# 1. 课题来源、选题依据、课题研究目的、工程应用价值

## 1.1 课题来源

本人在苏州意能通信息技术有限公司实习期间，发现对于用户意图识别功能，存在着用户口语表达文本集合与意图识别模型的训练文本集合存在一定差异，从而导致意图识别准确率下降。针对上述问题，本人希望利用文本蕴含关系识别的算法，通过识别用户表达的句子与目标句子的蕴含关系来检验两者之间是否有蕴含关系，以提高识别准确率。

## 1.2 选题依据

随着自然语言处理(Natural Language Processing, NLP)领域研究的不断深入，学者们的研究已经不只停留在词法解析、句法解析等表层信息的处理上了，让机器理解语义正渐渐成为学者们的研究目标，而自然语言推理作为自然语言理解的重要组成部分，也成为自然语言处理的一项研究热点。

自然语言推理(Natural Language Inference, NLI)通常是指识别文本间的蕴含关系，所以通常也称为文本蕴含关系识别(Recognize Textual Entailment, RTE)。

## 1.3 课题研究目的

## 1.4 工程应用价值

文本间的蕴含关系普遍存在于自然语言文本之间，所以文本蕴含关系识别在许多自然语言处理领域都存在着实际应用价值。本节将列出一些文本蕴含关系识别的应用[1]。

**问答系统** 一个理想的文本蕴含系统可以作为问答系统的答案验证模块，用来判断一个从知识库种抽取的候选答案是否能够推断出目标问题或是目标问题的陈述句形式。例如，某人提问“港珠澳大桥的总工程师是谁？”，通过信息检索技术从知识库中检索出候选答案“林鸣，自2010年12月起，担任港珠澳大桥岛隧工程项目总经理、总工程师”，利用文本蕴含系统可以验证是否可以从候选答案推断出目标问题。为检验文本蕴含系统的这种能力，欧洲跨语言评测平台(Cross Language Evaluation Forum, CLEF)在2006-2008举办了三届答案验证评测(Answer Validation Exercise, AVE)。在评测中，每一条数据由一个问题、一个候选答案和一篇支持文本组成，文本蕴含模型的目标就是在给定支持文本的情景下，针对目标问题，判断候选答案是否正确。

**语义检索** 语义检索的目标是基于检索语句的语义从大量的文档库中检索出目标文档的目标语句。如果将文本蕴含系统运用在语义检索中，它可以用来识别目标问题与目标源文档中各个语句的语义相似性。例如，用户搜索“人们示威反对自由贸易”，现存的基于关键字的搜索引擎将会返回包含“示威”、“自由贸易”等关键字的文档，如果文档中仅存在像“游行人员高举‘反对解除贸易壁垒协议’的标语”这样不含关键字的语句，那将不会被检索到，而造成信息遗漏，利用文本蕴含识别可以避免上述问题。

**自动摘要** 在自动摘要任务中，一个主要的挑战是怎样消除冗余。冗余现象在多文档摘要中特别明显，多文档摘要是融合了多篇源文档的内容，而多篇文档之间的内容可能存在语义相似的，从而造成自动生成的摘要中存在多条语句表达相似语义的问题。文本蕴含识别系统用来判断是否存在语句所表达的语义可以由其他摘要语句推理得到。自动摘要的另一个挑战就是正确性，也就是摘要应该准确地反映单个源文档或多个源文档的内容。文本蕴含识别系统可以通过判断源文档是否可以推理出摘要，以确保摘要的正确性。

**机器翻译的自动评估** 文本蕴含识别的一个相对比较新的应用就是自动评估机器翻译的性能。目前，机器翻译研究者主要利用BLEU作为翻译好坏的指标，BLEU是通过衡量机器翻译语句与人类翻译语句之间的重合程度来，并没有考虑语义层次的相似性。而文本蕴含识别可以判断机器翻译的语句是否可以推断出人类翻译的语句，若可以推断出，则可以认为这是一个好的翻译，即使机器翻译的句子与人类翻译的句子之间没有一个字是一样的。

