企业级案例丨深度学习在网易严选智能客服中的应用

（原文链接）

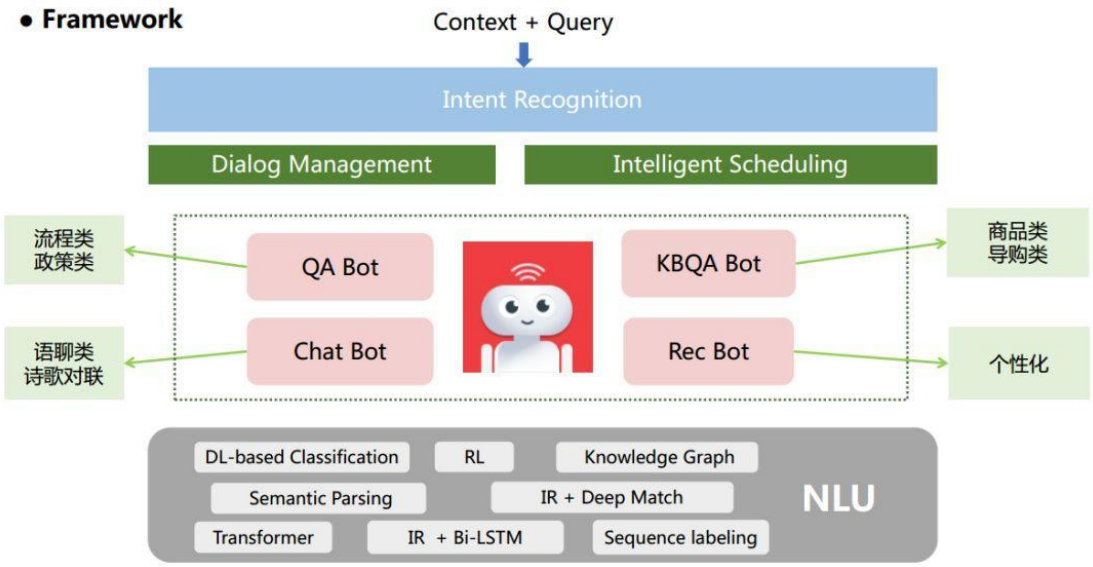
业务概要

随着自然语言处理技术的发展，智能客服作为电商领域内重要的业务场景，近年来受到的重视日益增加。因为在购物的全链路过程中，用户碰见问题或疑惑时都可能会转向客服，去寻求一些咨询或者支持。客服精准有效的回复，不仅会直接影响到用户的体验，也会对购买转化产生正面影响。如：

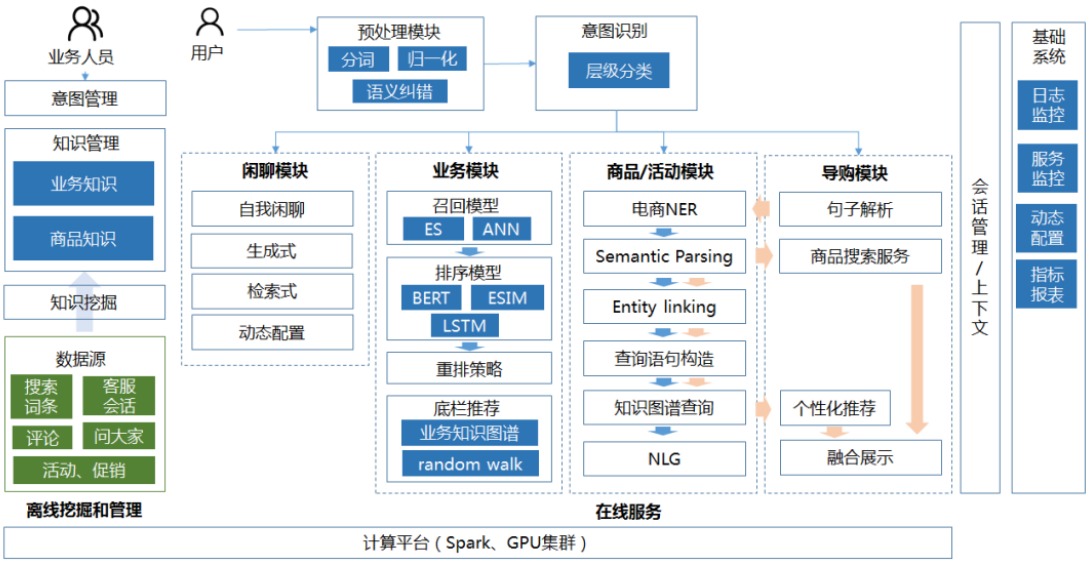
- 售前场景下，用户会针对感兴趣的商品或促销活动进一步提问
- 售后场景下，用户会询问与退换货、邮费、物流相关的问题

