

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

自我介绍



曾晓东

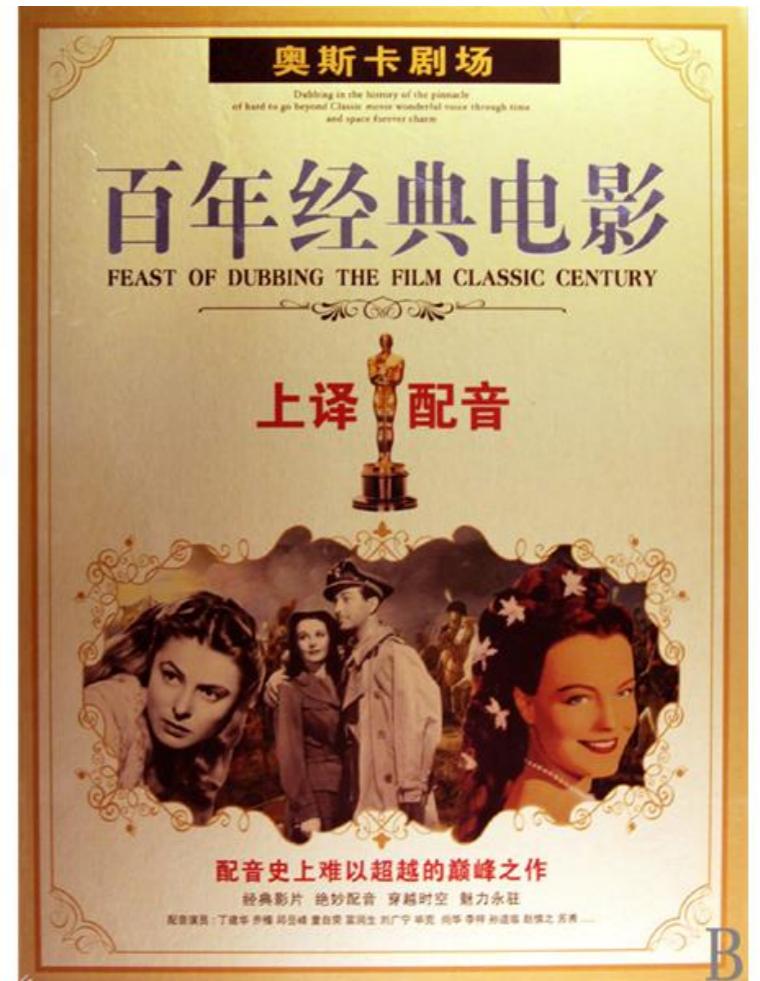
阿里巴巴集团 B2B技术部-翻译平台 技术专家

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

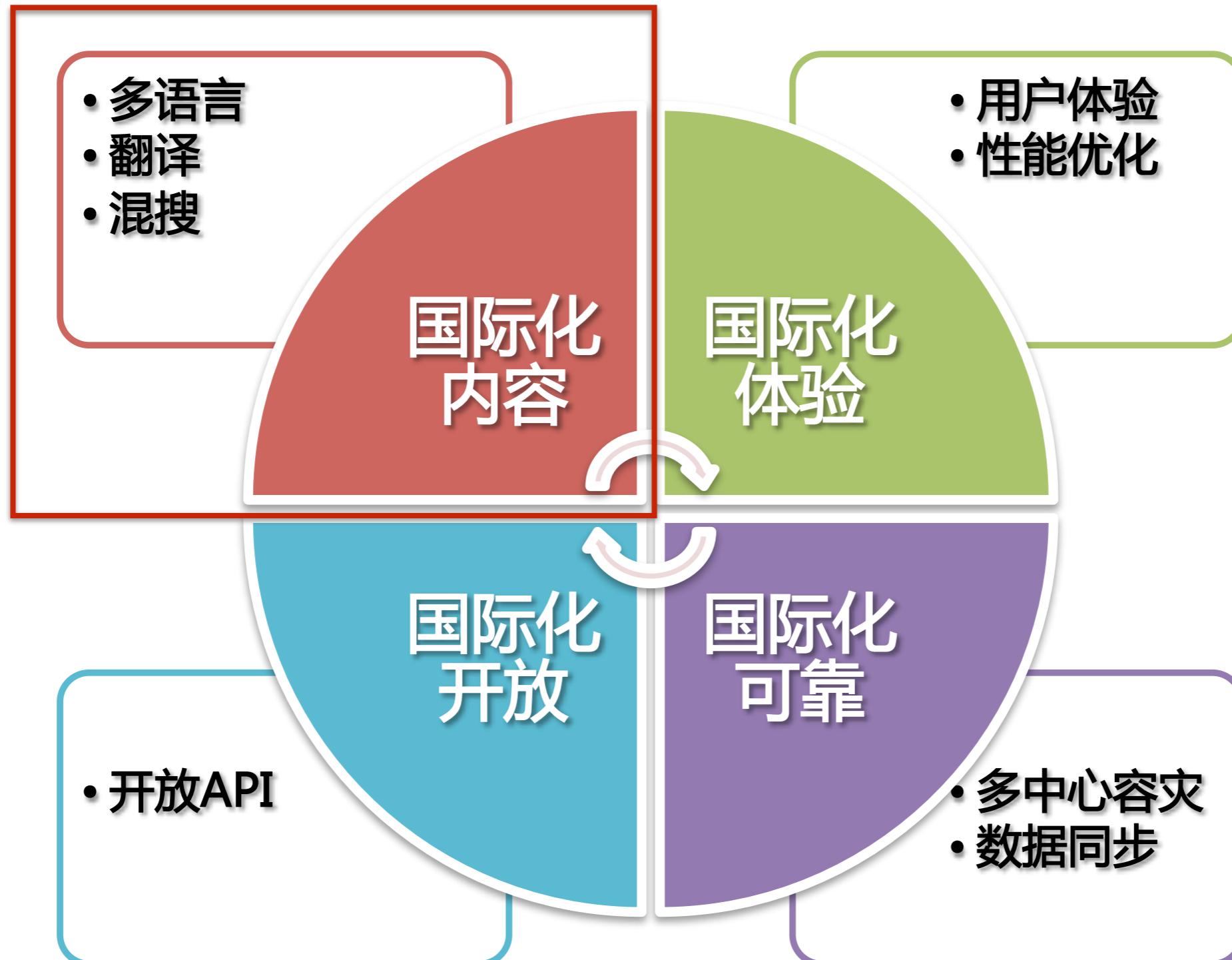
如何利用**翻译技术**帮助
电商网站的**国际化**

翻译



The Google Translate interface. At the top, there's a search bar with the word "Google". Below it, a red "Translate" button is followed by language selection boxes for English, Spanish, French, Detect language, and Arabic. To the right of these are two large, empty text input fields. At the bottom, a message says "Type text or a website address or translate a document."

网站国际化



目录

- 1 阿里巴巴电商国际化
- 2 机器翻译技术
- 3 人工(众包)翻译技术
- 4 经验总结

阿里巴巴全球化战略



Global Business(跨境贸易)

