基于统计机器翻译的知识库多语言自动扩展

李晓倩，曹海龙，赵铁军

（哈尔滨工业大学 机器智能与翻译实验室，黑龙江省 哈尔滨市 150001）

摘要：本文的研究任务为知识库的多语言自动扩展，并使用统计机器翻译的技术将知识库中实体标签翻译到另一种语言。为了解决知识库中特定词汇较多及特定领域双语平行语料缺少的问题，本文首先利用知识库中实体的源语言标签挖掘互联网中的双语例句。再次为了解决人名实体中未登录词的问题以提高知识库中人名实体翻译的正确性，本文添加了音译特征来翻译人名实体中的未登录词。此外，为了充分利用知识库中实体的属性信息，本文利用原知识库中人名的性别属性，帮助提高人名实体中未登录词的问题。本文使用电影领域知识库，并实现知识库实体标签从英文到中文的多语言自动扩展。实验表明相比基线系统本方法在 BLUE-2 及 BLUE 上分别提高了 1.3 和 0.9。

关键词：知识库；统计机器翻译；音译

Automatic Multilingual Expansion with SMT in Knowledge Base

Xiaoqian Li, Hailong Cao, Tiejun Zhao

(Machine Intelligence & Translation Lab, Harbin Institute of Technology,
Harbin, Heilongjiang 150001, China)

Abstract: Our research focuses on the automatic multilingual expansion of knowledge base. We present a statistical machine translation approach to translate the label of entity from source language to target language. Due to the fact that the knowledge base contains highly specific vocabulary and the lack of the parallel corpus in specific domain, we propose an approach to extract the parallel corpus from Internet with the source-side label of entity. And, we introduce a transliteration feature to translate the out-of-vocabulary (OOV) of name entity in knowledge base. Otherwise, in order to utilize the attribute of entity in knowledge, we additionally add a genre feature to improve the translation of OOV of name entity. We evaluate our approach on translation of a knowledge base in film domain. Our experiment shows an improvement overall 1.3 BLEU-2 and about 0.9 BLEU over our baseline.

Key words: Knowledge Base; Statistical Machine Translation (SMT); Transliteration

1 引言

知识库是是当前学术界和企业界的研究热点，其在自然界的应用也非常广泛。知识库在语义搜索、机器问答等实际的应用中有非常重要的意义。然而现有知识库大多以英文表示，Gracia[1]指出在 Watson 中有 80%的实体的标签是使用英文。在大型知识库 Freebase 中有 23M 实体，但是其中只有 1%的实体含有中文标签[2]。其他语种知识库的缺乏，大大限制了知识库在非英语自然语言上的应用，如中文等。因此知识库的多语言扩展是自然语言处理领域亟待解决的问题。

知识库的多语言扩展目的是实现知识库实体标签从源语言到目标语言的扩展。目前知识库的多语言扩展一般使用人工的方式，该方式不仅耗费大量的人力物力，而且需要大量的时间。因此如何实现知识库的多语言自动扩展成为许多科学家的研究重点。

本文提出一种使用统计机器翻译的方法对知识库进行自动扩展。此方法利用双语语料，并根据实体的相关属性，如类型、性别等，通过翻译实体的源语言端标签翻译得到目的语言标签。一般双语语料中含有大量的常有词汇，然而知识库中的标签多为特定词汇[3]，这些词汇或者短语往往稀缺，在通用双语语料中很少出现，如 leatherheads[1]、flamingos（火烈鸟）等。因此构建包含源语言端词汇的双语语料就显得格外重要。此外，在机器翻译中未登录词问题往往来自命名实体，而电影领域知识库中含有大量的人名

[1] 电影名《爱情达阵》，又名《傻瓜》

实体，因此解决人名实体的未登录(OOV, out-of-vocabulary)问题非常重要，且根据知识库中实体的类别标签其人名实体的性别标签，可以有效帮助解决人名实体 OOV 问题。

