(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109597994 B
(45)授权公告日 2023.06.06

(21)申请号 201811472838.5

(22)申请日 2018.12.04

(65)同一申请的已公布的文献号
申请公布号 CN 109597994 A

(43)申请公布日 2019.04.09

(73)专利权人 挖财网络技术有限公司
地址 310000 浙江省杭州市西湖区华星路96号第18层

(72)发明人 尤志强 潘琪

(74)专利代理机构 杭州裕阳联合专利代理有限公司 33289
专利代理师 姚宇吉

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/289(2020.01)
G06F 18/22(2023.01)

(56)对比文件
CN 107301225 A,2017.10.27
CN 106649868 A,2017.05.10
CN 108153737 A,2018.06.12
CN 106997376 A,2017.08.01
CN 106357942 A,2017.01.25
CN 104991891 A,2015.10.21
CN 107256228 A,2017.10.17
李晨星.“基于微博的意图识别”.《中国优秀硕士学位论文全文数据库 信息科技辑》.2018,I138-2166.
Dalin Zhang 等.“Ready for Use: Subject-Independent Movement Intention Recognition via a Convolutional Attention Model”.《CIKM ″18》.2018,

审查员 王咏冬

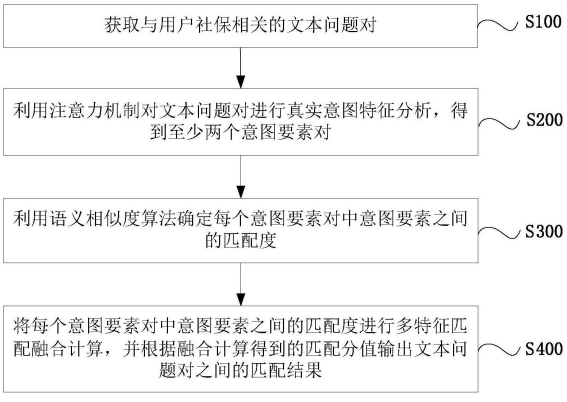
权利要求书2页 说明书18页 附图2页

(54)发明名称

短文本问题语义匹配方法和系统

(57)摘要

本发明公开了一种短文本问题语义匹配方法和系统,其中,方法包括:获取与用户社保相关的文本问题对;利用注意力机制对所述文本问题对进行真实意图特征分析,得到至少两个意图要素对;利用语义相似度算法确定每个所述意图要素对中意图要素之间的匹配度;将每个所述意图要素对中意图要素之间的匹配度进行多特征匹配融合计算,并根据融合计算得到的匹配分值输出文本问题对之间的匹配结果。本发明通过意图要素对确定句子的关键信息点,从而准确识别出句子表达的真实意图;使得本发明能够准确识别出句子微小的变化引起的巨大的意图差异,从而提升短文本问题语义匹配结果的准确性。



CN 109597994 B

1.一种短文本问题语义匹配方法,其特征在于,包括以下步骤;

获取与用户社保相关的文本问题对;

利用注意力机制对所述文本问题对中的第一文本问题进行真实意图特征分析;

具体为:根据分词关系表中的头部列表对第一文本问题的分词结果信息之间进行连续词识别,根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换,得到第一分词组合列表;

并依据所述分词关系表中的词身份列表、头部列表以及依存关系列表对第一文本问题中的句子成分的依存句法关系进行分析提取,得到至少两个第一意图要素;

利用注意力机制对所述文本问题对中的第二文本问题进行真实意图特征分析;

具体为:根据所述分词关系表中的头部列表对第二文本问题的分词结果信息之间进行连续词识别,根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换,得到第二分词组合列表;

并依据所述分词关系表中的词身份列表、头部列表以及依存关系列表对第二文本问题中的句子成分的依存句法关系进行分析提取,得到至少两个第二意图要素;

利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度;

将每个所述意图要素对中意图要素之间的匹配度进行多特征匹配融合计算,并根据融合计算得到的匹配分值输出文本问题对之间的匹配结果。

2.如权利要求1所述的短文本问题语义匹配方法,其特征在于,所述意图要素对至少包括意图主体对、意图动作对以及意图对象对中的两个。

3.如权利要求1所述的短文本问题语义匹配方法,其特征在于,还包括以下步骤;

在获取与用户社保相关的文本问题对后,对所述文本问题对进行文本问题预处理;

具体为:利用分词工具对所述文本问题对进行分词,对分词结果进行词性标注,得到词性标注结果;对分词结果进行依存句法分析,得到依存句法分析结果;

对所述词性标注结果和依存句法分析结果进行保存,生成分词关系表。

4.如权利要求3所述的短文本问题语义匹配方法,其特征在于,所述分词关系表包括文本问题对中的文本问题信息、和分别与每个文本问题信息对应的分词结果信息、词性标注信息、词身份列表、头部列表以及依存关系列表;

所述文本问题信息包括第一文本问题和第二文本问题。

5.如权利要求1所述的短文本问题语义匹配方法,其特征在于,所述利用语义相似度算法确定每个所述意图要素对中意图要素之间的匹配度,包括以下步骤;

根据预设的知识图谱,判断每个所述意图要素对中的意图要素之间是否等价;

若所述意图要素之间等价,则确定所述意图要素对中意图要素之间的匹配度;

若所述意图要素之间不等价,则通过预设的词向量模型训练对爬取的社保相关词汇进行训练,得到词汇和对应的词向量,将所述词汇和对应的词向量以键值对的形式存储成字典数据;

将每个所述意图要素对中的意图要素对所述字典数据进行查询,根据查询结果获取对应的词汇和词向量;

通过余弦相似度计算公式对与每个所述意图要素对查询得到的两个词向量进行意图要素之间的相似度计算,确定每个所述意图要素对中意图要素之间的匹配度。

6.如权利要求3所述的短文本问题语义匹配方法，其特征在于，还包括以下步骤；

在利用分词工具对所述文本问题对进行分词后，根据分词结果对所述文本问题对之间的重叠部分进行抽取，得到公共词列表；并将所述公共词列表按顺序排列成重叠词列表；所述文本问题对中的句子包括至少一个基础字符；并将所述重叠词列表与每个所述文本问题对中的文本问题进行句序分析，得到每个与文本问题对应的句序索引列表。

7.如权利要求6所述的短文本问题语义匹配方法，其特征在于，还包括以下步骤；

根据所述公共词列表、重叠词列表以及句序索引列表利用重叠词加权公式对所述文本问题对进行计算，得到所述文本问题对的加权指标；

根据所述句序索引列表利用衡量公式对所述文本问题对进行计算，得到所述文本问题对的一致性衡量值。

8.如权利要求7所述的短文本问题语义匹配方法，其特征在于，还包括以下步骤；

在进行多特征匹配融合计算时，基于意图主体对之间的匹配度、意图动作对之间的匹配度以及意图对象对之间的匹配度，并加入加权指标和一致性衡量值利用融合公式进行多特征匹配融合计算。

9.一种短文本问题语义匹配系统，其特征在于，包括获取模块、文本问题感受野模块、预处理模块、相似度计算模块以及融合计算模块；

所述获取模块，用于获取与用户社保相关的文本问题对；

所述预处理模块，用于在获取与用户社保相关的文本问题对后，对所述文本问题对进行文本问题预处理；

所述文本问题感受野模块，用于利用注意力机制对所述文本问题对进行真实意图特征分析，得到至少两个意图要素对，包括：利用注意力机制对所述文本问题对中的第一文本问题进行真实意图特征分析；

具体为：根据分词关系表中的头部列表对第一文本问题的分词结果信息之间进行连续词识别，根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换，得到第一分词组合列表；

并依据所述分词关系表中的词身份列表、头部列表以及依存关系列表对第一文本问题中的句子成分的依存句法关系进行分析提取，得到至少两个第一意图要素；

利用注意力机制对所述文本问题对中的第二文本问题进行真实意图特征分析；

具体为：根据所述分词关系表中的头部列表对第二文本问题的分词结果信息之间进行连续词识别，根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换，得到第二分词组合列表；

并依据所述分词关系表中的词身份列表、头部列表以及依存关系列表对第二文本问题中的句子成分的依存句法关系进行分析提取，得到至少两个第二意图要素；

所述相似度计算模块，用于利用语义相似度算法确定每个所述意图要素对中意图要素之间的匹配度；

所述融合计算模块，用于将每个所述意图要素对中意图要素之间的匹配度进行多特征匹配融合计算，并根据融合计算得到的匹配分值输出文本问题对之间的匹配结果。

短文本问题语义匹配方法和系统

技术领域

[0001] 本发明涉及自然语言处理技术领域，尤其涉及一种短文本问题语义匹配方法和系统。

背景技术

[0002] 社保即社会保险，是社会保障体系的重要组成部分，其在整个社会保障体系中居于核心地位。但由于各个地区经济发展水平的不同以及地方政策的差异，导致不同城市间的社保条款规定都存在差异，同一个社保相关的问题，在不同城市，可能就涉及不同的答案。因此，在社保问题的中短文本问题语义匹配具有重要的地位。而文本问题匹配算法在自然语言处理领域有着非常广泛的应用，在信息检索、问答中都有很广泛的应用场景。目前文本问题匹配算法主要分为两大类：监督式文本问题匹配算法以及无监督文本问题匹配算法。监督算法需要大量的标注数据，而涉及社保问题的文本问题匹配往往缺乏标注数据。

