Modern Linguistics 现代语言学, 2021, 9(4), 984-989
Published Online August 2021 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2021.94133

Hans汉斯

汉语－土耳其语句对齐自动校验方法研究

张贵林，易绵竹，李宏欣，陈靖博

信息工程大学洛阳校区，河南 洛阳
Email: lihongxin830@163.com

收稿日期：2021年6月16日；录用日期：2021年7月28日；发布日期：2021年8月4日

摘　要

通过互联网获取的句对齐平行语料常存在对齐错位或译文质量差的问题，针对这一问题，本文提出了一种基于反向翻译的汉语－土耳其语平行语料自动校验方法。该方法通过在线机器翻译系统获取反向翻译结果，并将译文作为中间语言构建词袋模型对句子相似度进行向量化表示，最后通过机器学习训练二分类模型的方法来判断句子是否对齐。实验结果显示，以汉语或土耳其语为中间语言时，系统能够获得较好的句对齐检验效果。

关键词

汉语－土耳其语，机器翻译，双语语料库，句子对齐

Research on Automatic Back-Translation Based Verification Method of Chinese-Turkish Sentence

Guilin Zhang, Mianzhu Yi, Hongxin Li, Jingbo Chen

Luoyang Campus of Information Engineering University, Luoyang Henan
Email: lihongxin830@163.com

Received: Jun. 16th, 2021; accepted: Jul. 28th, 2021; published: Aug. 4th, 2021

Abstract

The problem of misalignment or poor translation quality often exists in sentence-aligned parallel corpus obtained from the Internet. To solve this problem, the paper proposes an automatic back-translation based verification method for Chinese-Turkish parallel corpus. In this method, the back-translation results are obtained by online machine translation system, and the target

文章引用: 张贵林, 易绵竹, 李宏欣, 陈靖博. 汉语－土耳其语句对齐自动校验方法研究[J]. 现代语言学, 2021, 9(4): 984-989. DOI: 10.12677/ml.2021.94133

language is used as the intermediate language to construct a bag-of-words model to realize the vectorized representation of sentence similarity. With these vector values, a binary classification model is trained by machine learning to judge whether the sentences are properly aligned or not. The experimental results show that the system can achieve better sentence alignment verification results when Chinese or Turkish is used as the intermediate language.

Keywords

Chinese-Turkish, Machine Translation, Bilingual Corpus, Sentence Alignment

1. 引言

双语句对齐平行语料库是搭建统计和神经机器翻译系统的基本部件，相对于英语－汉语而言，在针对土耳其这种低资源的小语种构建机器翻译系统时，一般很难直接获得足够的大规模、高质量平行双语句对，因而，基于互联网海量数据进行多语言平行语料挖掘已经成为获取双语平行句对的一个重要途径。

通过互联网直接获取的原始数据大多会存在噪声大、污染重等问题，这会给后续的机器翻译任务带来很多麻烦，因此，需要对这些数据进行清洗，以尽量保证句对齐双语平行语料的真实可用性。针对上述问题，本文提出了一种基于反向翻译的汉语－土耳其语平行语料自动校验方法，通过谷歌翻译获取双语伪译文、中文分词、土耳其语词形还原、词袋模型取交集及占比关系计算和相似度判断等多个步骤，检验原有土耳其语－汉语平行语料句子对齐的质量。

文章第一部分简要介绍了句对齐检验的常见方法；第二部分详细阐述了实验系统的整体框架和具体实现过程；第三部分在实验数据上验证了本文所提方法的有效性，并对四种不同情况下句对齐检验效果进行了对比分析；第四部分对全文进行总结，并对今后的研究方向进行了展望。