国内外学者针对知识库的多语言自动扩展做了非常多的研究。在 2008 年，Suarez-Figueroa 等[2]提出本体库的本土化概念（ontology localization），并给出了本体库的本土化的定义，“the adaptation of an ontology to a particular language and culture”，并且提出概念——本体库的翻译（ontology translation），“the activity of changing the representation formalism or language of an ontology from one to another”。Cimiano 等[5]提出本体库的本体化分为两层，词汇层及概念层，其认为本土化过程是以某本体库为输入，生成另一个本体库，该本题库可能可以是一个新的知识库，或者在原知识库上对其实体标签添加了另一种语言表述形式的多语言知识库。Espinoza 等[6]提出了系统 LabelTranslator，该系统通过自动获取翻译候选并对翻译候选进行排序，最终得到实体标签在另一种语言下的表述。然而这种使用目标语言词典去代替源语言实体标签的方式虽然有很高的准确率但是召回率明显偏低。Li 等[2]旨在翻译电影领域知识库，其方法首先选取了大量翻译候选集，包括统计机器翻译模型翻译结果，维基百科，百度百科等，然后利用原知识库构建翻译图，对各翻译候选进行打分，最终得到翻译结果。以上这些方法的基本思想都是构建翻译候选集，并从中选取最优候选翻译。

机器翻译[7]技术是一种将句子从源语言自动翻译为目标语言的技术方法，一般用于对自然语句的自动翻译。许多研究者使用机器翻译方法来实现知识库的多语言自动扩展。Fu 等人[8]提出方法 CLOM，该方法使用免费可得的机器翻译工具将知识库中的实体标签翻译为目标语言。Espinoza 等[9]总结了在翻译时可能遇到的问题，并对知识库标签进行了歧义消除。Arcan 等人[10]研究翻译特定领域的知识库，特定领域的知识库的实体标签往往含有特殊词汇，且由于实体标签较短，往往还有很少的上下文信息，为此提出一种特定领域的知识库翻译方法，该方法从大规模性双语语料中选取领域内语料，并使用已有大规模知识库对该领域知识库的翻译进行补充。本文的方法不同于该方法，根据知识库中的实体源语言端标签从互联网中爬取相关双语语料。

本研究针对电影领域知识图谱，并实现实体的标签从英文到中文的翻译。本文随机从该知识库中分别选取了 1000 个实体作为开发集和测试集，并标注了其中文标签。

不同于以上方法，本文考虑到特定领域的双语语料的缺乏，首先根据知识库中实体的源语言标签，从网页中抽取了大量相关双语语料，考虑到电影领域有一定量的人名，而双语语料中人名的缺乏导致人名实体翻译错误，提出基于音译特征的知识库多语言自动扩展方法此外为了充分利用实体的属性特征，如性别，本文在上述方法的基础上引入了性别特征。

本文的组织结构如下：第二部分介绍本研究的方法，第三部分介绍实验及结果分析，第四部分介绍结论及未来工作。

2 基于机器翻译的知识库多语言扩展

2．1 基于实体标签的平行语料获取

为了提高机器翻译的性能，往往需要双语平行语料的规模较大且质量较好。双语平行语料是机器翻译任务的前提，双语平行语料的好坏将直接影响机器翻译任务的效果。而随着互联网的快速发展，网络数据也随之增多，其中存在了大量的双语语料[11],这使得许多研究者从事于从互联网上获取双语平行语料。

和一般机器翻译相同，语义知识库翻译往往对双语平行语料的质量和规模有非常大的依赖性。如通用的英汉双语语料的规模就非常大，且语料质量较好。然而这些大规模通用语料虽然含有大量常见词汇，但对特定领域的词汇，如 flamingos 等很可能为未登录词。然而对于特定领域的知识库，其领域内双语语料规模往往较少，对齐质量也相对较差。因此构建知识库翻译的双语语料非常重要。

考虑到不同于一般的机器翻译系统，待翻译的双语语料是完全未知的，也即目标语言句子和源语言句子均未知。而在此任务中知识库中实体的源语言标签是已知的，且考虑到特定领域知识库中常常含有一些特殊的词汇，因此可以根据知识库中实体的源语言标签集去网上爬取相关双语语料。