[0003] 目前的无监督的文本问题匹配算法，主要有：基于字典的向量空间模型，再利用余弦相似度算法计算两个文本问题间的相似度。基于编辑距离的文本问题相似度匹配算法，是将两个字符串的相似度问题，归结为将其中一个字符串转化成另一个字符串所要付出的代价。转化的代价越高，说明两个字符串的相似度越低。通常可以选择的转化方式包含插入，替换以及删除。上述传统的算法仅仅考虑了关键词匹配或者字符串匹配的程度，而没有能力识别出语句的关键信息点，即不具备句子分析能力，无法识别是句子中表达的真实意图。

发明内容

[0004] 本发明提供的短文本问题语义匹配方法和系统，其主要目的在于克服传统的算法不具备句子分析能力，无法识别是句子中表达的真实意图的问题。

[0005] 为解决上述技术问题，本发明采用如下技术方案：

[0006] 一种短文本问题语义匹配方法，包括以下步骤；

[0007] 获取与用户社保相关的文本问题对；

[0008] 利用注意力机制对所述文本问题对进行真实意图特征分析，得到至少两个意图要素对；

[0009] 利用语义相似度算法确定每个所述意图要素对中意图要素之间的匹配度；

[0010] 将每个所述意图要素对中意图要素之间的匹配度进行多特征匹配融合计算，并根据融合计算得到的匹配分值输出文本问题对之间的匹配结果。

[0011] 作为一种可实施方式，所述意图要素对至少包括意图主体对、意图动作对以及意图对象对中的两个。

[0012] 作为一种可实施方式，还包括以下步骤；

[0013] 在获取与用户社保相关的文本问题对后，对所述文本问题对进行文本问题预处理；

[0014] 具体为:利用分词工具对所述文本问题对进行分词,对分词结果进行词性标注,得到词性标注结果;对分词结果进行依存句法分析,得到依存句法分析结果;

[0015] 对所述词性标注结果和依存句法分析结果进行保存,生成分词关系表。

[0016] 作为一种可实施方式,所述分词关系表包括文本问题对中的文本问题信息、和分别与每个文本问题信息对应的分词结果信息、词性标注信息、词身份列表、头部列表以及依存关系列表;

[0017] 所述文本问题信息包括第一文本问题和第二文本问题。

[0018] 作为一种可实施方式,所述利用注意力机制对所述文本问题对进行真实意图特征分析,包括以下步骤;

[0019] 利用注意力机制对所述文本问题对中的第一文本问题进行真实意图特征分析;

[0020] 具体为:根据所述分词关系表中的头部列表对第一文本问题的分词结果信息之间进行连续词识别,根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换,得到第一分词组合列表;

[0021] 并依据所述分词关系表中的词身份列表、头部列表以及依存关系列表对第一文本问题中的句子成分的依存句法关系进行分析提取,得到至少两个第一意图要素;

[0022] 利用注意力机制对所述文本问题对中的第二文本问题进行真实意图特征分析;

[0023] 具体为:根据所述分词关系表中的头部列表对第二文本问题的分词结果信息之间进行连续词识别,根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换,得到第二分词组合列表;

[0024] 并依据所述分词关系表中的词身份列表、头部列表以及依存关系列表对第二文本问题中的句子成分的依存句法关系进行分析提取,得到至少两个第二意图要素。

[0025] 作为一种可实施方式,所述利用语义相似度算法确定每个所述意图要素对中意图要素之间的匹配度,包括以下步骤;

[0026] 根据预设的知识图谱,判断每个所述意图要素对中的意图要素之间是否等价;

[0027] 若所述意图要素之间等价,则确定所述意图要素对中意图要素之间的匹配度;

[0028] 若所述意图要素之间不等价,则通过预设的词向量模型训练对爬取的社保相关词汇进行训练,得到词汇和对应的词向量,将所述词汇和对应的词向量以键值对的形式存储成字典数据;

[0029] 将每个所述意图要素对中的意图要素对所述字典数据进行查询,根据查询结果获取对应的词汇和词向量;

[0030] 通过余弦相似度计算公式对与每个所述意图要素对查询得到的两个词向量进行意图要素之间的相似度计算,确定每个所述意图要素对中意图要素之间的匹配度。

[0031] 作为一种可实施方式,还包括以下步骤;

[0032] 在利用分词工具对所述文本问题对进行分词后,根据分词结果对所述文本问题对之间的重叠部分进行抽取,得到公共词列表;并将所述公共词列表按顺序排列成重叠词列表;所述文本问题对中的句子包括至少一个基础字符;

[0033] 并将所述重叠词列表与每个所述文本问题对中的文本问题进行句序分析,得到每个与文本问题对应的句序索引列表。

[0034] 作为一种可实施方式,还包括以下步骤;

[0035] 根据所述公共词列表、重叠词列表以及句序索引列表利用重叠词加权公式对所述文本问题对进行计算，得到所述文本问题对的加权指标；

[0036] 根据所述句序索引列表利用衡量公式对所述文本问题对进行计算，得到所述文本问题对的一致性衡量值。

[0037] 作为一种可实施方式，还包括以下步骤；

[0038] 在进行多特征匹配融合计算时，基于意图主体对之间的匹配度、意图动作对之间的匹配度以及意图对象对之间的匹配度，并加入加权指标和一致性衡量值利用融合公式进行多特征匹配融合计算。

[0039] 相应的，本发明还提供一种短文本问题语义匹配系统，包括获取模块、文本问题感受野模块、预处理模块、相似度计算模块以及融合计算模块；

[0040] 所述获取模块，用于获取与用户社保相关的文本问题对；

[0041] 所述预处理模块，用于在获取与用户社保相关的文本问题对后，对所述文本问题对进行文本问题预处理；

[0042] 所述文本问题感受野模块，用于利用注意力机制对所述文本问题对进行真实意图特征分析，得到至少两个意图要素对；

[0043] 所述相似度计算模块，用于利用语义相似度算法确定每个所述意图要素对中意图要素之间的匹配度；

[0044] 所述融合计算模块，用于将每个所述意图要素对中意图要素之间的匹配度进行多特征匹配融合计算，并根据融合计算得到的匹配分值输出文本问题对之间的匹配结果。

[0045] 与现有技术相比，本技术方案具有以下优点：

[0046] 本发明提供的短文本问题语义匹配方法和系统，能够利用注意力机制对与用户社保相关的文本问题对进行真实意图特征分析，得到至少两个意图要素对；通过意图要素对确定句子的关键信息点，从而准确识别出句子表达的真实意图；再利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度；最后将每个匹配度进行多特征匹配融合计算从而输出匹配结果。使得本申请能够准确识别出句子微小的变化引起的巨大的意图差异，从而提升短文本问题语义匹配结果的准确性。

附图说明

[0047] 图1为本发明实施例一提供的短文本问题语义匹配方法的流程示意图；

[0048] 图2为本发明实施例一中第一文本问题的依存关系矩阵示意图；

[0049] 图3为本发明实施例一中第二文本问题的依存关系矩阵示意图；

[0050] 图4为本发明实施例二提供的短文本问题语义匹配系统的结构示意图。

[0051] 图中：100、获取模块；500、预处理模块；200、文本问题感受野模块；300、相似度计算模块；400、融合计算模块。

具体实施方式

[0052] 以下结合附图，对本发明上述的和另外的技术特征和优点进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明的部分实施例，而不是全部实施例。

[0053] 请参阅图1和图4，本发明实施例一提供的短文本问题语义匹配方法，包括以下步

骤；

[0054] S100、获取与用户社保相关的文本问题对；

[0055] S200、利用注意力机制对文本问题对进行真实意图特征分析，得到至少两个意图要素对；

[0056] S300、利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度；

[0057] S400、将每个意图要素对中意图要素之间的匹配度进行多特征匹配融合计算，并根据融合计算得到的匹配分值输出文本问题对之间的匹配结果。

[0058] 需要说明的是，本申请的目的是计算两个社保问题的相似匹配度，因此模型输入必须是社保问题的文本问题对，即计算第一文本问题与第二文本问题的相似匹配程度。也就是本申请的文本问题对是以两个文本问题成对的形式出现。比如，文本问题A为“社保卡怎么补办”；文本问题B为“社保卡丢了如何挂失补办”；那么文本问题对即为“社保卡怎么补办”和“社保卡丢了如何挂失补办”。上述的文本问题对主要会涵盖医疗保险、养老保险、工伤保险、生育保险、失业保险5类险种，以及社保办理查询相关的内容。于本实施例中，“对”是成对出现的意思。比如，意图要素对中具有两个意图要素，也可以理解为与文本问题对中每个文本问题对应的意图要素。对于本申请中其他出现的“对”也是相同的意思，在此就不一一举例。