阿里巴巴全球化战略



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

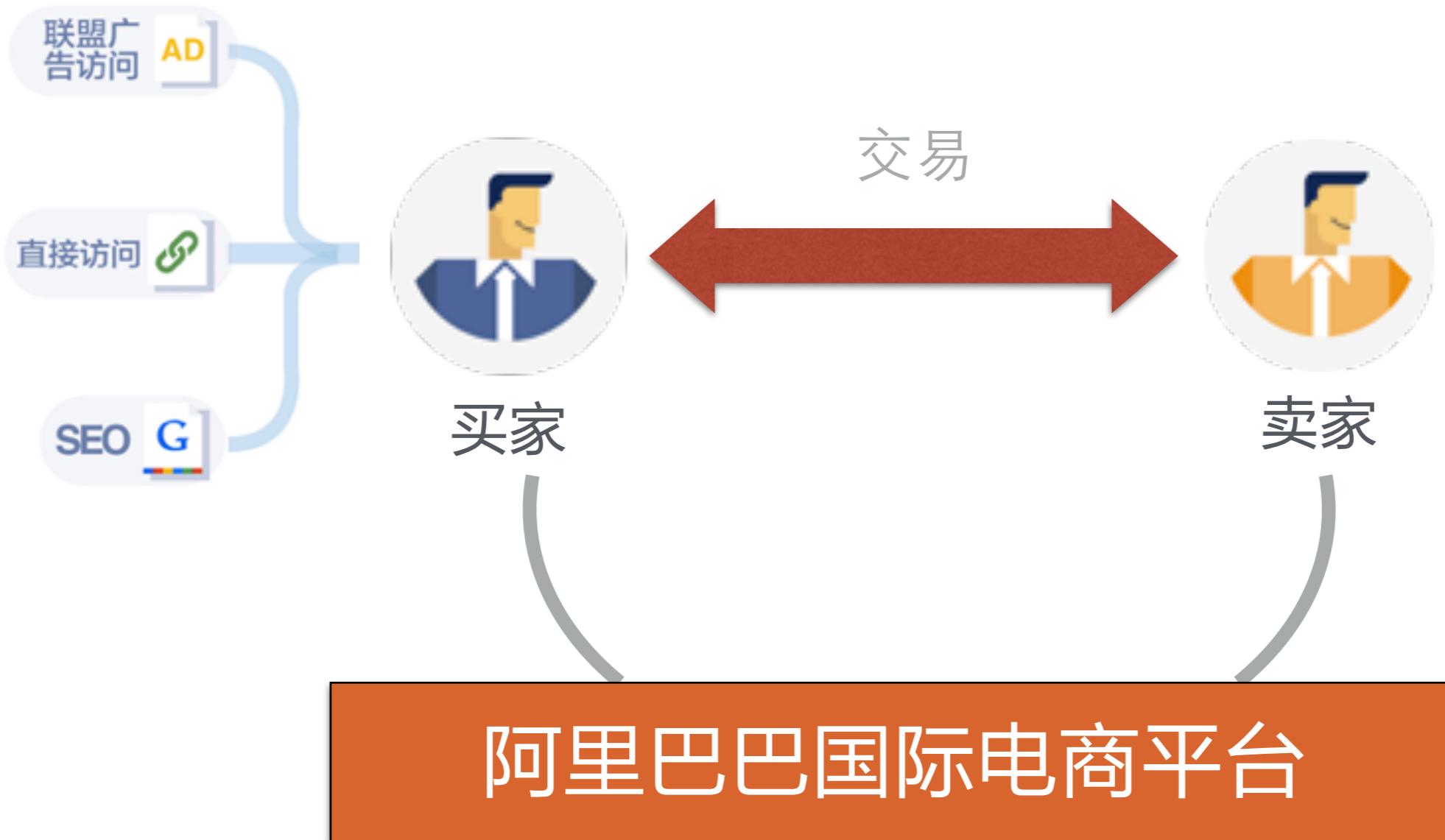
探索全球美好生活

1688 全球货源
in.1688.com

Alibaba.com®
Global trade starts here.™

Ali**express**

为什么本地化很重要



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

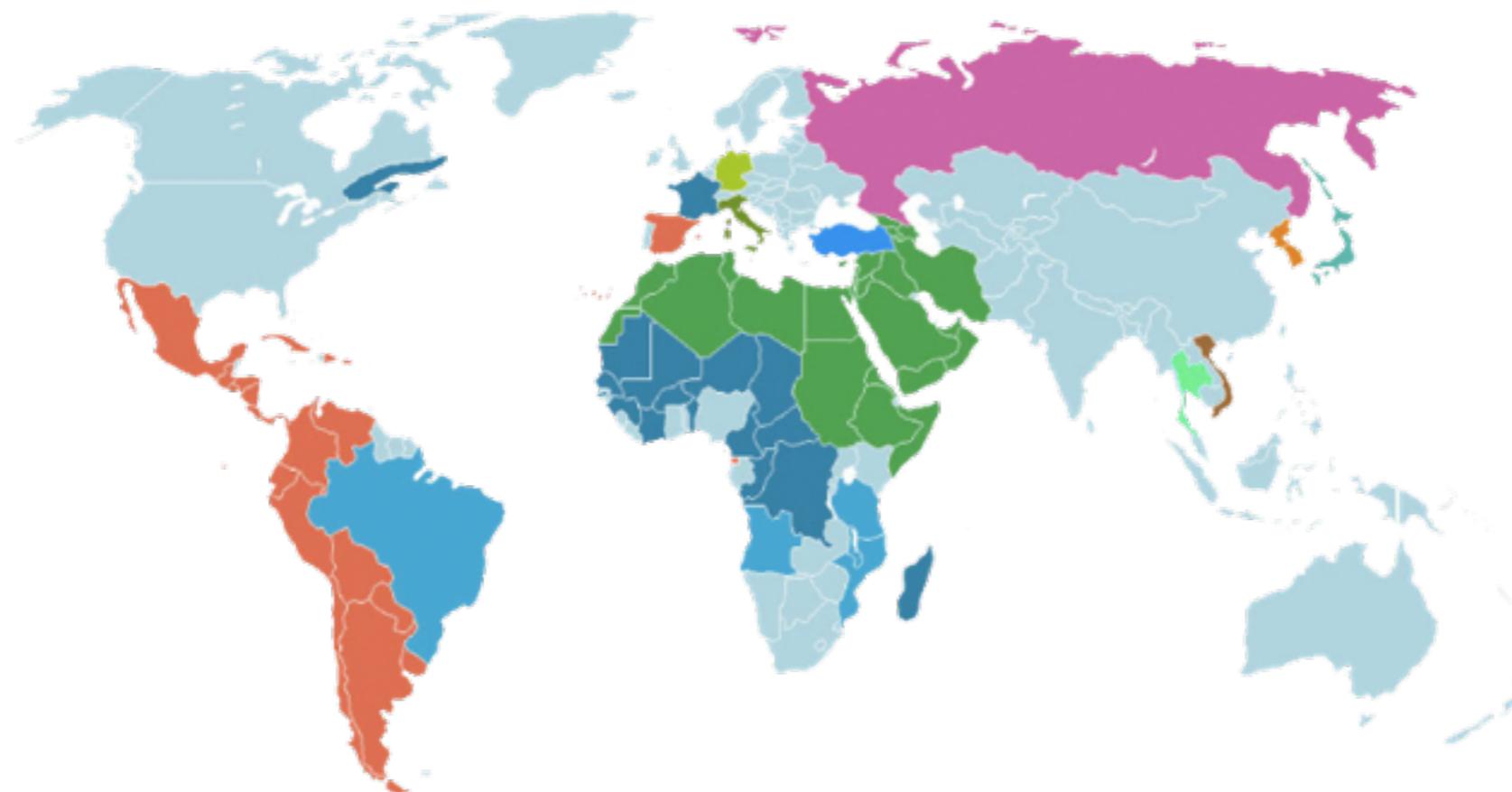
Alibaba.com®
Global trade starts here.™

AliExpress

为什么本地化很重要



买家



卖家

■ 西班牙语 ■ 土耳其语 ■ 日语 ■ 德语 ■ 韩语 ■ 意大利语
■ 葡萄牙语 ■ 阿拉伯语 ■ 俄语 ■ 法语 ■ 泰语 ■ 越南语

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

- 西班牙语 ■ 土耳其语 ■ 日语 ■ 德语 ■ 韩语 ■ 意大利语
- 葡萄牙语 ■ 阿拉伯语 ■ 俄语 ■ 法语 ■ 泰语 ■ 越南语

来自世界各地，说着“不同的语言”

AliExpress.com - Online 

www.aliexpress.com

卖家入口 Buyer Protection Help Mobile Ship to USA / USD / Select Language

AliExpress Smarter Shopping, Better Living!

I'm shopping for... All Categories  Cart (14) Wish List Sign in | Join My AliExpress (4)

CATEGORIES See All >

- Women's Clothing
- Men's Clothing
- Phones & Accessories
- Computer & Office
- Consumer Electronics
- Jewelry & Watches
- Home & Garden
- Bags & Shoes
- Toys, Kids & Baby
- Sports & Outdoors
- Health & Beauty
- Automobiles & Motorcycles
- Home Improvement

SuperDeals Featured Brands AliExpress Collections New Bestselling Trending Styles Tech Discovery

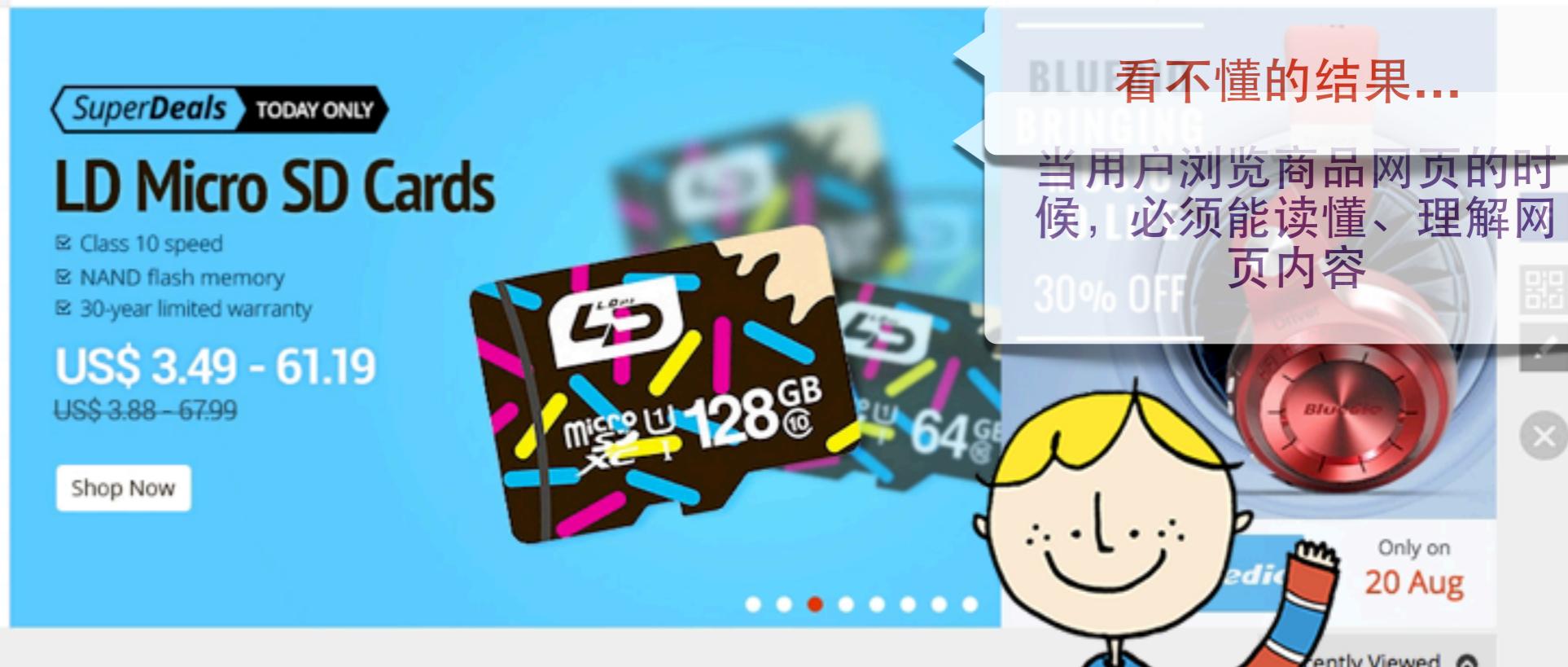
SuperDeals TODAY ONLY

LD Micro SD Cards

Class 10 speed
NAND flash memory
30-year limited warranty

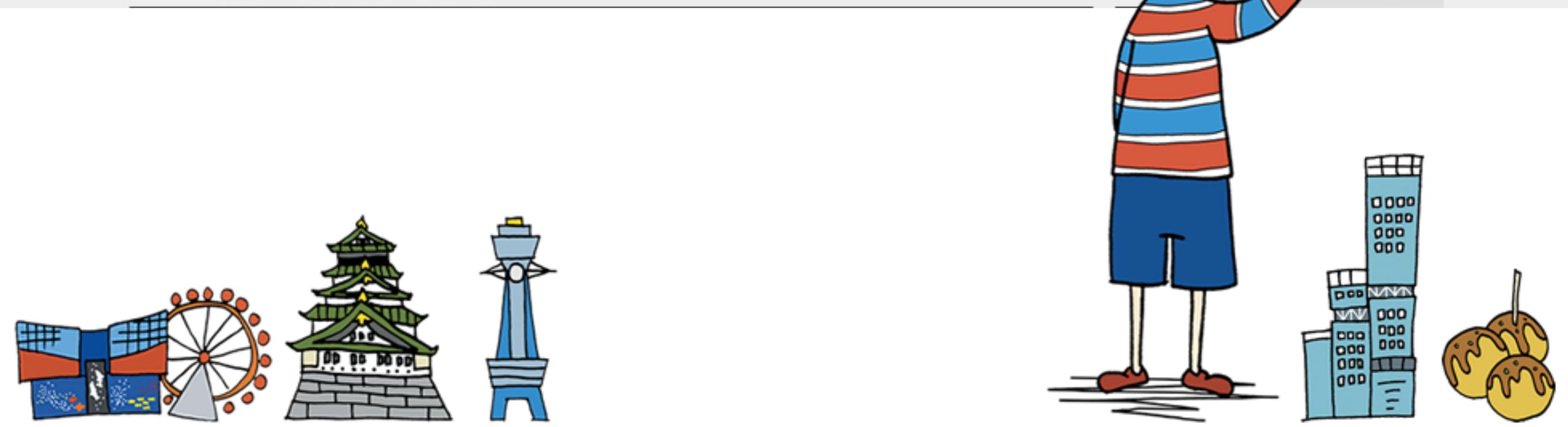
US\$ 3.49 - 61.19
US\$ 3.88 - 67.99

Shop Now

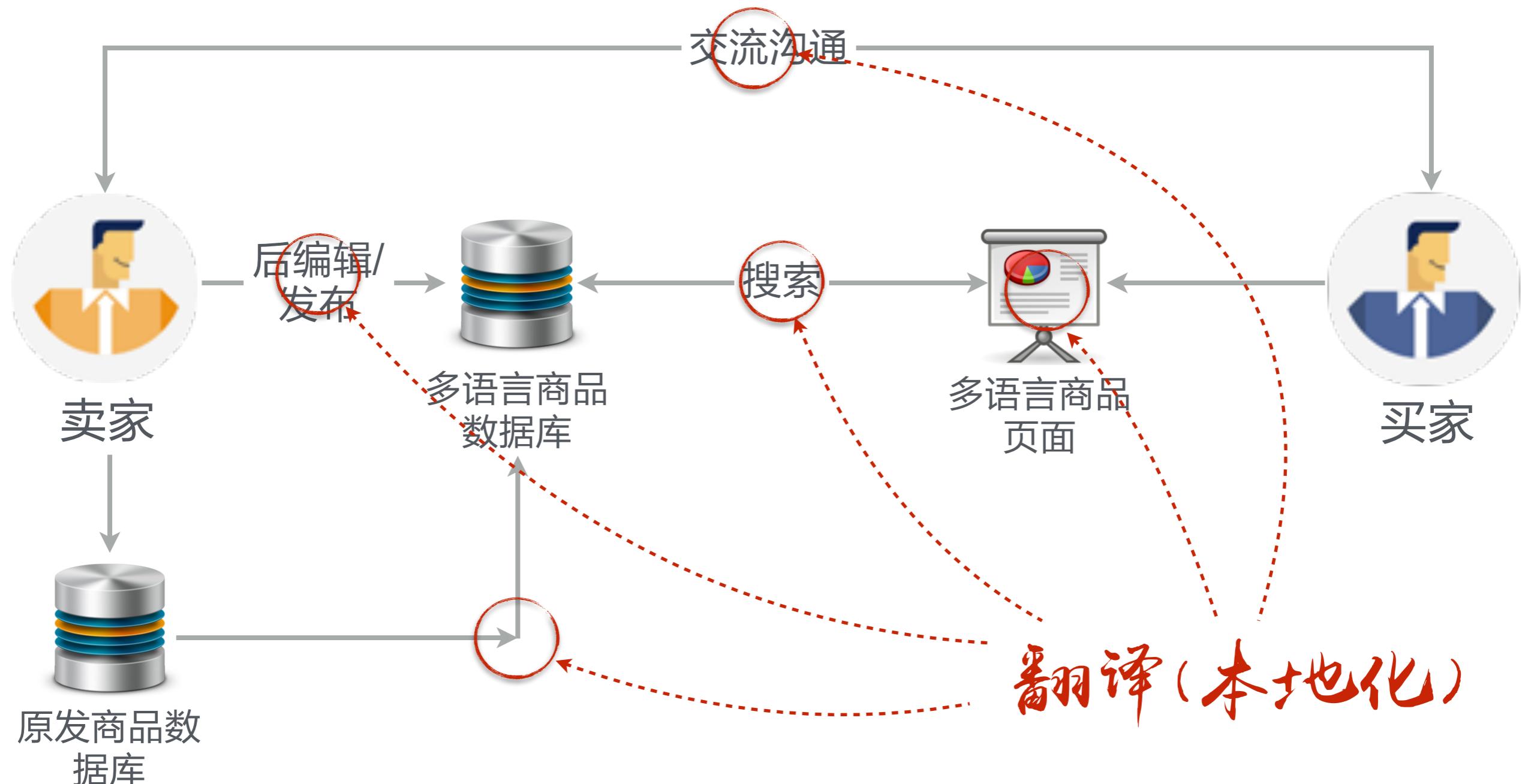


看不懂的结果...

