。

作者简介： 王亚娟(1983—)，女，讲师、博士研究生，主研方向为多语种信息处理；李 晓，研究员、博士生导师；杨雅婷，副研究员、博士；米成刚，助理研究员、博士。

收稿日期： 2018-01-29 **修回日期：** 2018-03-27 **E-mail：** wangyajuan@ms.xjb.ac.cn

近年来,国内外学者将系统融合应用于机器翻译领域[6-8],如基于混淆网络解码的词汇级别融合方法[9]。在词汇级别系统融合方法中,翻译假设词对齐影响系统融合性能[10],其代表方法主要有基于编辑距离的翻译假设词对齐[11]、基于语料库的翻译假设词对齐[12]、基于语言学知识的翻译假设词对齐[13]和基于 Meteor 的翻译假设词对齐[14-15],但将上述方法应用于汉语翻译假设词对齐时,无法找到汉语翻译假设中互为释义的词或短语。释义是表达相同信息的替代方法[16],在文本生成和文本摘要研究中已被证明可以用来生成更加流利和多样性的文本,且能够找到输入文档中的重复信息[17-19],同时可以提高机器翻译质量。

本文结合释义在机器翻译中的应用以及维汉机器翻译的自身特性,提出一种机器翻译系统融合方法。利用维汉双语平行语料提取汉语端的释义信息,并将其引入系统融合的翻译假设词对齐阶段,通过改善词对齐质量来提升系统融合效果。

1　相关工作

目前,系统融合已成为机器翻译领域的重要研究方向。根据研究对象的不同,现有系统融合主要从句子级别[20]、短语级别[21]和词汇级别[9]这 3 个层面进行研究。基于编辑距离的翻译假设词对齐[11]在各种大型机器翻译任务中性能优异,但是在搜索最佳对齐时,其使用表面形式匹配来计算编辑距离,因此不支持同义词匹配且对非单调词排序的建模比较粗糙。基于语料库的翻译假设词对齐方法[12]将各系统的翻译假设当成是平行语料,用双语对齐模型翻译假设的双向词对齐,但该对齐方法的对齐质量依赖语料库大小,在语料缺乏时容易导致数据稀疏,且同样不支持除表面形式外的其他形式匹配。基于语言学知识的翻译假设词对齐方法[13]通过引入语言学知识如 WordNet 来提高翻译假设词对齐质量,同时增加同义词和词干匹配,但过度依赖语言学知识,对缺少语言学知识资源的语言并不适用,而且不支持释义信息匹配。基于 Meteor 翻译假设词对齐[14-15]是目前词汇级别系统融合中广泛使用的翻译假设词对齐方法,它在表面形式匹配、词干、同义词匹配基础上增加释义匹配模块,但仅支持部分语言的释义匹配,因此在进行维汉机器翻译系统融合研究时,没有可用的汉语释义信息,需要单独提取汉语释义信息,并将其用于维汉机器翻译的系统融合过程中。

释义提取方法主要有人工收集、利用现有的词汇资源(如 WordNet)提取和基于语料库的释义提取[22]3 种。其中,根据语料库类型不同,基于语料库的释义提取方法可以分为从单语语料库中提取、从单语可比语料库中提取、从单语平行语料库中提取和从双语平行语料库中提取[22]等方法。释义提取方法主要通过双语语料库生成释义[23],只要有可用的双语语料资源和双语短语表,一种语言中短语的释义对就可以通过使用另一种语言中的短语作为枢轴来推断。

2　基于释义信息的系统融合

在实现维汉机器翻译的系统融合时,本文利用双语语料资源提取汉语端释义信息,并修正翻译假设词对齐结果。在此基础上,通过构建和解码混淆网络获得维汉机器翻译系统的融合结果。

2.1　汉语释义表提取

在系统融合时,各系统翻译假设间的单语词对齐至关重要,其直接影响融合翻译选取的好坏。当汉语翻译假设进行单语词对齐时,现有工具只支持单词表面形式的精准匹配(如果 2 个翻译假设中出现的词完全相同,则认为这 2 个词匹配),会导致具有相同匹配数且能组成比较完整信息的对齐丢失。文献[24]通过引入释义信息使现有对齐工具支持多个词的短语匹配,如果一个短语被释义数据库认为是另一个短语的释义,那么释义匹配器就会匹配该短语。但加入释义匹配功能的对齐工具仅支持英语、阿拉伯语、捷克语、法语、德语和西班牙语的释义匹配,由于没有汉语的释义信息,因此无法利用对齐工具直接进行汉语释义匹配。

本文从双语平行语料中提取释义,从维汉双语语料库中提取汉语释义表[22],在提取释义表时利用维汉短语表和维汉双语训练语料自动学习释义表。具体学习过程为:对于短语表中的每个汉语短语 e_1,首先确定翻译成 e_1 的维吾尔语短语 f,然后查找所有可能翻译成 f 的汉语短语 e_2,每个可以翻译成 f 的汉语短语 e_2 被认为是具有概率 $p(f|e_1)p(e_2|f)$ 的 e_1 释义,e_2 释义 e_1 的总概率计算公式如下:

$$p(e_2|e_1)=\sum_f p(f|e_1)\cdot p(e_2|f) \tag{1}$$

为提高释义精确性,要进行一系列过滤。为避免释义中只包含“啊”“吧”等停用词,丢弃 $p(f|e_1)\cdot p(e_2|f)<0.001$ 的释义,然后丢弃 e_1、f 和 e_2 中包含任何标点符号的释义,最终为避免释义表影响词对齐效率,只保留 $p(e_2|e_1)>0.01$ 的释义条目。

