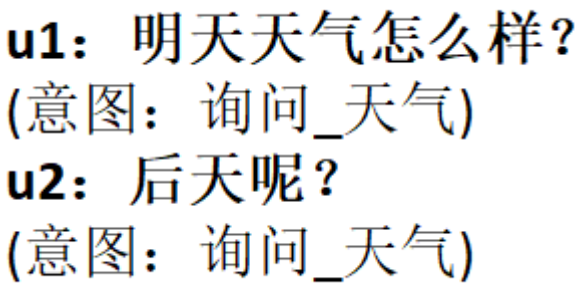
多轮对话下的多标签分类

摘要

Spoken language understanding (SLU)是对话系统的核心组成部分，其中包括领域分类，意图分类和槽填空[1]。意图分类属于多标签文本分类问题，对话内容之中可能存在多个意图，对多个意图的准确识别是对话系统中关键的一个环节。对话系统按处理粒度可以分为单轮对话和多轮对话。单轮对话下的意图分类存在两个较大的问题，第一点：一个意图是很难通过一句话完成，这使得单轮的意图分类变得困难。第二点：一个意图可能和之前的意图有关系，这使得单独的一轮对话很难进行准确的意图识别。我们将提出一种模型来处理上面的两个问题，基于多轮的对话情景来完成意图分类，识别对话的多个意图。首先利用存储器来保存上下文的信息，利用attention机制筛选有用信息，并利用门控制是否利用上下文，使得模型适应于更多的应用场景。基于多轮的对话情景来完成意图分类，识别对话的多个意图。

介绍

对话系统是人工智能和自然语言处理的研究人员长久以来倍感兴趣研究但是应用效果不佳的一个应用研究方向。对话系统可以用来解放大量的客户服务人员，如帮助预定电影票，解决售后预约，充当导游等。多轮对话下的意图识别是多标签文本分类的一个应用，目前是热门研究的方向。微软的小冰，百度的度秘，苹果的Siri等聊天机器人，以及一些任务驱动的多轮对话机器人都把Spoken language understanding (SLU)中的意图分类作为最基础最重点的研究之一。所谓的意图包括说话者的语言行为和相关属性。意图的单轮对话的意图分类定义为：给定当前对话内容ui，目标是预测出ui的意图。多轮对话的意图分类定义为：给定当前对话ui和之前的聊天内容vi，目标是预测出ui的意图。如下图1所示，u1的意图仅靠单轮对话就可以判断，但是这里如果没有上文的信息，u2的意图分类却比较模糊。

