Geekbang 极客邦科技 InfoQ

QCon 全球软件开发大会【北京站】2016

“跨越语言的鸿沟”
—电商系统中的多语言翻译技术

曾晓东 阿里巴巴

Intemational Software Development Conference

自我介绍

曾晓东

阿里巴巴集团 B2B技术部–翻译平台 技术专家

澳门大学计算机硕士，2014年加入阿里，担任联盟搜索翻译算法团队的技术专家，主要负责阿里机器翻译算法设计与优化，同时也负责多语言自然语言处理技术的构建。在加入阿里之前，曾担任澳门INESC-MACAU与澳门自然语言处理与葡中机器翻译实验室的助理研究员。有超过7年的自然语言处理、机器翻译研究经验，其多项研究成果发表在国际顶级会议与期刊中。

今天讲些什么

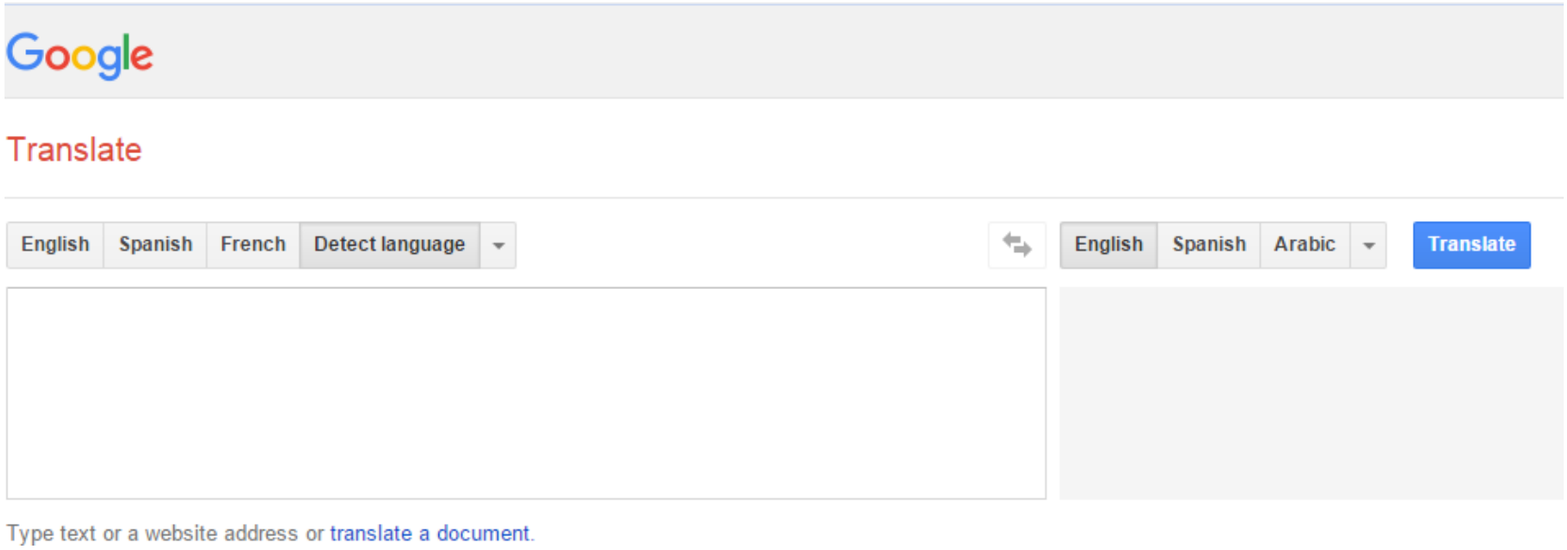
如何利用翻译技术帮助

电商网站的国际化

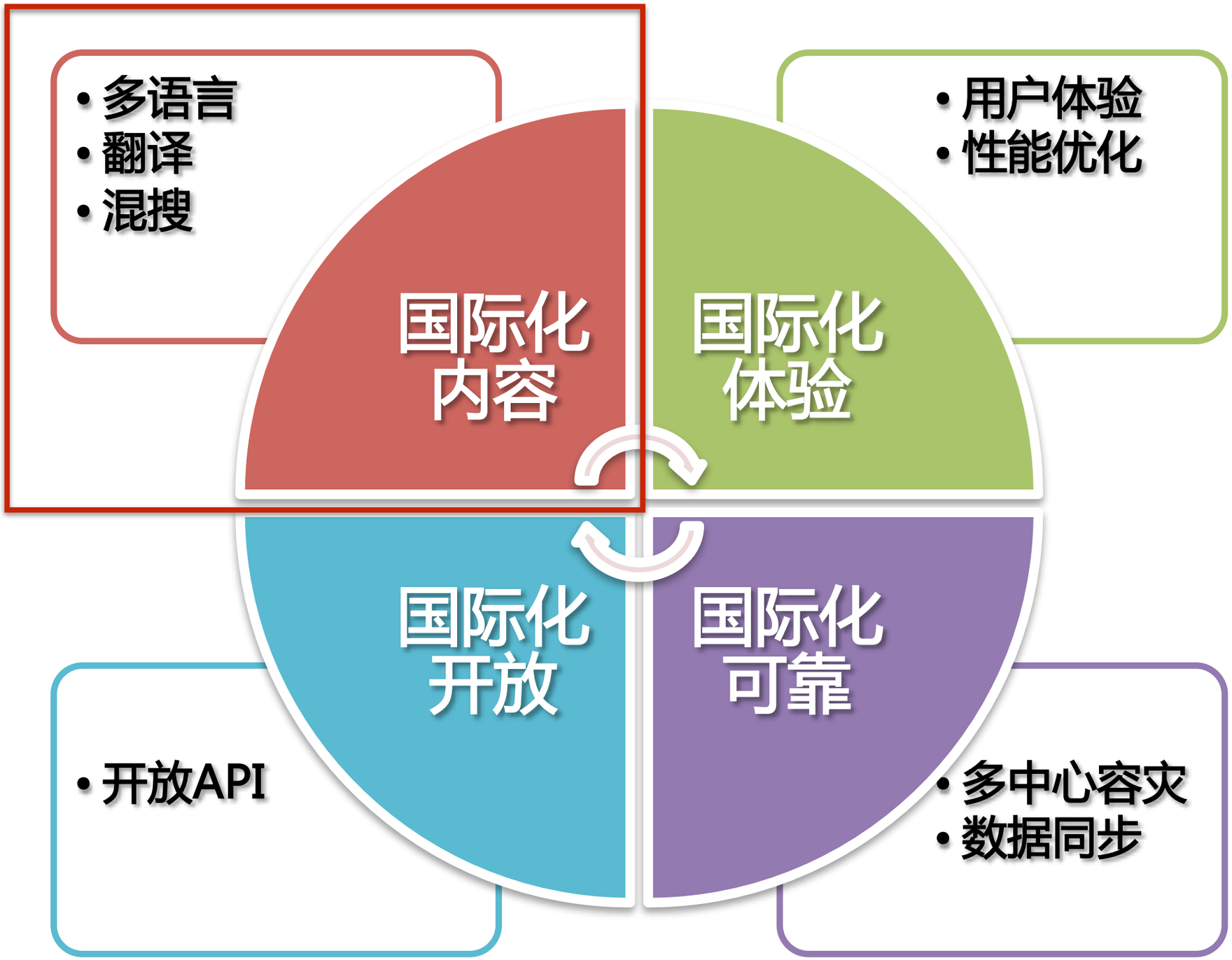
翻译


翻 译


奥斯卡剧场
百年经典电影
FEAST OF DUBBING THE FILM CLASSIC CENTURY
上译配音
配音史上难以超越的巅峰之作
经典影片 绝妙配音 穿越时空 魅力永驻
B


Google
Translate
English
Spanish
French
Detect language
English
Spanish
Arabic
Translate
Type text or a website address or translate a document.

网站国际化



目录

阿里巴巴全球化战略

Global Business(跨境贸易)