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 🔍 Cart 0 Wish List Sign in | Join My AliExpress

[Back to search results](#) | Home > All Categories > Women's Clothing & Accessories > Dresses

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

Recently Viewed

Product ID: 32298823720

share:

什么需要进行翻译

The screenshot shows a product listing on AliExpress for a "New 2015 Womens Celebrity Elegant Vintage Pinup Bow Ruch Tunic Business Casual Cocktail Party Business Bodycon Dress 266". The page is displayed in five different languages, each with its own title and description. The languages are: English, Russian, Portuguese, Korean, and Japanese. The English title is "New 2015 Womens Celebrity Elegant Vintage Pinup Bow Ruch Tunic Business Casual Cocktail Party Business Bodycon Dress 266". The Russian title is "Новый 2015 женщин знаменитости элегантные старинные кинозвезды с бантом рух туника бизнес свободного покроя коктейль ну вечеринку бизнес Bodycon платье 266". The Portuguese title is "Novo 2015 Womens celebridade elegante do Vintage Pinup Bow Ruch túnica negócios Cocktail Party Casual negócios Bodycon vestido 266". The Korean title is "새로운 여자 2015 유명 우아한 빈티지 핀업 활 루흐 옷 비즈니스 캐주얼 칵테일 파티 비즈니스 266 bodycon 드레스". The Japanese title is "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". The page also shows the price range (US \$14.99 - 19.99), sizes (XXXL, 4XL, 5XL, S, M, L, XL, XXL), shipping information (US \$3.15 to United States via USPS), and a top-rated seller badge.

俄语

葡萄牙语

韩语

日语

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

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

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

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

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 🔍 Cart 0 Wish List Sign in | Join My AliExpress

[Back to search results](#) Home > All Categories > Women's Clothing & Accessories > Dresses

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

Recently Viewed

Product ID: 32298823720

share:

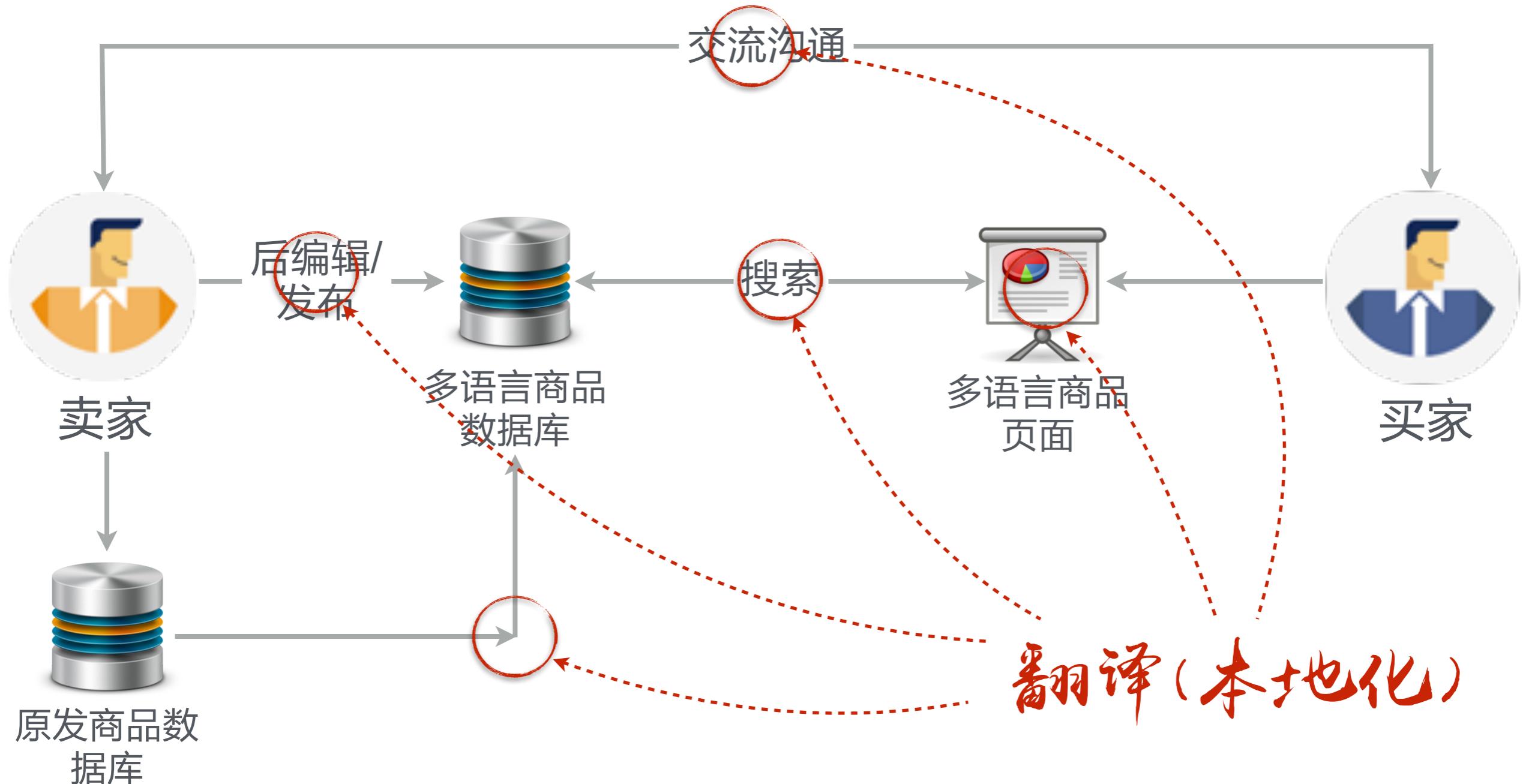
什么需要进行翻译

The screenshot shows a product page on AliExpress for 'New 2015 Women's Celebrity Elegant Vintage Pinup Bow Ruch Tunic Business Casual Cocktail Party Business Bodycon Dress 266'. A large white arrow points from the Spanish search term 'vestidos' in the search bar at the top left to the English translation 'dresses' in the product title on the right. The page includes details like price (\$14.99 - 18.99), seller information ('Top-rated Seller'), and shipping options.