2.2 基于释义信息的翻译假设词对齐

在对汉语翻译假设进行词对齐时,由于语料缺乏且翻译假设间存在词序不一致、同义词、同根词和同源词等比较难处理的情况,因此本文使用支持单字匹配和释义匹配等多种匹配方式的单语词对齐方法来实现汉语翻译假设间的对齐。对齐处理过程具体描述如下:

1)翻译假设预处理。对齐只支持句对的对齐,不支持多个翻译假设的同时对齐,而在系统融合时翻译假设的输入为M个(M为参与融合单个系统的数量),且在进行词对齐时需要选择一个翻译假设作为主要假设,其他假设均与该主要假设对齐。考虑到不同翻译假设的不同词序,在对齐时让每个系统假设都当一次主要假设,剩余系统假设与当前主要假设对齐。对齐过程逐句进行,假设有M个翻译系统,每个翻译系统的翻译假设均由N个句子组成,那么最后经过组合共生成$M\times(M-1)\times N$个待对齐的句对。

2)根据不同匹配方式找出所有可能的匹配集合。表 1 给出待对齐的翻译假设句对,以翻译假设 2 为主要假设,翻译假设 1 为次要假设且与翻译假设 2 对齐,在对齐时首先找出两句话中所有的精准匹配,在寻找匹配时将 2 个字符串分别划分为单个单词,循环遍历翻译假设 1 中的所有单词,并和翻译假设 2 中所有单词比较,如果 2 个单词完全一样,则认为这 2 个单词匹配,并将该匹配加入候选匹配集合。其中,“主体”匹配信息如表 2 所示。由于引入释义匹配信息后,对齐单位可能从单词变为短语,因此不仅要记录匹配的开始位置信息,还要记录匹配的长度信息。“主体”是翻译假设 2 中索引为 3 的词,“3:1 3:1”表示翻译假设 2 中索引为 3 的词与翻译假设 1 中索引为 3 的词匹配,且匹配长度均为 1。“主体”一词在翻译假设 1 中出现了 3 次,它们分别与翻译假设 2 中的“主体”匹配,将所有可能的匹配列出就出现了 3 组匹配信息,分别是 3:1 3:1、3:1 6:1、3:1 12:1。

表 1 待对齐的翻译假设

翻译假设	句子
翻译假设 1	企业是市场主体和投资主体以及经济贸易合作的主体。
翻译假设 2	企业是市场主体和投资主体,也是经贸合作的主体。

表 2 “主体”精准匹配结果

索引	词	匹配信息
3	主体	3:1 3:1 3:1 6:1 3:1 12:1

在精准匹配后进行释义匹配,需借助外部的释义信息,在查找释义匹配时分别以翻译假设 1、翻译假设 2中的词为索引,2 个方向逐个查找在释义表中以该词索引开始的释义条目,如果存在就保存并继续查找该索引后跟字符串中第 2 个词的释义条目是否存在,以此类推,直到找到 2 个字符串中同时出现且互为释义的短语或单词并将其匹配信息保存,经过释义匹配后,“经贸”匹配集合结果如表 3 所示。匹配集合结合中多了一个匹配信息“10:1 8:2”,这表示翻译假设 2 中第 10 个词与翻译假设 1 中的第 8 个和第 9 个词匹配(匹配长度为 2),“经贸”一词与短语“经济 贸易”匹配,即释义匹配找到新的匹配信息。

表 3 “经贸”释义匹配结果

索引	词	匹配信息
10	经贸	10:1 8:2

3)确定最终对齐结果。按照每个句子中的每个单词都会映射到另一个句子中 0 个或 1 个单词的原则,确定所有可能的匹配集合,将匹配集合组合可以得到多个对齐子集,找到得分最高的对齐子集,并将其作为最终对齐结果输出,其中,图 1 给出对齐子集实例。具体描述如下:

(1)如果获得多个对齐子集,则选择 2 个句子覆盖单词数量最多的子集。如图 1(a)和图 1(b)所示,对齐子集 1 覆盖翻译假设 2 中的单词数量为 11,对齐子集 2 覆盖翻译假设 2 中的单词数量为 12,因此选择覆盖单词数量较多的对齐子集 2。

(2)选择“块”数量最小的对齐(“块”指在 2 个句子中连续且顺序相同的一系列匹配短语)。如图 1(c)和图 1(d)所示,对齐子集 3 的块数量为 4,对齐子集 4 的块数量为 2,因此选择块数量较小的对齐子集 4。

(3)如果获得多个对齐子集,且 2 个子集覆盖单词的数量相同,则选择 2 个句子中匹配开始索引之间绝对距离最小的子集,即更偏向于选择在 2 个句子中出现在相似位置的短语对齐。如图 1(e)和图 1(f)所示,对齐子集 5 和对齐子集 6 所覆盖的单词数量和块数量都相同,但对齐子集 6 的匹配开始索引之间绝对距离比对齐子集 5 的短,因此选择距离较短的对齐子集 6。