阿里巴巴全球化战略

to C

全球购 G.TAOBAO.COM 探索全球美好生活

天猫国际 TMALL.HK 买进口，上天猫国际

AliExpress

进口

出口

1688 全球货源 in.1688.com

Alibaba.com® Global trade starts here.™

to B

为什么本地化很重要

买家

卖家

阿里巴巴国际电商平台

天猫国际 TMALL.HK 买进口，上天猫国际

全球购 G.TAOBAO.COM 探索全球美好生活

1688 全球货源 in.1688.com

Alibaba.com® Global trade starts here.™

AliExpress

为什么本地化很重要

阿里巴巴国际电商平台

天猫国际 TMALL.HK 买进口，上天猫国际

全球购 G.TAOBAO.COM 探索全球美好生活

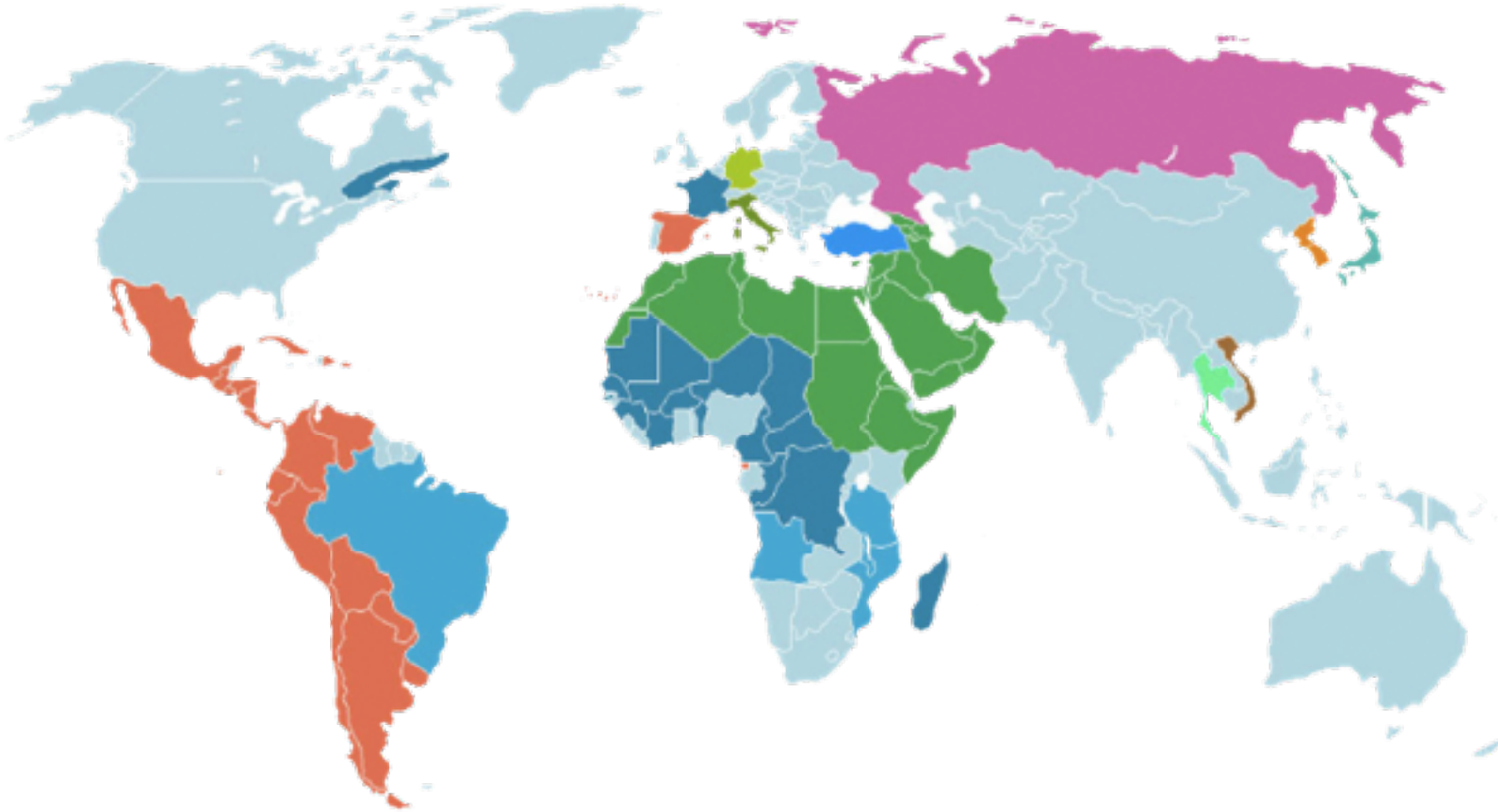
1688 全球货源 in.1688.com

Alibaba.com® Global trade starts here.™

AliExpress

为什么本地化很重要

买家



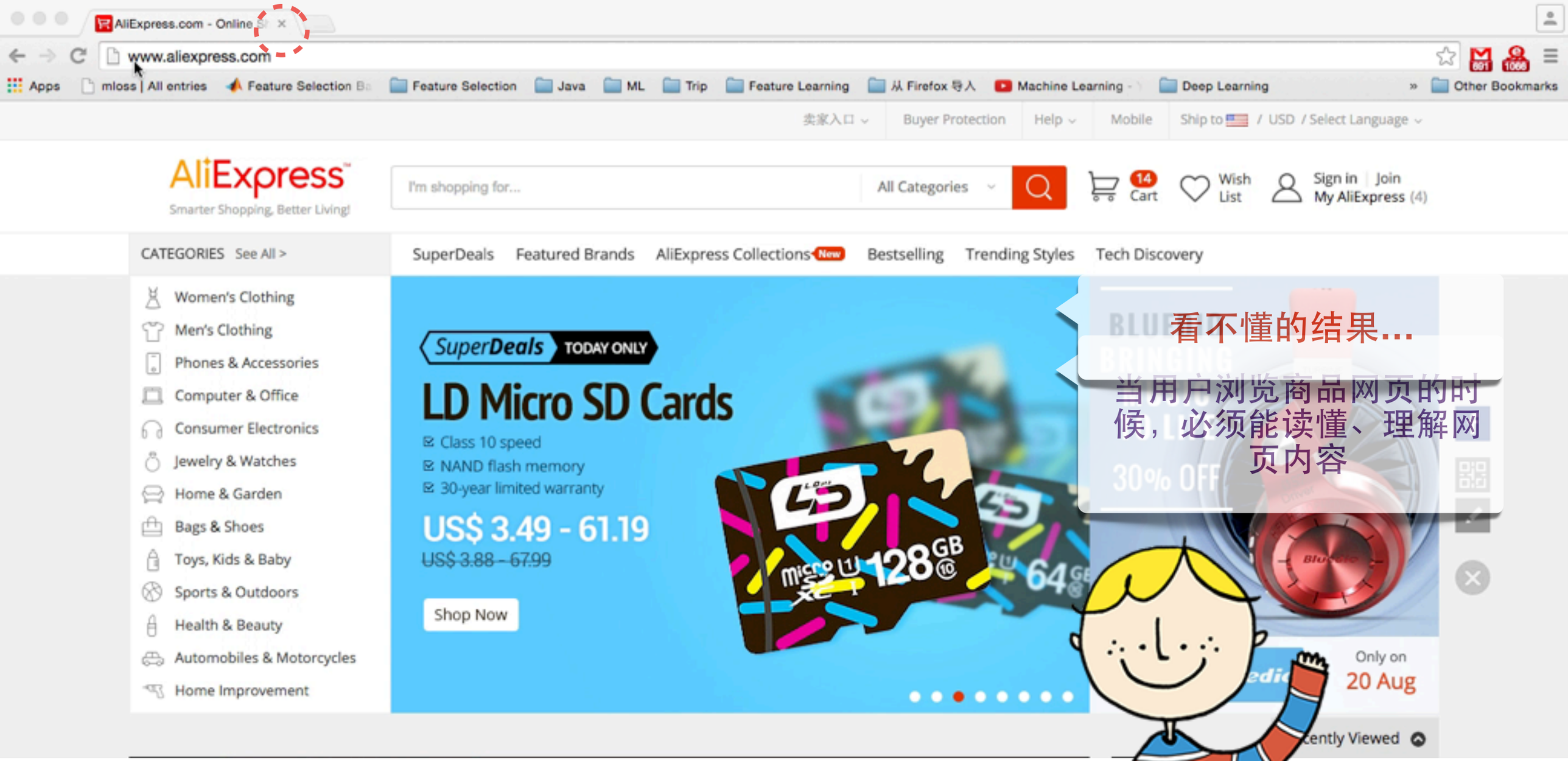
卖家

来自世界各地，说着“不同的语言”

为什么本地化很重要

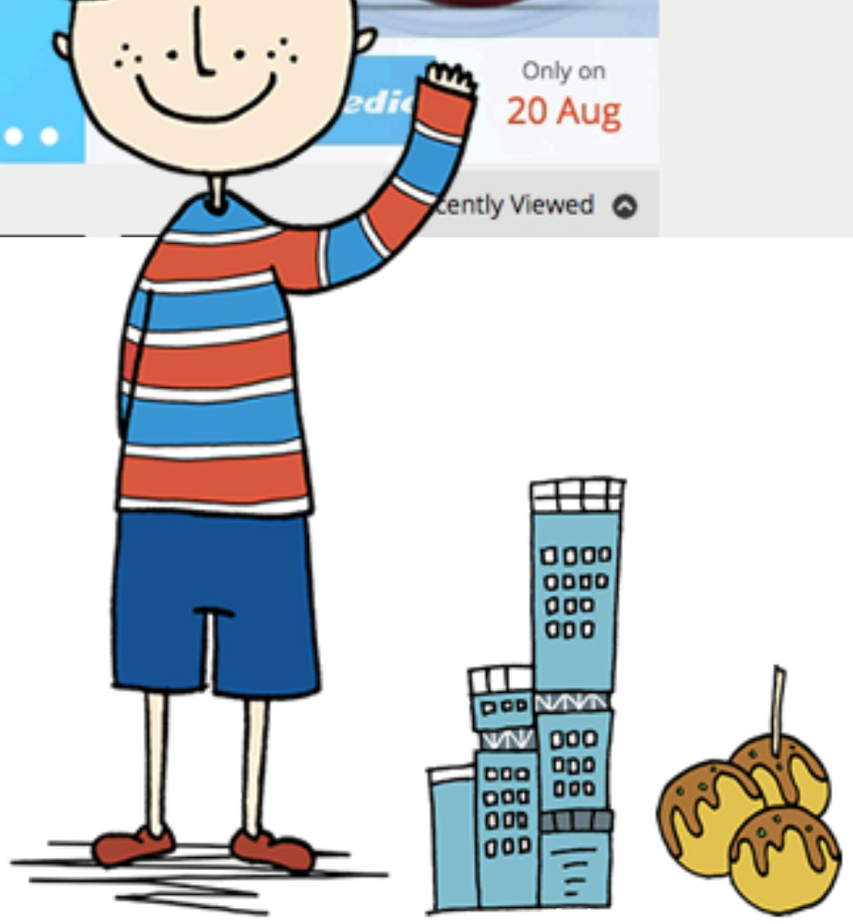
语言是跨境电子商务的障碍

来自世界各地，说着“不同的语言”