多语言翻译成英文

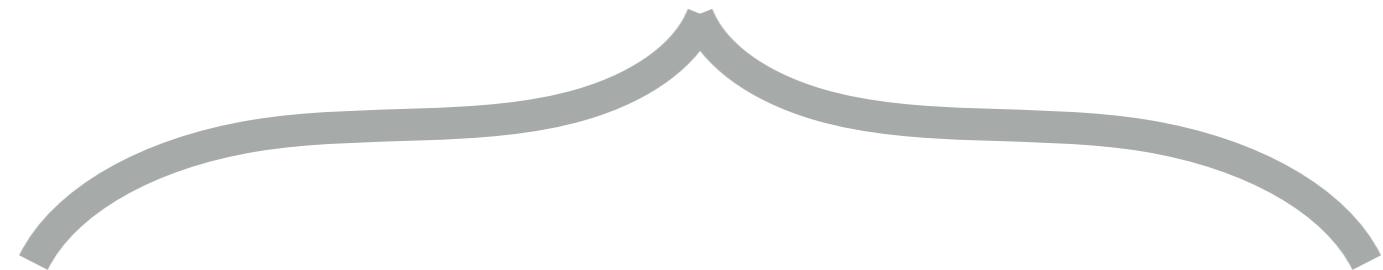
vestidos → dresses

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

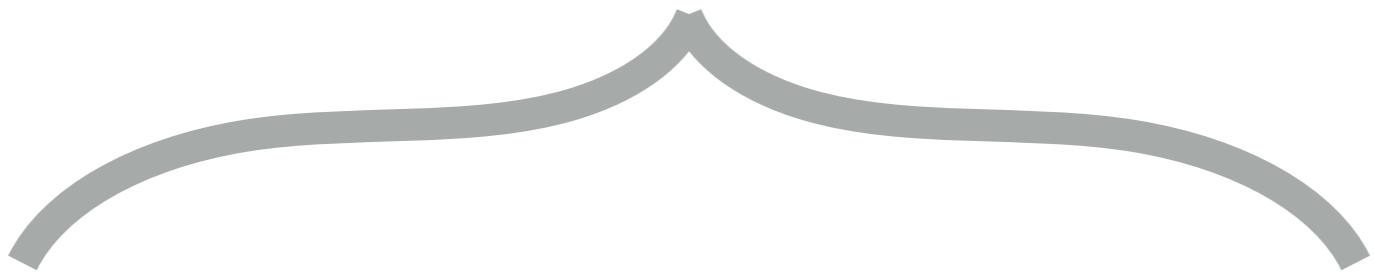
速度要求高

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

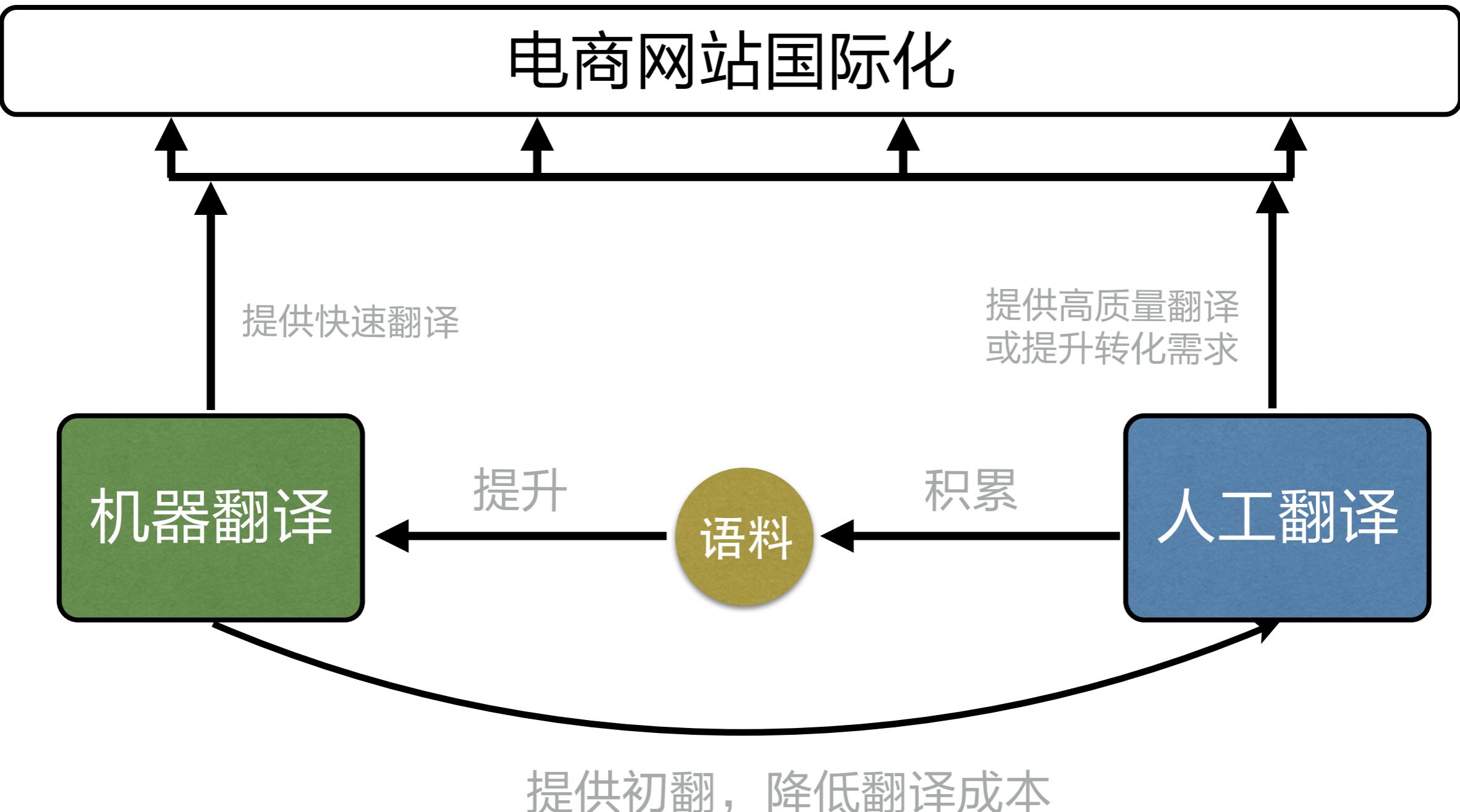
量级小

精度要求高

机器翻译

人工翻译

如何进行翻译



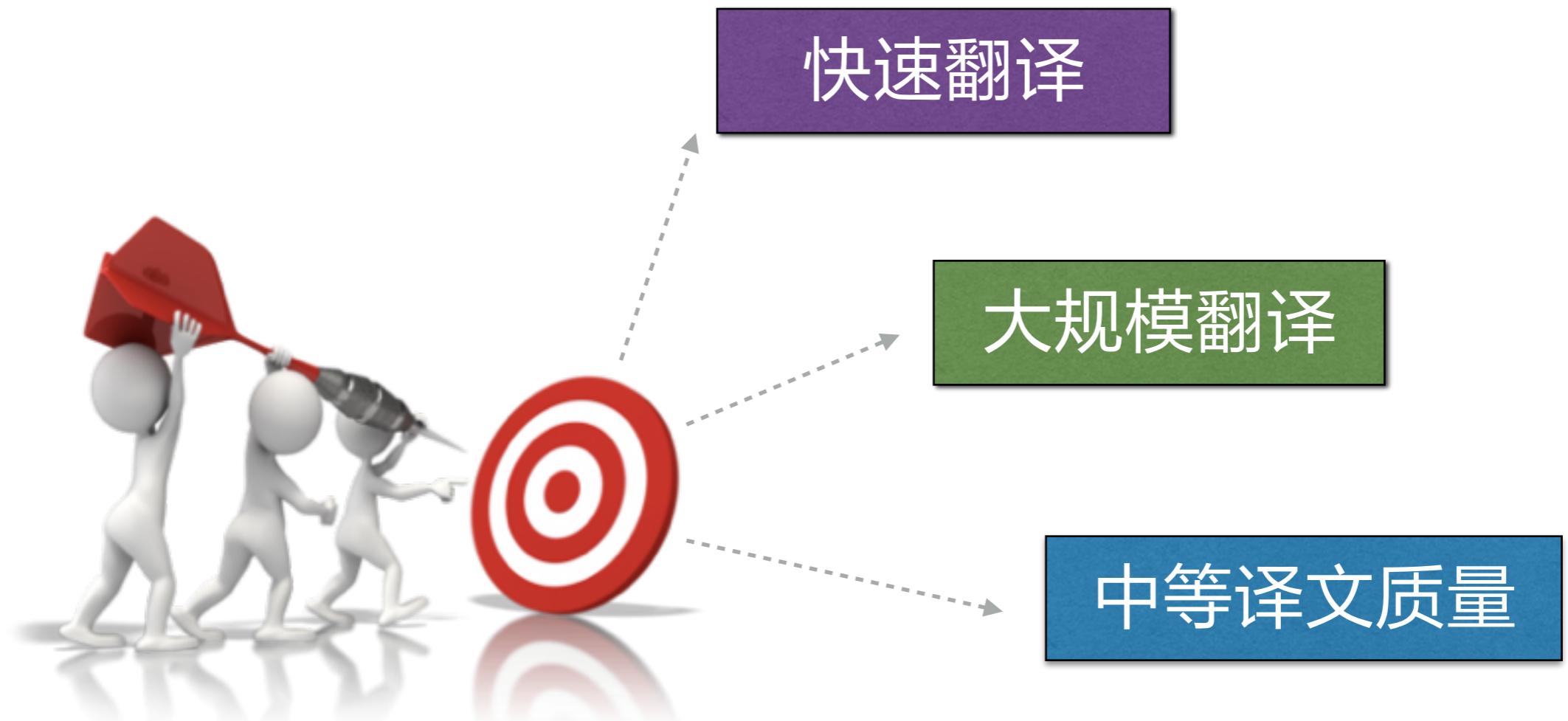
目录

- 1 阿里巴巴电商国际化
- 2 机器翻译技术
- 3 人工(众包)翻译技术
- 4 经验总结

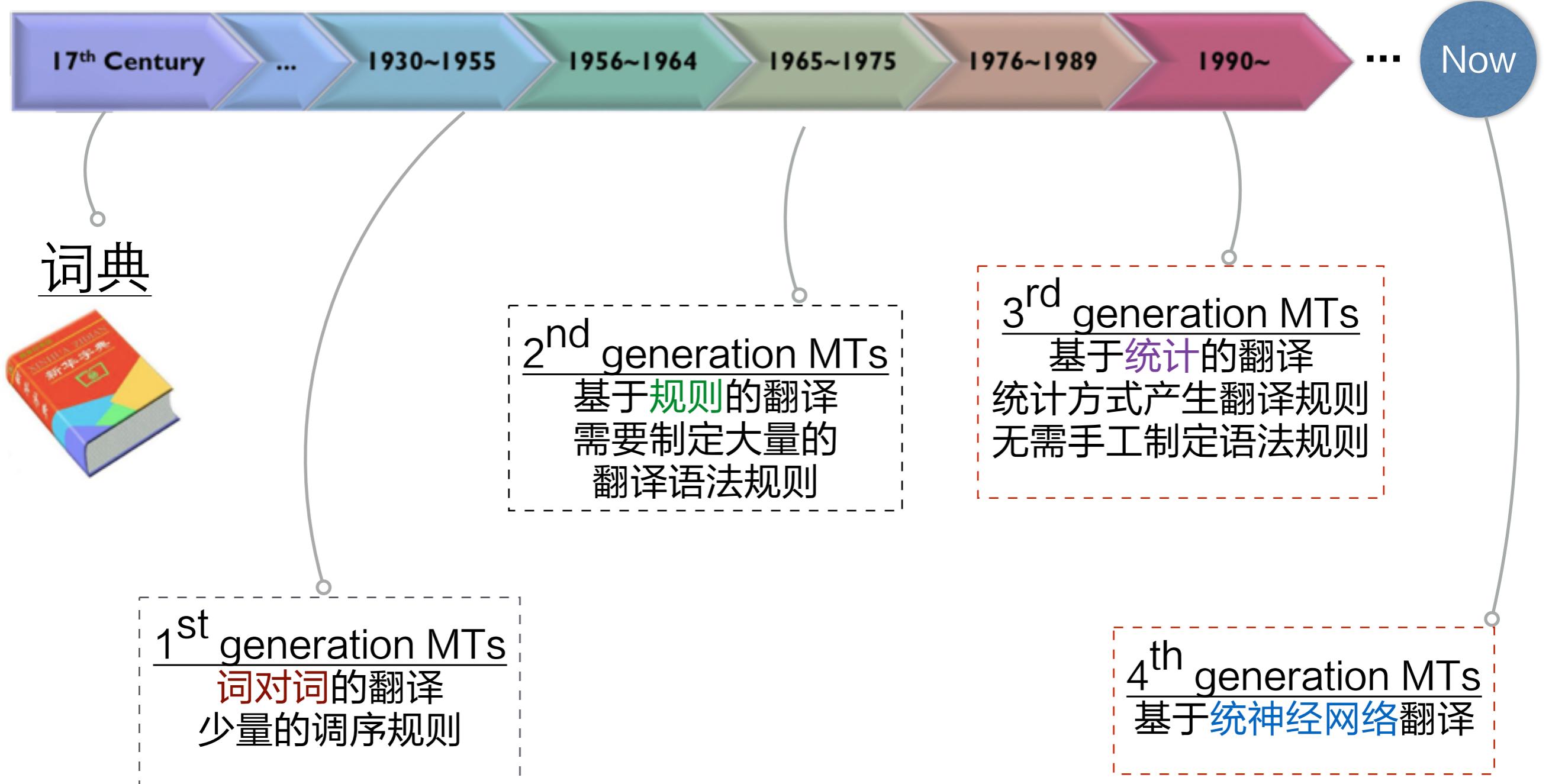
机器翻译，你怎么看？



机器翻译能做什么？



主流机器翻译技术



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 .

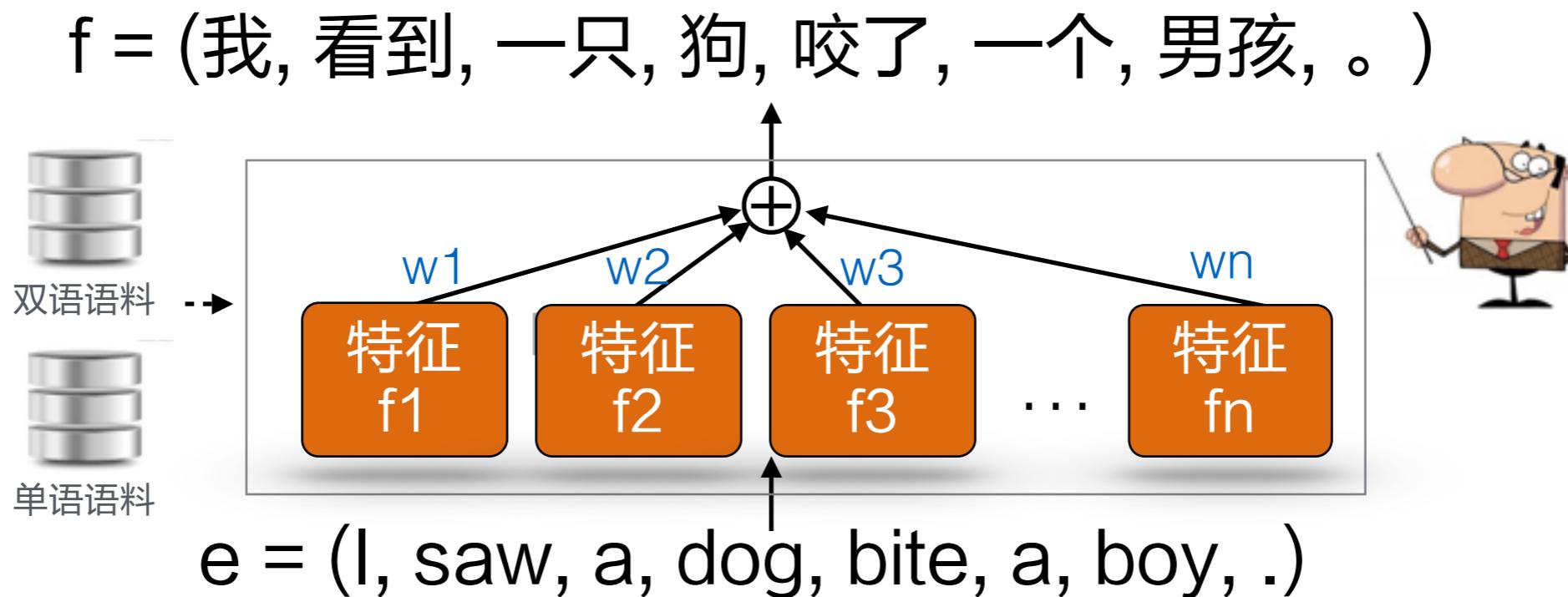
| | |
|---------|--------|
| 我是一个男孩。 | 0.5634 |
| 我是一个小子。 | 0.2325 |
| 我是一男子。 | 0.1231 |
| 我是一个小伙。 | 0.0233 |
| ... | ... |