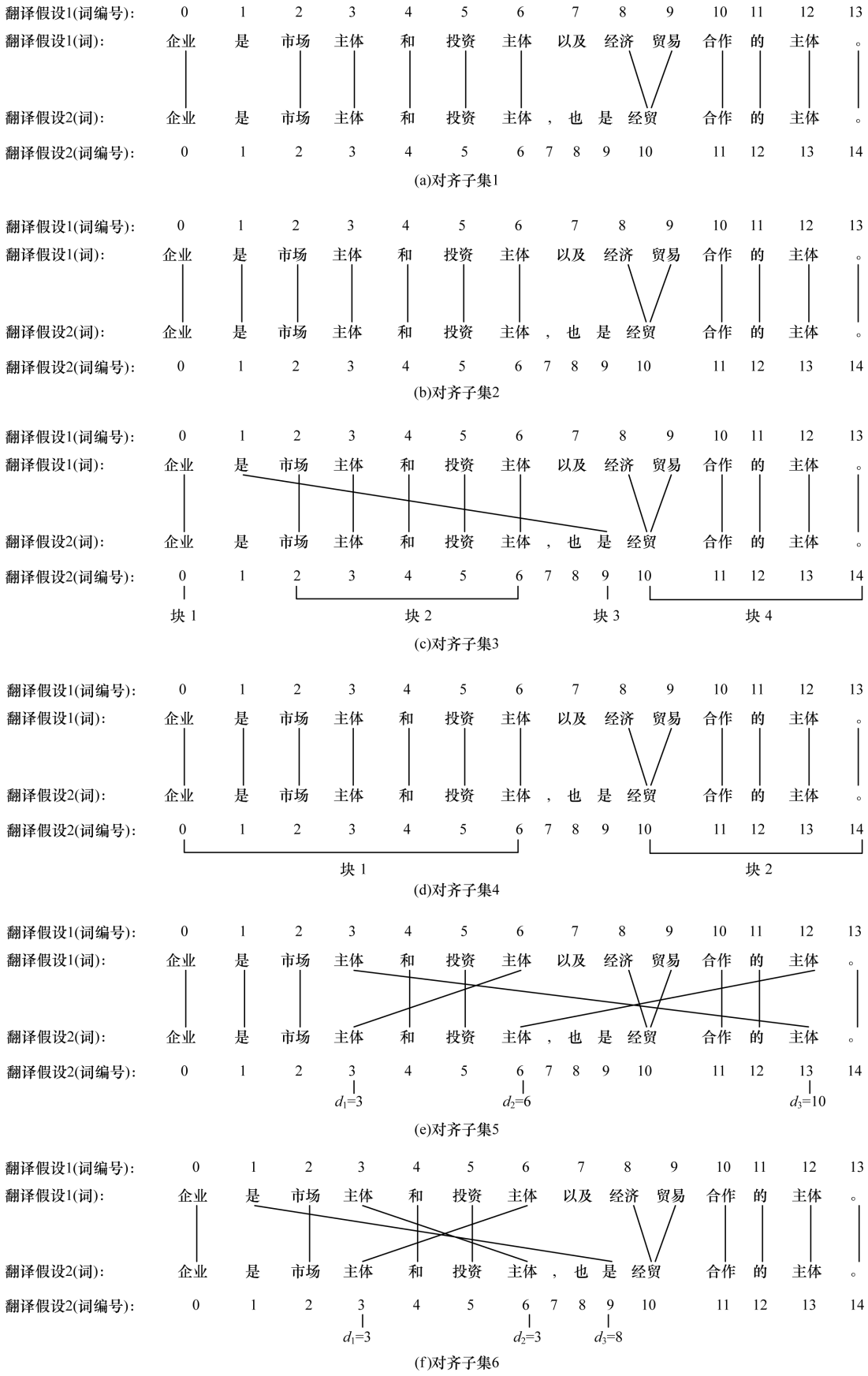
翻译假设1(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13
翻译假设1(词): 企业 是 市场 主体 和 投资 主体 以及 经济 贸易 合作 的 主体 。
翻译假设2(词): 企业 是 市场 主体 和 投资 主体 , 也 是 经贸 合作 的 主体 。
翻译假设2(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
(a)对齐子集1
翻译假设1(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13
翻译假设1(词): 企业 是 市场 主体 和 投资 主体 以及 经济 贸易 合作 的 主体 。
翻译假设2(词): 企业 是 市场 主体 和 投资 主体 , 也 是 经贸 合作 的 主体 。
翻译假设2(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
(b)对齐子集2
翻译假设1(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13
翻译假设1(词): 企业 是 市场 主体 和 投资 主体 以及 经济 贸易 合作 的 主体 。
翻译假设2(词): 企业 是 市场 主体 和 投资 主体 , 也 是 经贸 合作 的 主体 。
翻译假设2(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
块 1
块 2
块 3
块 4
(c)对齐子集3
翻译假设1(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13
翻译假设1(词): 企业 是 市场 主体 和 投资 主体 以及 经济 贸易 合作 的 主体 。
翻译假设2(词): 企业 是 市场 主体 和 投资 主体 , 也 是 经贸 合作 的 主体 。
翻译假设2(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
块 1
块 2
(d)对齐子集4
翻译假设1(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13
翻译假设1(词): 企业 是 市场 主体 和 投资 主体 以及 经济 贸易 合作 的 主体 。
翻译假设2(词): 企业 是 市场 主体 和 投资 主体 , 也 是 经贸 合作 的 主体 。
翻译假设2(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
d_1=3
d_2=6
d_3=10
(e)对齐子集5
翻译假设1(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13
翻译假设1(词): 企业 是 市场 主体 和 投资 主体 以及 经济 贸易 合作 的 主体 。
翻译假设2(词): 企业 是 市场 主体 和 投资 主体 , 也 是 经贸 合作 的 主体 。
翻译假设2(词编号): 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
d_1=3
d_2=3
d_3=8
(f)对齐子集6