[0059] 且文本问题对是与用户社保相关的，是通过爬虫抓取与用户社保相关的社保数据，并对社保数据进行有效解析，能将爬取到的社保资讯新闻、政策规定等长篇文章进行问答对的提取，解析得到短文本问题的问题-长文本问题的答案的形式数据，需要使用到的问题是短文本问题的形式。也就是说处理后的社保数据包括两个字段为问题和答案，而本申请是对其中的问题对的匹配度计算，不涉及答案，即文本问题对为有效解析后社保数据中的问题对。于本实施例中，获取文本问题对可以通过多少方式实现，可以是用户输入的、用户浏览的、用户选取的、基于用户操作生成的等，本申请对此并不进行限制。

[0060] 注意力机制是一种资源分配方案，将计算资源分配给更重要的任务决定需要关注输入的哪部分，分配有限的信息处理资源给重要的部分，以获得最核心的信息。于本实施例中，注意力机制用于对短文本问题语句的理解，聚焦句子中的核心意图成分，识别出语句中最核心的信息，忽略无效信息，来进行匹配度的计算；从而准确识别出句子表达的真实意图。也就是意图要素对直接能够体现短文本问题的核心信息，从而忽略噪音及无效信息。而意图要素对可以至少包括意图主体对、意图动作对以及意图对象对中的两个。并不是每隔文本问题都会具有这三者意图要素的，可能只会存在其中的两个。再利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度，那么在多特征匹配融合计算时，计算对象起码为两个意图要素对之间的匹配度。使得融合计算得到的匹配分值能够表征文本问题对之间的匹配结果，一般来说，分值越大，表明两个问题意思越相近，越匹配；使得最终的结果对句子之间的匹配度具备合理的、有效的、准确的刻画。

[0061] 本发明提供的短文本问题语义匹配方法和系统，能够利用注意力机制对与用户社保相关的文本问题对进行真实意图特征分析，得到至少两个意图要素对；通过意图要素对确定句子的关键信息点，从而准确识别出句子表达的真实意图；再利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度；最后将每个匹配度进行多特征匹配融合计算从而输出匹配结果。使得本申请能够准确识别出句子微小的变化引起的巨大的意图差异，

从而提升短文本问题语义匹配结果的准确性。

[0062] 下面对各步骤的具体过程进行详细说明：

[0063] 在步骤S200后，还包括对文本问题对进行文本问题预处理。可以是利用开源的分词工具，比如ltp、jieba、hanlp、百度分词api等工具对文本问题对进行分词处理；对此并不进行限制。对分词得到的结果可以有两部分处理方式，以适用于不同的后续处理。一部分是，对分词得到的结果，进行词性标注和依存句法分析。

[0064] 具体的词性标注和依存句法分析包括以下步骤；在利用分词工具对文本问题对进行分词后，对分词结果进行词性标注，得到词性标注结果；对分词结果进行依存句法分析，得到依存句法分析结果；对词性标注结果和依存句法分析结果进行保存，生成分词关系表。

[0065] 文本问题对中的文本问题信息包括第一文本问题和第二文本问题，对于文本问题对来说，是对文本问题对中的第一文本问题和第二文本问题分别进行的，即利用注意力机制对文本问题对中的第一文本问题进行真实意图特征分析；利用注意力机制对文本问题对中的第二文本问题进行真实意图特征分析。因此，第一文本问题和第二文本问题的真实意图特征分析是相互独立的过程，只是对于每个文本问题的分析过程都是一样的。那么分词关系表包括文本问题对中的文本问题信息、和分别与每个文本问题信息对应的分词结果信息、词性标注信息、词身份列表、头部列表以及依存关系列表。每个文本问题的分析结果组成一个分词关系表。

[0066] 下面举例说明词性标注和依存句法分析：

[0067] 比如，文本问题信息的第一文本问题：北京社保怎么办理；

[0068] 分词结果：['北京','社保','怎么','办理']；

[0069] 词性标注：['ns','n','r','v']；

[0070] 词id列表：['1','2','3','4']；

[0071] head列表：['2','4','4','0']；

[0072] 依存关系列表：['ATT','SBV′,'ADV','HED']。

[0073] 上述即为第一文本问题的分词关系表。依存句法分词数据结果，由于依存关系必然涉及到两个对象，比如“怎么”和“办理”之间的关系为ADV，也就是表示“怎么”对“办理”造成的是ADV的关系影响，即“怎么”是“办理”的状语成分。

[0074] 在这里，首先对分词会进行设置id(身份)区分，由于分词结果只有4个对象，因此：词id列表为[1,2,3,4]，id为1的表示“北京”这个词，id为2的表示“社保”，以此类推，id是4的表示“办理”。

[0075] 其次，会为每一个id的词，保存一个head(头)列表，head列表与词id(身份)列表位置是一一对应的，这个head列表表示的就是与该对应位置id词存在依存句法关系的词的id。结合对应的依存关系列表。以该例子进行说明，比如head列表的第一个为2，即表示id为2的词，当前所处位置对应到id列表中也是第一位，即指id为1的词，那么说明id为1的词“北京”与id为2的词“社保”存在ATT关系。同理，id为2的词“社保”与id为4的词“办理”存在SBV关系，id为3的词“怎么”与id为4的词“办理”存在ADV关系，而id为4的词“办理”的head对应的是id为0，0表示不对应其他词，对应的是根节点，其是整个句子的核心。整个句子阐述的核心意图是“办理”。

[0076] 通过下表1对词性符号进行解释：

[0077] 表1词性符号含义

[0078]

词性	含义	词性	含义	词性	含义	词性	含义
Ag	形语素	g	语素	ns	地名	u	助词
a	形容词	h	前接成分	nt	机构团体	vg	动语素
ad	副形词	i	成语	nz	其他专名	v	动词
an	名形词	j	简称略语	o	拟声词	vd	副动词
b	区别词	k	后接成分	p	介词	vn	名动词
c	连词	l	习用语	q	量词	w	标点符号
dg	副语素	m	数词	r	代词	x	非语素字
d	副词	Ng	名语素	s	处所词	y	语气词
e	叹词	n	名词	tg	时语素	z	状态词
f	方位词	nr	人名	t	时间词	un	未知词

[0079] 通过下表2对依存句法符号进行解释：

[0080] 表2依存句法符号含义

[0081]

依存关系	含义	依存关系	含义	依存关系	含义	依存关系	含义
APP	同位关系	QUN	数量关系	COO	并列关系	ATT	定中关系
POB	介宾关系	ADJ	附加关系	VOB	动宾关系	DC	依存分句
TMP	时间关系	SBV	主谓关系	SIM	比拟关系	WP	标点
DI	“地”字结构	LOC	处所关系	DE	“的”字结构	IS	独立结构
BA	“把”字结构	DEI	“得”字结构	SUO	“所”字结构	VNV	叠词关系
CMP	动补结构	BEI	“被”字结构	ADV	状中结构	IC	独立分句
CS	关联结构	DBL	兼语结构	CNJ	关联词	YGC	一个词
HED	核心	MT	语态结构	VV	连谓结构		
TOP	主题	FOB	前置宾语	DOB	双宾语		

[0082] 那么对于文本问题对中第一文本问题“请问一下,北京养老保险怎么缴费”和第二文本问题“北京怎么缴纳养老保险,都不知道怎么操作”的实例为：

[0083] 第一文本问题：

[0084] 第一文本问题:请问一下,北京养老保险怎么缴费;分词结果:['请问','一','下',',','北京','养老保险','怎么','缴费'];词性标注:['n','m','q','w','ns','n','r','v'];词id列表:[1,2,3,4,5,6,7,8];head列表:[8,3,1,1,6,8,8,0];依存关系列表:['IS','QUN','CMP','WP','ATT','TOP','ADV','HED']。

[0085] 第二文本问题：

[0086] 第二文本问题:北京怎么缴纳养老保险,都不知道怎么操作;分词结果:['北京','怎么','缴纳','养老保险',',',',','都','不知道','怎么','操作'];词性标注:['ns','r','v','n','w','d','v','r','v'];词id列表:[1,2,3,4,5,6,7,8,9]head列表:[3,3,0,3,3,7,3,9,7];依存关系列表:['LOC','ADV','HED','VOB','WP','ADV','IC','ADV','VOB']。

[0087] 进一步的,注意力机制的分析是微观上进行匹配性衡量,也是局部感受野。之所以引入注意力机制,是因为发现在短文本的匹配相似度计算上,即使两句话几乎一模一样,但因为一字之差,可以让所表达的意思天差地别。比如“养老保险补缴条件”与“养老保险补缴

基数”，可以看到这两句话80％的内容都是一样，但是显然如果关注其真实意图，可以发现一个是问“条件”，一个是问“基数”，表达的意思完全不一样，匹配度应该为0。再比如“养老保险补缴所需材料”与“养老保险缴费所需材料”仅仅一字之差，也导致真实意图的完全不一样。这就要求能够从微观层面发现这个细微的差异。基于这样的背景，因此引入了注意力机制，也可以理解为从微观层面，聚焦到问题的关键部位，进行一致性检验。