搜寻空间

寻找概率最大候选翻译

概率计算

统计机器翻译



- 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

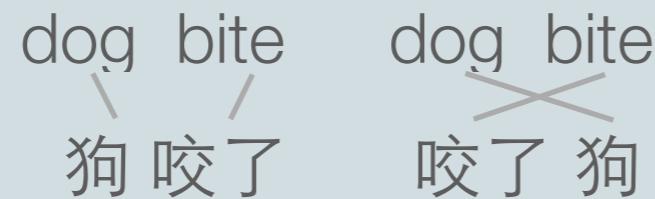
翻译模型



语言模型

$$p(s) = p(\text{狗} \mid \text{一只}) \times p(\text{咬了} \mid \text{狗}) \dots$$

调序模型



MERT调参

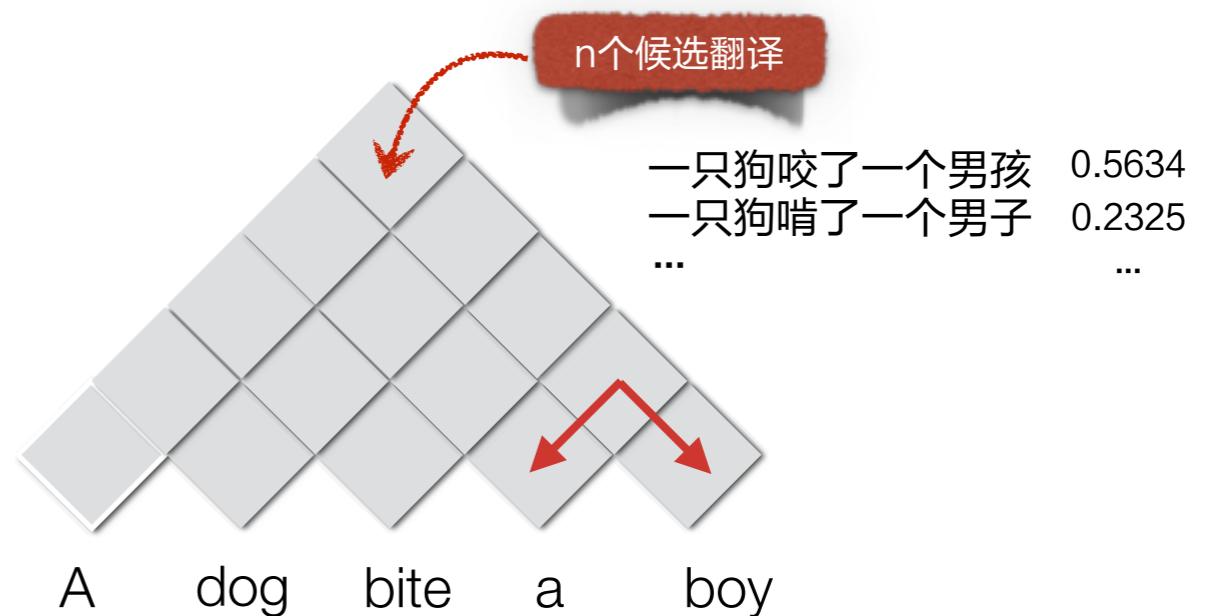
翻译模型 语言模型 调序模型

寻找“最优”权重： 0.53, 0.12, 0.31

翻译(解码)

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

权重
特征



神经网络机器翻译

$f = (\text{我, 看到, 一只, 狗, 咬了, 一个, 男孩, 。})$

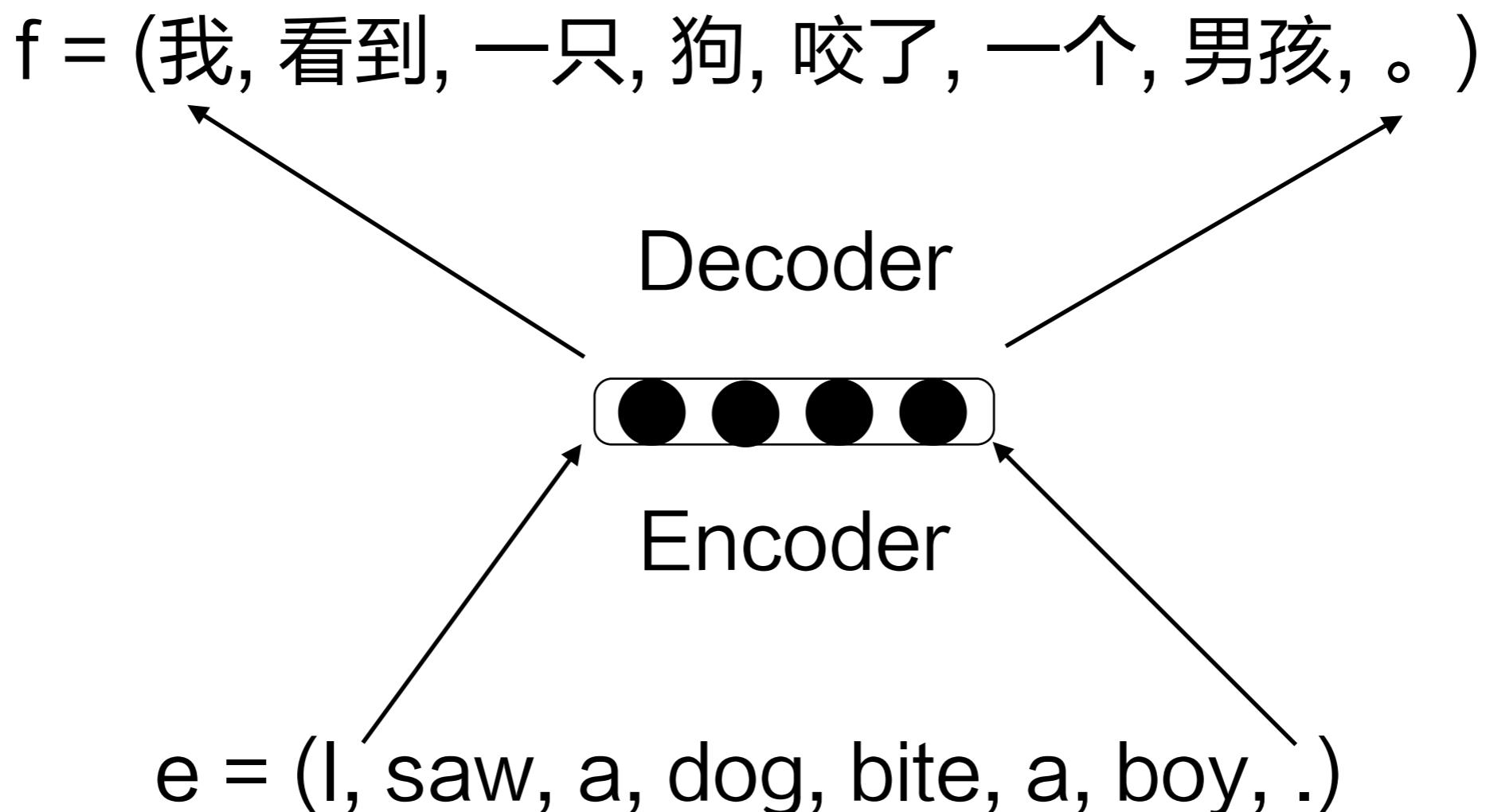
: 语言生成



: 理解

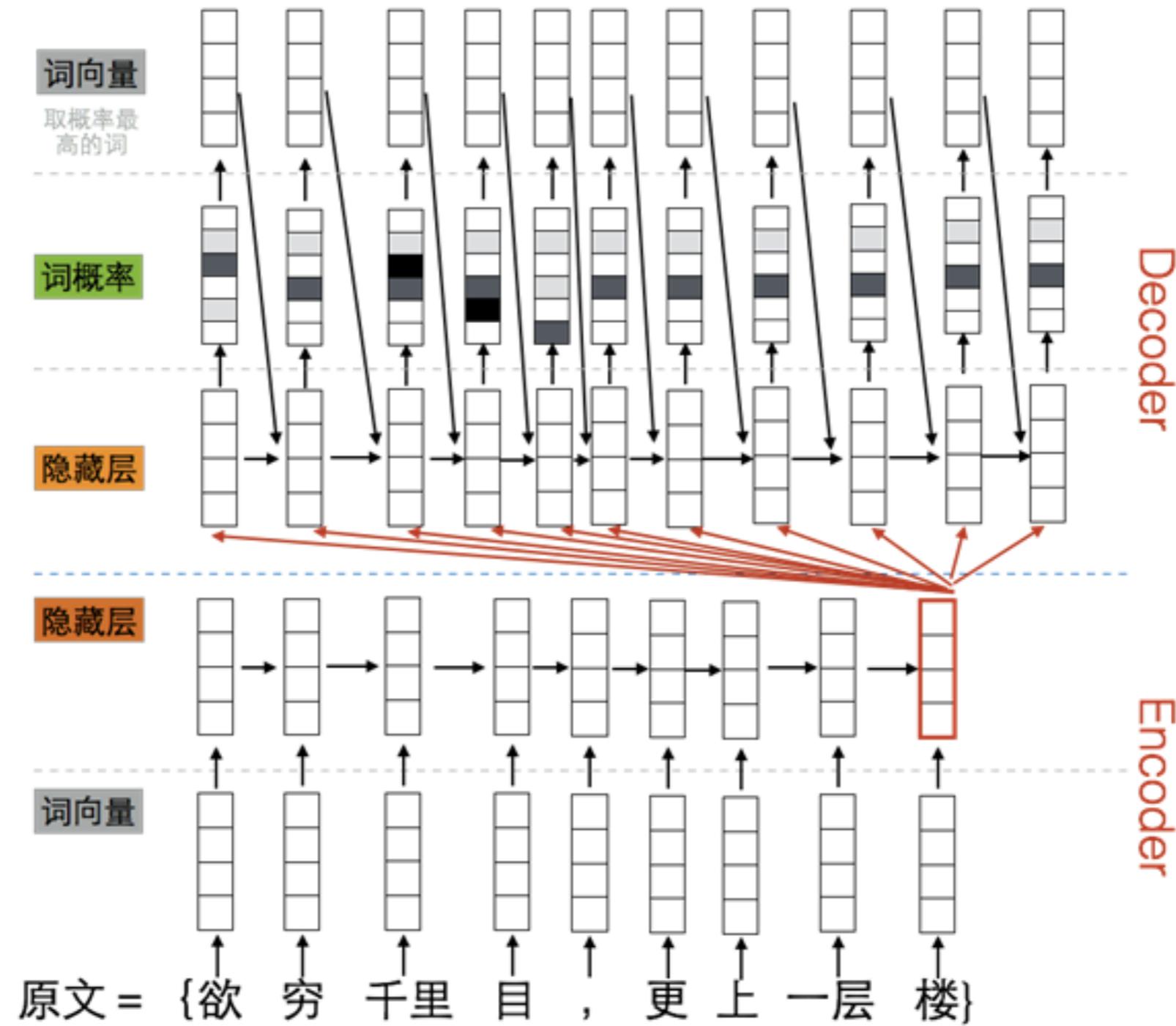
$e = (\text{l, saw, a, dog, bite, a, boy, .})$