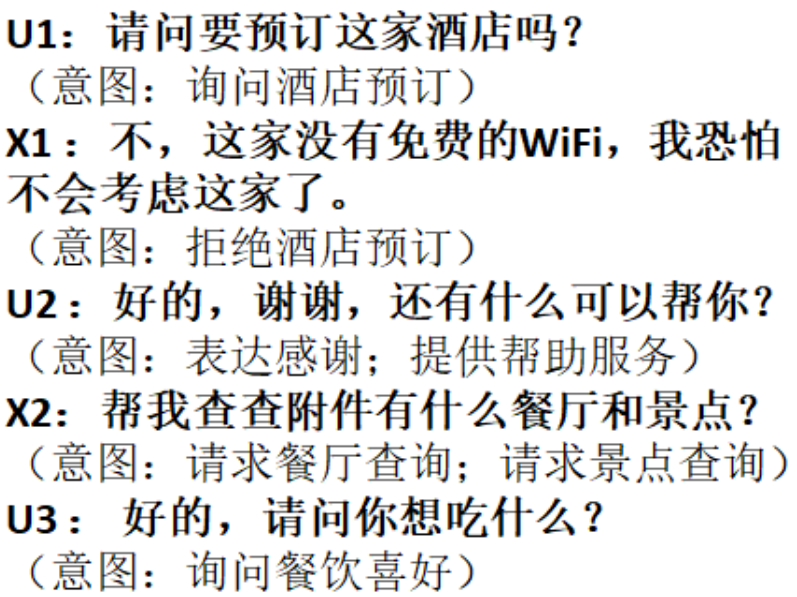


传统的文本分类方法有支持向量机（SVM）和最大熵等，过去常被用来做新闻等长文本分类，其应用到对话系统上，在对话片段的领域分类上，依旧有不错的表现[2][3][4]。但应用于对话系统的意图分类这类短文本分类上，由于短文本本身较短，这类方法难以抽取有效的特征信息，导致分类效果较差。随着深度学习的发展，深度神经网络（DNNs）因为其强大的特征抽取能力，被应用于意图分类[5][6][7]。最近Ravuri et al.提出一个RNN框架来解决意图分类问题[8]；Xu et al.提出利用CNN框架结合CRF模型同时解决意图分类和槽填空问题[9]； Hakkani-Tur et al.提出一个基于RNN框架的多个领域多个任务联合学习的模型[10]。然而这些模型都聚焦于单轮的对话系统，每个对话的意图都是相互独立的。这导致模型在很多应用场景中受到限制，不能被广泛应用于现实场景中。

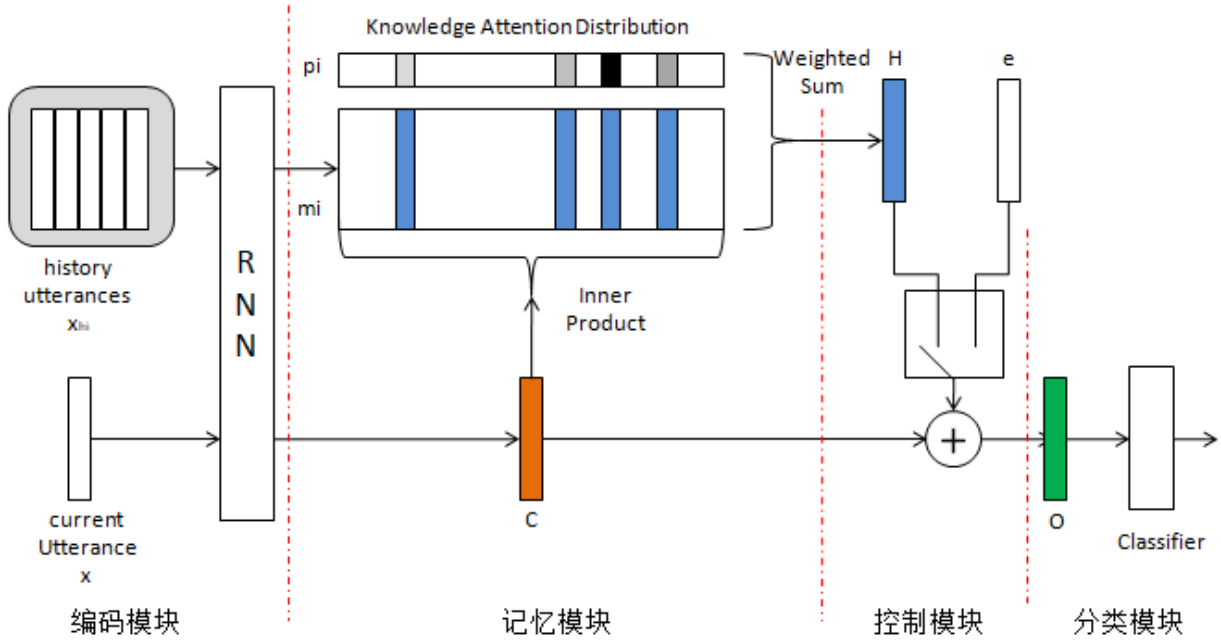
上下文的信息对于理解当前对话的意图有着重要的作用，如上文提到的多轮对话的例子，u2的意图被预测为询问_天气需要依靠u1对话中的信息，这种省略在口语对话中非常常见。在这种情况下，需要利用上文对话提供必要的信息才能判断当前意图。 Hakkani-Tur et al.提出SVM-HMMs从上文的对话中抽取信息联合当前对话来实现槽填空和意图识别[11]；Xu et al.提出利用RNNs框架结合上文的对话信息来实现领域分类和意图识别[12]。然而过去的这些