在网易严选业务实际运转时，会产生和沉淀大量的商品属性、活动运营、售后政策等多维度知识，并有对应的复杂逻辑。智能客服即是依赖这些数据信息，自动化或者辅助人工客服解答用户问题的智能对话系统。

但电商领域的业务点细多繁杂，与此同时用户的输入具备口语化、多样化的特性，这些难点为语义的正确理解带来了新的挑战。为此我们在通用的客服场景下，结合网易严选的业务，设计了如下的一套系统框架，如下图所示。



网易严选智能客服框架图

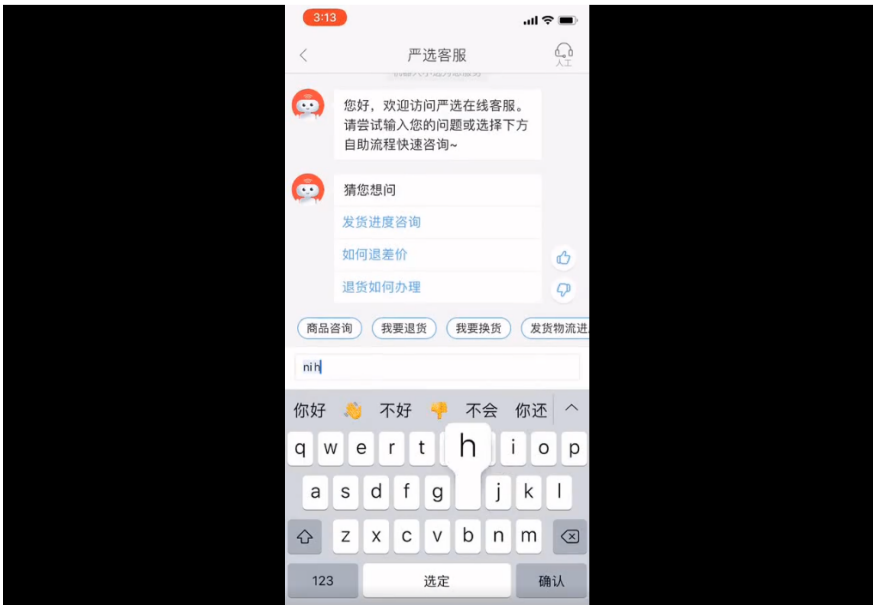


（网易严选智能客服架构图）

- 当用户实时输入时，输入的文本会与上下文信息首先一起送入意图识别模块
- 意图识别模块准确的解析出用户的多层次意图，进而分发到不同的子模块
- 不同的子模块负责更具针对性的业务问答，不同的子模块也采用了不同的技术选型

可以看到，在整个框架中，我们使用了很多个深度学习模型。因为依靠 NLP 领域中的深度学习模型，我们可以获取到更泛化、多粒度的语义信息。同时成熟的 TensorFlow 工具为业务的落地开展提供了一整套的方案，包含模型构建、训练并部署至线上使用。