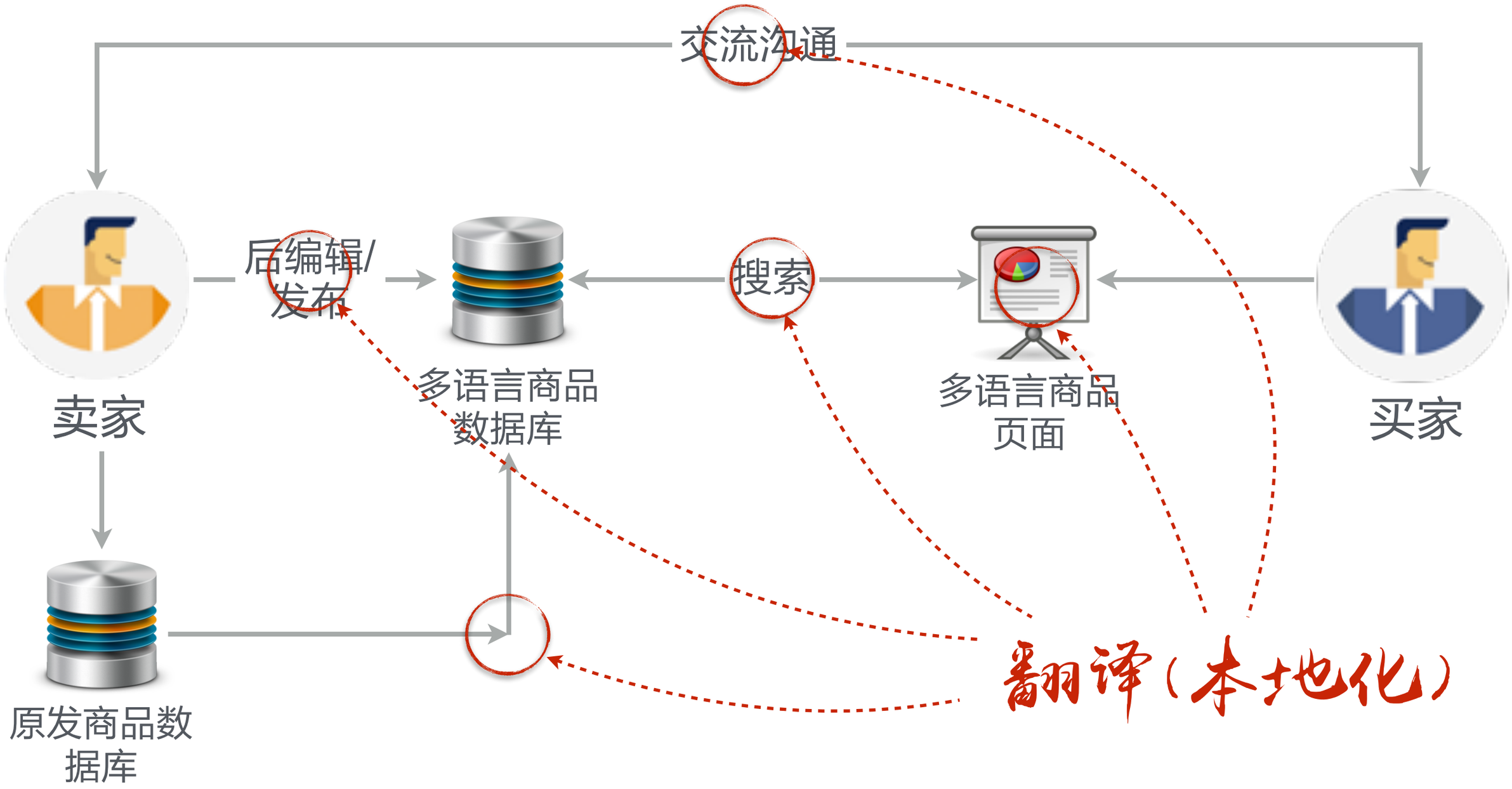


看不懂的结果...

当用户浏览商品网页的时候，必须能读懂、理解网页内容



什么需要进行翻译



什么需要进行翻译

举个“栗子”

什么需要进行翻译

AliExpress™ | vestidos | All Categories | 0 Cart | Wish List | Sign in | Join | My AliExpress

Back to search results | Home > All Categories > Women's Clothing & Accessories > Dresses

Mouse over to zoom in

Product ID: 32298823720

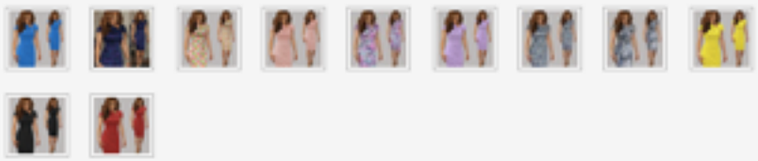
share:

New 2015 Womens Celebrity Elegant Vintage Pinup Bow Ruch Tunic Business Casual Cocktail Party Business Bodycon Dress 266

98.5% of buyers enjoyed this product! (1695 votes) | 2588 orders

Price: US $14.99 - 19.99 / piece

Color:



Size: XXXL | 4XL | 5XL | S | M | L | XL | XXL

Ships From: United States | China

Enjoy Domestic Delivery & 7-Day Easy Returns Learn more

Shipping: US $3.15 to United States via USPS

Estimated Delivery Time: 2-7 days (ships out within 3 business days)

Quantity: 1 piece (584 pieces available)

Total Price: Depends on the product properties you select

Buy Now | Add to Cart

Top-rated Seller

Sold By

Valuefashionshop

China (Mainland)

26605

99.6% Positive feedback

Detailed seller ratings

Visit Store

Add to My Favorite Stores (21337 Adds)

Contact Seller

Contact Now

Recently Viewed

什么需要进行翻译

New 2015 Womens Celebrity Elegant Vintage Pinup Bow Ruch Tunic Business Casual Cocktail Party Business Bodycon Dress 266

俄语

Новый 2015 женщин знаменитости элегантные старинные кинозвезды с бантом рух туника бизнес свободного покроя коктейль ну вечеринку бизнес Bodycon платье 266

葡语

Novo 2015 Womens celebridade elegante do Vintage Pinup Bow Ruch túnica negócios Cocktail Party Casual negócios Bodycon vestido 266

韩语

새로운 여자 2015 유명 우아한 빈티지 핀업 활 루흐 옷 비즈니스 캐주얼 칵테일 파티 비즈니스 266 bodycon 드레스

日语

新しい2015女性セレブのエレガントなヴィンテージルーシュピンナップ弓チュニックビジネスカジュアルなカクテルパーティビジネス266bodyconドレス

翻译成多国语言

什么需要进行翻译

AliExpress™ | vestidos | All Categories | 0 Cart | Wish List | Sign in | Join | My AliExpress

Back to search results | Home > All Categories > Women's Clothing & Accessories > Dresses

Mouse over to zoom in

Product ID: 32298823720

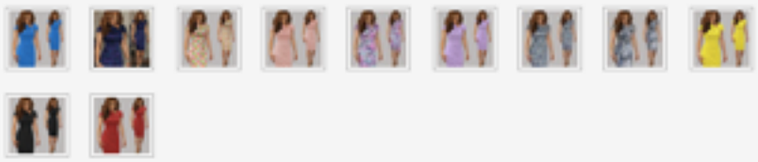
share:

New 2015 Womens Celebrity Elegant Vintage Pinup Bow Ruch Tunic Business Casual Cocktail Party Business Bodycon Dress 266

★★★★★ **98.5%** of buyers enjoyed this product! (1695 votes) | **2588** orders

Price: **US $14.99 - 19.99** / piece

Color:



Size: XXXL | 4XL | 5XL | S | M | L | XL | XXL

Ships From: United States | China

Enjoy Domestic Delivery & 7-Day Easy Returns Learn more

Shipping: **US $3.15** to United States via USPS

Estimated Delivery Time: **2-7** days (ships out within 3 business days)

Quantity: 1 piece (584 pieces available)

Total Price: Depends on the product properties you select

Buy Now | Add to Cart

Top-rated Seller

Sold By

Valuefashionshop
China (Mainland)

26605

99.6% Positive feedback

Detailed seller ratings

Visit Store

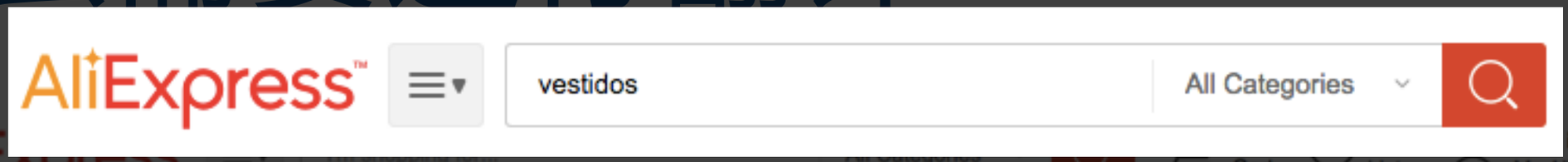
Add to My Favorite Stores
(21337 Adds)