方法仅仅利用上文对话内容抽取信息，忽略了长依赖和主次信息，这将导致上文对话信息的利用率降低。

最近，有一些计算模型会使用存储器和注意力概念来提升模型效果[13][14][15]。对于很多神经网络模型，缺乏了一个长时记忆的组件方便读取和写入。作为RNN，lstm和其变种gru都使用了一定的记忆机制。然而这些记忆都太小了，因为把状态及其权重全部都嵌入到一个低维空间，把这些知识压缩成一个稠密的向量，丢失了不少信息。Memory Networks 通过添加一个存储器，实现了长期记忆（大量的记忆），并且易于实现从长期记忆中读取和写入。存储器结合神经网络构成连续的表示模型，将编码的知识信息通过读和写的操作更新存储器中的信息。注意力机制可以将从存储器中读取的信息分为主要信息和次要信息，从而使得历史信息的利用率提高，提高模型效果。Hakkani-Tur et al.提出一个END-To-END的RNNs框架利用存储器和注意力机制实现多轮对话的意图分类[15]。然后模型解决了当前意图和上文相关的问题，但是对于出现的新的意图，与上文不相关的情况下，模型会因为引入的历史信息而引入大量的噪声。我们提出一个新的改进模型，在模型读取存储器的时候，为模型添加一个门，来控制在判断当前对话意图的时候，是否需要引入历史信息。如果意图与历史信息相关，门通过读取历史信息，来帮助当前对话意图的识别。如果意图与历史信息不相关，门拒绝读取历史信息，仅仅依靠当前对话识别意图。这样做，可以综合考虑多个意图的出现的多种情况，防止在某些情况下引入噪声。如下图2所示场景，U1和X1，X2和U3之间的意图有承接关系，判断X1和U3的意图时，需要通过门，访问存储器中的上文数据。另外在解决用户的查询餐厅的意图后，还需要历史数据中X2提及的另一个意图信息，已更好的提供服务。而U1，U2和X2的意图与上下文不相关，可以直接通过本次对话内容，直接判断意图，不需要访问存储器中的历史数据。



模型



向量表示

把文本表达为可以让计算机来理解的形式，所谓文本表示，文本向量化。文本向量化，可以分为词语的向量表达，短文本的向量表达，长文本的向量表达, 因为不同的情景需要用到不同的方法和处理方式。对话系统的输入是短文本，采用短文本的向量表达。文本向量化是通过embedding模块来完成。embedding模块目前常用的有四种方法，1.embedding_rand:所有的词向量都随机初始化，在模型训练过程中优化参数；2.embedding_static:所有的词向量直接使用无监督学习方法word2vec预先训练好词向量，训练过程中不在变化；3.embedding_non_static:所有的词向量直接使用无监督学习方法word2vec预先训练好词向量，训练过程中通过微调优化。本实验中，由于训练的对话数据集较小，embedding_rand方法会导致测试时中出现大量的未登陆词，而embedding_non_static方法容易使得模型过拟合，所以选择embedding_static作为embedding模块的方法。利用大量的维基百科数据训练词向量模型，再本模型训练过程中，不在调节词语的词向量表示。