[0088] 也就是说，对于第一文本问题和第二文本问题，都将句子的真实意图的组成结构分为三部分主要素：意图主体、意图动作、意图对象。另外，根据具体的匹配所需，可以扩展加上“状态”这一副要素。在这里，主要关注三部分主要素。比如，“养老保险补缴条件”的三要素：养老保险（意图主体），补缴（意图动作），条件（意图对象），而“养老保险补缴基数”的三要素：养老保险（意图主体），补缴（意图动作），基数（意图对象），很显然发现这两个问句的意图对象不一致，尽管主体和动作是匹配的，因此仍然判定这两句话是不匹配的。同理，也能很容易发现“养老保险补缴所需材料”与“养老保险缴费所需材料”的意图动作是不一致的，“补缴”与“缴费”的意义是不同的，也会被判定句子意图为不匹配。

[0089] 基于上述的背景介绍及例子介绍，可以看到引入注意力机制，将注意力聚焦到意图三要素：主体、动作、对象。问题对之间的匹配度衡量，细分成这三要素的细粒度对比。首先注意到“主体”，如果主体不一致，显然是不同的问题。其次注意到“动作”，如果动作不一致，即使主体匹配度很高，仍然是属于不同的问题。最后，聚焦到意图对象进行对比。只有通过三要素的一致性检验，都具有很高的匹配度，那么所进行比较的问题对，才会获得较高的匹配分值。

[0090] 而在实际的问题匹配计算中，会遇到很多结构的句子，三要素的结构，会稍显复杂，比如“职工生育险怎么报销”与“职工如何报销生育险”。在这两个问题句子中，可以发现意图主体是“职工”，动作为“怎么报销/如何报销”，意图对象为“生育险”。因此要使注意力机制真正发挥价值并得到准确的结果，设计出一套能够应对复杂结构的意图三要素算法至关重要。

[0091] 因此，提出一种基于依存句法分析和词性标注的意图三要素提取算法。具体为：对于利用注意力机制对文本问题对中的第一文本问题进行真实意图特征分析，包括以下步骤；根据分词关系表中的头部列表对第一文本问题的分词结果信息之间进行连续词识别，根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换，得到第一分词组合列表；并依据分词关系表中的词身份列表、头部列表以及依存关系列表对第一文本问题中的句子成分的依存句法关系进行分析提取，得到至少两个第一意图要素。

[0092] 同样的，利用注意力机制对文本问题对中的第二文本问题进行真实意图特征分析，包括以下步骤；根据分词关系表中的头部列表对第二文本问题的分词结果信息之间进行连续词识别，根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换，得到第二分词组合列表；并依据分词关系表中的词身份列表、头部列表以及依存关系列表对第二文本问题中的句子成分的依存句法关系进行分析提取，得到至少两个第二意图要素。

[0093] 下面，以第一文本问题“请问一下，北京养老保险怎么缴费”与第二文本问题“北京怎么缴纳养老保险，都不知道怎么操作”为例，进行阐述：

[0094] 首先引入中心词概念，会基于依存句法分析得到的head列表，进行连续词识别，比

如修饰词与名词可以构成连续的词组形式。所谓中心词，形如“怎么缴费”这个词组的中心词即为“缴费”，而“怎么”是修饰这个中心词的，作为补充，是副词。中心词的识别，可以让意图三要素表达上更加完善，因为有些词就是需要放在一起才能表达出更完整的意思，可为之后的三要素提取进行词组的准备。

[0095] 在实践过程中，发现中心词的模式主要有：

[0096] head列表中连续id之间的差值为1。

[0097] head列表中前一个id＝后一个位置索引+1(特别说明，在这里索引是从0开始计数，而不是从1开始)，对于这类情况又可以细分出多种子情况，比如后一个位置如果head id为0或者前后id都为同id。子情形可以细分出4类，这里不做具体阐述。

[0098] 以对第一文本问题进行中心词识别为例：分词列表：['请问','一','下',',','北京','养老保险','怎么','缴费']；head列表为：[8,3,1,1,6,8,8,0]。

[0099] 可以看到，在head列表中连续的id 3,1满足中心词(2)的模式：第一个id为1的索引为2，那么id 3＝2+1，即3＝id为1索引+1。组合后，从原来的“一”、“下”变成词组”一下”。同理，发现连续的id 8,0也满足中心词(2)的模式：第一个id为0的索引为7，那么id 8＝7+1，即8＝id为0索引+1。组合后，从原来的“怎么”，“缴费”变成词组“怎么缴费”。那么原来的分词列表就转换为：['请问',",'一下',',','北京','养老保险',",'怎么缴费']。

[0100] 再以第二文本问题进行中心词识别为例：分词列表：['北京','怎么','缴纳','养老保险',',',',','都','不知道','怎么','操作']；head列表：[3,3,0,3,3,7,3,9,7]。

[0101] 以同样的中心词处理方式，对第二文本问题进行处理，得到第二文本问题新的分词列表：['北京','怎么缴纳',",'养老保险',',',",','都不知道',",'怎么操作']。

[0102] 本申请提出一种基于依存句法分析及词性标注的注意力机制算法，也就是能够对短文本句子进行聚焦到核心的细节进行微观分析，在本文中即是进行意图三要素抽取分析，准确提取出句子表达的真实意图。本文涉及的意图三要素为意图主体、意图动作、意图对象，可以对句子进行关键信息分析，聚焦到核心，忽略噪音及无效信息，使用这种局部感受野(注意力机制)，能够准确有效捕捉短文本句子之间的微小差异，而不会被全局信息给掩盖。另外，所提出的注意力机制识别方法，能够应对复杂的不同表达形式的语句，算法能够有效抵御句子结构调整带来的识别差异，在应对各种语义结构的句子匹配度计算上，具备较高的鲁棒性。

[0103] 进一步的，为了提高关系搜索的效率。依据分词关系表中的词身份列表、头部列表以及依存关系列表建立关系矩阵，根据关系矩阵对文本问题对中的每个文本问题信息的核心词进行分析提取。

[0104] 也就是说，基于词id列表与head列表、依存关系列表进行二维矩阵表示，可以将这三者关系放到一起进行使用，提高关系搜索的效率。基于该矩阵，对句子成分的依存句法关系进行分析，进行意图三要素的模式识别提取。

[0105] 先介绍矩阵的表示方式：矩阵的结构为M X M，M表示行数或者列数，在这里行数与列数是相同的，行数等于分词列表中的词的个数，行数序号从上往下依次增大，在实施例中，第一行的行索引为1，第二行为2，依次类推，这个行索引序号等同于词id列表中的id值，因此可以使用行序号来表征某一个词的id。比如第一行的行序号就可以表示词id为1。同样的，列与行等同，从左到右列序号依次增加，第一列的列索引为1，第二列列索引为2，依次类

推。矩阵中的元素就可以表示某一行代表的id对应的词与某一列代表的id对应的词的依存关系。第一文本问题与第二文本问题的依存关系矩阵如下所示，可以看到第一文本问题中id为5的词与id为6的词之间的依存关系为"ATT"，即"北京"是修饰"养老保险"的，也就是表示这个句子讲的是"北京的养老保险"，而不是其他城市的。同理，可以看到第二文本问题中id为2的词与id为3的词之间的依存关系为"ADV"，即"怎么"是修饰"缴纳"，也就是表达了"怎么缴纳"的意思，如图2所示的第一文本问题的依存关系矩阵示意图和图3所示的第二文本问题的依存关系矩阵示意图。

[0106] 依存句法关系中有一种关系符号为HED，表示句子"核心"，比如"社保怎么缴费"中的"缴费"会被标注为HED，并且HED在实施例中一个短文本社保问句中一般只有一个，而且HED仅涉及与root(根)的关系，由于root是虚拟的，HED一般为单个词。或者也可以说是HED自己与自己形成的关系，仅自身形成了HED关系。因此。它是整个依存句法关系结构的最重要的部分，是一个句子的中心思想。在做意图三要素的模式识别抽取的时候，以HED为切入点。因此会选择HED所在的行作为模式提取的重点，其中该行中非空的关系，即表示该位置索引对应的词与HED的词是产生关系的。比如第一文本问题中，HED对应的行索引为8，其所在的行表示的就是id为8的词与其他列索引对应的词产生的关系。首先可以看到HED自身对应的id为8的词，而自身与自身形成的是HED的关系，即id为8的词就是这句话的核心。然后id为7的词与HED对应的词形成了"ADV"的关系，id为6的词与HED对应的词形成了"TOP"的关系，id为1的词与HED对应的词形成"IS"的关系。

[0107] 由于HED为单个词，因此可以对该词按照词性划分出多种不同的情况。比如HED可以为动词，系动词，名词，"的"字结构短语等常见的几类。这里举例子分别说明一下HED为对应词性的情况：1、北京养老保险办理地点，HED为"地点"，词性为"名词"。2、社保卡是医保卡吗?HED为"是"，词性为"系动词"。3、上海外来人口养老保险怎么转入?HED为"转入"，词性为"动词"。4、养老账户余额怎么计算出来的?HED为"怎么计算出来的"，为"的"字结构短语。