Contact Seller

Contact Now

Recently Viewed

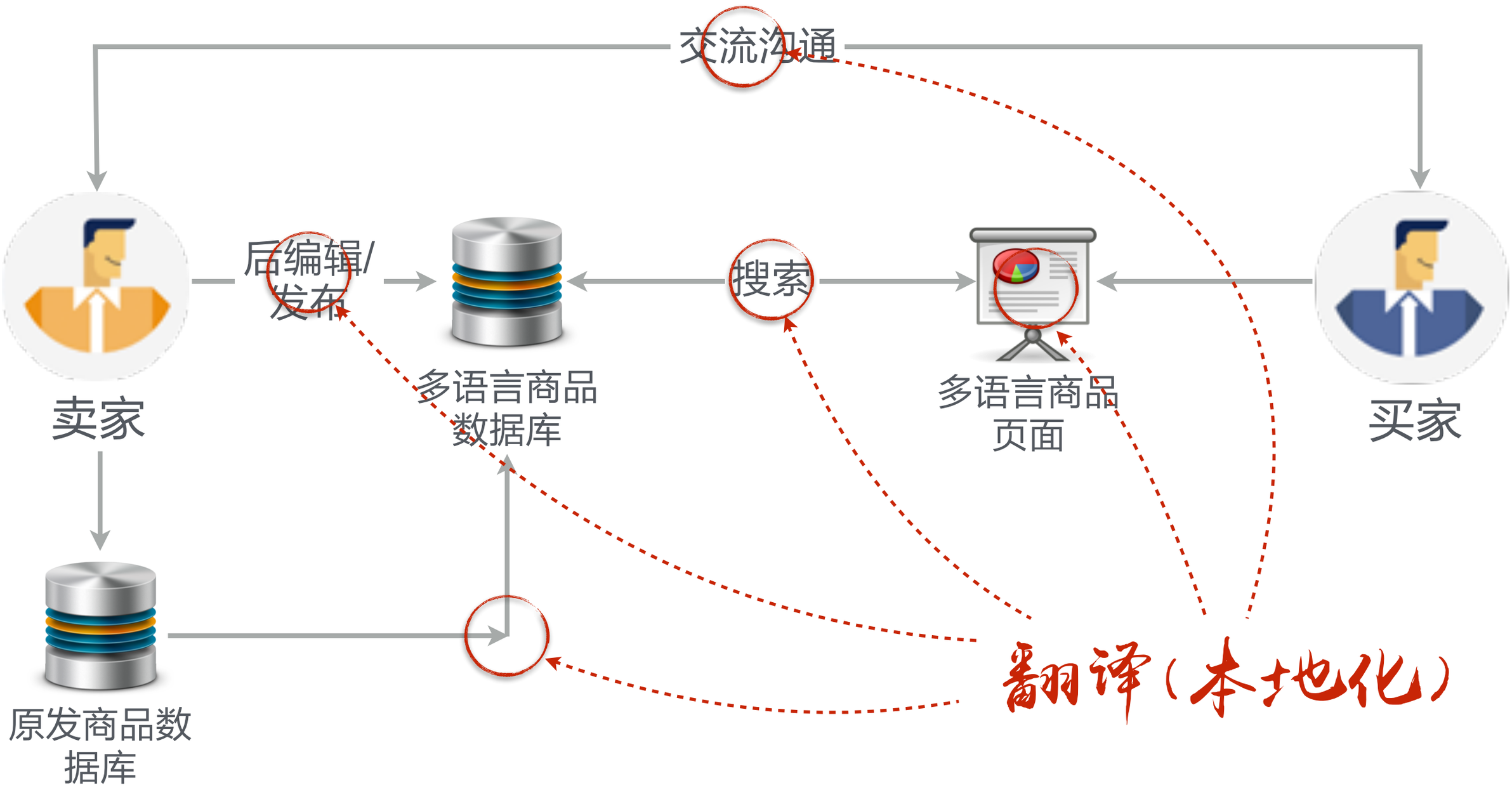
什么需要进行翻译



多语言 翻译成 英文

vestidos → dresses

进行英文搜索



翻译场景还有很多…

翻译场景还有很多…

量级大

速度要求高

量级小

精度要求高

分为2种类型

翻译场景还有很多…

量级大

速度要求高

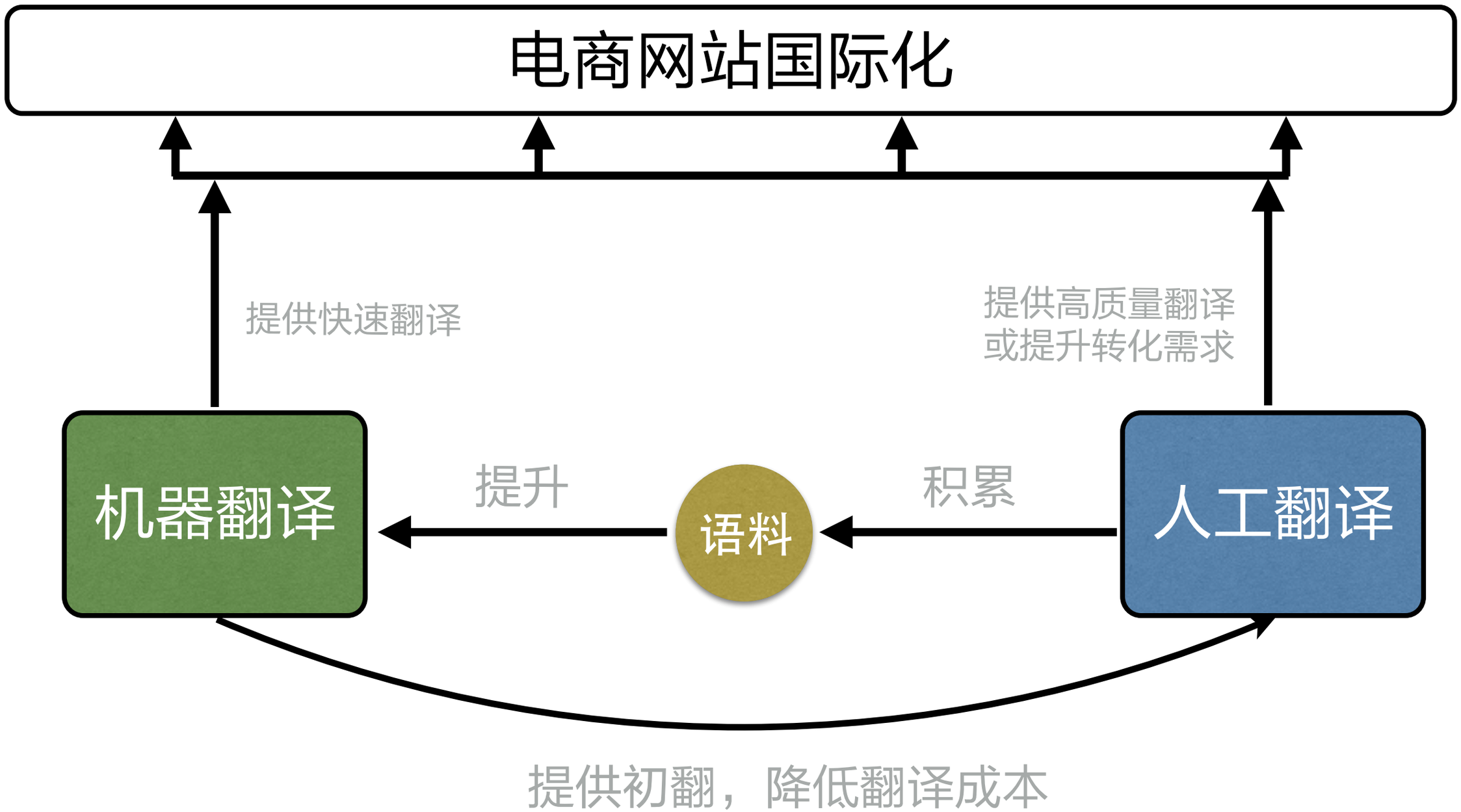
机器翻译

量级小

精度要求高

人工翻译

如何进行翻译


电商网站国际化
提供快速翻译
提供高质量翻译
或提升转化需求
机器翻译
提升
语料
积累
人工翻译
提供初翻，降低翻译成本

目录

机器翻译，你怎么看?


Translate server error
餐 厅


世纪联华 CENTURYmart
一次性用品
A TIME SEX THING


芳草萋萋 踏之何忍
Fangcaoqiqi riding the Heren


带好孩子
Take the child
小心落水
Fall into water carefully
www.engrish.com

机器翻译能做什么?

快速翻译

大规模翻译

中等译文质量

主流机器翻译技术

词典

1st generation MTs
词对词的翻译
少量的调序规则

2nd generation MTs
基于规则的翻译
需要制定大量的
翻译语法规则

3rd generation MTs
基于统计的翻译
统计方式产生翻译规则
无需手工制定语法规则

4th generation MTs
基于统神经网络翻译

统计机器翻译

I am a boy .

我是一个男孩。

寻找概率最大候选翻译

统计机器翻译

I am a boy .

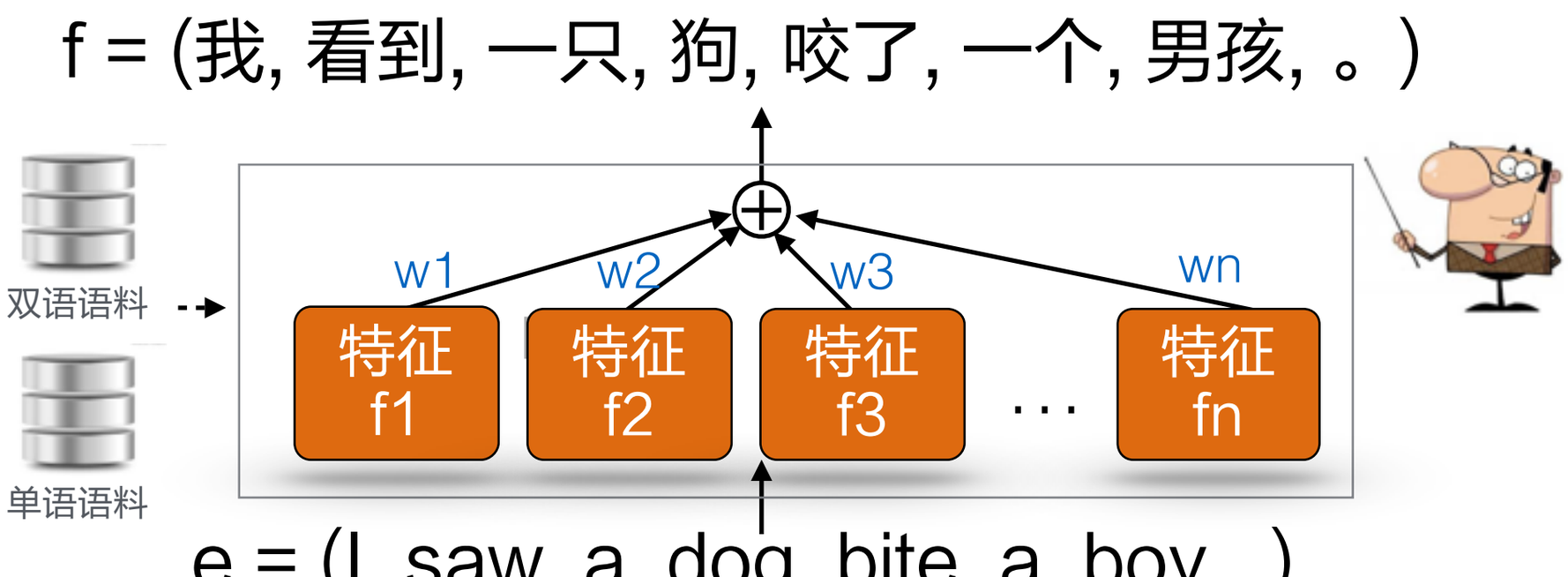
我是一个男孩。 0.5634
我是一个小子。 0.2325
我是一男子。 0.1231
我是一个小伙。 0.0233
……

搜寻空间

概率计算

寻找概率最大候选翻译

统计机器翻译



$$\log p(f|e) \approx \sum_i w_i f_i(e, f) + C$$

- 翻译系统选型为Log-linear Model,
- 融合大量的文本翻译特征，支持传统的Feature Engineering方式
- Maximize another metric, e.g., BLEU

统计机器翻译

训练

翻译模型

语言模型

p(s)=p(狗 | 一只) x
p(咬了 | 狗)..