2. 研究方法概述

句子对齐检验是提高双语平行句对齐语料库的一种重要手段，常见检验方法大致可分成三类：基于句子相对长度、基于词汇共现和基于混合策略的检验方法[1]。基于句子相对长度的句对齐检验方法是一种较为简单的规则方法，基本依据是源语言原文与目标语言译文的长度通常会满足一定的比例关系，通过设定译文长度阈值可筛除部分过译、略译和少译的双语平行句对[2]。该方法的优点是简单有效，缺点是未考虑词汇层因素，不能有效检验译文错误或译文质量较差的句对。基于词汇共现的方法主要思想是将句子视为一组词，句子相似度的计算只依赖于单词相似度，通过双语词典和同义词典来统计句子中同现的单词数，可有效检验双语句对的语义相似程度[3] [4]。基于词的句对齐检验方法虽然清晰、易于实现，但忽略了词序、句法和上下文等句子结构信息。混合策略的句对齐检验多为上述两种方法的结合，一些方法会增加句法、词序和局部结构信息，考虑单词重叠相似性和句法依赖的多个特征，通过神经网络学习句子特征的向量分布，可实现句子对齐的有效检验[5] [6] [7] [8]。该方法能够最大化地融合不同方法的优点，但当约束条件较多时，容易出现检验筛除过度的情况。

3. 基于反向翻译的句对齐校验方法

基于反向翻译的双语平行句对齐校验是一种基于词汇同现的句对齐检验方法，由于反向翻译译文的

词序、句法和上下文结构均不能确保绝对的准确性，即机器翻译的语句多有错误，因而，该方法主要通过伪译文与原译文间词汇语义相似度的判断来实现句对齐检验。我们首先利用谷歌在线机翻系统爬取待检验双语句对的对照伪译文，再使用汉语分词和土耳其语词形还原工具分别完成汉语分词和土耳其语词形还原，通过停止词筛选后，制作双语句对的词袋模型，分别对相应语言的词袋模型取交集，并分别计算出交集在各自对应词袋模型中所占的比例。最终，根据这两两间比例关系再进行二分类，判断上述句子之间的相似度，从而检验土汉双语平行句对是否对齐。系统整体框架如图 1 所示：