[0108] 在识别出HED之后，就以HED所在的行中非空关系，进行提取相应的意图主体、意图对象，以及完善意图动作等。在实施例中，SBV、VOB、FOB、TOP、VV等关系可以提取出相应的意图主体及意图对象，比如第一文本问题中TOP关系，可以提取得到该问题的意图主体为"养老保险"，同样的在第二文本问题中的HED所在行的VOB关系，可以提取得到该问题的意图对象为"养老保险"。ATT、ADV、CMP可以用于完善意图动作、意图主体、意图对象，比如第一文本问题的意图动作通过ADV关系完善可得到"怎么缴纳"。而LOC,TMP,IS,IC可以被用于识别句子的状态，比如第二文本问题通过HED所在行的TOP关系，得到状态为"都不知道怎么操作"。在这里需要特别指出的是：如果某个依存关系对应的词已经与HED对应的词在中心词识别的时候组成了词组，将不再对该依存关系的词做处理，而直接使用词组。比如前面提到在第一文本问题中进行中心词识别得到的"怎么缴纳"，"怎么"已经与"缴纳"构成了连续词块，会被作为整体使用，将不再对该ADV的关系进行处理，而只会识别提取"IS"和"TOP"两个关系涉及的词。

[0109] 于本实施例中，还引入了交叉检验，即当一个句子中存在某个实体词，而另一个句子识别出来的实体词不存在对应的词，会检验没有该实体词的句子整体，检测是否存在该实体词，如果存在的话会进行补足。这能一定程度上弥补少数情况下在模式抽取时候的漏掉某个主体词的情况。

[0110] 在这里，需要特别指出，一个完整的意图不一定非要三要素全部满足，有时候满足其中两个也是能够表达完整意思的。比如“养老保险怎么缴费”，意图主体为“养老保险”，意图动作为“怎么缴费”，而没有意图对象。又比如“怎么缴纳医保”，意图动作为“怎么缴纳”，意图对象为“医保”，而没有意图主体。虽然缺失了某一个要素，但并不影响意图完整性。为了应对一个句子有主体无对象，一个句子有对象无主体，会进行“意图主体”与“意图对象”的匹配度计算，作为意图主体的匹配度的Ssim值及意图对象的匹配度的Tsim值。以第一文本问题“请问一下，北京养老保险怎么缴费”与第二文本问题“北京怎么缴纳养老保险，都不知道怎么操作”为例，可以识别得到：第一文本问题的意图主体为“养老保险”、意图动作为“怎么缴费”；第二文本问题的意图对象为“养老保险”、意图动作为“怎么缴纳”。

[0111] 进一步的，利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度，包括以下步骤；

[0112] 根据预设的知识图谱，判断每个意图要素对中的意图要素之间是否等价；

[0113] 若意图要素之间等价，则确定意图要素对中意图要素之间的匹配度；

[0114] 若意图要素之间不等价，则利用词向量相似度算法确定每个意图要素对中意图要素之间的匹配度。

[0115] 需要说明的是，但是由于词向量模型自身的局限，以及相似分的波动原因，往往会引入不合适的结果。比如“缴纳”和“欠缴”在社保问题中有明显的差别，又比如“生育”与“再生育”是两个不同的概念。另外，社保当中的明确规定，也是词向量无法应对的，比如“医保生育”说的是医疗保险里的概念，而不是“生育保险”中的概念。“职工”与“居民”是在社保概念定义上是严格区分的两个群体。因此单靠词向量，会引入错误判断，导致知识的错误匹配，同时没有专家经验的使用，也造成了概念上的区分混乱。

[0116] 因此，引入了基于知识图谱的远程监督方案。如果两个对象词在知识图谱中存在不等价关系，那么包含该两个对象的子模块也能表示这种关系。在知识图谱中可以设置了几百条不等价关系以及等价关系，通过下表3中罗列出了一些不等价及等价关系的展示：

[0117] 表3知识图谱中实体对象关系

[0118]

对象1	对象2	关系
医保	失业	不等价
医保	工伤	不等价
儿童	新生儿	不等价
职工	居民	不等价
参保	退保	不等价
转入	转出	不等价
…	…	…
挂失	激活	不等价
缴纳	补缴	不等价
材料	指南	等价
老人	一老	等价
…	…	…
津贴	待遇	等价

医保生育	医保	等价

[0119] 对根据预设的知识图谱，判断每个意图要素对中的意图要素之间是否等价举例进行说明：比如，第一文本问题“医保怎么报销”与第二文本问题“失业怎么报销“进行匹配度计算时：第一文本问题的意图主体为“医保”，第二文本问题的意图主体为“失业”。由于“医保”与“失业”在知识图谱中的关系为“不等价”。因此，两句话的意图主体的匹配度直接返回“0”，无需再进行词向量相似度的计算。因此可知，“医保怎么报销”与“失业怎么报销”是不匹配的。也就是说先根据预设的知识图谱进行远程监督校验，如果两个对象在知识图谱中不存在关系，也就是不存在等价关系，才会进行词向量的相似度计算。如果是存在等价关系，则说明击中了知识图谱中的关系，直接返回等价(1)或者不等价(0)的结果。引入远程监督之后，能够有效且灵活的使用社保专家的经验以及应对词向量无法应对的问题。进一步提高了算法匹配的准确度，而且很容易进行更多关系的扩展。

[0120] 进一步的，利用词向量相似度算法确定每个意图要素对中意图要素之间的匹配度，包括以下步骤；

[0121] 通过预设的词向量模型训练对爬取的社保相关词汇进行训练，得到词汇和对应的词向量，将词汇和对应的词向量以键值对的形式存储成字典数据；

[0122] 将每个意图要素对中的意图要素对字典数据进行查询，根据查询结果获取对应的词汇和词向量；

[0123] 通过余弦相似度计算公式对与每个意图要素对查询得到的两个词向量进行意图要素之间的相似度计算，确定每个意图要素对中意图要素之间的匹配度。

[0124] 需要说明的是，在使用词向量进行相似度计算之前，需要预先训练出词汇的分布式表达，也就是一个词会被表示成一个低维度的向量形式，比如“社保”会被表示成[0,0.1,0.12,0,0.4,0.5,0.13,0.55]8维的向量。维数的大小是在训练过程中可以控制，一般会设置100-300之间的大小。这个低维的向量，代表的是“社保”这个词在8维空间中的所处位置。意思相近的词，在空间中的位置也会更近。而意思相差较大的词，在空间中的距离会更远。比如“社保”与“医保”在空间中的距离会近，而与“电脑”这个词就会相距较远。这也是能够利用词向量进行词与词之间相似度计算的原因，用空间中的距离来刻画词之间的相似程度。为了能够尽可能完整且准确表征每一个社保相关的词汇，需要爬取大量的社保相关的数据进行训练。

[0125] 在预先爬取了各城市社保官网，第三方社保网站的社保政策规定、新闻资讯、事例解答等长短文本数据后。对这些数据进行清洗，包括去掉特殊符号、去掉HTML标签、分词，去除停用词等。再利用词向量使用word2vec工具进行训练。主要包含两个模型：跳字模型(skip-gram)和连续词袋模型(continuous bag of words，简称CBOW)，以及两种高效训练的方法：负采样(negative sampling)和层序softmax(hierarchical softmax)。word2vec词向量可以较好地表达不同词之间的相似和类比关系。在本实施例中，使用的是其中的CBOW算法，负采样训练，另外模型其他参数：window为5，min_count为2，向量size为300。在这里，使用的是python编程语言的一个第三方包gensim，其提供了word2vec的训练功能。

[0126] 在训练完成后，可以获得词汇以及对应的词向量字典数据。格式为“词：词向量”。这里举例展示一下词向量识别近义词的效果，比如要查询“缴纳”的近义词：1、缴纳：1.0；2、交纳：0.76575；3、缴交：0.723778；4、缴费：0.651198；5、缴：0.642742；6、缴付：0.636492；7、

代缴:0.579413;8、欠缴:0.541146;9、缴满:0.52438。可以看到,返回了9个近义词,其中分值就是该近义词与“缴纳”的相似度。但从中也能看到当相似度分值越低,其近似词含义与“缴纳”还是产生了一定的差异。比如“缴纳”和“欠缴”、“缴满”还是有差别的。

[0127] 在这里两个单词的相似分值的计算,使用的是余弦相似度计算公式,值越趋近于1,代表两个词越相似。

[0128] 于本实施例中,余弦相似度计算公式具体为:

[0129] $$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|_2\|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$。

[0130] 式中,A为单词A;B为单词B;单词A的词向量为(x_1,y_1);单词B的词向量为(x_2,y_2)则余弦相似度计算公式变换为:

[0131] $$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\times\sqrt{x_2^2 + y_2^2}}$$。

[0132] 下面举一个实际例子进一步说明:

[0133] 比如,“断缴”的词向量为[0.5,0.3,0.4],“断交”的词向量为[0.45,0.31,0.39],“异地”的词向量为[0.1,0.9,0.1],那么:

[0134] cos(“断缴”,“断交”)

[0135] =(0.5*0.45+0.3*0.31+0.4*0.39)/(sqrt(0.5*0.5+0.3*0.3+0.4*0.4)*sqrt(0.45*0.45+0.31*0.31+0.39*0.39))

[0136] =0.998。

[0137] 同理,计算出cos(“断缴”,“异地”)=0.558。因此,从数值上可以看出来,(“断缴”,“断交”)更相似。可以看到单个词与单个词是很容易使用余弦相似度对两个单词的词向量进行计算相似度的。但是有时候需要比较多个词与多个词之间的相似度,比如,问题的主体涉及多个单词的时候,这个时候就需要进行相应处理。假设要比较[“养老”、“保险”]与[“失业金”]之间的相似度。