调序模型

MERT调参

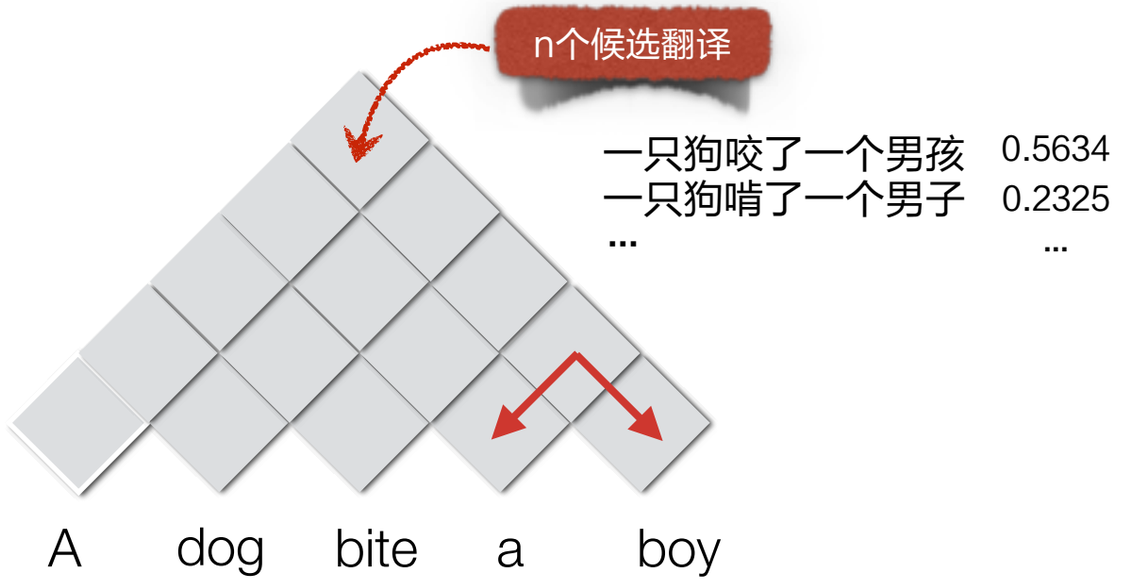
翻译模型 语言模型 调序模型

寻找“最优”权重：0.53, 0.12, 0.31

翻译(解码)

$$t^* = \arg\max_t p(t \mid s) = \arg\max_t \frac{\exp \sum_{i=1}^{m} w_i f(s,t)}{\sum_{t'} \exp \sum_{i=1}^{m} w_i f(s,t')}$$

译文 原文 权重 特征



神经网络机器翻译

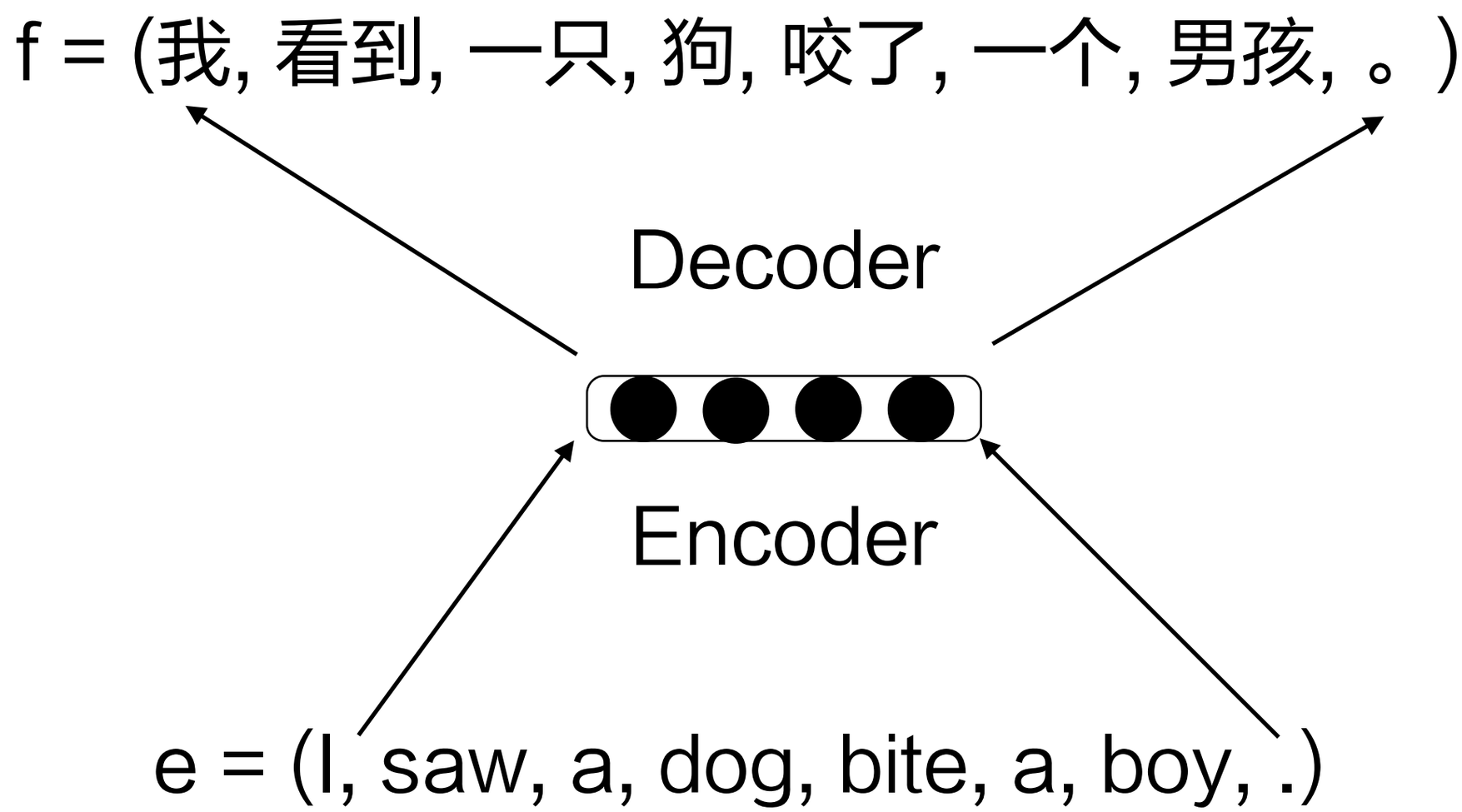
f = (我, 看到, 一只, 狗, 咬了, 一个, 男孩, 。)

语言生成

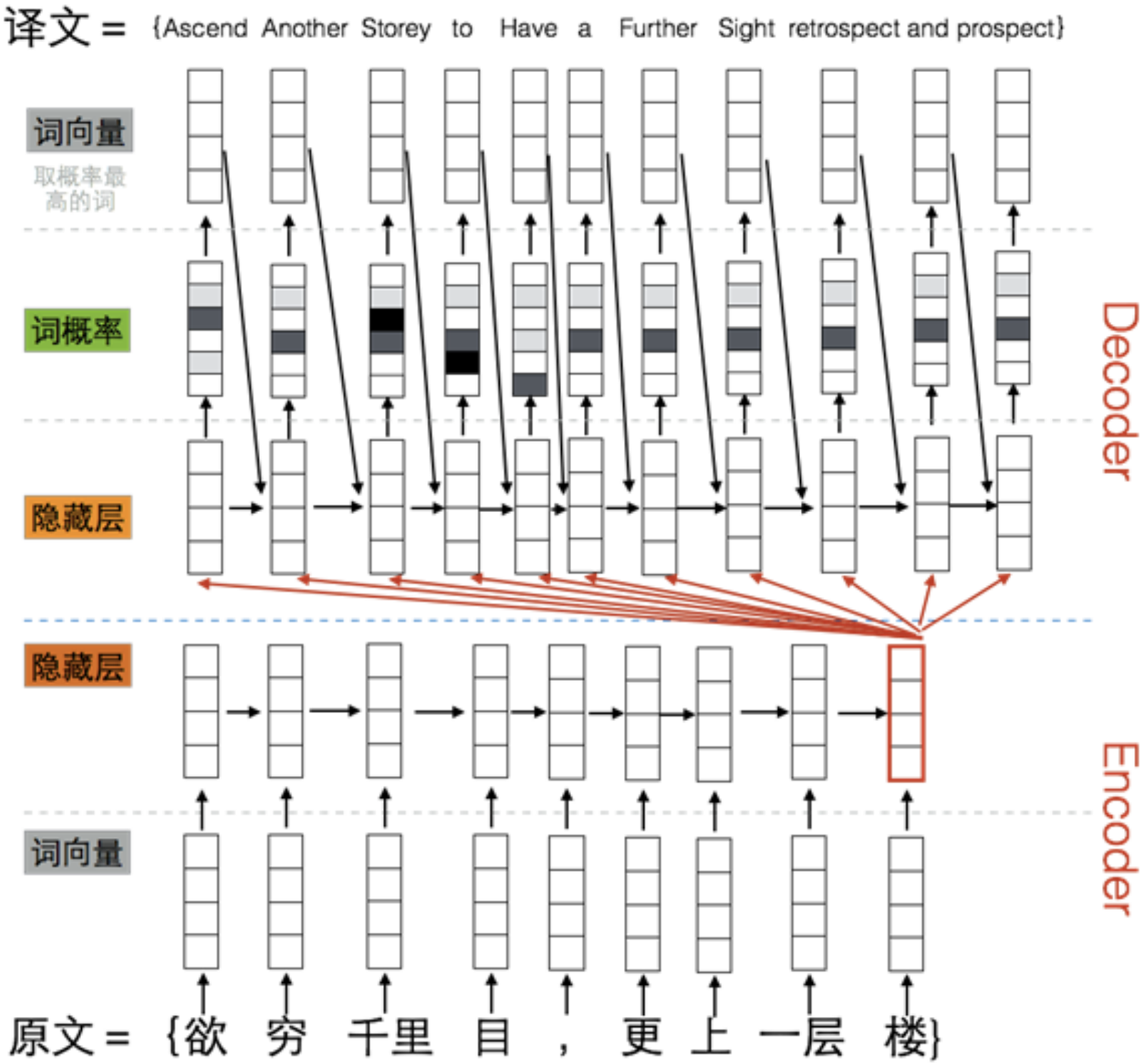
理解

e = (I, saw, a, dog, bite, a, boy, .)