embedding模块实现介绍：

输入：不定长字符串，即为每个对话的内容。

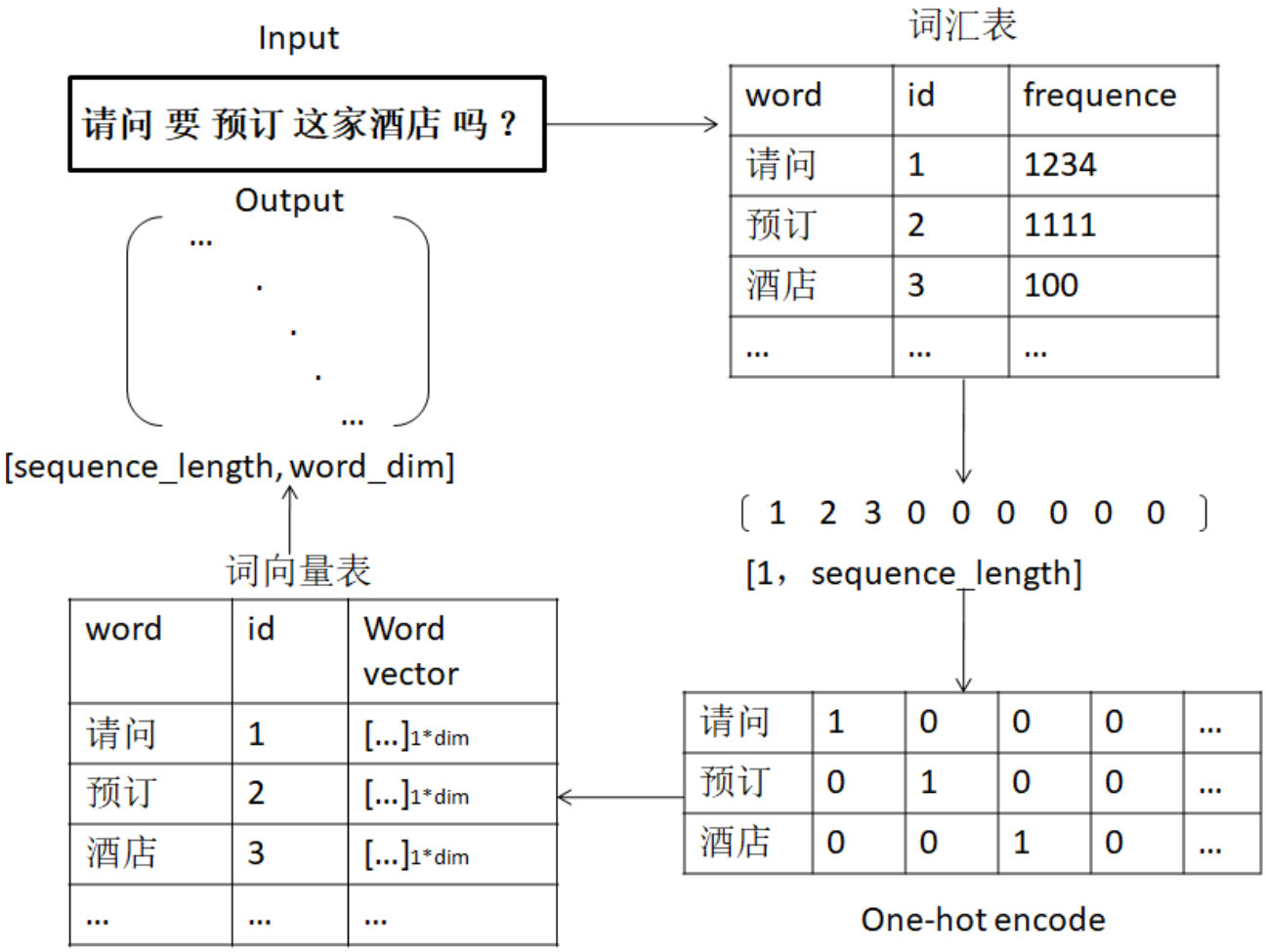
词汇表：先将文本数据集中不重复的单词提取出来，得到一个词汇表。如果文本数据集太大，那么得到的词汇表中可能存在几万个单词，建立的索引表会很大。因此，我们通常在计算词汇表的时候，可以排除那些出现次数太少的单词，得到一个大小为V的词汇表。将词汇表中的词按照在数据集中出现的词频顺序排序，对每个词汇表中的词对应排序位置建立索引。