本文对包含有双语例句的网页进行了爬取，尤其是有道翻译[2]这种提供翻译服务的网页。其中收集了来自现代汉英综合大词典、柯林斯例句及其他网页的爬取到的双语例句。

本文利用 python 编写了网页爬取工具，并使用 python 的 BeautifulSoup 工具包对爬取的网页进行解析，分析出其中的双语相关例句，其存储格式如表 1 所示。从表 1 中可以看出针对实体标签“Battlestar Galactica”，可以爬取到其相关双语语句。

表 1 双语例句示例

No.	句子
1	He's the star of "**Battlestar Galactica.**" 他是《太空堡垒卡拉狄加》的主演.
2	How do Tyrol and his deck crew honor Adama on the day of the **Galactica**'s decommissioning? 在卡拉狄加计划退役的那天，泰罗尔和他甲板的队员如何对阿达玛表示敬意?
3	I can see the influence of **Battlestar Galactica** and Gene Roddenbery's Andromeda. 我能在书中看到《太空堡垒卡拉狄加》和《星舰复国记》的影响.

2．2 基于音译特征的知识库未登录词翻译

和一般的机器翻译任务中常常遇到未登录词问题相同，在知识库翻译过程中也会遇到未登录词的问题。且由于特定领域知识库中的实体标签往往含有特定词汇或短语，因此在知识库翻译中未登录词的问题更加严重。在知识库中往往含有大量的人名实体，而人名实体是未登录词中比较重要的一部分。因此利用知识库中的实体类别属性，对人名实体中出现未登录词的问题，本文引入音译特征来解决，以期提高特定领域知识库翻译的准确性。

在机器翻译过程中，除按短语表创建翻译候选集外，对人名实体源语言标签中的未登录词，利用音译模型生成候选翻译。一般来说人名音译是指利用源语言及目标语言发音规则的异同将人名从源语言形式翻译成目标语言形式。如人名 Isabel，翻译为中文为“伊莎贝尔”。Sherif 等人[12]受到机器翻译模型的启发，使用基于短语的机器翻译模型训练音译模型，并取得了不错的实验效果。因此本文采用基于短语的统计机器翻译实现英汉人名音译系统。

本方法主要分为如下几步：1、生成人名对齐词典 2、训练人名音译模型 3、在机器翻译系统中添加音译特征。

1 生成人名对齐词典

为了训练人名音译模型，需要构建人名对齐词典。本文首先从 Freebase 中抽取含有英种实体标签的人名实体得到双语人名语料，格式如表 2 所示。

表 2 英中双语人名示例

英文名	中文名
Lan Tianye	蓝天野
Andrej Kiska	安德烈·基斯卡
Nikolai Gogol	**尼古拉·瓦西里耶维奇·果戈里**
Alfred Brendel	阿尔弗雷德·布伦德尔

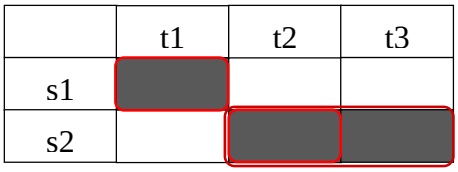
为了抽取其中的人名对齐词典，首先使用对齐算法对双语人名语料进行对齐，在 Och[13]短语对抽取方法基础上做了一定调整，本实验只考虑正向短语抽取，并且只抽取短语中正向短语长度为 1 的短语。抽取到的一个短语对可以用一个四元组表示

$$H(i,l_s,j,l_t) \qquad (1)$$

来表示。其中，i、j 分别表示源语言与目标语言中词的起始位置，而 l_s、l_j 分别表示源语言与目标语言中等价单位的长度。如图 1 所示，在该图中共抽取到 3 个短语，分别为短语 $H(1,1,1,$ 及短语 $H(2,2,1,$ 及 $H(2,2,2,1)$。