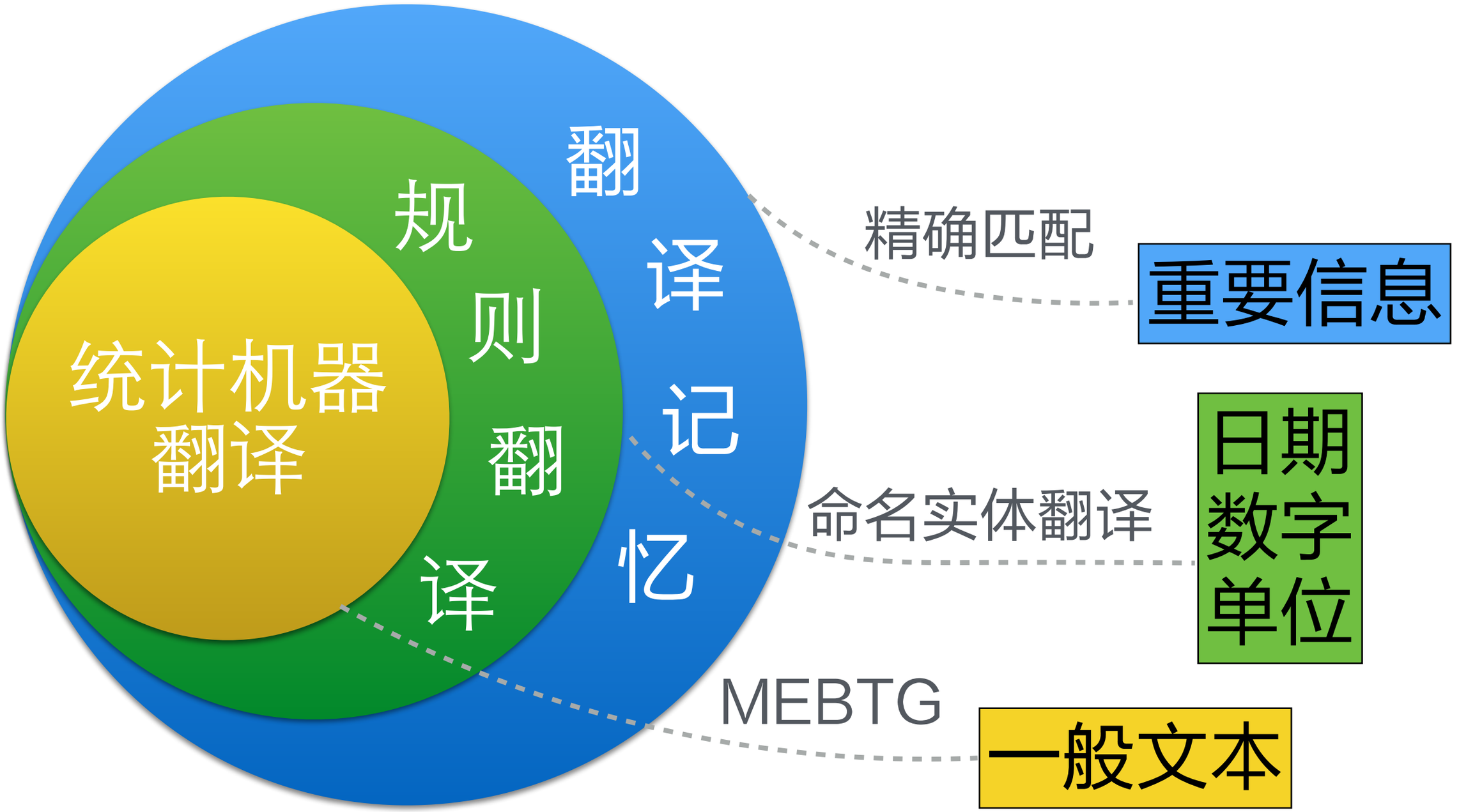
神经网络机器翻译



神经网络机器翻译


译文 = {Ascend Another Storey to Have a Further Sight retrospect and prospect}
词向量
取概率最高的词
词概率
隐藏层
Decoder
隐藏层
词向量
Encoder
原文 = {欲 穷 千里 目 ， 更 上 一层 楼}

面向电商的阿里机器翻译引擎



面向电商的阿里机器翻译引擎

训练

翻译模型

语言模型

p(s)=p(狗 | 一只) x
p(咬了 | 狗)..

调序模型

MERT调参

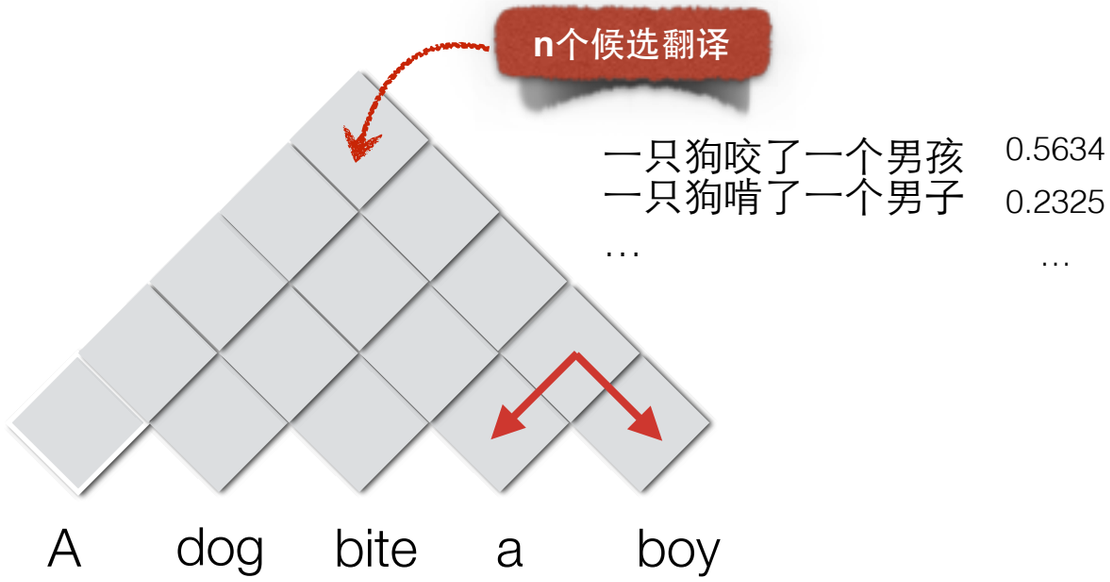
翻译模型 | 语言模型 | 调序模型

寻找“最优”权重：0.53, 0.12, 0.31

翻译(解码)

$$t^* = \arg\max_t p(t \mid s) = \arg\max_t \frac{\exp \sum_{i=1}^{m} w_i f(s,t)}{\sum_{t'} \exp \sum_{i=1}^{m} w_i f(s,t')}$$

译文 原文 权重 特征



搭建电商领域的机器翻译引擎

数据驱动系统

训练、翻译可能会很慢

领域相关性强

搭建电商领域的机器翻译引擎

数据驱动系统

要什么样的数据？ 数据从哪里来？

要什么样的数据

电商领域的双语语料

电商领域的单语语料

电商专业词表

电商品牌词表

电商高频短语翻译

通用领域单语语料

通用领域双语语料

要什么样的数据

网络抓取

人工翻译

要什么样的数据

行业化数据

高质量数据

原始数据

行业化高质量双语

行业化高质量双语

行业化分类管理

高质量双语

高质量单语

双语质量评估过滤

单语质量评估过滤

原始双语平行语料

原始单语语料

web数据

人工翻译数据

搭建电商领域的机器翻译引擎

数据驱动系统

训练、翻译可能会很慢

领域相关性强

训练&翻译的效率

机器翻译人员最大的一项技能...

等待

原来...

语料

训练

翻译模型 --> 48~72小时

语言模型 --> 6小时

调序模型 --> 4小时

参数优化 --> 1小时

约3天

原来...