图 1　对齐子集实例

根据上述规则对所有可能的对齐子集进行排序,并选出得分最高的子集,即为该句对最终对齐结果,如图 2 所示。该句对最终对齐结果如图 2(a)所示,释义匹配的其他情况如图 2(b)和图 2(c)所示。其中,●表示不使用释义信息的对齐点,★代表引入释义信息后新增的对齐点。在进行释义匹配后,不仅可以匹配“经贸”和“经济 贸易”这一对释义短语,还可以匹配“拓宽”和“扩大”这一对同义词(图 2(b)),以及同时匹配“情况”和“状况”、“稳定”和“平稳”这两对同义词(图 2(c))。因此,引入释义信息既可以实现单个词的同义词匹配,也可以完成多个词的短语匹配。

	企业	是	市场	主体	和	投资	主体	,	也	是	经贸	合作	的	主体	。
企业	●														
是		●													
市场			●												
主体				●											
和					●										
投资						●									
主体							●								
以及															
经济											★				
贸易											★				
合作												●			
的													●		
主体														●	
。															●

(a)释义短语匹配

	第二	,	进一步	拓宽	合作	领域	。
第二	●						
,		●					
进一步			●				
扩大				★			
合作					●		
领域						●	
。							●

(b)单个同义词匹配

	福岛	核电站	事故	机组	情况	稳定	。
福岛	●						
第一							
核电站		●					
事故			●				
趋于							
平稳						★	
机组				●			
状况					★		
。							●

(c)多个同义词匹配

图 2 词对齐结果

2.3 混淆网络构建与解码

2.3.1 混淆网络构建

在确定词对齐后,所有向主要假设对齐的次要假设中的词均按照主要假设词序进行调序,然后生成 $M-1$ 个一对一的对齐结果。对齐结果实例如表 4所示,其中,加粗表示系统 1 的假设为主要假设,系统 2、系统 3 和系统 4 翻译假设与系统 1 翻译假设对齐。对齐用“|”符号表示,“|”右边为主要假设,左边为次要假设。“$”符号代表空词,其中,“行|旅”表示次要假设中的词“行”与主要假设中的词“旅”对齐,“$|此”表示没有词与主要假设中的词“此”对齐,“由|$”表示次要假设中的词“由”没有与主要假设中的任何词对齐。

表 4 系统假设及对齐实例

系统假设及对齐	句子
系统假设 1	**中国 藏学家 代表团 此 次 欧洲 之 旅 国务院 新闻 办公室 组织 的 。**
系统假设 2	中国 藏学家 代表团 由 国务院 新闻 办公室 组织 的 这 次 欧洲 之 行 。
系统假设 3	由 国务院 新闻 办公室 组织 的 中国 藏学家 代表团 此 次 欧洲 之 。
系统假设 4	中国 藏学家 代表团 此 次 欧洲 之 行,国务院 新闻 办公室 组织 的 。
对齐结果 1	中国|中国 藏学家|藏学家 代表团|代表团 由|$ $|此 次|次 欧洲|欧洲 之|之 行|旅 国务院|国务院 新闻|新闻 办公室|办公室 组织|组织 的|的 这|$ 。|。
对齐结果 2	中国|中国 藏学家|藏学家 代表团|代表团 此|此 次|次 欧洲|欧洲 之|之 $|旅 由|$ 国务院|国务院 新闻|新闻 办公室|办公室 组织|组织 的|的 。|。
对齐结果 3	中国|中国 藏学家|藏学家 代表团|代表团 此|此 次|次 欧洲|欧洲 之|之 行|旅,|$ 国务院|国务院 新闻|新闻 办公室|办公室 组织|组织 的|的 。|。

混淆网络是一种加权有向图,从起始节点到终端节点的每个路径都经过所有其他节点。在创建混淆网络时,从初始状态 0 开始,从左向右开始处理主要假设,每出现一个词,就为该词创建一个状态,然后从先前状态到该状态为与该词对齐的所有词(包括空词)创建弧。如果该词后面有新词,为该新词创建状态和弧,直至所有词结束。混淆网络的构建实例如图 3 所示。