[2] http://dict.youdao.com/

图 1 对齐图中的短语示例



通过该方法我们可以得到人名对齐词典，如表 3 所示。通过该表可以看出对于英文人名 Abbott，有多种音译方式，如亚伯特及阿伯特等。

表 3 人名对齐词典

英文	中文
abbott	亚伯特
abbott	阿伯特
abbott	阿博特
oumarou	乌尔德

2 训练音译模型

本文采用基于短语的统计机器翻译实现英汉人名音译,主要思路为:首先,对人名对齐词典进行分词等预处理，并选取一部分数据作为测试集，一部分数据作为开发集；使用统计机器翻译模型对测试集进行训练得到音译模型;最后,采用开发集对该音译模型进行调参。

本文使用开源工具 Moses[14]来训练音译模型。在数据预处理时，本实验根据音节划分的方式的不同，采用两种方式划分：一种是以字母或者音节作为基本单元；对目的端语料划分时，直接将每个中文词语作为基本单元。

对于以音节作为音译单位的音译模型，定义其音节划分的规则[15,16,17]，首先对音节字母定义如下：

定义 a,o,u,e,i 为元音；

定义 y 为元音，如果其跟在非元音字母之后；

定义 m,n 为鼻音;

定义其他所有字符为辅音；

给出音节的定义后根据英语的翻译规则，定义音节划分规则，如表 4 所示。

本文音译模型训练时，考虑的特征如下：短语的正向翻译概率 $p(e|f)$；短语的反向翻译概率 $p(f|e)$；正向词权重 $lex(e|f)$；反向词权重 $lex(f|e)$；目标语言的语言模型；短语长度惩罚特征等。

表 4 音节划分规则

No.	英文中字符连续出现的情况	音节划分初始方式
1	连续的辅音	划分为多个独立的辅音
2	连续的元音	合并作为组合元音
3	辅音+元音	(辅音+元音)
4	任何独立的元音/辅音	作为独立的音节
5	元音+鼻音+元音	元音+(鼻音+元音)
6	元音+鼻音+辅音/结束符	(元音+鼻音)+辅音/结束符
7	c/s/z/t/p/w+h	(c/s/z/t/p/w+h)合并为辅音
8	元音+r+元音	元音+(r+元音)
9	元音+r+辅音/结束符	(元音+r)+辅音/结束符

利用从人名对齐词典中随机抽取 1000 个词作为开发集对其进行参数调整。参数调整采用 MERT[18](最小化错误率)。解码时采用束搜索算法[19]。

3 在机器翻译系统中添加音译特征

此方法主要针对的是在知识库翻译中的未登录词为人名实体的问题。此方法的思想是在机器翻译过程中如果发现一个未登录词，则调用音译模型，使用该模型生成翻译结果补充候选翻译，并添加音译特征。

形式化表示如下：给定一个未登录词 f，使用音译模型得到最优候选翻译 $\tilde{e}$，

$$\tilde{e} = \arg\max_{e} P(e|f) \qquad (2)$$

其音译特征值记为 $\tilde{p}$,如公式 2 所示。

$$\tilde{p} = P(\tilde{e}|f) \qquad (3)$$

2．3 性别特征

早在 2007 年，Li 等人[20]在训练模型将语料按照人名的语言来源及性别分别进行训练，然后将人名的语言来源、性别等语义

特征融合到音译系统,大大改善了音译系统的性能。如音节“na”,当在女性名称中时，常常翻译为“娜”；当在男性名称中，常常翻译为“纳”。

考虑到在知识库中有大量的实体属性信息，充分利用这些属性信息非常重要，如对于人名实体，一般含有性别属性及国籍或出生地等属性，利用实体的属性信息可以提高音译模型的准确性。因此本文利用知识图谱中的人名实体的性别特征，来提高人名未登录词的翻译效果。该方法在上节所述方法的基础上，添加了音译特征，对不同性别的人名分别训练音译模型，且根据未登录词的性别来选择对应的音译模型。此外训练一个通用模型，如果人名实体未标注其性别属性时，调用该通用模型。