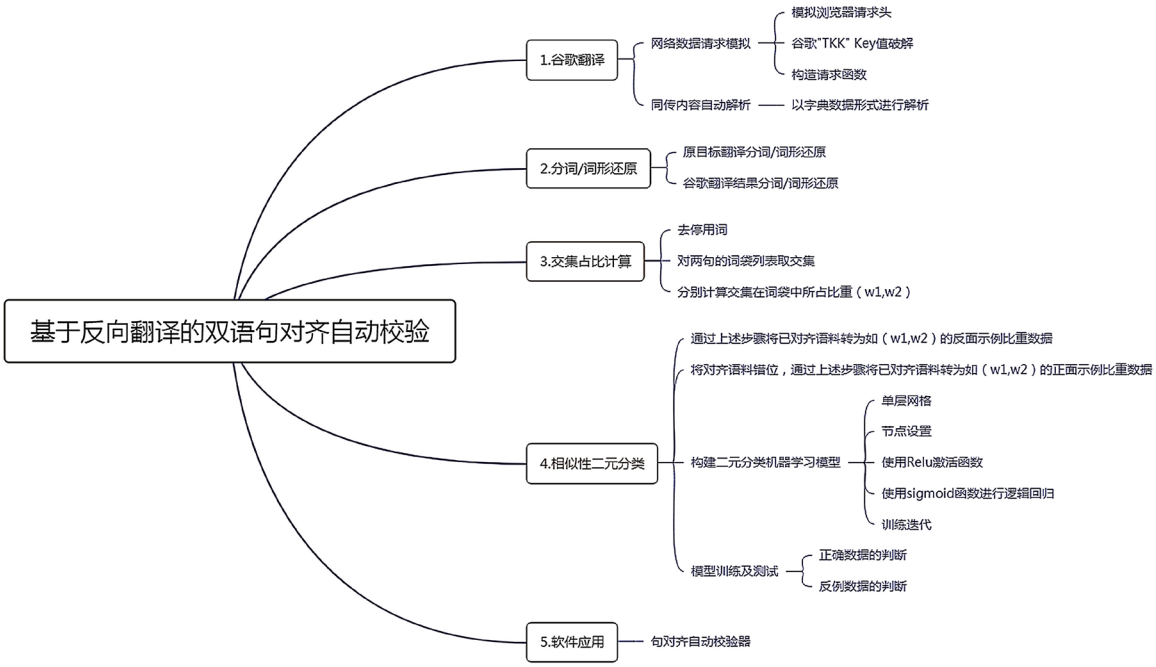

Figure 1. Overall framework of automatic verification system
图 1. 系统整体框架图

3.1. 谷歌翻译的获取

谷歌翻译获取部分，我们编写爬虫模拟浏览器对请求并爬取下谷歌翻译反馈的网页中翻译部分的结果。

具体过程为：首先模仿浏览器构造 https 请求头，然后对 data 部分进行构建。谷歌翻译网络请求的 data 部分关键性参数主要有四个，“sl”“tl”“tk”以及“q”。其中“sl”和“tl”分别为原始语言和翻译语言类别，tk 是一个验证 key 值。q 为我们上传的翻译语句。我们通过 python 调用 js 代码从模拟浏览器运行，根据 q 及 TKK 值计算出 tk 值，完成整个请求数据构造。返回的数据为 json 形式的字符串数据，我们先将其转换为 json 型数据，再提取其中的谷歌翻译反馈。

3.2. 汉语分词和土耳其语词形还原

汉语分词部分，我们采用 python 第三方开发库 jieba 分词，分别对原始汉语语句和谷歌翻译伪汉语语句进行分词。之后，通过比照汉语停用词表将过于常见的词汇去除。这样就分别构建了原汉语语句和伪汉语语句的词袋模型。土耳其语词形还原部分，我们采用自建形态分析词典，分别对原始土耳其语语句和谷歌翻译伪土耳其语语句进行形态分析。之后，通过比照土耳其语停用词表，筛除过于常见的词根和形态分析序列，进而构建原土耳其语语句和伪土耳其语语句的词袋模型。

3.3. 基于词袋模型的相似度向量表示

基于词汇同现的相似度计算方法，亦称为单词重叠或“词袋”方法，常用于各类信息检索系统。该方法的主要思想是，给定两个句子(A,B)，通过计算两个句子中同时出现的单词个数来获得两个句子之间的相似度。这一技术依赖于这样一个假设，即更加相似的句子会共享更多相同的单词，但通常两个句子均具有足够长度时，这种方法才会更加有效[7]。

基于上述假设，本文提出采用同现词占比作为向量来表示句子间的相似度，句子相似度的具体计算方法为：通过对比原译文和谷歌翻译伪译文语句的词袋模型，找出两者共享的同现词，形成同现词列表，将同现词列表中的单词数与源语言语句词列表的单词数做比，计算得出w_1；将同现词列表中的单词数与谷歌翻译伪译文语句词列表的单词数做比，计算得出w_2。此时，(w_1,w_2)可作为代表原译文和谷歌翻译伪语句之间的相似度量化数值，用于语句关联度的计算。根据上述方法，句子相似度的向量形式可表示如下：

$$SIM(A,B)=(w_1,w_2)=\left(\frac{n\cap m}{n},\frac{n\cap m}{m}\right)$$

表达式中，m、n 分别表示句子 A、句子 B 中的词袋列表或筛除停止词之后的词袋列表，$n\cap m$ 表示句子A和句子B中共享的同现词，w_1、w_2分别为共享词在句子A、句子B中所占的比重。如果我们将(w_1,w_2)作为二维坐标平面上的点来看待，那么所有句子对所组成的向量值则会形成一个点群。如果这些句子都是对齐语料的句子对，那么这个点群就可以代表对齐的句子对的量化值所处的范围区，我们可以通过画一条曲线来圈出这个范围区，将其作为判断一个句子对是否对齐的依据。

3.4. 基于相似度的二分类

为了找出对齐语句相似度向量的分布规律，我们将现有对齐的双语平行语料做一个错位处理，这样可以获得对齐语句的反面案例。根据上述方法步骤，分别对正面的对齐平行语料和反面的非对齐平行语料进行处理，进而可以得到一组代表对齐案例的$(w_{11},w_{12})(w_{21},w_{22})(w_{31},w_{32})\cdots(w_{i1},w_{i2})$的数组列表和一组代表反面案例的$(w_{11},w_{12})(w_{21},w_{22})(w_{31},w_{32})\cdots(w_{j1},w_{j2})$的数组列表。如果我们将$(w_1,w_2)$视为二维坐标轴上的点，那么就可以得到对齐案例与反面案例数据在几何空间上的分布，如下图 2 所示：