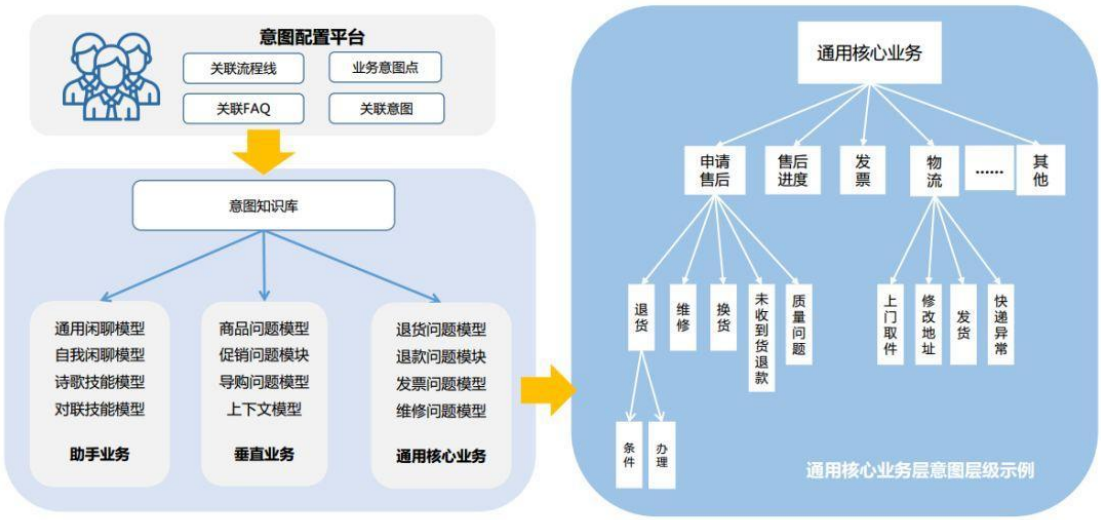
下图给出了真实对话场景下，小选机器人的回复效果示意图。项目中涉及到与深度学习相关的技术点，以下将分模块来作以介绍。
（线上会话示例）



模型构建

意图模块 — 多层级分类模型

伴随着用户输入的文本，我们基于当前输入、上下文语境以及用户的历史行为，利用 TensorFlow 构建了多层级多分类的意图识别模型。具体的一级意图可分为四大类：售前类商品问题、售后类问题、语聊问题、或其他类。在售后类问题中，针对核心常见的政策问题，归纳总结出更细致的子层级意图。下图给出了意图业务的示例。



意图层级关系示例

意图识别本质上可以视为分类问题。在搭建分类体系时，我们采用 ABL（Attention Bi-LSTM）模型结构作为初版 baseline，后续引入融入浅层语意特征和用户行为数据，并借鉴 Transformer 中的 positional encoding (位置编码) 与原始输入进行融合，最终使模型的准确率提升三个点。除此之外，我们还基于 BERT 模型，采用 fine-tune 的方式，在更小的标注数据集上训练分类模型，最终的效果与 ABL - based 模型相差不多。预训练模

型可以采用较少的样本就可以获得较好的泛化能力，减少人工标注的成本，但值得注意的是，这同时也需要付出更多的计算资源。

FAQ 模块 — 文本匹配模型 FAQ 业务问答是智能客服的核心功能之一，该模块由召回、重排两部分组成。

- **召回阶段**，采用了词粒度的离散检索，以及基于稠密句向量的语义检索
- **重排阶段**，通过 TensorFlow 构建了文本匹配模型，对召回的候选问答进行重排序
- **辅佐策略融合**，得到最终的答案

在自动问答领域，文本匹配技术通常被应用在句子相似度判别、问答语句相关性判别任务中。从最基础的 Siamese-LSTM 网络，到 InferNet、Decomposable Attention、ESIM，再到 BERT，匹配模型也经历了一系列的结构变迁。多数模型集中在两个维度发力：单句表征的 encode 方式，及利用各式 attention 捕捉句间的交互语义。

业务层面，我们采取了多样化的问句匹配方案：

1. 输入问题 Q 与答案 A 进行关联匹配；
2. 输入问题 *Q*1 与标准问题 *Q*2 进行相似匹配；
3. 输入问题 *Q* 与标准问题 *Qs* 的相似问题匹配。