# 2. 国内外研究现状

## 2.1 文本蕴含识别研究现状

## 2.2 深度学习在自然语言理解领域的研究现状

**神经语言模型(Neural Language Model)** 深度学习在自然语言处理领域的第一次影响深远的运用是利用神经网路构建语言模型。语言模型的任务是在已知一句话的前几个词情况下，来预测下一个词。语言模型在自然语言处理领域具有非常重要的地位，在一些简单的自然语言处理实践中被广泛应用，例如：智能拼写、拼写纠错，邮件回复建议。传统的语言模型主要是基于n元语法，并利用数据平滑算法对未登录词进行处理[2]。2001年 Bengio 等人[3]第一次提出利用前馈神经网络构建语言模型，模型结构如图1所示：

图1：神经语言模型(Bengio 等人于2001年发表)

该模型将一段文本的前n个词在词表中的索引作为输入，然后在词表C中查找索引所对应的词向量，这种技术现在输出为一个与词表长度一样长的向量，向量的每一位代表对应词是第n+1个词的概率。之后，Mikolov 等人[4]尝试利用循环神经网代替全连接神经网为语言模型建模，Graves 等人在此基础上又将循环神经网替换为长短记忆神经网（Long short-term memory neural network，LSTM）[5]。直到现在，虽然神经语言模型一直在发展，但是基于LSTM的建模方式依然具有很强的竞争力[6]。

神经语言模型虽然功能很简单，但是对后续的很多自然语言处理领域的发展具有深远意义。在词向量研究方面，词向量的训练过程就是一个语言模型的简化。在序列到序列模型中，模型的输出是每个时刻预测一个词，最后形成一串序列，这样的形式也是受到语言模型的启发。在之后的预训练语言模型中，这些预训练的模型的预训练过程都是在训练一个语言模型，以求能够从中理解语义，然后进行迁移学习。

**多任务学习**

**词嵌入(Word Embedding)** 自然语言处理领域在过去很长时间里都是应用词袋模型，利用稀疏向量对文本进行表示。2001年，学者们开始利用稠密向量表示词。2013年，Mikolov 等人[7][8]提出两种利用神经网络训练词向量的模型，分别是CBOW和Skip-gram，这两种模型更加有效词向量其开源工具word2vec

**神经网络在自然语言处理中的应用**

循环神经网络(Recurrent Neural Network)、卷积神经网络(Convolution Neural Netword)及递归神经网络(Recursive Neural Network)

**Seq2seq 模型**

**注意力机制(Attention)**

**记忆网络(Memory-based Networks)**

**预训练模型**

# 3. 关键理论和技术

# 4. 关键点及难点

# 5. 解决方案及可行性分析

# 参考文献

[1] MACCARTNEY B. Natural Language Inference[D]. 2009: 179.

[2] KNESER R, NEY H. Improved Backing-off for N-gram Language Modeling[C]//International Conference on Acoustics, Speech, and Signal Processing. 1995: 181–184 vol.1.

[3] BENGIO Y, DUCHARME R, VINCENT P. A Neural Probabilistic Language Model[C]//2001.

[4] INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010[M]. KOBAYASHI T, HIROSE K, NAKAMURA S. ISCA, 2010.

[5] GRAVES A. Generating Sequences With Recurrent Neural Networks[J]. CoRR, 2013, abs/1308.0850.

[6] MELIS G, DYER C, BLUNSOM P. On the State of the Art of Evaluation in Neural Language Models[J]. CoRR, 2017, abs/1707.05589.

[7] MIKOLOV T, CHEN K, CORRADO G, 等. Efficient Estimation of Word Representations in Vector Space[J]. 2013: 1–12.

[8] MIKOLOV T, SUTSKEVER I, CHEN K, 等. Distributed Representations of Words and Phrases and their Compositionality[J]. 2013: 1–9.