词向量表：根据预训练好的词向量模型和词汇表，建立大小为V的词向量表。词汇表和词向量表之间通过id相互对应。词向量的维度大小为word_dim。

定长：为了批量化处理，可以将不定长的输入补充到指定长度sequence_length。通常可以在原输入前面或者后面补充0，所以词汇表和词向量表的id从1开始计数。此时，输入被表示为维度为（1，sequence_length）的2D张量。

one-hot：用一个V维的向量来表示输入文本的每个词汇表中的词汇，向量中的第d个维度上的1表示词汇表中的第d个单词出现在文本中，0表示单词未出现在文本中。

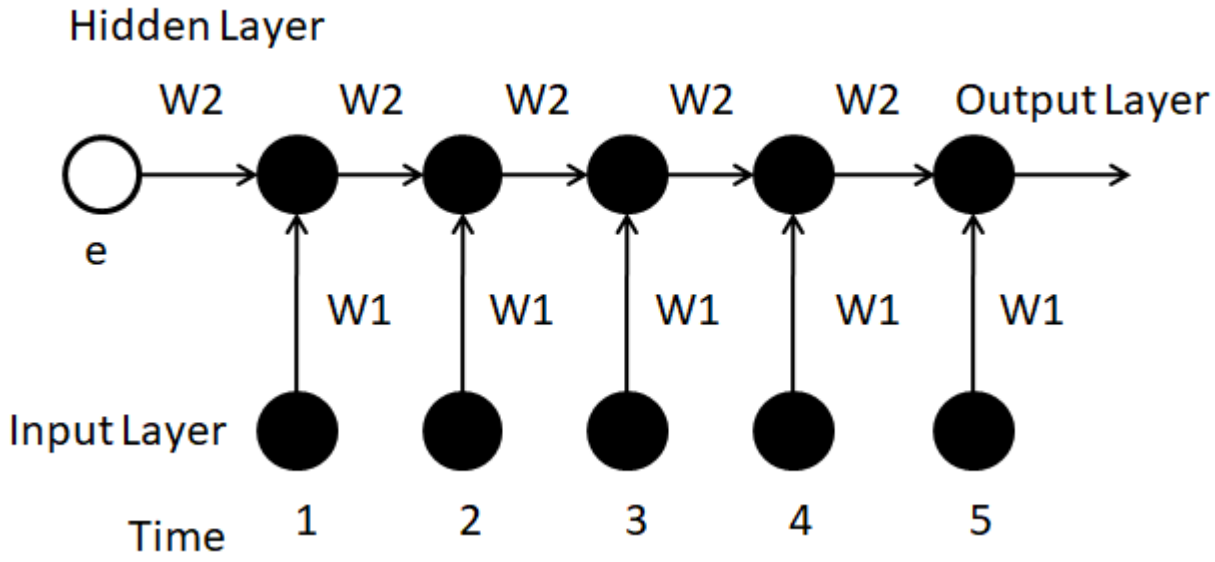
输出：利用词向量表和词汇表的映射关系，将输出表示为维度（sequence_length，word_dim）的2D张量，即输入的每个词替换为了对应词向量。对于为了保持指定长度sequence_length而补充的0，对应位置的词向量用全0向量填充。



句子编码

Sentence encoder模块在整个系统中是最基础的模块，该模块将对话中的语义信息抽取出来，是保证系统能够理解对话，确保识别对话意图的基础。Sentence encoder模块常用的模型有CNNs，RNNs和改进后的LSTMs。CNNs模型可以捕捉局部语义信息，考虑了局部词汇间的时序问题。当我们去尝试预测"I grew up in France...I speak fluent *French*"的最后一个单词，最近的信息表明下一个单词应该是语言的名字，但是如果我们想缩小语言的范围，看到底是哪种语言，我们需要France这个在句子中比较靠前的上下文信息。相关信息和需要预测的点的间隔很大的情况是经常发生的。因而CNNs仅仅考虑局部时序是无法解决长依赖问题的。LSTMs模型通过复杂的链式结构和内部组件的相互交互，有效的解决了时序问题和长依赖问题。对于本模型LSTMs应该是最合适的模型，但是由于LSTM模型复杂，需要大量的训练数据，若没有充足的训练数据，容易造成过拟合。标注大量的训练数据需要耗费大量的人力物力，且对话文本的内容并不是很长，可以折中考虑采用RNNs模型作为句子编码模型。