语料

翻译模型

语言模型

调序模型

参数优化

训练

离线批量翻译

线上调用翻译

翻译

阿里翻译在云端

语料

翻译模型

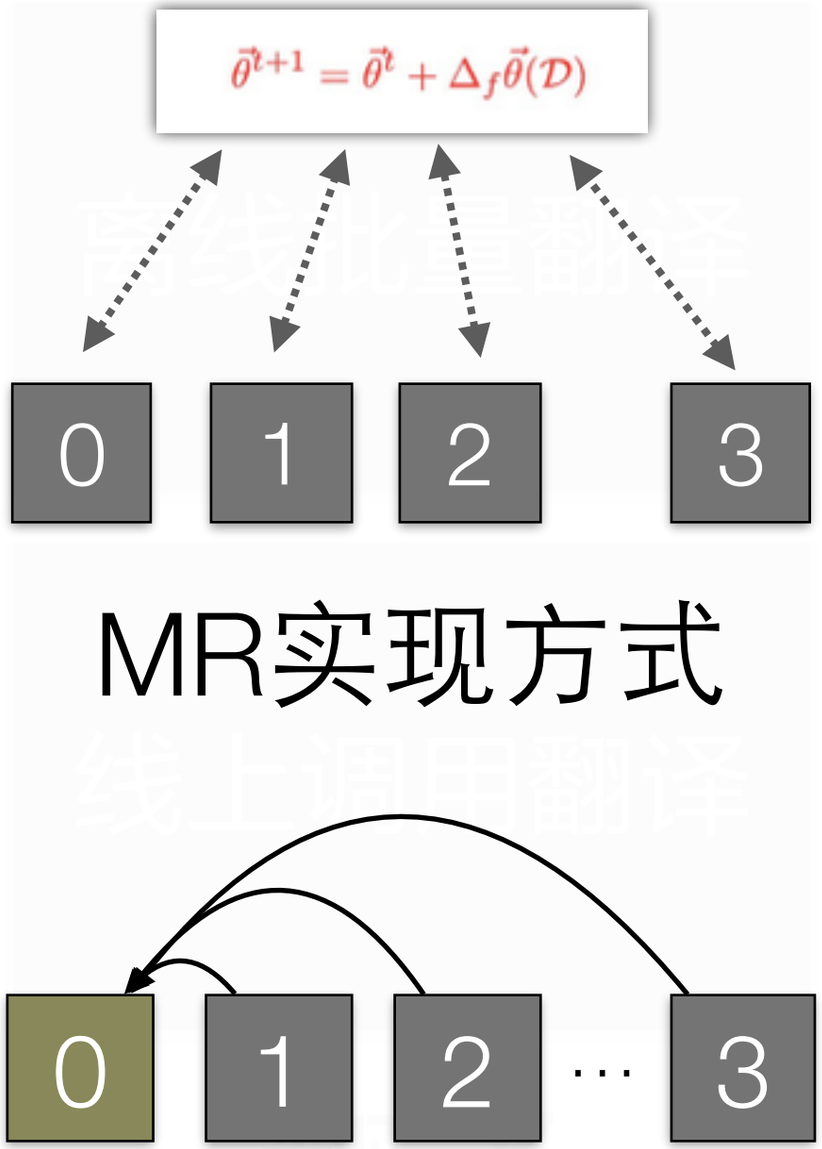
语言模型

调序模型

3天

6小时

训练



阿里翻译在云端

数据分片 ◀- - - - - 20天→1天

分布式集群 ◀- - - - - 200台→40台

搭建电商领域的机器翻译引擎

数据驱动系统

训练、翻译可能会很慢

领域相关性强

搭建电商领域的机器翻译引擎

领域相关性强

如何适应电商领域翻译?

电商领域翻译

适应电商文本翻译技术体系

电商领域翻译

数据

单/双语质量自动评估技术

领域语料自动筛选技术

语料运营平台

web数据

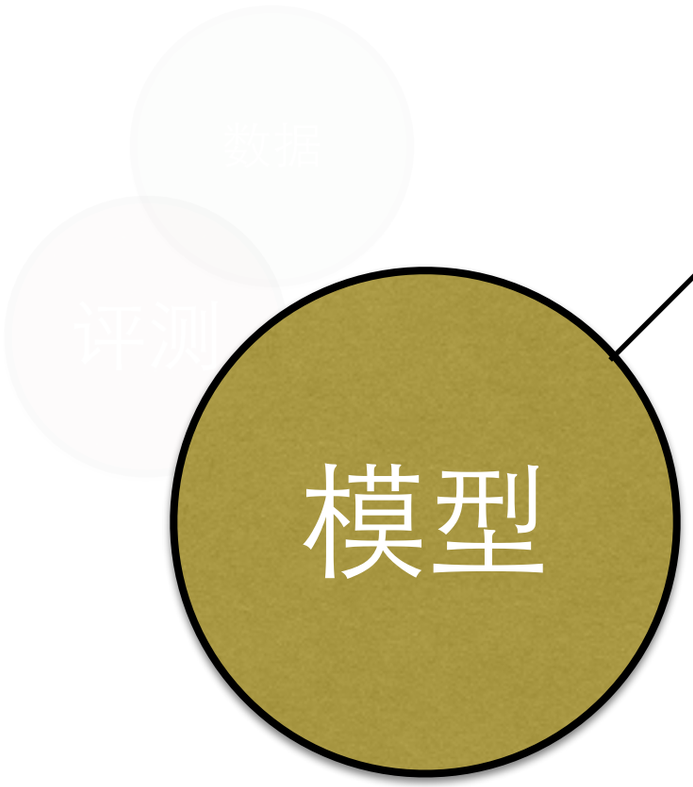
人工翻译数据

运营

算法

领域数据

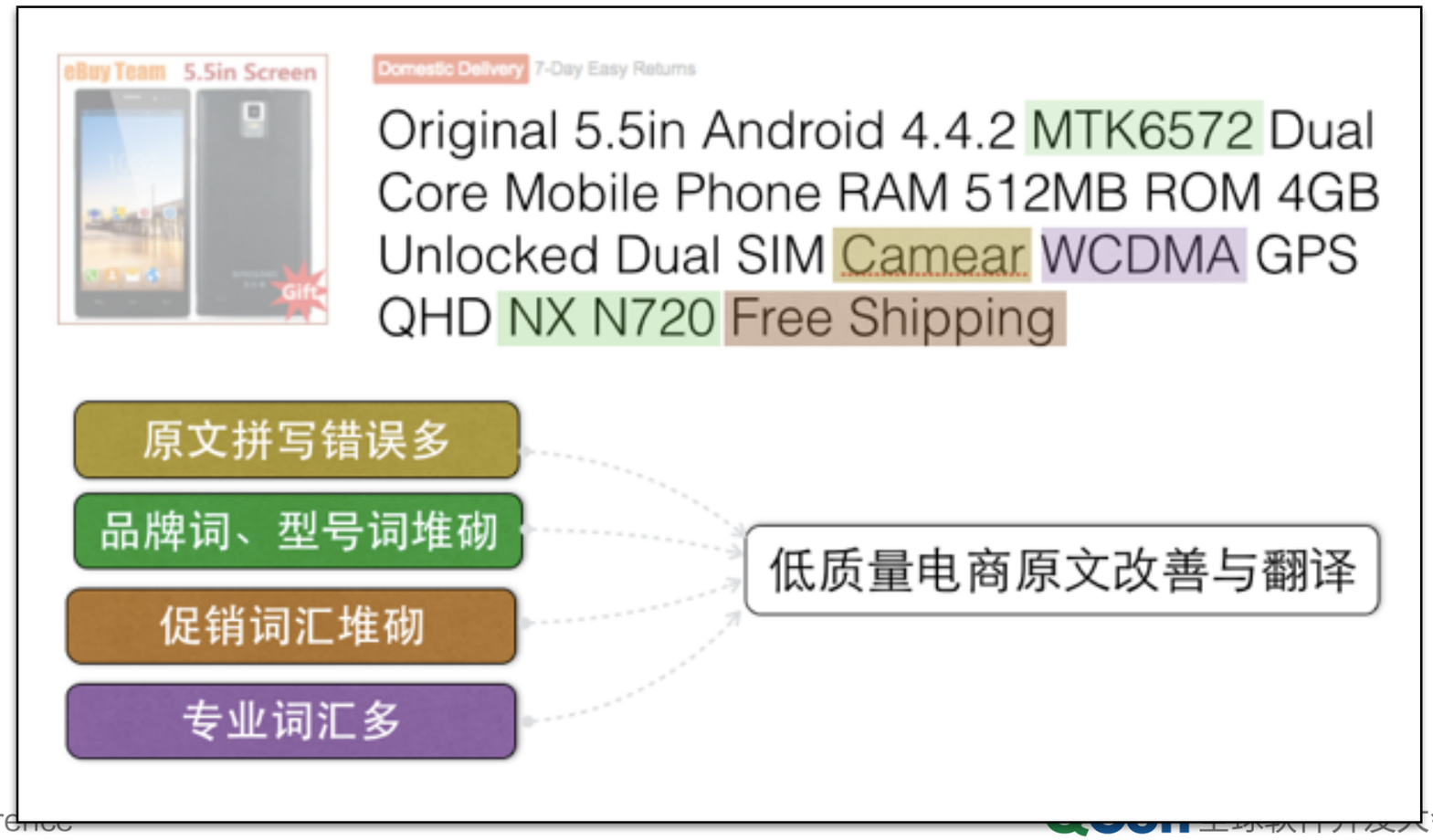
电商领域翻译



电商原文优化

领域特征自动挖掘方法

添加领域特征



电商领域翻译

数据

评测

模型

电商文本优化

领域特征自动挖掘方法

添加领域特征

是否含hot word
是否含query
关键词的位置
…

模型拟合

CTR预估模型

根据用户线上数据挖掘文本翻译特征

电商领域翻译

模型

电商文本优化

领域特征自动挖掘方法

添加领域特征

不同行业（类目）专业词的翻译错误

原文：...black nuts (黑螺母)…

译文：… Ч е р н ы й о р е х и (黑核桃)