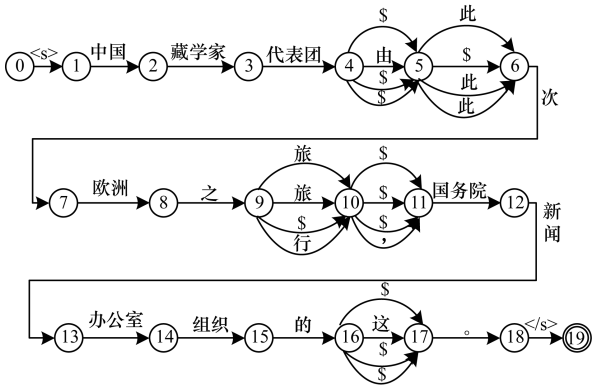

图 3 混淆网络示意图

2.3.2　混淆网络解码

混淆网络解码就是抽取融合翻译的过程,为避免单一“主要假设”为融合翻译带来错误的词序,将多个混淆网络组合形成多混淆网络,图 3 为单混淆网络示意图,如果分别使用系统 2、系统 3 和系统 4 假设作为主要假设,其余 3 个系统的假设与之对齐,每个句子会得到 4 个混淆网络,将这 4 个混淆网络组合形成一个多混淆网络,最终解码就是在多混淆网络中寻找一条从起始节点到终端节点且遍历所有中间节点的最优路径。传统抽取融合翻译方法是在混淆网络每个位置计算词在该位置上的概率,融合翻译抽取过程就是词序列的排列,在每个位置选择概率最高的词,将这些词连接起来得到最终融合翻译。

由于选词时不排除空弧的参与,因此可能会从多混淆网络中抽取多个相同的词片段。本文将相同片段的概率相加,同时利用 n-gram 语言模型对多个混淆网络的集合进行重新打分。该语言模型是在翻译假设集合上训练的一个特殊语言模型,用于指导融合翻译的选择,在使用 n-gram 语言模型重新打分时需要对多混淆网络中的路径进行变换,将空弧移除,只计算非空词的概率,然后将词概率和语言模型进行对数线性插值,为每个可能的路径计算得分,得分最高为最终输出结果。具体计算公式如下:

$$(\hat{L},\hat{e}_1^{\hat{L}}) = \underset{L,e_1^L}{\operatorname{argmax}} \left\{ \alpha^L \prod_{i=1}^{L} (p(e_i \mid F) \cdot p_{LM}^{\lambda}(e_i \mid e_{i-1})) \right\} \tag{2}$$

其中,$p(e_i \mid F)$ 为单词 e 出现在位置 i 的概率,i 从 1 ~ L,L 是变换后单个完整路径的长度,λ 是语言模型的比例参数,α 是词惩罚。首先 λ、α 被分配一个初始值,然后在开发集上利用最小错误率训练算法对 2 个参数进行优化,以找到最佳得分的翻译作为共识翻译。

3　实验结果与分析

3.1　实验设置

由于维汉机器翻译公开可用的翻译系统数量不多、翻译系统种类较少且无法满足系统融合的要求,因此本文实验所用到的 4 个单个翻译系统(PBWTW、HPWTW、PBSTW、HPSTW)均使用公开的维汉双语语料在 Moses 平台[25]单独训练得到。在进行系统融合实验时,汉语单语词对齐使用的工具是美国卡内基·梅隆大学语言技术研究所开发的自动机器翻译评价系统 Meteor[24],词对齐是 Meteor 评价过程中的中间结果,混淆网络解码器使用德国亚琛工业大学开发的开源统计机器翻译工具包 Jane 中的解码器[14]。在训练单个翻译系统时所使用的语料是 CWMT2013 评测中的一组公开的维汉双语新闻领域的语料。在搭建单个机器翻译系统时,本文将语料分为训练集、开发集和测试集。数据样本情况如表 5 所示。

表 5　维汉双语语料数据样本

类别	句对数
训练集	109 485
开发集	700
测试集	1 000

基于短语翻译模型训练的维语词到汉语词的维汉机器翻译系统(PBWTW)和基于层次短语翻译模型训练的维语词到汉语词的维汉机器翻译系统(HPWTW)是维汉机器翻译领域常用的 2 个翻译系统,它们均使用维、汉双语语料的最小划分单位为词,训练翻译模型时分别采用基于短语的翻译模型和基于层次短语的翻译模型。除此之外,文献[26]发现在维汉机器翻译中将维吾尔语切分为更小粒度的词干可以进一步提高维汉机器翻译质量,在维吾尔语词干语料基础上,分别训练基于短语模型的维吾尔语词干到汉语词的翻译系统(PBSTW)和基于层次短语模型的维吾尔语词干到汉语词的翻译系统(HPSTW),这 2 个系统均使用词干级别维吾尔语端的语料,汉语端使用词作为最小划分单位。4 个单个翻译系统均采用 SRILM 语言模型工具包[27]训练一个五元语言模型并使用同一个语言模型;双语词对齐训练由 GIZA ++ 工具包[28]完成,词对齐采用“grow-diag-final-and”策略;2 个基于短语模型的翻译系统需要进行单独的调序,调序的策略为“wbe-msd-bidirectional-fe-allff”;4 个系统使用同样的开发集进行最小错误率训练来优化模型参数,并采用相同的测试集测试翻译效果,翻译效果使用 BLEU 值[29]来评估。4 个系统的翻译 BLEU 值如表 6 所示。

表 6　不同翻译系统 BLEU 值对比　%

系统	BLEU 值
PBWTW 系统	33.93
HPWTW 系统	35.35
PBSTW 系统	35.03
HPSTW 系统	35.97

本文设置 3 组实验进行对比,分别是:

1)基线系统融合。使用不带释义表的单字匹配进行翻译假设的词对齐,按顺序完成混淆网络的搭建和解码,最后在测试集上进行测试。

2)带释义表的对齐 + 系统融合。利用维汉双语语料训练释义表,使用单字匹配和释义匹配结合的方式进行翻译假设的词对齐,按顺序完成混淆网络的搭建和解码,最后在测试集上进行测试。

3)带过滤后释义表的对齐 + 系统融合。将释义表释义概率的阈值提高,过滤掉部分释义表,使用单字匹配和释义匹配(过滤后的释义表)结合的方式进行翻译假设的词对齐,按顺序完成混淆网络的搭建和解码,最后在测试集上进行测试。

3.2 结果分析

为更加细致地观察释义表信息的引入是否会影响维汉机器翻译系统融合质量,在进行翻译假设词对齐时:首先观察释义表的引入是否会影响词对齐的质量;其次观察在影响翻译假设词对齐的基础上对最终的融合效果是否有影响;最后通过对释义表的过滤,观察释义表信息的减少对系统融合的效果是否有影响。具体分析如下:

1)释义信息对翻译假设词对齐的影响

本文对不加入释义表和加入释义表时的对齐结果进行对比,如表 7 所示。

表 7 引入释义表前后翻译假设词对齐结果

对齐任务	精确率	召回率	Meteor 得分
无释义表的翻译假设词对齐	0.82	0.82	0.71
带释义表的翻译假设词对齐	0.83	0.83	0.79

从表 7 可以看出,在进行翻译假设间的词对齐时,通过使用维汉双语语料抽取的汉语释义表,精确率和召回率以及 Meteor 得分均有提高,Meteor 得分提高 0.08,结果表明,在对齐阶段引入汉语释义信息后,使得 Meteor 在实现单字匹配的同时开启释义匹配的功能,释义的匹配使对齐的粒度从词级别扩展到短语级别,在一定程度上可以提高对齐质量。

2)释义表对系统融合结果的影响

在验证释义表对翻译假设词对齐质量有所提高后,将引入释义表前后的对齐结果分别放入系统融合的过程中,观察释义信息对系统融合效果的影响,系统融合结果如表 8 所示,其中,HPSTW 是单个最好系统。

表 8 引入释义表前后系统融合结果 %

系统	BLEU 值
HPSTW 系统	35.97
基线系统融合	36.55
带释义表的对齐 + 系统融合	36.71

从表 8 可以看出,系统融合的确是提高维汉机器翻译质量的有效途径,基线系统融合后得到融合翻译,其 BLEU 值比单个最好系统提高 0.58,基线系统融合已经获得比单个翻译系统较好的结果,在此基础上将带有汉语释义信息的对齐结果引入到系统融合后发现,BLEU 值在基线系统融合基础上又提高 0.16,与单个系统相比 BLEU 值提高了 0.74,结果表明,在系统融合过程中引入汉语释义信息后,在提高翻译假设的词对齐质量的同时提高了系统融合质量。但是,通过观察最终得到的融合结果后会发现,虽然引入释义表后最终融合结果中的句子大都比原始翻译假设的翻译结果好,但也有部分句子翻译效果不理想,这是因为释义表中存在部分错误释义信息,为对齐结果加入噪声信息进而影响系统融合的效果。因此,本文设置第 3 组实验,通过提高释义表中释义概率的过滤阈值来筛选释义表,观察过滤后释义表对系统融合结果的影响。

3)释义表过滤对系统融合结果的影响

为使释义表中既不包含全是停用词的条目,又能涵盖尽可能多的释义信息,释义表释义概率的过滤阈值为 0.01,但是通过观察释义表发现,释义表中确实存在许多错误的释义信息。为保留正确的释义信息,筛选错误的释义信息,本文通过改变过滤阈值方法来调整释义表。释义表过滤阈值步长设置的过大会将许多有用的释义信息直接过滤掉,无法找到合适的过滤阈值;释义表过滤阈值步长设置的过小,会使实验工作量加大进而降低实验的效率,因此本文将过滤阈值从 0.02 ~ 1 按照步长为 0.02 分别过滤释义表,然后用过滤后的释义表进行翻译假设词对齐,并将不同的词对齐结果引入系统融合来观察过滤后释义表对系统融合效果的影响。通过释义表过滤,系统融合结果如图 4 所示。

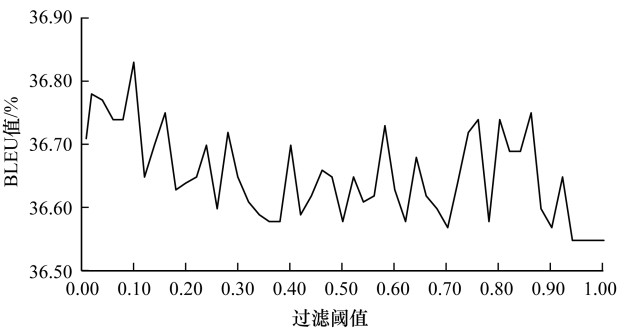

图 4 维汉机器翻译系统融合结果

从图 4 可以看出,用不同阈值过滤释义表,将过滤后的释义表引入翻译假设词对齐后,系统融合的结果均优于单个系统,但还是会上下浮动,这是因为将阈值设置过低会引入错误的释义信息,阈值设置的过高又过滤掉有用的释义信息,因此要为释义表设置合适的阈值。当阈值为 0.1 时,系统融合结果最好;当阈值为 0.01 时,释义表的条目数为 2 806 958 行;当阈值提高至 0.1 时,释义表的条目数仅剩 214 530 行,即仅使用 7.6% 的释义信息,就可以达到较好的效果。但释义信息不是越多越好,有用的释义信息才能帮助提高系统融合效果。引入 2 个释义表后系统融合对比结果如表 9 所示。