神经网络机器翻译

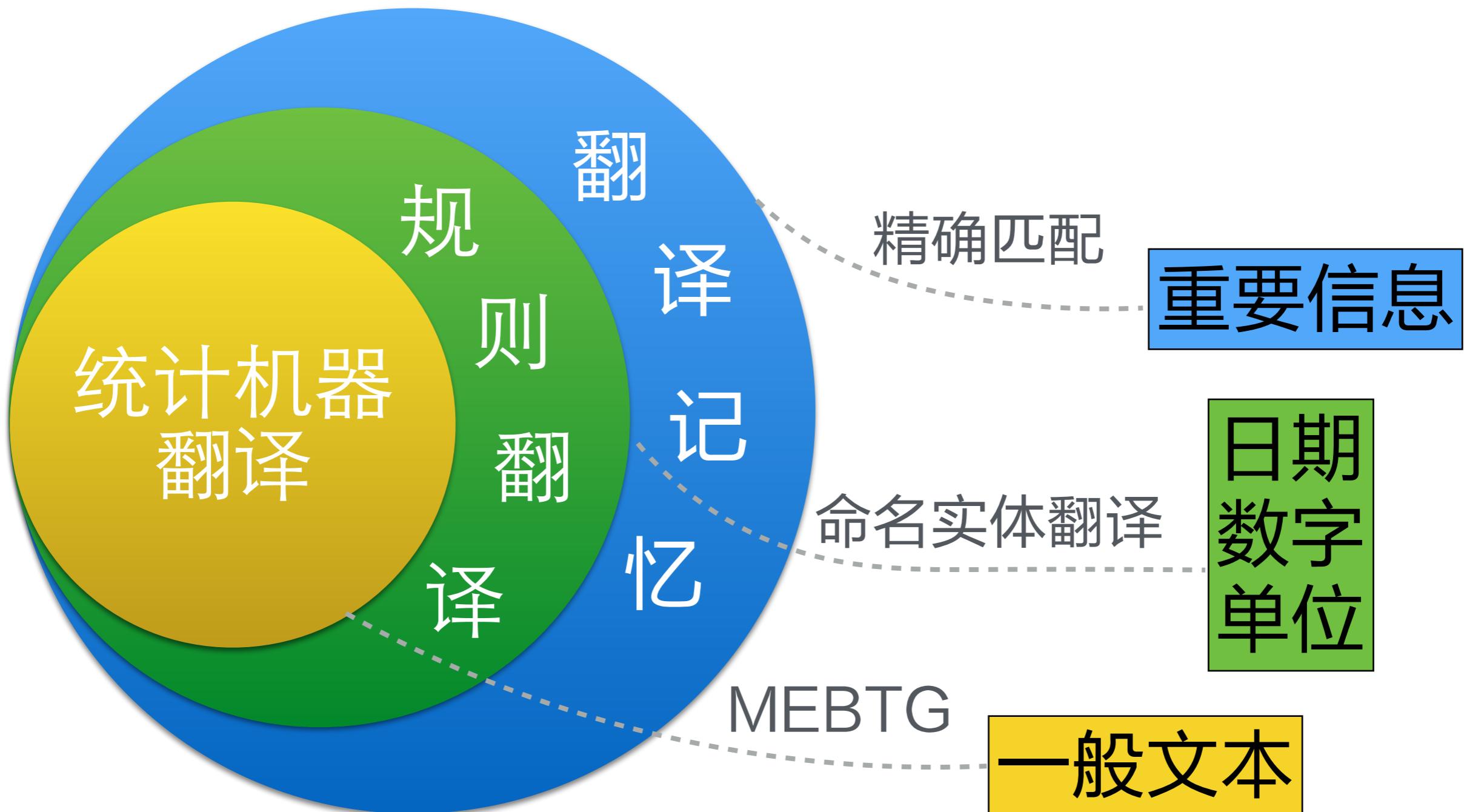


神经网络机器翻译

译文 = {Ascend Another Storey to Have a Further Sight retrospect and prospect}

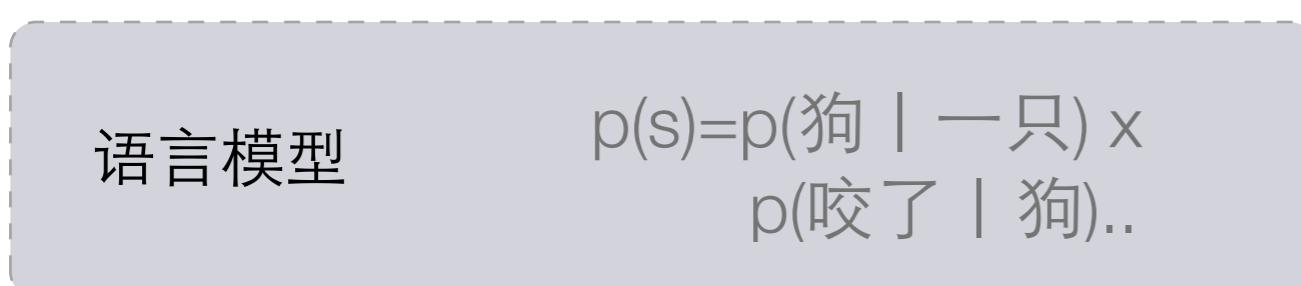


面向电商的阿里机器翻译引擎



面向电商的阿里机器翻译引擎

训练

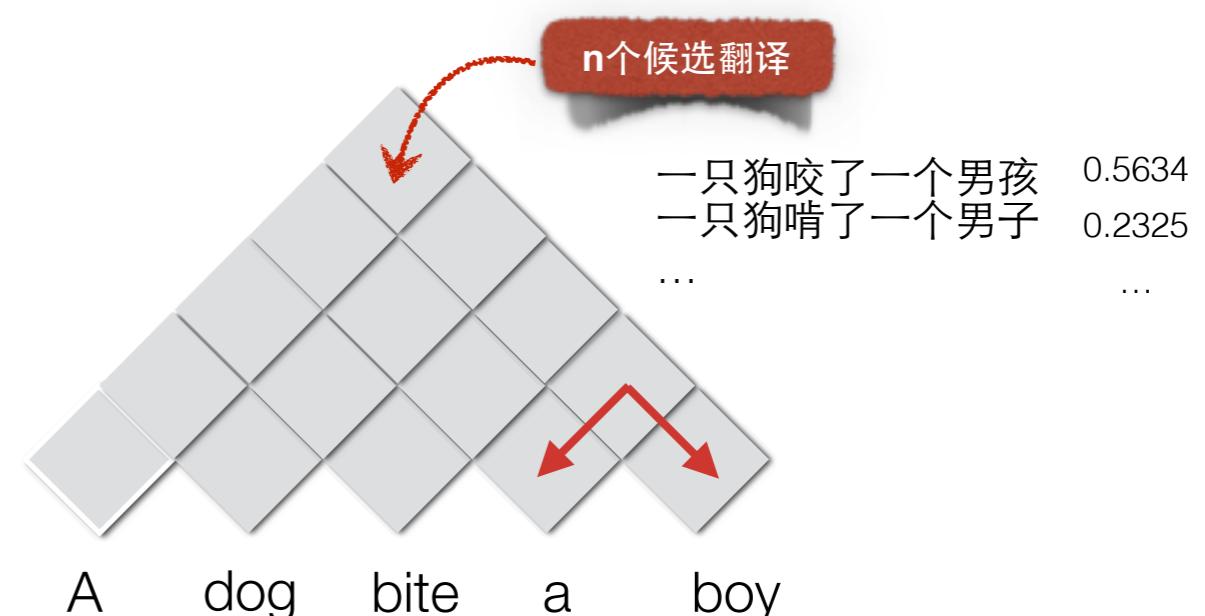


翻译(解码)

译文 原文

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

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**

电商专业词表

电商高频短语翻译

通用领域**双语语料**

电商领域的**单语语料**

电商品牌词表

通用领域**单语语料**



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

行业化高质量双语

行业化高质量双语

高质量数据

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

原始数据



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

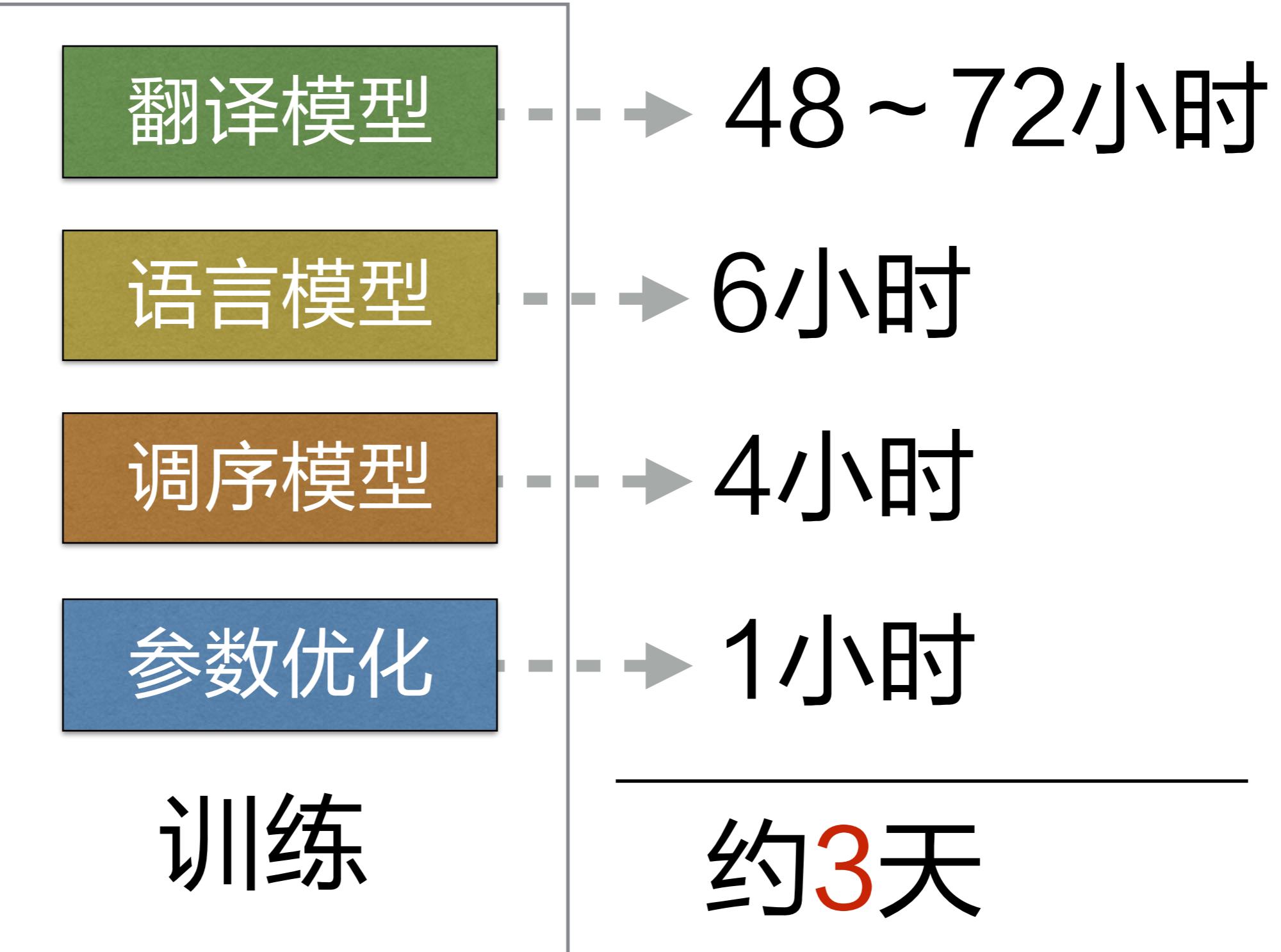
机器翻译人员最大的一项技能...



等待

原来....