[0138] 首先,会对[“养老”、“保险”]的词向量进行处理。第一步会利用词向量分别找出“养老”的近义词列表,“保险”的近义词列表。

[0139] 然后,将“养老”和“保险”的词向量分别与其近义词列表的每一个词的词向量进行逐个元素地相加取平均,分别得到“养老”和“保险”这个词的新的词向量。最后将“养老”与“保险”两个词的新的词向量按等位相加,得到最终的[“养老”、”保险”]的合成词向量。比如,“养老”的词向量为[0.1,0.3],养老的近义词为“退休”,词向量为[0.2,0.4]。那么”养老”的新的词向量为[(0.1+0.2)/2,(0.3+0.4)/2]=[0.15,0.35]。“保险”的词向量为[0.3,0.2],保险的近义词为“保障”,词向量为[0.2,0.25]。那么“保险”的新的词向量为[(0.3+0.2)/2,(0.2+0.25)/2]=[0.25,0.225]。最后,将“养老”的新词向量与“保险”的新词向量进行等位相加得到最终结果。最终的[“养老”、”保险”]合成词向量为:[0.15+0.25,0.35+0.225]=[0.4,0.575]。然后用这个合成的新的词向量与失业金的词向量做余弦相似度计算,得到[“养老”、“保险”]与[“失业金”]之间的相似度。

[0140] 因此,提出在词向量相似度计算基础上,融合基于知识图谱的远程监督模块。既可以保证通用性,具备相近词的自动识别能力,能够通过衡量两个词在空间中的距离远近来刻画语义上的近似程度,具备对判别的对象的语义理解能力,又能够应对词向量无法学习的专家经验,以及修正词向量的识别误差问题,对语义理解上的偏差进行纠正。通过远程监督的方式,领域知识以及专家经验能够以高扩展性、高灵活性,结合到整体算法框架中,使得算法可以更加完备。两种方式的有效结合,既保证了泛化识别能力,又提高了算法判别的准确性。

[0141] 而且基于远程监督以及词向量相似度计算两种方法进行的多特征匹配融合计算,包括以下步骤;基于意图主体对之间的匹配度、意图动作对之间的匹配度以及意图对象对之间的匹配度利用融合公式进行多特征匹配融合计算。

[0142] 对注意力机制识别得到的“意图主体”、“意图动作”、“意图对象”分别计算匹配度计算,分别得到对应的匹配度。即意图主体对之间的匹配度Ssim,意图动作对之间的匹配度Vsim以及意图对象对之间的匹配度Tsim。于本实施例中,多特征匹配融合计算是基于3类相似度特征或该3类相似度特征中的2类进行的。

[0143] 对分词得到的结果的另外一部分处理方式是对文本问题对之间的重叠部分与句子表达顺序进行预先抽取。具体步骤为:在利用分词工具对文本问题对进行分词后,根据分词结果对文本问题对之间的重叠部分进行抽取,得到公共词列表;并将公共词列表按顺序排列成重叠词列表;文本问题对中的句子包括至少一个基础字符;并将重叠词列表与每个文本问题对中的文本问题进行句序分析,得到每个与文本问题对应的句序索引列表。

[0144] 对于文本问题对的重叠部分抽取方式:需要满足,1.两个句子必须满足值不为空,且句子长度大于0。2.对句子进行分词处理,再利用动态规划算法识别出句子的公共词列表CW_list。CW_list(公共词列表)可以包含单个字、词、词组等形式。

[0145] 举例说明重叠部分的抽取结果:假设有文本问题对为“上海女职工怎么报销医疗保险”与“上海女职工医疗保险报销流程”:1.首先这两个问题满足“值不为空且句子长度大于0”,进行分词处理。2.分词后分别为“[上海,女,职工,怎么,报销,医疗保险]”和“[上海,女,职工,医疗保险,报销,流程]”。3.使用动态规划算法识别出公共词列表,这里如果两个公有词连续,会进行组合成词组,那么结果为:CW_list=[上海,女职工,报销,医疗保险]。

[0146] 对于文本问题对的句序分析的句子子部分抽取。需要满足,1.句序分析只针对重叠部分进行。利用CW_list(公共词列表),以单个词为单位,获取的到两个句子的重叠词列表SW_list。2.顺序遍历第一文本问题,标记出SW_list(重叠词列表)的词wi在第一文本问题中的索引I1,同一词在第一文本问题中多次出现,只记录首次索引。3.顺序遍历第二文本问题,标记SW_list的词wj在“第一文本问题”中的索引I2,同一词在第二文本问题中多次出现,仅记录首次索引。

[0147] 举例说明句序分析部分的抽取结果:还是以文本问题对“上海女职工怎么报销医疗保险”与“上海女职工医疗保险报销流程”为例:1.假设已经通过重叠部分识别获得了CW_list=[上海,女职工,报销,医疗保险]。2.由于句序部分是以单个词为单位,因此重叠部分的单词列表为SW_list=[上海,女,职工,报销,医疗保险]。3.顺序遍历第一文本问题,获得SW_list的词wi在第一文本问题中的索引列表,[1,2,3,4,5]。4.顺序遍历第二文本问题,获得SW_list的词wj在“第一文本问题”中的索引列表,[1,2,3,5,4]。

[0148] 另外在这里展示第一文本问题“请问一下,北京养老保险怎么缴费”与第二文本问题“北京怎么缴纳养老保险,都不知道怎么操作”的示例:CW_list＝[‘北京’,‘养老保险’,‘怎么’];SW_list＝[‘北京’,‘养老保险’,‘怎么’];SW_list的词在第一文本问题中的句序索引＝[1,2,3];SW_list的词在第二文本问题中的句序索引＝[1,3,2]。

[0149] 为了在注意力机制得到的结果上,能够补充更多的辅助信息。能够对一些特殊情况,进行有效的准确的匹配度比较,使得算法更加完备。这里引入全局感受野利用全局范围进行相关特征的提取,即宏观上方面进行匹配性衡量,达到最优的匹配度衡量结果。这里举两个个特殊情况的例子:

[0150] 例子1,假设需要为第一文本问题,在2,3中寻找最匹配的问句,即第二文本问题与问题3,到底谁与第一文本问题更加匹配。第一文本问题为养老保险费怎么缴纳?第二文本问题为养老保险费应该怎么缴纳?问题3为怎么缴纳养老保险费?

[0151] 首先看:第一文本问题与第二文本问题在全局感受野下:CW_list为[养老保险费,怎么,缴纳];SW_list为[养老保险费,怎么,缴纳];SW_list的词在第一文本问题中的句序索引I1＝[1,2,3];SW_list的词在第二文本问题中的句序索引I2＝[1,2,3]。

[0152] 第一文本问题与问题3在全局感受野下:CW_list为[养老保险费,怎么,缴纳];SW_list为[养老保险费,怎么,缴纳];SW_list的词在第一文本问题中的句序索引I1＝[1,2,3];SW_list的词在问题3中的句序索引I2＝[2,3,1]。

[0153] 可以看到第一文本问题与第二文本问题在公有词数量上是相同的,而且在句序上也是一致的。而第一文本问题与问题3虽然在公有词数量上是相同的,但是在句序上不一致。因此会判定第一文本问题与第二文本问题更匹配,而问题3其次。

[0154] 例子2:假设需要为第一文本问题,在2,3中寻找最匹配的问句,即第二文本问题与问题3,到底谁与第一文本问题更加匹配。第一文本问题为养老保险如何转入本地?第二文本问题为养老保险转入本地的流程?问题3为养老保险转移流程?

[0155] 首先看:第一文本问题与第二文本问题在全局感受野下:CW_list为[养老保险,转入,本地];SW_list为[养老保险,转入,本地];SW_list的词在第一文本问题中的句序索引I1＝[1,2,3];SW_list的词在第二文本问题中的句序索引I2＝[1,2,3]。

[0156] 第一文本问题与问题3在全局感受野下:CW_list为[养老保险];SW_list为[养老保险];SW_list的词在第一文本问题中的句序索引I1＝[1];SW_list的词在问题3中的句序索引I2＝[1]。

[0157] 可以看到第一文本问题与第二文本问题在公有词数量上要更多,而且在句序上也是一致的。而第一文本问题与问题3在公有词数量上只有一个,只有1个词的时候句序是无意义的。因此会判定第一文本问题与第二文本问题更匹配,而问题3其次。

[0158] 从以上两个例子可以看到,虽然第一文本问题与第二文本问题、问题3表达的意思意图都是一致的,但是在判断谁的匹配度更高的时候,认为在保证意图一致的情况下,重叠度更高,共有词更多,且句序一致性更高的句子,获得的匹配分要更高,需优先选择。因此,引入了全局感受野的信息,使得算法更加的完备合理,可以应对特殊情况。

[0159] 在该步骤,直接使用识别得到的CW_list,SW_list,句序索引列表。还是以第一文本问题“请问一下,北京养老保险怎么缴费”与第二文本问题“北京怎么缴纳养老保险,都不知道怎么操作”为例,得到以下结果:CW_list＝[‘北京’,‘养老保险’,‘怎么’];SW_list＝