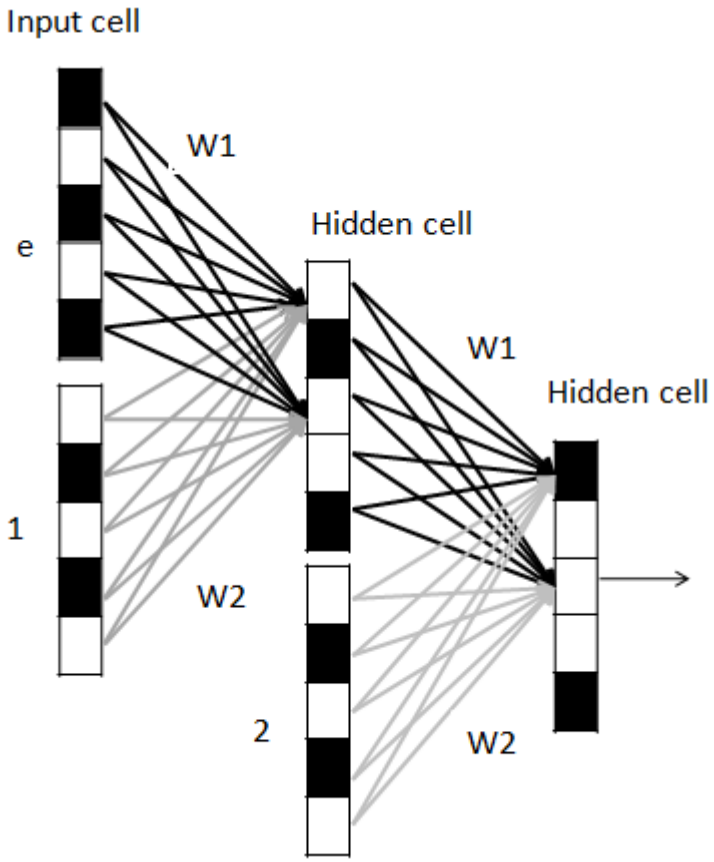
RNN模型分解图如下：



其中，e是补充的开始向量，通常可以设置为维度为（1，word_dim）的全0的2D张量，1，2，3，4，5为输入层的词向量，它们的维度为（1，word_dim）的2D张量。W1，W2是RNN模块的参数，它们的维度为（word_dim，word_dim）的2D张量。最后将迭代的最后一次的隐藏层输出向量作为该模块的输出向量，输出向量的维度为（1，word_dim）的2D张量。

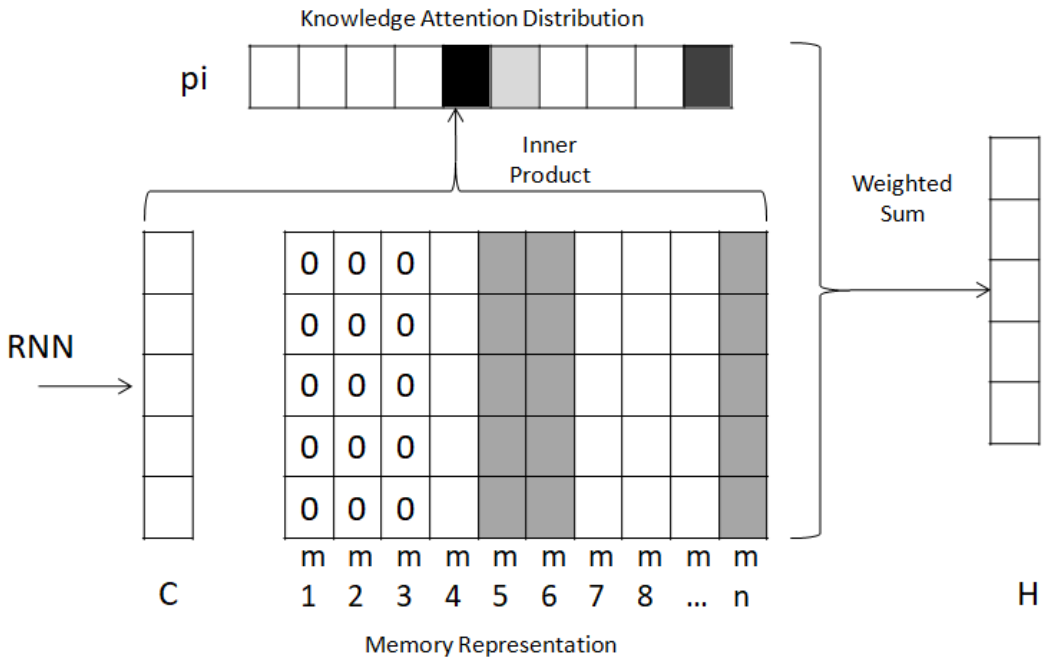
$$RNN(i) = g(W1 * i + W2 * RNN(i-1)), RNN(0) = e$$