语
料



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

3天

调序模型

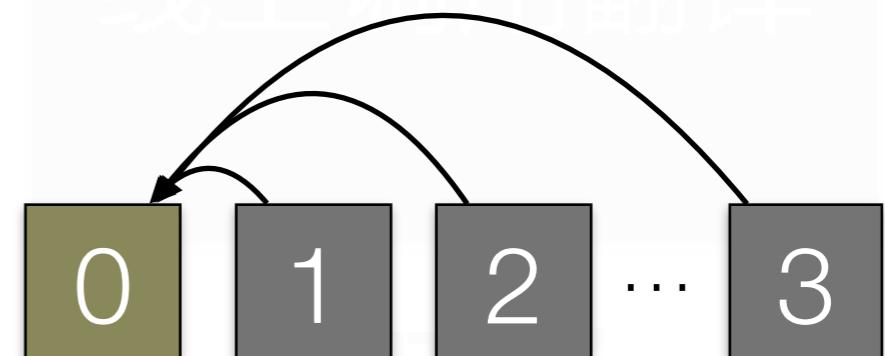
6小时

训练

$$\vec{\theta}^{t+1} = \vec{\theta}^t + \Delta_f \vec{\theta}(\mathcal{D})$$

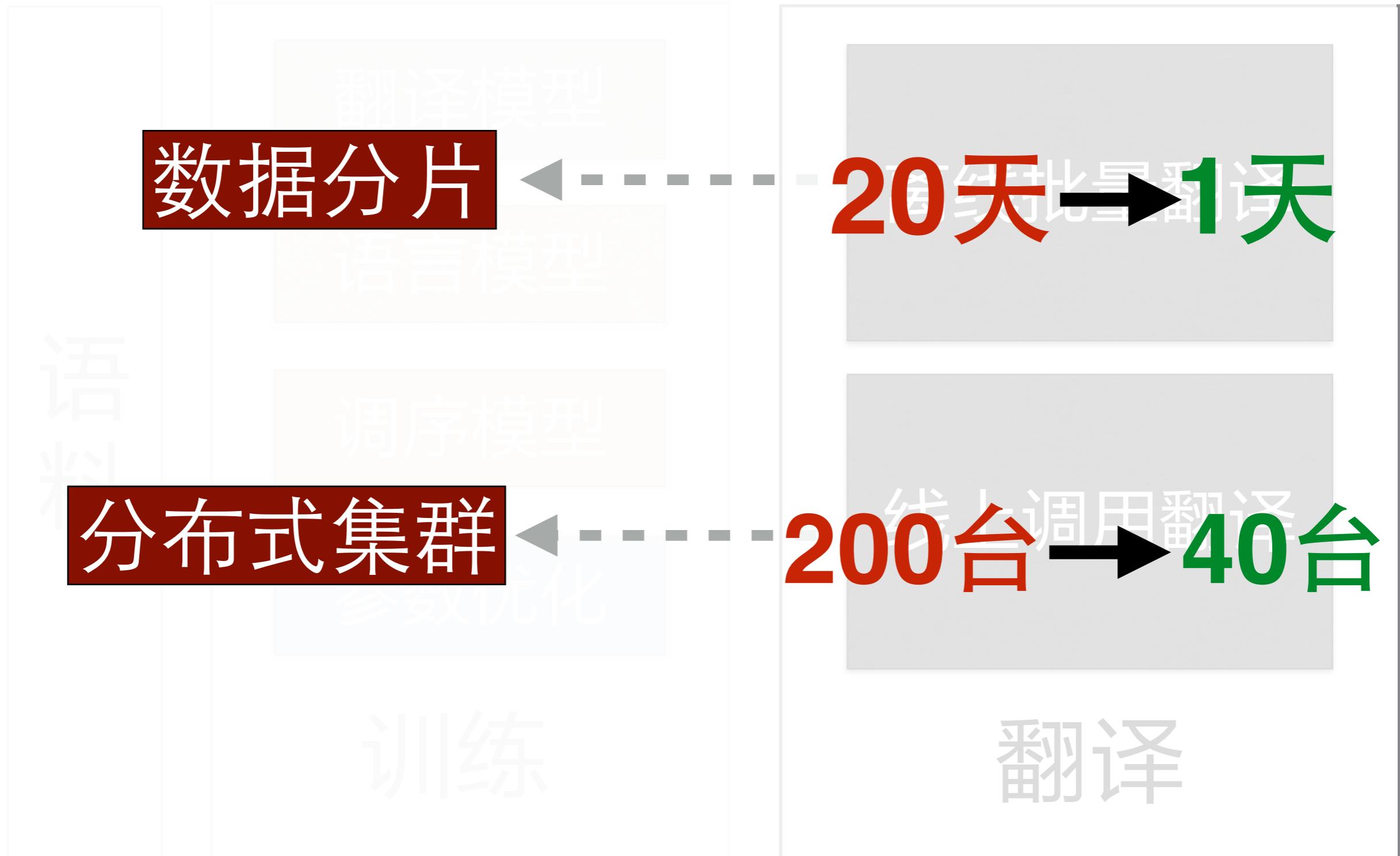


MR实现方式



BSP实现方式

阿里翻译在云端



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

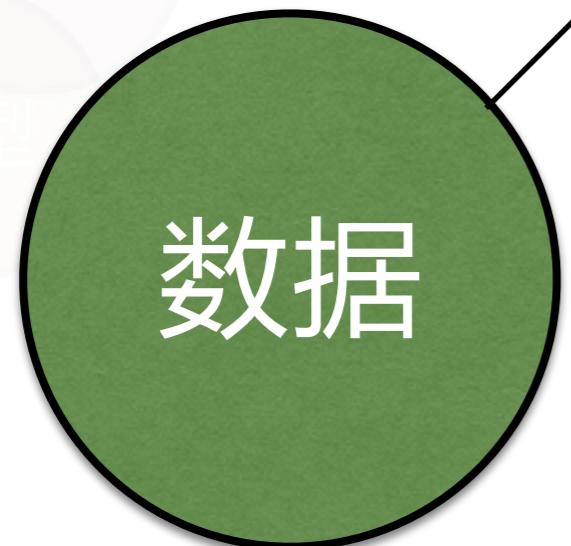
如何适应电商领域翻译？

电商领域翻译

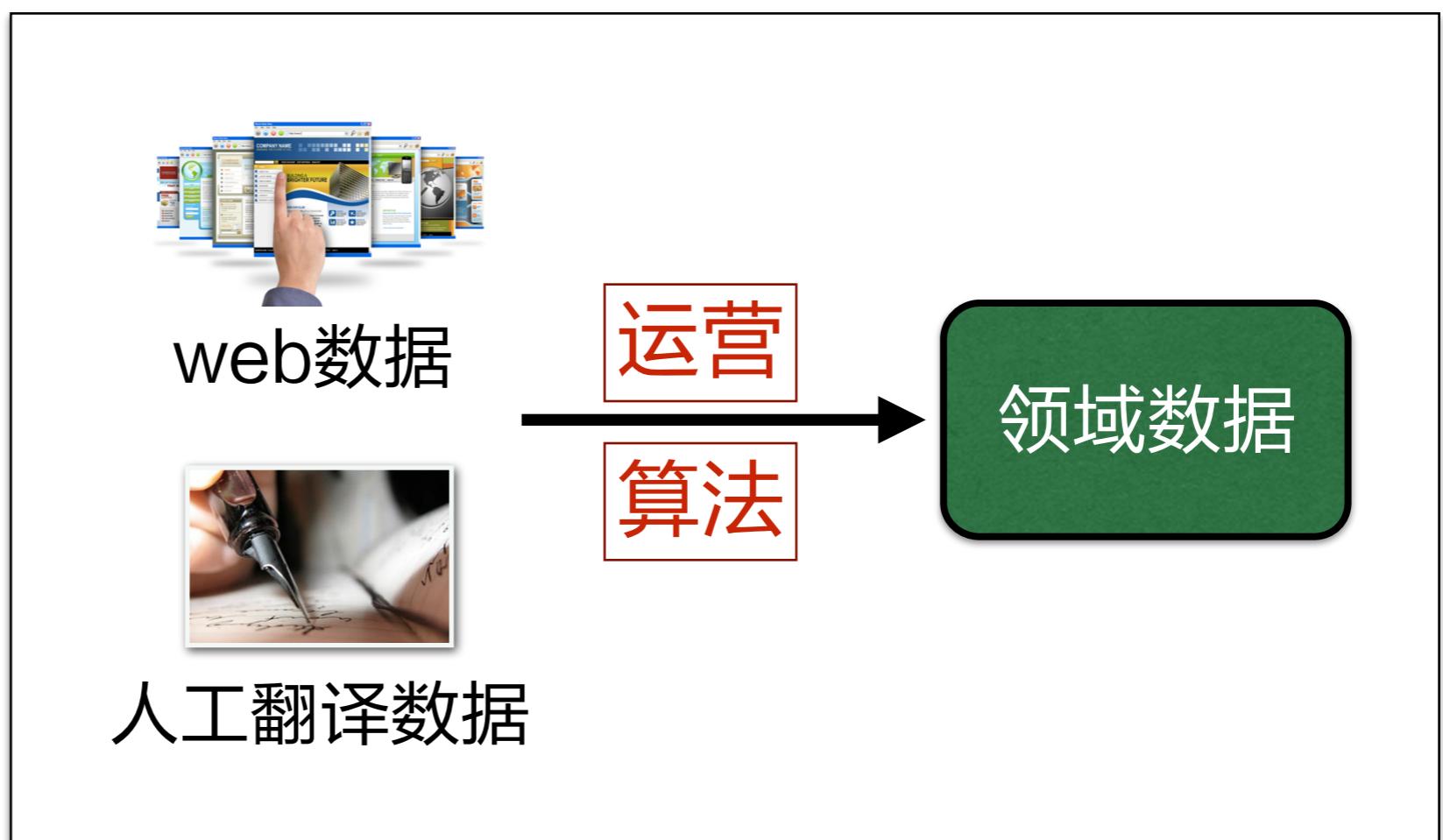


适应电商文本翻译技术体系

电商领域翻译



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



电商领域翻译



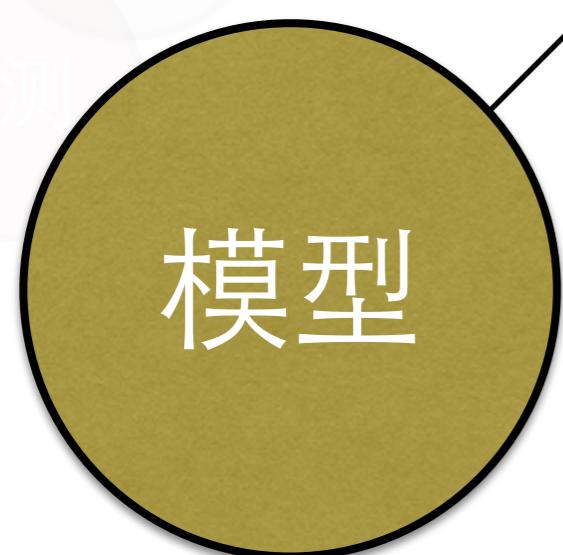
电商原文优化
领域特征自动挖掘方法
添加领域特征

The screenshot shows a product listing for a smartphone. At the top left is a small image of the phone with the text "eBuy Team 5.5in Screen". To its right are two small text boxes: "Domestic Delivery 7-Day Easy Returns" and "Original 5.5in Android 4.4.2 MTK6572 Dual Core Mobile Phone RAM 512MB ROM 4GB Unlocked Dual SIM Camear WCDMA GPS QHD NX N720 Free Shipping". Below the screenshot is a list of four issues with arrows pointing to a central box labeled "低质量电商原文改善与翻译".

- 原文拼写错误多
- 品牌词、型号词堆砌
- 促销词汇堆砌
- 专业词汇多

低质量电商原文改善与翻译

电商领域翻译



电商文本优化
领域特征自动挖掘方法
添加领域特征

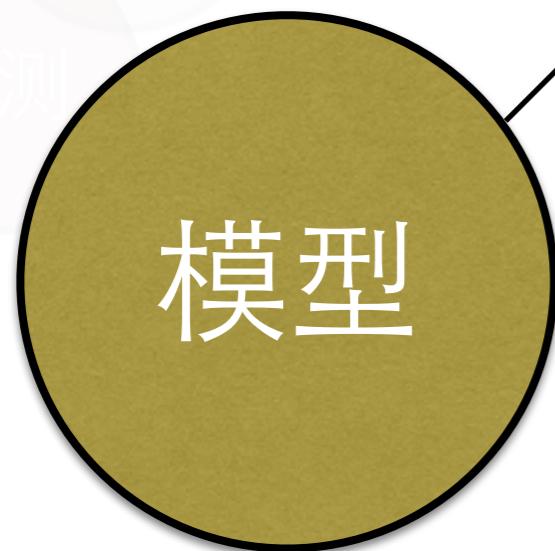
是否含hot word
是否含query
关键词的位置
...

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



电商文本优化
领域特征自动挖掘方法
添加领域特征

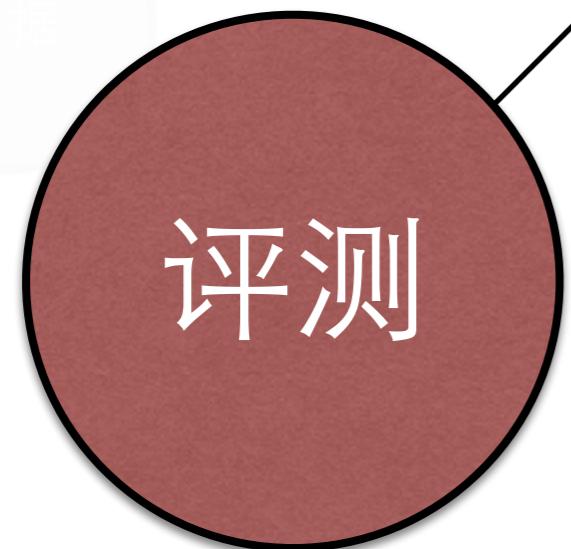
不同行业（类目）专业词的翻译错误

原文: ...black nuts (黑螺母)...