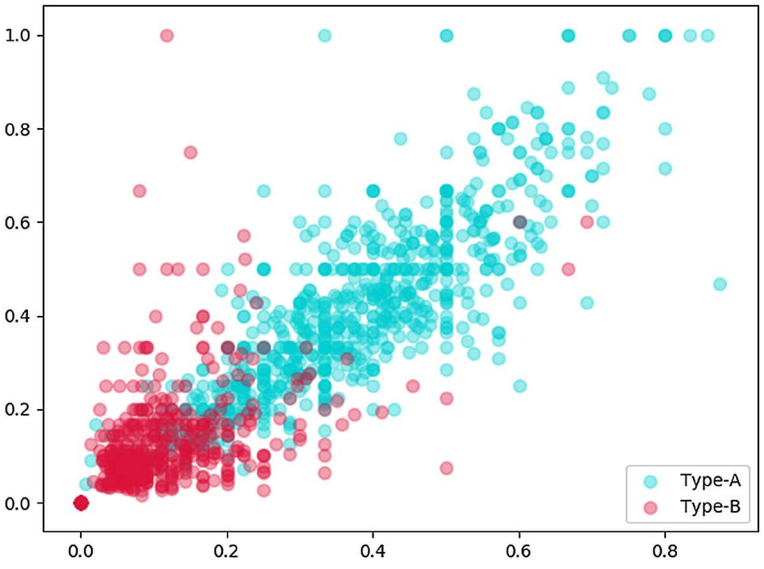

Figure 2. Geometric distribution of aligned and non-aligned statements
图 2. 对齐语句和非对齐语句的几何分布

上图显示的是测试用土汉双语句对齐语料正面和反面案例的几何分布，其中红色为反面案例，蓝色为对齐案例。可以看出，两种数据相对来说泾渭分明。因此我们可以使用数学手段二分类器对其进行分类。

我们的二分类器通过机器学习训练获得。模型构建部分我们使用单层神经网络进行训练，设置 50 个节点，使用 relu 激活函数和 sigmoid 逻辑回归函数。这样我们可以将训练数据映射到平滑空间并且使返回数据映射到(0,1]的值域内，以便对预测可能性进行判断。整个模型训练过程我们进行了 1000 次迭代，最终获得的模型测试效果较好，在以汉语为中间语言时，对正确数据的判断准确率为 89.7%，对反例数据的判断准确率为 91.4%，如图 3 所示。

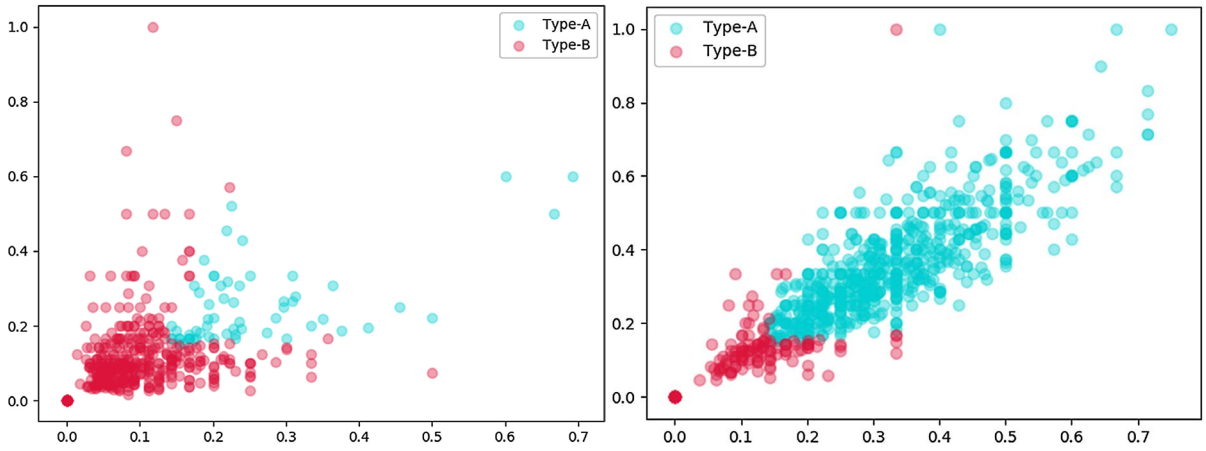

Figure 3. Similarity judgment results of 1000 positive and negative instances
图 3. 1000 个正反实例的相似度判断结果