具体展开如下图：



记忆模块

memory模块是实现多轮对话下意图识别的关键，该模块将对话中抽取的语义信息存放起来，保证在需要这些信息的时候可以进行便捷的读取，还要保证现有数据可以方便写入。记忆方式分为两种，内部记忆和外部记忆。传统的深度学习模型（RNN、LSTM、GRU等）使用隐藏状态或者Attention机制作为他们的记忆功能，但是依靠内部记忆方法产生的记忆太小了，无法精确记录一段话中所表达的全部内容，也就是在将输入编码成稠密向量的时候丢失很多信息。并且内部记忆不易从记忆中读取需要的信息，经过编码处理后，所有信息都被高度抽象，无法灵活方便提取信息。而外部记忆则对于本系统更加友好，对于多轮对话而言，并不是所有的上下文对话信息都对本轮对话的意图判断有影响，我们仅仅需要那些相关的信息。这就要求系统对于历史信息的筛选具有一定能力。外部记忆的存储结合attention机制可以很容易实现相关历史信息的读取，并且处理之后写入时，可以保证信息的时序。



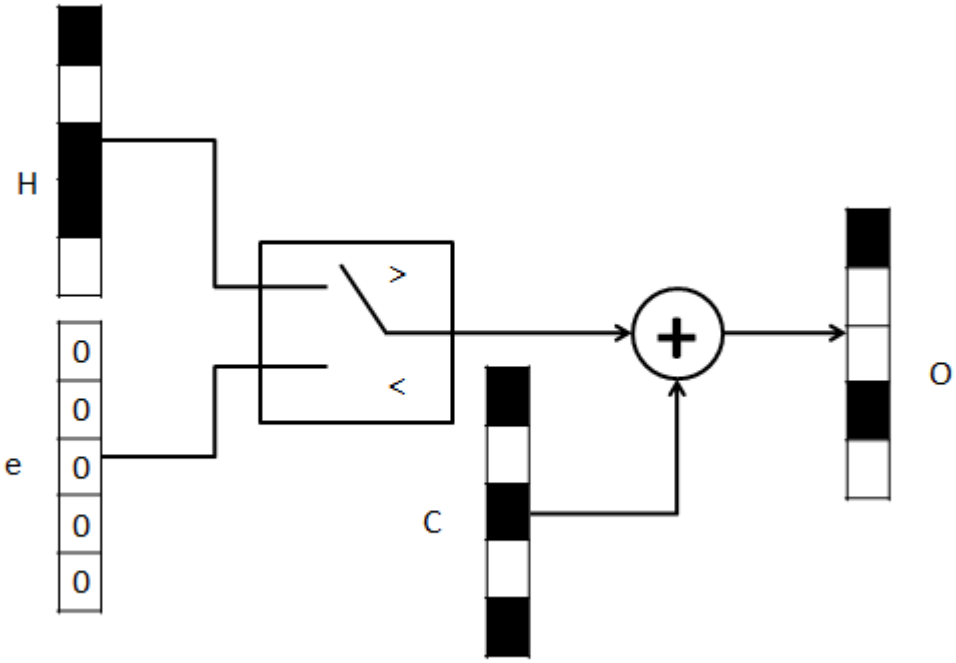
其中C，m_i 和H是维度为（1，word_dim）的2D张量，p_i 是常数。$RNN(x)$是当前输入向量，即当前对话信息，$RNN(x_{hi})$是存储器中存储的历史向量，即历史对话信息。

$$m_i = RNN(x_{hi})\ C = RNN(x)\ p_i = softmax(C^T m_i), softmax(z_i) = e^{z_i} / \sum_j e^{z_j}\ H = \sum_i p_i m_i$$