译文: ...Черный орехи (黑核桃)

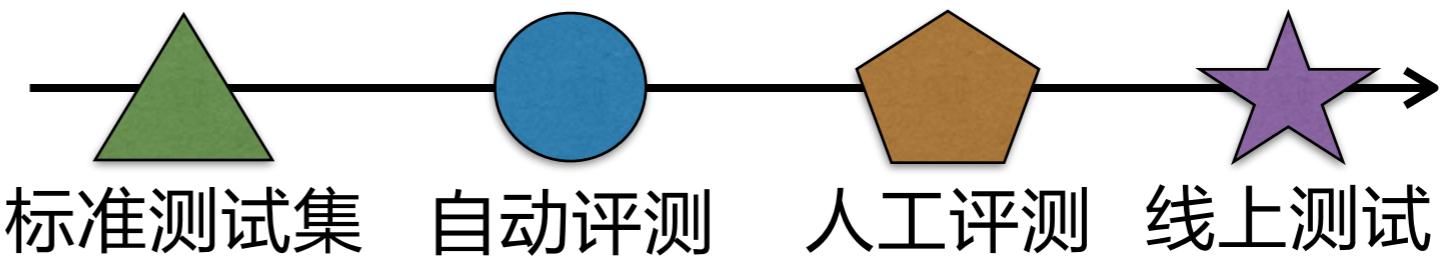


电商领域翻译



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

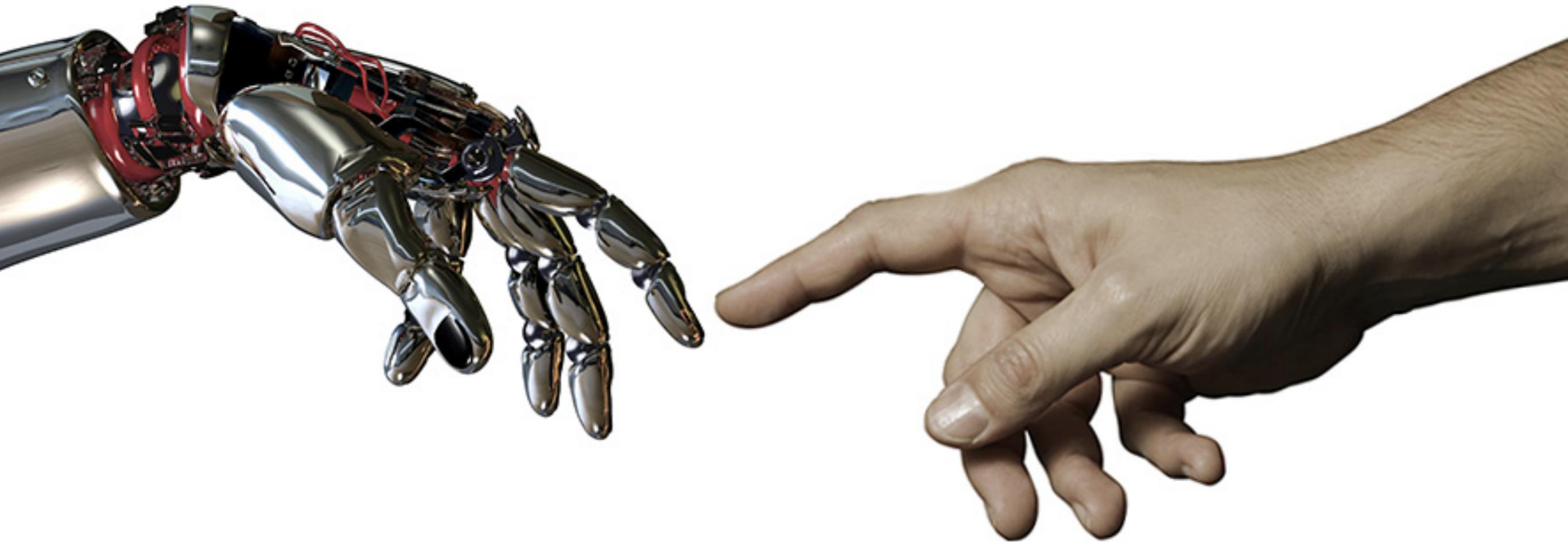


目录

- 1 阿里巴巴电商国际化
- 2 机器翻译技术
- 3 人工(众包)翻译技术
- 4 经验总结



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



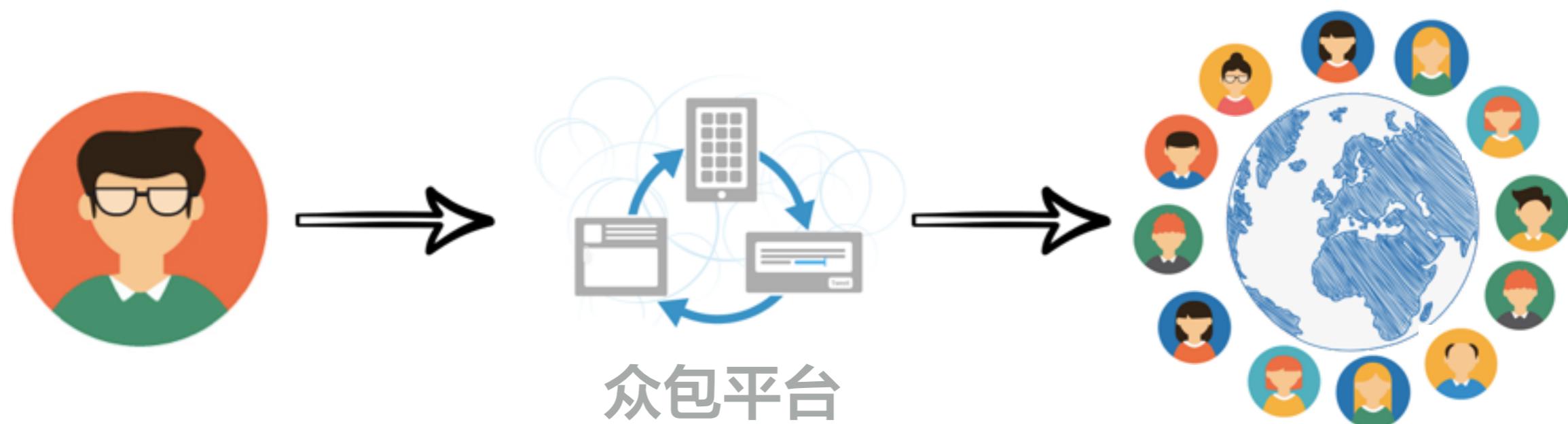
翻译效率



众包翻译

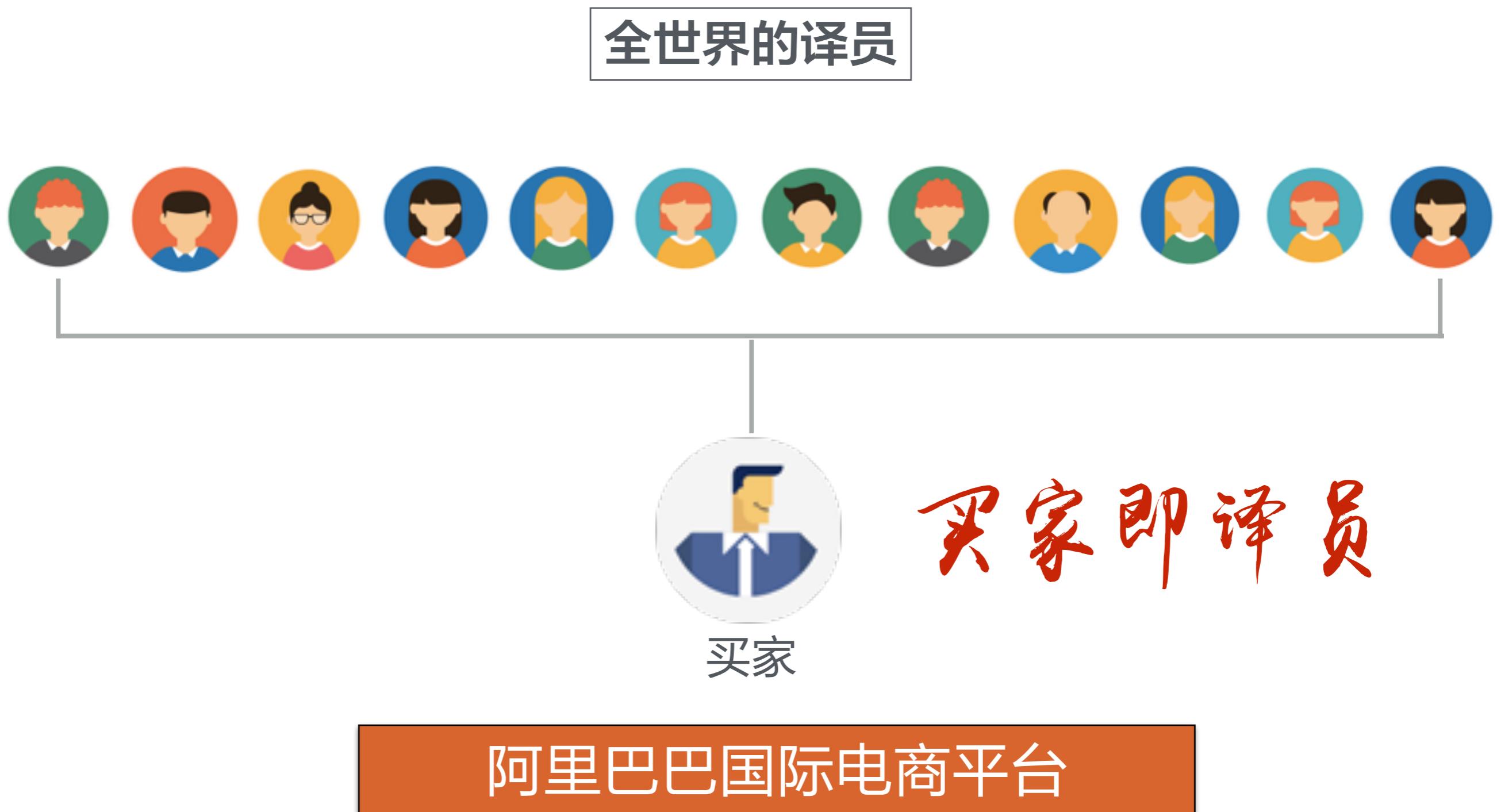
翻译需求

全世界的译员



“独翻译，不如众翻译”

众包翻译



众包翻译



[Back to AliExpress](#) | [Hi test, Sign Out](#) | English

Crowdsourcing translation platform

My Tasks

| Translation | Source Text: (English) | Translation: (Spanish) | Time Left: 23 hours 59 minutes | Submit | Skip |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------|------------------------|--------------------------------|--------|------|
| | Bead piece sequins embroidery patch Pineapple ice cream following deserve to act the role of clothing decorative laminated bag | | | | |
| See Product Page | | | | | |



买家

阿里巴巴国际电商平台

众包翻译



[Back to AliExpress](#) | [Hi test, Sign Out](#) | English

Crowdsourcing translation platform

My Tasks

| Vote | Source Text: (English) | Translation: (Spanish) | Time Left: | Action Buttons |
|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | Rear Foot Peg Footrest for BMW K1200R 2004-2008 & K1200S 2003-2008 & F800R 2005-2013 & R1200S HP2 2004-2006 | Задняя ножка Peg подножка для BMW K1200R 2004 - 2008 и K1200S 2003 - 2008 и F800R 2005 - 2013 и R1200S HP2 2004 - 2006 | 16 hours 20 minutes |  Good  Bad  Skip |

 See Product Page

 My Tasks

 Task Management

 Points & Ranking

 My Account

 Help Center



买家

阿里巴巴国际电商平台

目录

- 1 阿里巴巴电商国际化
- 2 机器翻译技术
- 3 人工(众包)翻译技术
- 4 经验总结

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