三类方式从不同角度分别提供问句的相关性召回效果、问答关联匹配效果，在 match 及 rank 阶段可以用策略灵活的加权判别。

模型层面，我们搭建 Siamese-LSTM 作为 baseline 模型，后续模型迭代方案包括以下：

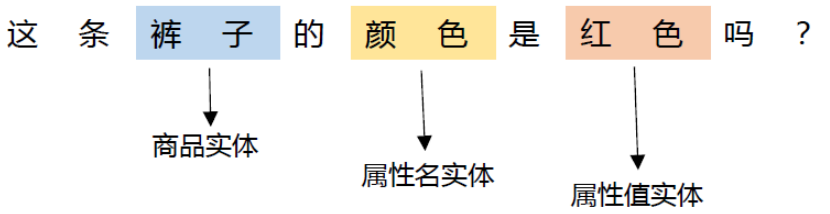
- 将 LSTM 子单元更换为 Transformer 的 encoder，并将余弦距离表征模块，替换为句子对向量特征 $(\vec{u},\vec{v},|\vec{u}-\vec{v}|,\vec{u}\cdot\vec{v})$ 接上 MLP 层
- 融合 ELMO 特征的 ESIM 模型
- 基于 BERT 的 fine-tune

实验下来，几个模型都有一定的提升，如 Transformer 的 encoder 最终在任务（1）和（3）中获取了更好的准确率，效果提升了近五个点。

另外我们发现，不需要额外的特征构造及技巧，BERT 便能获取稳定优秀的匹配效果。究其原因，BERT 在预训练阶段的目标任务之一就是预测两个句子之间是否是上下文的关系，可以学习到句间关系的知识；此外，自注意力机制更加擅长捕捉深层的语意，可以获得句子 A 中单词和句子 B 中任意单词的细粒度的匹配结果，这在文本匹配任务中是至关重要的。

KBQA 模块 — NER 模型

在商品知识问答及导购模块中，我们基于 TensorFlow 构建了电商领域的命名实体识别（NER）模型，用来识别用户问句中的商品名、商品属性名、商品属性值等关键的商品信息（如下图所示），并将实体词送至后续模块，结合知识图谱问答技术生成最终的答案。



（电商实体识别示例）

NER 算法常用的模型是双向的 LSTM 加上条件随机场 CRF，前者可以抓取对话文本的前后特征、理解语境、充分提取上下文信息，后者则注重于当前对话文本的局部特征与全局特征构成的概率转移，有效挖掘当前对话文本的语义信息。网易严选基于 Bi-LSTM 加 CRF 构建的词粒度的 baseline 模型，服务于线上的智能客服系统。后续通过实验尝试了基于 BERT 的特征集成（feature extraction）和微调（fine-tune）的模型。

使用方式	Precison	Recall	F1	单条响应时间
特征集成 Bi-LSTM + CRF	0.97	0.88	0.92	> 100ms
微调 多层特征融合	0.94	0.88	0.9	< 10ms
微调 高层特征	0.94	0.84	0.88	< 10ms

基于 BERT 模型的 NER 效果

在网易严选数据集上，得到的实验效果如上表所示，从中我们获取以下经验：

- 特征集成的方式效果优于 fine-tune。在特征集成模型中，在 Bi-LSTM 抽取的语义及结构信息之外再引入 BERT 的特征等，使我们拥有更为丰富的语义及结构表征。附加额外参数带来的效果增益（即特征集成的方式）要明显大于普通的 fine-tune。
- 多层特征融合的效果优于仅使用高层特征。原因是对于序列标注的任务，我们不仅需要考虑语义的表征，还需要融合句子其它粒度的表征，如句法结构信息等。
- 在响应时间上，特征集成（附加额外参数）的方式满足不了线上的需求，适用于离线服务，而微调的模型可以满足线上服务时效的需求。