3 实验

3.1 实验数据

1 训练数据

使用 2.1 节所述的方法，构建训练集。如表 4 所示训练集平行语料中句对数大约 1M，其中源语言端词典个数为 69k，目标语言端词典个数为 99k。

表 4 实验数据

		英语	中文
训练集	句子数	106373	
	词典数	68229	99274
开发集	句子数	1000	1000*2
	单词数	2406	5009
	词典数	1753	2066
测试集	句子数	1000	1000*2
	单词数	2419	5542
	词典数	1699	1974

2. 开发集及测试集

在一般句子级别的机器翻译系统中，为了能正确评估机器翻译模型的准确性，往往添加多个翻译候选。和句子级别机器翻译模型相同，一个实体可能有多个名称。在人名实体中该现象尤为明显，往往因为音译方式等的不同，使得实体有多个名称，如 Judy Holliday 好莱坞著名女演员，其中文名称常有茱蒂·霍利德及朱迪·霍利德。在电影实体中，对于外国电影常常采用根据片名直译、意译及音译，如著名科幻动画电影 WALL-E，中文常见译名为机器人总动员，还有一种常见音译名称为瓦力。

因此本工作从电影领域知识库中分别随机选取了 1000 个实体作为开发集合测试集，并对其标注翻译候选集。人工从百度百科维基等网站搜索实体对应的中文名称，并构建了两个翻译候选集。如表 4 所示，从开发集中实体的源语言标签的平均长度为 2.41，目标语言标签的平均长度为 2.50，在测试集中实体的源语言标签的平均长度为 2.42，实体的目标语言标签的平均长度为 2.77。且开发集和测试集中实体源语言标签中分别平均每 1.37 或 1.42 个单词就有一个新单词，可以看出知识库中特定词汇较多，而常用词汇非常少。

3 Freebase 英中双语人名

知识库 Freebase[3]中有部分人名实体既有英文标签又有中文标签，这些实体个数为 61324，其中人名实体为女性的为 9928 个，为男性的为 45711，没有标记性别属性的为 5685。

表 5 英中人名实体

人名实体	英中双语人名实体对个数
all	61324
female	9928
male	45711
Null	5685

3．2 实验设置

本研究使用机器翻译工具包 Moses[14]，词对齐使用使用开源工具 GIZA++[21]并采用 grow-diag-final-and 启发式策略产生双语词对齐。此外我们使用 SRILM[22]语言模型训练四元语言模型，并使用改进 Kneser-key 平滑算法[23]。对于目标语言使用 Stanford 分词工具进行分词。采用 MERT[17]对参数进行调参。本文使用 BLUE 对实验结果进行评估。

3.3 实验结果及分析

为了验证本文提出方法的有效性，本文设计了四组系统。

实验 1：使用 3.1 所述的数据，并采用

[3] https://en.wikipedia.org/wiki/Freebase

最基本的基于短语的机器翻译模型。

实验 2：在实验 1 的基础上，添加音译特征，未登录词按字母进行划分。

实验 3：在实验 1 的基础上，添加音译特征，未登录词按音节进行划分。

实验 4：在实验 3 的基础上，添加性别特征，对于不同性别的未登录词调用不同的音译模型。

表 6 实验结果

实验	BLEU	BLEU-2
实验 1	24.45	35.1
实验 2	24.96	35.9
实验 3	25.14	36.2
实验 4	25.31	36.4

如表 6 所示，通过实验 2,3,4 和实验室 1 的对比，说明本文提出的基于音译特征的多语言知识库扩展方法可以有效的提高知识库翻译系统的正确性。由于知识库中实体的在开发集及测试集上目标语言标签平均长度为大约为 2.50 或 2.77，因此对 BLEU-2 的提高也非常有说明性，最终本方法，整体 BLEU 值取得了 0.86 的提升，在 BLUE-2 上有 1.3 的提高。

4 结论及未来工作

本实验针对特定领域双语语料不足的及特定领域知识库中特定词汇或短语较多的问题，根据知识库中实体的源语言标签爬取双语对齐语料，在一定程度上提高了领域内双语语料的规模。为了解决在知识库翻译过程的未登录词问题，本文充分利用知识库中的类型属性及性别属性，添加音译特征，对人名未登录词使用音译模型构建起翻译候选。本文的方法不仅适用于电影领域的知识库多语言扩展，同样适用于其他领域，而不限于扩展语言为电影领域。