['北京','养老保险','怎么'];SW_list的词在第一文本问题中的句序索引I1=[1,2,3];SW_list的词在第二文本问题中的句序索引I2=[1,3,2]。

[0160] 为了使得在进行多特征匹配融合计算时输出的匹配结果准确性更高,引入结构相似度。结构相似度计算包含两部分:(1)重叠词加权指标;(2)句序一致性衡量。具体的步骤为;根据公共词列表、重叠词列表以及句序索引列表利用重叠词加权公式对文本问题对进行计算,得到文本问题对的加权指标;根据句序索引列表利用衡量公式对文本问题对进行计算,得到文本问题对的一致性衡量值。

[0161] 也就是以公共词列表CW_list、重叠词列表SW_list、句序索引列表为基础计算重叠词加权指标。从文本问题对的重叠词个数、重叠词在两个句子中的连续性和句序一致性三方面来测算结构相似度。

[0162] 对于重叠词加权指标;重叠词加权公式具体为:

[0163] $$Csim=(A,B)=\sum_{i=1}^{n}N(c_i)^{\omega}$$ 。

[0164] 式中,A为第一文本问题;B为第二文本问题;w为加权系数;于本实施例中,w设置为1.1,C_i为CW_list中的词,$N(C_i)$表示C_i中包含的词的个数。当C_i为独立单词时,$N(C_i)$为1,加权后仍为1。当C_i为连续词组成的词组时,$N(C_i)$大于1,加权系数起到加权的作用。

[0165] 假设有文本问题对为"上海女职工怎么报销医疗保险"与"上海女职工医疗保险报销流程":CW_list为[上海,女职工,报销,医疗保险];SW_list为[上海,女,职工,报销,医疗保险];Csim=1+2^1.1+1+1=5.144。因为,"女职工"是由"女"和"职工"连续词组成的词组,N(女职工)=2,所以加权起到作用。也就是说连续的公共词越多,Csim的得分越高。

[0166] 对于句序一致性衡量:基于对文本问题对进行了句序分析,抽取出了SW_list分别在第一文本问题中的句序索引I_1和在第二文本问题中的句序索引I_2。由于I_1是顺序的,只需要计算I_2中索引的词序,对不是正常顺序的索引进行惩罚,便可以得到第一文本问题与第二文本问题的重叠部分的句序一致性程度。

[0167] 衡量公式Osim具体为:

[0168] $$Osim(SW_n,SW_{n-1})=\begin{cases}\prod_{p=1}^{|Q|}\delta, & \text{if } I_{2_n}-I_{2_{n-1}}<0\\ 1, & \text{if } I_{2_n}-I_{2_{n-1}}>0\end{cases}$$ 。

[0169] 式中,当I_2中第n个索引比第n-1个索引值小时,对词序一致性进行惩罚,其中Q表示索引之间的差值,δ表示惩罚因子,值域为0到1之间,于本实施例中设置为0.8。当I_2中第n个索引比第n-1个索引大时,说明该词在第二文本问题中出现的顺序与第一文本问题中保持一致,记语序相似度(一致性值)值为1。计算完成后,会将所有的(SWn,SWn-1)的值进行求和,并除以(索引列表I_2中元素的个数-1)。比如,以第一文本问题"请问一下,北京养老保险怎么缴费"与第二文本问题"北京怎么缴纳养老保险,都不知道怎么操作"为例进行说明:基于以下结果:CW_list=['北京','养老保险','怎么'];SW_list=['北京','养老保险','怎么'];SW_list的词在第一文本问题中的句序索引I1=[1,2,3];SW_list的词在第二文本问题中的句序索引I2=[1,3,2]。显然$Osim(SW_2,SW_1)$=1,因为3-1>0;$Osim(SW_3,SW_2)$=0.8,因为2-3<0;Osim=(Osim(SW_2,SW_1)+Osim(SW_3,SW_2))/(3-1)=0.9。

[0170] 最后,在进行多特征匹配融合计算时,可以是依据5类相似度特征进行的。包括以

下步骤;在进行多特征匹配融合计算时,基于意图主体对之间的匹配度、意图动作对之间的匹配度以及意图对象对之间的匹配度,并加入加权指标和一致性衡量值利用融合公式进行多特征匹配融合计算。

[0171] 需要说明的是,在实际计算社保问题之间匹配度的时候,会先把地址信息进行提取,只有保证地址信息匹配的情况下,再进行5类特征相似度的计算。因为地址一致是社保问题匹配的前提。而5类特征相似度指的是,意图主体对之间的匹配度Ssim,意图动作对之间的匹配度Vsim,意图对象对之间的匹配度Tsim,重叠部分的加权指标Csim,句序一致性的一致性衡量值Osim。

[0172] 需要对这5类特征进行有效融合,计算出最终的问题对的匹配相似度。也就是说考虑了两个方面:1、意图主体、意图动作、意图宾语三者的相似度越高,表明两个问题对之间的相似度更高。2、连续的重叠词越多,且顺序一致性越高,两个问题对之间的相似度更高。

[0173] 多特征匹配融合计算的融合公式为:

[0174] $s_1=(1+Csim\times Osim)*(Ssim)$;

[0175] $s_2=\alpha\times(Vsim+Tsim)$;

[0176] $$Sim=\frac{(1+\beta^2)\times(s_1\times s_2)\times\lambda}{\beta^2\times s_1+s_2}$$。

[0177] 式中,α,β,λ为可调系数,于本实施例中分别为4,2,28。Sim为最终的两个问题的匹配度分值,分值越大,表明两个问题意思越相近,越匹配。

[0178] 通过融合局部注意力信息以及全局信息,以及在词向量基础上结合基于知识图谱的远程监督,使得本申请所提出的短文本匹配算法能够应对复杂结构的语句,对语句中细微的变化也能够很好的捕捉,能够理解句子表述的中心思想,同时也能够应对特殊的社保问题知识,具有良好的扩展性,得到更加鲁棒的无监督文本语义匹配算法。而且能够将意图主体相似性、意图动作相似性、意图对象相似性、重叠部分加权指标以及句序一致性指标有效融合,使得最终的结果对句子之间的匹配度具备合理的、有效的、准确的刻画。

[0179] 本申请能够综合考虑局部感受野(注意力机制)及全局感受野两种范围的信息,既能够从微观角度,准备捕捉到句子之间存在的微小差异,又能够利用全局结构信息,对结果进行优化增强。同时,利用基于知识图谱的远程监督以及词向量相似度计算模块300,能够同时对一般情况和特殊情况的文本相似度匹配都进行有效应对,并且具有很好的知识扩展性。最后通过多特征相似度的有效融合,得到鲁棒的准确度高的文本匹配度。

[0180] 基于同一发明构思,本发明实施例还提供一种短文本问题语义匹配系统,该系统的实施可参照上述方法的过程实现,重复之处不再冗述。

[0181] 如图4所示,是本发明实施例二提供的短文本问题语义匹配系统的结构示意图,包括获取模块100、预处理模块500、文本问题感受野模块200、相似度计算模块300以及融合计算模块400;获取模块100,用于获取与用户社保相关的文本问题对;预处理模块500,用于在获取与用户社保相关的文本问题对后,对文本问题对进行文本问题预处理;文本问题感受野模块200,用于利用注意力机制对文本问题对进行真实意图特征分析,得到至少两个意图要素对;相似度计算模块300,用于利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度;融合计算模块400,用于将每个意图要素对中意图要素之间的匹配度进行多特

征匹配融合计算，并根据融合计算得到的匹配分值输出文本问题对之间的匹配结果。

[0182] 本发明能够利用注意力机制对与用户社保相关的文本问题对进行真实意图特征分析，得到至少两个意图要素对；通过意图要素对确定句子的关键信息点，从而准确识别出句子表达的真实意图；再利用语义相似度算法确定每个意图要素对中意图要素之间的匹配度；最后将每个匹配度进行多特征匹配融合计算从而输出匹配结果。使得本申请能够准确识别出句子微小的变化引起的巨大的意图差异，从而提升短文本问题语义匹配结果的准确性。

[0183] 进一步的，意图要素对至少包括意图主体对、意图动作对以及意图对象对中的两个。

[0184] 预处理模块500被配置为，利用分词工具对文本问题对进行分词，对分词结果进行词性标注，得到词性标注结果；对分词结果进行依存句法分析，得到依存句法分析结果；对词性标注结果和依存句法分析结果进行保存，生成分词关系表。分词关系表包括文本问题对中的文本问题信息、和分别与每个文本问题信息对应的分词结果信息、词性标注信息、词身份列表、头部列表以及依存关系列表；文本问题信息包括第一文本问题和第二文本问题。

[0185] 文本问题感受野模块200被配置为；利用注意力机制对文本问题对中的第一文本问题进行真实意图特征分析；利用注意力机制对文本问题对中的第二文本问题进行真实意图特征分析。

[0186] 文本问题感受野模块200还被配置为；根据分词关系表中的头部列表对第一文本问题的分词结果信息之间进行连续词识别，根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换，得到第一分词组合列表；并依据分词关系表中的词身份列表、头部列表以及依存关系列表对第一文本问题中的句子成分的依存句法关系进行分析提取，得到至少两个第一意图要素。