语聊模块 — 生成模型

作为一个独立完整的客服机器人，帮助用户解答疑难问题至关重要，同时，也需具备闲聊的能力，才能更加彰显机器人人性化智能化的一面。

我们为小选机器人赋能，构建了负责日常闲聊打趣的语聊模块。语聊模块的核心模型有两个：检索式 QA 及生成式 QA。

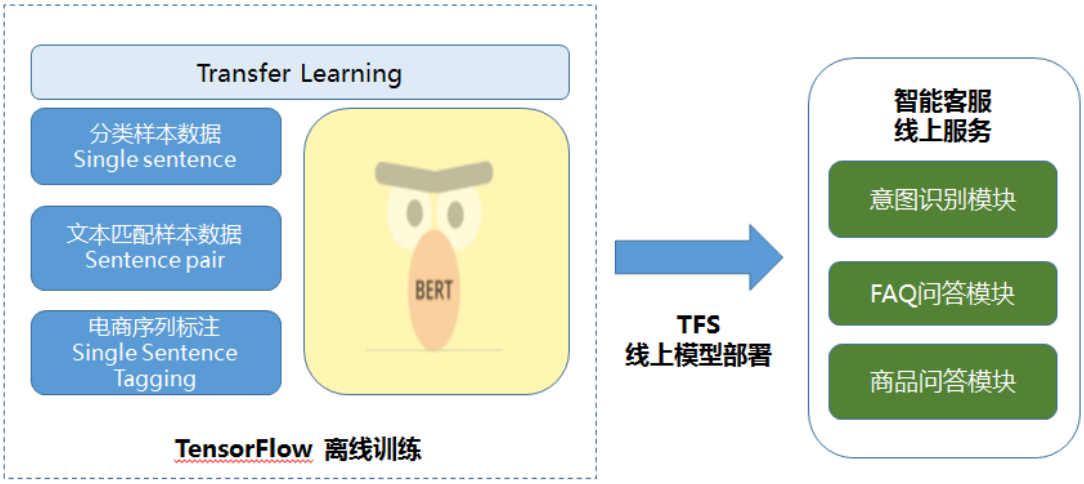
- 检索式 QA，先从预备好的语料库中进行召回，再利用文本匹配模型进行 answer 的重排序；

- 生成模式 QA，直接采用 TensorFlow 中的 tensor2tensor 模块训练 Transformer 生成模型，通过端对端的方式生成用户的回复。

但由于纯端对端的方式生成的回复不利于控制，线上服务中我们最终采纳了两者模型的融合，保障更可靠的回复。

模型部署

下图为基于 BERT 模型的线上业务流转示意图。得益于 BERT 等语言模型开源的 TensorFlow 版本，在实际业务中，可仅利用少量的标注样本，有效的搭建高准确率的各类文本模型，并利用 GPU 加速计算满足线上 QPS 的要求，最后基于 TF Serving 快速部署上线。正是因为 TensorFlow 提供的一系列支持，我们才能稳健高效的进行线上业务的部署与迭代。



（基于 BERT 模型的线上业务流转示意图）

总结

随着深度学习技术的发展，在自然语言处理领域，新的突破性模型一直层出不穷。学术界新兴的研究成果，能够平滑的应用至工业领域中，并获取优秀的业务成果，这一切都离不开 TensorFlow 的功劳。在网易严选业务场景下， TensorFlow 提供了灵活细致的 api，供给算法人员敏捷开发尝试各类模型，极大地方便了算法模型的迭代工作。

- **除了上述案例，你还可以阅读 TensorFlow 的其他案例：**

  想了解 TensorFlow 的前端应用，可以阅读《前端案例丨零基础也能在小程序上实现机器学习》，了解机器学习如何与小程序进行结合。

如果你希望加深了解 TensorFlow 在端侧的应用，建议阅读《移动端案例丨使用 TFLite 在移动设备上优化与部署风格转化模型》，我们在其中分享了 TensorFlow Lite 针对移动设备优化的预训练风格转化模型，你可以依据模型创造出自己的应用；

同时，也欢迎大家用微信端打开 TensorFlow 案例库，了解更多精彩案例。
配合官网阅读，体验更佳：**https://tensorflow.google.cn/**

阅读案例以后，你可以通过以下方式持续进阶：

- **你还可以加入 TFUG 社区，认识更多优秀开发者，在社区中进步。**
  TFUG，欢迎你的加入！

  我们为专业的 TensorFlow 开发者提供正式认证和证书，它不仅能够证明你的学习能力，同时也助力你的职业发展点亮 LinkedIn 技能。

- **关注 TensorFlow 官方微信公众号，回复“认证”，即可获得《 TensorFlow 开发者认证候选人手册》，助你在机器学习道路上更进一步：**

期待你顺利踏上 TensorFlow 之旅！