4. 实验和结果分析

实验所用的语料为本文通过互联网直接获取的土耳其语－汉语句对齐平行语料，选取的句子总条数为 1000，汉语停用词个数为 507，土耳其语停用词个数为 706。实验中，分别以土耳其语和汉语伪译文作为中间语言，对未使用停止词和使用停止词的正反实例相似度向量分布进行计算，实验结果如表 1 所示：

Table 1. Experimental results
表 1. 实验结果

类别	未筛除停止词 汉语	筛除停止词 汉语	未筛除停止词 土耳其语	筛除停止词 土耳其语
对齐	97.9%	89.7%	82.5%	86.2%
反例	58.6%	91.4%	94.3%	96.7%

可以看出，在以汉语为中间语言且不使用停用词时，因数据的离散度较高，导致反例数据的判断准确率仅为 58.6%，远小于对齐数据的 97.9%，不能有效地反映出句子的对齐程度；在以土耳其语为中间语言且不使用停用词时，反例数据的判断准确率达 94.3%，较汉语判断结果高出约 35 个百分点，对齐数据的判断准确率为 82.5%，较汉语判断结果下降约 15 个百分点，此时对齐和反例相似度判断结果相对泾渭分明，基本能够达到句对齐检验的效果。在筛除停止词之后，结果显示汉语数据的离散问题得到了很好的解决，对齐和反例数据的判断准确率均为 90%左右，显示出了很好的平衡性；土耳其语对齐和反例数据的判断准确率则分别提高了大约 3.7 个百分点和 2.4 个百分点。相比较而言，停止词筛除与否可直接决定汉语数据检验的有效性，对于土耳其语数据，筛除停止词则可有效提高句子对齐检验的准确性。采用

土耳其语或汉语作为中间语言进行句对齐检验，均可达到不错的效果。

5. 结论

针对构建机器翻译系统时对高质量、大规模土耳其语－汉语双语句对齐平行语料库的现实需求，本文结合句子相似度计算的相关技术，提出了一种基于反向翻译的土汉双语句对齐校验方法。在大小规模为1000句的互联网领域土汉双语句对齐平行语料中进行实验，结果显示，句子相似度向量分布可作为判断句子对齐与否的一种可靠依据，基于上述方法构建的句对齐检验系统，可自动筛除部分粗糙、低质量的双语平行语句，对土汉双语句对齐平行语料库的建设具有较强的实用意义。在实验中，停止词表的设置对检验系统整体性能的提升起到了关键作用，因汉语和土耳其语的句子结构相对比较复杂，如何优化停止词表以提高系统性能，是将来有待进一步研究的一项重要内容。

参考文献

[1] 路琦. 基于跨语言词向量的句子对齐方法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2020.

[2] Canhasi, E. (2013) Measuring the Sentence Level Similarity. *Advances in Architecture and Engineering*, **1**, 35-42

[3] 丁颖. 基于词对和词典的句子对齐研究[D]: [硕士学位论文]. 苏州: 苏州大学, 2019.

[4] Wali, W., Gargouri, B. and Hamadou, A.B. (2017) Sentence Similarity Computation Based on Word Net and Verb Net. *Computación y Sistemas*, **4**, 627-635. https://doi.org/10.13053/cys-21-4-2853

[5] 黄佳跃. 基于神经网络的句对齐研究及应用[D]: [硕士学位论文]. 苏州: 苏州大学, 2020.

[6] 李玉龙. 基于神经网络的句子相似度计算研究[D]: [硕士学位论文]. 湘潭: 湖南科技大学, 2020.

[7] 彭晓娅, 周栋. 跨语言词向量研究综述[J]. 中文信息学报, 2020, 34(2): 1-16.

[8] Deng, H., Zhu, X. and Li, Q. (2017) Sentence Similarity Calculation Based on Syntactic Structure and Modifier. *Mathematical Problems in Engineering*, **43**, 240-244.