表 9　释义表过滤前后系统融合对比结果　%

系统	BLEU 值
HPSTW 系统	35.97
基线系统融合	36.55
带释义表的对齐 + 系统融合	36.71
带过滤后释义表的对齐 + 系统融合	36.83

从表 9 可以看出,将释义表的阈值设置为 0.1 后,系统融合结果在释义表阈值 0.01 基础上 BLEU 值提高 0.12,比基线系统融合高 0.28,与单个最好系统相比,BLEU 值总体提高 0.86,过滤后系统融合结果 BLEU 值比单个最好系统提升 2.4% 。因此,在维汉机器翻译中应用系统融合的方法可行,释义信息的引入能够有效地提升机器翻译的质量。

4　结束语

本文提出一种基于释义信息的维汉机器翻译系统融合方法,将系统融合技术应用于维汉机器翻译,单独提取汉语释义信息,并将其引入维汉机器翻译系统融合过程中。实验结果表明,基于释义信息的维汉机器翻译系统融合可以显著提高维汉机器翻译的质量。但目前的系统融合技术仅采用语言模型和全局系统权重作为混淆网络解码的主要参数,且对释义表的过滤只是采用简单地提高阈值的方法。下一步将从维吾尔语语言特性出发,寻找更适合维汉机器翻译的系统融合新方法。

参考文献

[1] 李晓,蒋同海,周喜,等. 维汉机器翻译关键技术研究概述[J]. 网络新媒体技术,2016,5(1):19-25.

[2] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Stroudsburg, USA: Association for Computational Linguistics, 2003: 48-54.

[3] CHIANG D. Hierarchical phrase-based translation [J]. Computational Linguistics,2007,33(2):201-228.

[4] DURRANI N, FRASER A, SCHMID H, et al. Can Markov models over minimal translation units help phrase-based SMT? [C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics,2013:399-405.

[5] 李晓,蒋同海,周喜,等. 面向复杂形态语言机器翻译的多模型融合词性标注研究[J]. 网络新媒体技术,2014,3(1):60-64.

[6] PETER J T,ALKHOULI T,NEY H,et al. The QT21/HimL combined machine translation system [C]//Proceedings of the 1st Conference on Machine Translation. Stroudsburg, USA: Association for Computational Linguistics, 2016: 344-355.

[7] DURRANI N, DALVI F, SAJJAD H, et al. QCRI machine translation systems for IWSLT 16 [EB/OL]. [2017-12-20]. https://arxiv.org/pdf/1701.03924.pdf.

[8] BANGALORE B, BORDEL G, RICCARDI G. Computing consensus translation from multiple machine translation systems[C]//Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding. Washington D. C., USA:IEEE Press,2001:351-354.

[9] FREITAG M,PETER J T,PEITZ S,et al. Local system voting feature for machine translation system combination [C]// Proceedings of the 10th Workshop on Statistical Machine Translation. Stroudsburg, USA: Association for Computational Linguistics,2015:467-476.

[10] ROSTI A V I, HE X, KARAKOS D, et al. Review of hypothesis alignment algorithms for MT system combination via confusion network decoding[C]//Proceedings of the 7th Workshop on Statistical Machine Translation. Stroudsburg, USA: Association for Computational Linguistics, 2012: 191-199.

[11] GONZÁLEZ-RUBIO J,CASACUBERTA F. Minimum Bayes' risk subsequence combination for machine translation [J]. Pattern Analysis and Applications,2015,18(3):523-533.

[12] ZHANG H,WANG D H,LIU C L. Character confidence based on N-best list for keyword spotting in online Chinese handwritten documents [J]. Pattern Recognition, 2014,47(5):1880-1890.

[13] KARAKOS D, EISNER J, KHUDANPUR S, et al. Machine translation system combination using ITG-based alignments [C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics,2008:81-84.

[14] FREITAG M, HUCK M, NEY H. Jane: open source machine translation system combination[C]//Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg,USA:Association for Computational Linguistics, 2014:29-32.

[15] HEAFIELD K, LAVIE A. Combining machine translation output with open source:the carnegie mellon multi-engine machine translation scheme [J]. The Prague Bulletin of Mathematical Linguistics,2010,93(1):27-36.

[16] 张丽林,李茂西,肖文艳,等. 机器翻译自动评价中领域知识复述抽取研究[J]. 北京大学学报(自然科学版),2017,53(2):230-238.

[17] AL-SMADI M, JARADAT Z, AL-AYYOUB M, et al. Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features [J]. Information Processing and Management,2017,53(3):640-652.

[18] SERAJ R M, SIAHBANI M, SARKAR A. Improving statistical machine translation with a multilingual paraphrase database[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics,2015:1379-1390.