然而通过实验发现本文获取的双语语料的规模仍然较小，下一步工作将考虑从大规模通用语料中抽取语料来对双语语料进行扩展。此外在知识库翻译中仍有很多属性信息可以使用。如出生地，根据其出生地可以大约判定其人名来源，从而提高知识库翻译性能。此外，研究中发现，由于知识库中的实体在源语言上可能有多个名称，下一步将考虑利用此信息，提高机器翻译的性能。

参考文献

[1] Gracia J, Montiel-Ponsoda E, Cimiano P, et al. Challenges for the multilingual web of data[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2012, 11: 63-71.

[2] Li Q, Liu S, Lin R, et al. Entity Translation with Collective Inference in Knowledge Graph[M]//Natural Language Processing and Chinese Computing. Springer International Publishing, 2015: 49-63.

[3] Chandrasekaran B, Josephson J R, Benjamins V R. What are ontologies, and why do we need them?[J]. IEEE Intelligent systems, 1999 (1): 20-26.

[4] Suarez-Figueroa M C, Gómez-Pérez A. First attempt towards a standard glossary of ontology engineering terminology[J]. 2008

[5] Cimiano P, Montiel-Ponsoda E, Buitelaar P, et al. A note on ontology localization[J]. Applied Ontology, 2010, 5(2): 127-137.

[6] Espinoza M, Gómez-Pérez A, Mena E. Enriching an ontology with multilingual information[M]. Springer Berlin Heidelberg, 2008.

[7] Koehn P. Statistical machine translation[M]. Cambridge University Press, 2009.

[8] Fu B, Brennan R, O'Sullivan D. Cross-lingual ontology mapping–an investigation of the impact of machine translation[M]//The Semantic Web. Springer Berlin Heidelberg, 2009: 1-15.

[9] Mejía M E, Montiel-Ponsoda E, de Cea G A, et al. Ontology localization[M]//Ontology Engineering in a Networked World. Springer Berlin Heidelberg, 2012: 171-191.

[10] Arcan M, Turchi M, Buitelaar P. Knowledge portability with semantic expansion of ontology labels[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. ACL, Beijing, China. 2015

[11] 刘飞. 双语平行句对的获取与语料库的建立[D]. 哈尔滨工业大学, 2013.

[12] Sherif T, Kondrak G. Substring-based transliteration[C]//Annual Meeting-Association for Computational Linguistics. 2007, 45(1): 944.

[13] Och F J, Ney H. The alignment template approach to statistical machine translation[J]. Computational

linguistics, 2004, 30(4): 417-449.

[14] 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. Association for Computational Linguistics, 2007: 177-180.

[15] 王丹丹. 英汉人名音译的研究[D]. 大连理工大学, 2014.

[16] Li L, Wang P, Huang D, et al. Mining english-chinese named entity pairs from comparable corpora[J]. ACM Transactions on Asian Language Information Processing (TALIP), 2011, 10(4): 19.

[17] Zhang C, Li T, Zhao T. Syllable-based machine transliteration with extra phrase features[C]//Proceedings of the 4th Named Entity Workshop. Association for Computational Linguistics, 2012: 52-56.

[18] Och F J. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 160-167.

[19] Koehn P. Pharaoh: a beam search decoder for phrase-based statistical machine translation models[M]//Machine translation: From real users to research. Springer Berlin Heidelberg, 2004: 115-124.

[20] Li H, Sim K C, Kuo J S, et al. Semantic transliteration of personal names[C]//Annual Meeting-Association for Computational Linguistics. 2007, 45(1): 120.

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

[22] Stolcke A. SRILM-an extensible language modeling toolkit[C]//INTERSPEECH. 2002, 2002: 2002.

[23] Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling[C]//Proceedings of the 34th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1996: 310-318.