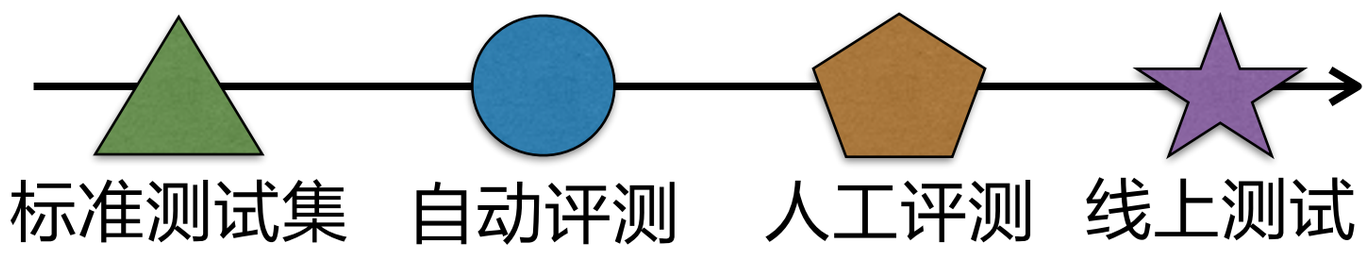
特殊词汇的自动挖掘 + 翻译运营平台 + 类目主题模型

电商领域翻译

评测

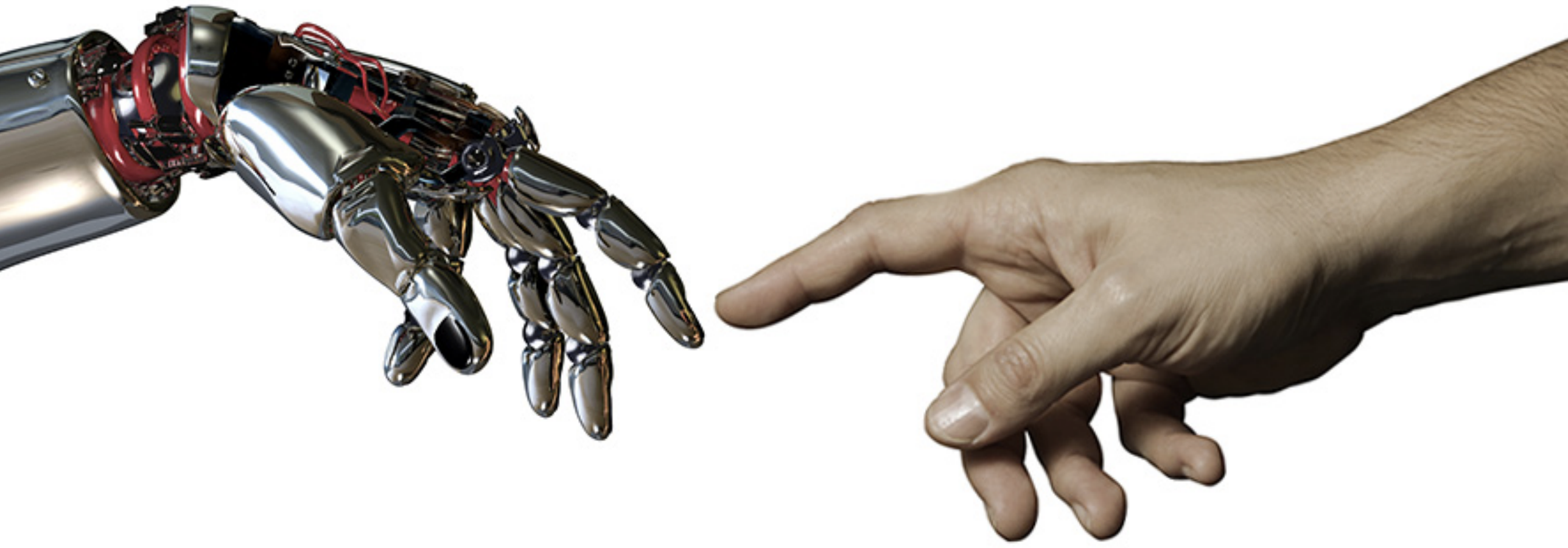
电商文本翻译的评测方法

电商翻译不仅仅只是语言学问题



目录

机器翻译&人工翻译



翻译质量上还是有相当大的距离

传统人工翻译的优缺点

翻译质量 高

翻译效率 低

众包翻译

翻译需求

众包平台

全世界的译员

“独翻译，不如众翻译”

众包翻译

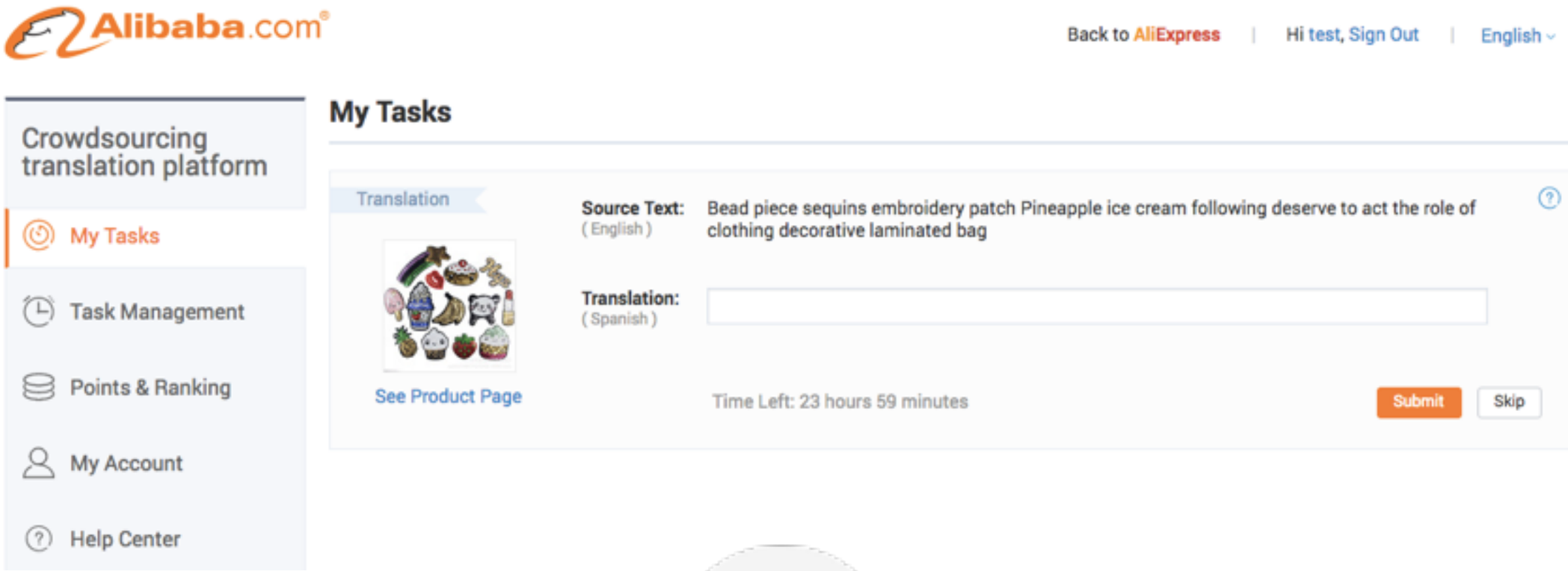
全世界的译员

买家

买家即译员

阿里巴巴国际电商平台

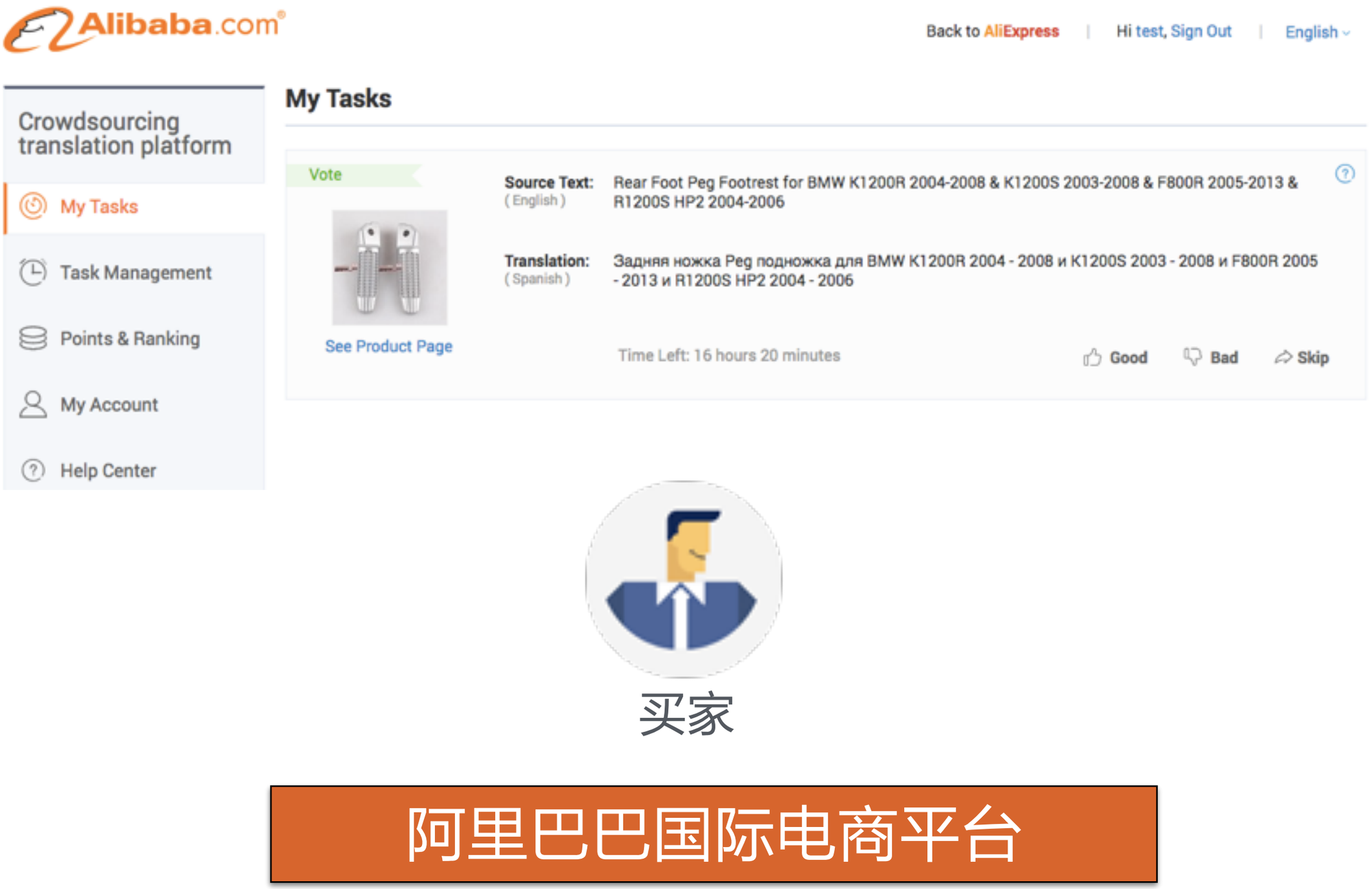
众包翻译



买家

阿里巴巴国际电商平台

众包翻译



买家

阿里巴巴国际电商平台

目录

经验1: 机器翻译充当网站国际化的主要角色，为你的领域搭建专属的机翻系统

经验2: 人工翻译其实可以做更多事情

经验3: 永远别指望翻译模型解决所有问题，可以更加关心翻译数据的累积

Thanks!

Q&A