[19] MA W Y,MCKEOWN K. System combination for machine translation through paraphrasing [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg,USA:Association for Computational Linguistics,2015:1053-1058.

(下转第 301 页)

[4] TSIOURIS K M,MARKOULA S,KONITSIOTIS S,et al. A robust unsupervised epileptic seizure detection methodology to accelerate large EEG database eval uation[J]. Biomedical Signal Processing and Control, 2018,40:275-285.

[5] 张栋,陈东伟,游雅,等. 基于自适应 Lempel-Ziv 复杂度的情感脑电信号特征分析[J]. 计算机应用与软件,2014,31(9):162-165.

[6] LIN Y P,JUNG T P. Exploring day-to-day variability in EEG-based emotion classification[C]//Proceedings of IEEE International Conference on Systems, Man and Cybernetics. Washington D. C. ,USA:IEEE Press,2014: 2226-2229.

[7] LIN Y P,HSU S H,JUNG T P. Exploring day-to-day variability in the relations between emotion and EEG signals[C]//Proceedings of Foundations of Augmented Cognition. Berlin,Germany:Springer,2015:461-469.

[8] JAO P K,LIN Y P,YANG Y H,et al. Using robust principal component analysis to alleviate day-to-day variability in EEG based emotion classification[C]// Proceedings of Engineering in Medicine and Biology Society. Washington D. C. , USA: IEEE Press, 2015: 570-573.

[9] LIN Y P,JAO P K,YANG Y H. Improving cross-day EEG-based emotion classification using robust principal component analysis [J]. Frontiers in Computational Neuroscience,2017,11:64.

[10] ARVANEH M,GUAN C,ANG K K,et al. EEG data space adaptation to reduce intersession nonstationarity in brain-computer interface[J]. Neural Computation,2013, 25(8):2146-2171.

[11] CHENG M M,LU Z H,WANG H X. Regularized common spatial patterns with subject-to-subject transfer of EEG signals[J]. Cognitive Neurodynamics, 2017, 11 (2): 173-181.

[12] SONG X M, YOON S C. Improving brain-computer interface classification using adaptive common spatial patterns[J]. Computers in Biology and Medicine,2015, 61(1):150-160.

[13] 吕甜甜,王心醉,俞乾,等. 基于脑电和肌电多特征的自动睡眠分期方法[J]. 计算机工程,2017,43(10): 283-288.

[14] 尚允坤,段锁林,潘礼正. 基于 ERS/ERD 的二级共空间模式的运动想象脑电信号特征提取[J]. 计算机工程与科学,2017,39(7):1385-1390.

[15] 吴林彦,鲁昊,高诺,等. 基于 CSP 算法与小波包分析方法的运动想象脑电信号特征提取性能的比较[J]. 生物医学工程研究,2017,36(3):224-228.

[16] SAMEK W,KAWANABE M,MULLER K R. Divergence-based framework for common spatial patterns algorithms[J]. IEEE Reviews in Biomedical Engineering,2014,7:50-72.

[17] LIU Y,SOURINA O,NGUYEN M K. Real-time EEG-based emotion recognition and its applications[M]// GAVRILOVA M L, KENNETH C J. Transactions on Computational Science XII. Berlin, Germany: Springer, 2011:256-277.

编辑　陆燕菲

(上接第 295 页)

[20] GHANNAY S, BARRAULT L. Using hypothesis selection based features for confusion network MT system combination [C]//Proceedings of the 3rd Workshop on Hybrid Approaches to Translation. Stroudsburg, USA: Association for Computational Linguistics,2014:1-5.

[21] RIKTERS M, SKADINA I. Combining machine translated sentence chunks from multiple MT systems[C]//Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics. Berlin, Germany: Springer, 2016: 27-37.

[22] YIMAM S M, ALONSO H M, RIEDL M, et al. Learning paraphrasing for multiword expressions[C]// Proceedings of the 12th Workshop on Multiword Expressions. Stroudsburg, USA: Association for Computational Linguistics,2016:1-10.

[23] PAVLICK E, RASTOGI P, GANITKEVITCH J, et al. PPDB2.0:better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 425-430.

[24] DENKOWSKI M, LAVIE A. Meteor universal: language specific translation evaluation for any target language[C]// Proceedings of the 9th Workshop on Statistical Machine Translation. Stroudsburg,USA:Association for Computational Linguistics,2014:376-380.

[25] KOEHN P,HOANG H,BIRCH A,et al. Moses:open source toolkit for statistical machine translation[C]//Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Stroudsburg, USA: Association for Computational Linguistics,2007:177-180.

[26] TURSUN E, GANGULY D, OSMAN T, et al. A semisupervised tag-transition-based markovian model for uyghur morphology analysis[J]. ACM Transactions on Asian and Low-Resource Language Information Processing,2016, 16(2).

[27] STOLCKE A. SRILM——an extensible language modeling toolkit[C]//Proceedings of International Conference on Spoken Language Processing. Stroudsburg,USA:Association for Computational Linguistics,2002:901-904.

[28] OCH F J. Giza++:training of statistical translation models[EB/OL]. [2017-12-25]. http://www. fjoch. com/GIZA++. html.

[29] PAPINENI K,ROUKOS S,WARD T,et al. BLEU:a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg,USA:Association for Computational Linguistics,2002:311-318.

编辑　赵　辉