[0187] 文本问题感受野模块200还被配置为；根据分词关系表中的头部列表对第二文本问题的分词结果信息之间进行连续词识别，根据预设的核心词规则和连续词识别结果对分词结果信息进行识别转换，得到第二分词组合列表；并依据分词关系表中的词身份列表、头部列表以及依存关系列表对第二文本问题中的句子成分的依存句法关系进行分析提取，得到至少两个第二意图要素。

[0188] 进一步的，预处理模块500还被配置为；依据分词关系表中的词身份列表、头部列表以及依存关系列表建立关系矩阵，根据关系矩阵对文本问题对中的每个文本问题信息的核心词进行分析提取。

[0189] 进一步的，相似度计算模块300被配置为；根据预设的知识图谱，判断每个意图要素对中的意图要素之间是否等价；若意图要素之间等价，则确定意图要素对中意图要素之间的匹配度；若意图要素之间不等价，则利用词向量相似度算法确定每个意图要素对中意图要素之间的匹配度。

[0190] 相似度计算模块300还被配置为；通过预设的词向量模型训练对爬取的社保相关词汇进行训练，得到词汇和对应的词向量，将词汇和对应的词向量以键值对的形式存储成字典数据；将每个意图要素对中的意图要素对字典数据进行查询，根据查询结果获取对应的词汇和词向量；通过余弦相似度计算公式对与每个意图要素对查询得到的两个词向量进行意图要素之间的相似度计算，确定每个意图要素对中意图要素之间的匹配度。

[0191] 融合计算模块400被配置为;基于意图主体对之间的匹配度、意图动作对之间的匹配度以及意图对象对之间的匹配度利用融合公式进行多特征匹配融合计算。

[0192] 预处理模块500还被配置为;在利用分词工具对文本问题对进行分词后,根据分词结果对文本问题对之间的重叠部分进行抽取,得到公共词列表;并将公共词列表按顺序排列成重叠词列表;文本问题对中的句子包括至少一个基础字符;

[0193] 并将重叠词列表与每个文本问题对中的文本问题进行句序分析,得到每个与文本问题对应的句序索引列表。

[0194] 相似度计算模块300还被配置为;根据公共词列表、重叠词列表以及句序索引列表利用重叠词加权公式对文本问题对进行计算,得到文本问题对的加权指标;

[0195] 根据句序索引列表利用衡量公式对文本问题对进行计算,得到文本问题对的一致性衡量值。

[0196] 融合计算模块400还被配置为;在进行多特征匹配融合计算时,基于意图主体对之间的匹配度、意图动作对之间的匹配度以及意图对象对之间的匹配度,并加入加权指标和一致性衡量值利用融合公式进行多特征匹配融合计算。

[0197] 本发明虽然已以较佳实施例公开如上,但其并不是用来限定本发明,任何本领域技术人员在不脱离本发明的精神和范围内,都可以利用上述揭示的方法和技术内容对本发明技术方案做出可能的变动和修改,因此,凡是未脱离本发明技术方案的内容,依据本发明的技术实质对以上实施例所作的任何简单修改、等同变化及修饰,均属于本发明技术方案的保护范围。

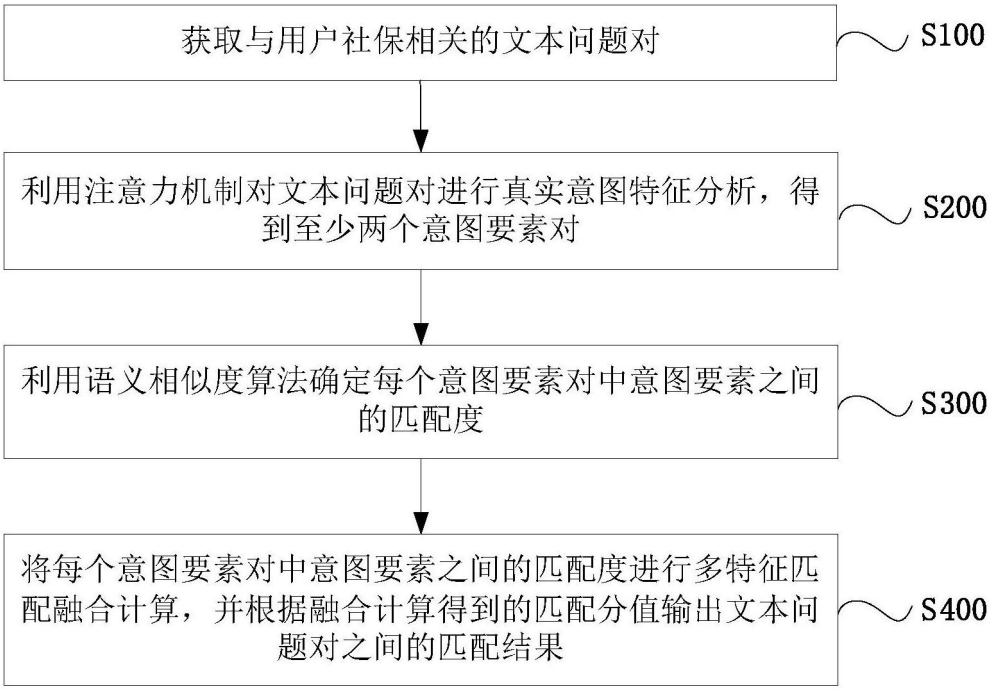


图1

$$\begin{bmatrix} '' & '' & 'CMP' & 'WP' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' \\ '' & 'QUN' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & 'ATT' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' \\ 'IS' & '' & '' & '' & '' & 'TOP' & 'ADV' & 'HED' \end{bmatrix}$$

图2

$$\begin{bmatrix} '' & '' & '' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' & '' \\ 'LOC' & 'ADV' & 'HED' & 'VOB' & 'WP' & '' & 'IC' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & 'ADV' & '' & '' & 'VOB' \\ '' & '' & '' & '' & '' & '' & '' & '' & '' \\ '' & '' & '' & '' & '' & '' & '' & 'ADV' & '' \end{bmatrix}$$

图3

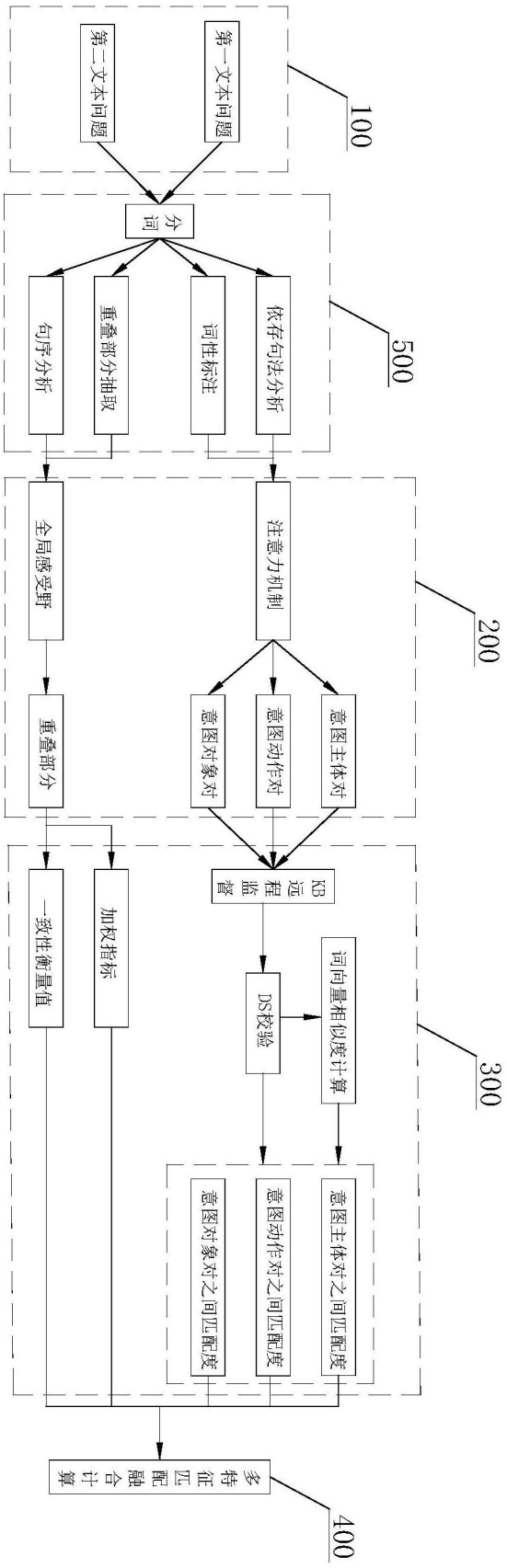

第二文本问题
第一文本问题
100
分词
句序分析
重叠部分抽取
词性标注
依存句法分析
500
全局感受野
注意力机制
200
重叠部分
意图对象对
意图动作对
意图主体对
KB远程监督
一致性衡量值
加权指标
DS校验
词向量相似度计算
300
意图对象对之间匹配度
意图动作对之间匹配度
意图主体对之间匹配度
多特征匹配融合计算
400

图4