所以该模块的输入为（1，word_dim）的2D张量C和（memory_length，word_dim）的2D张量M，memory_length是记忆模块的存储器大小。初始化时，存储器内存储着全0的2D张量，存储器类似于计算机中的定长的队列。按距离当前对话的时间远近顺序，将历史信息编码后的向量由远及近的插入队列中。该模块输出一个（1，word_dim）的2D张量H。

控制模块

多轮对话下的意图识别并不是每轮对话都需要多轮的历史信息。控制模块就是保证系统既可以处理依靠历史信息，有可以摆脱历史信息的干扰。对于需要历史信息才能准确理解对话内容意图的时候，从记忆中抽取出相关的信息，可以帮助对话的理解。而对于新出现的意图，其独立于历史信息之外，尤其在意图转换的时候，如果这时系统强行抽取无关的信息干扰对话意图的判断，将会导致系统效果大大降低。所以控制模块就是一个读入阀门，如果当前对话与历史信息相关，则读取历史信息的相关信息；如果当前对话与历史信息无关，则读取空信息。



该模块的输入为H，e和C和p ，H，e和C均为维度为（1，word_dim）的2D张量，p为上一个记忆模块的参数 $p = max(p_i), (0 < i < n)$ 。当p大于设置的阈值，则$O = W_{kg}(H + C)$ ，否则$O = W_{kg}(e + C)$ ，W_{kg}为联合参数，维度为（word_dim，word_dim）。该模块的输出为维度为（1，word_dim）的2D张量O。

分类模块

分类模块采用多层全连接结合softmax层构成，该模块输入为维度为（1，word_dim）的2D张量O，输出为维度为（1，label_dim）的2D张量，label_dim是标签集合的元素个数。

参考文献

[1] Gokhan Tur and Renato De Mori. 2011. Spoken language understanding: Systems for extracting semantic information from speech. John Wiley & Sons.

[2] P. Haffner, G. Tur, and J. H. Wright, “Optimizing svms for complex call classification,” in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP), vol. 1. IEEE, 2003, pp. I–632.

[3] C. Chelba, M. Mahajan, and A. Acero, “Speech utterance classification,” in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP), vol. 1. IEEE, 2003, pp. I–280.

[4] Y.-N. Chen, D. Hakkani-Tur, and G. Tur, “Deriving local relational surface forms from dependency-based entity embeddings for unsupervised spoken language understanding,” in 2014 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2014, pp. 242–247.

[5] R. Sarikaya, G. E. Hinton, and B. Ramabhadran, “Deep belief nets for natural language call-routing,” in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 5680–5683.

[6] G. Tur, L. Deng, D. Hakkani-Tur, and X. He, “Towards deeper un- ¨ derstanding: Deep convex networks for semantic utterance classification,” in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012, pp. 5045–5048.

[7] R. Sarikaya, G. E. Hinton, and A. Deoras, “Application of deep belief networks for natural language understanding,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 778–784, 2014.

[8] S. Ravuri and A. Stolcke, “Recurrent neural network and lstm models for lexical utterance classification,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[9] P. Xu and R. Sarikaya, “Convolutional neural network based triangular CRF for joint intent detection and slot filling,” in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2013, pp. 78–83.

[10] Yun-Nung Chen, Dilek Hakkani-Tur, Gokhan Tur, Jianfeng Gao, Li Deng and Ye-Yi Wang, “Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM,” In Proceedings of The 17th Annual Meeting of the International Speech Communication Association. pages 715–719.

[11] A. Bhargava, A. Celikyilmaz, D. Hakkani-Tur, and R. Sarikaya, “Easy contextual intent prediction and slot detection,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2013, pp. 8337–8341.

[12] P. Xu and R. Sarikaya, “Contextual domain classification in spoken language understanding systems using recurrent neural network,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 136–140.

[13] J. Weston, S. Chopra, and A. Bordesa, “Memory networks,” in International Conference on Learning Representations (ICLR), 2015.

[14] S. Sukhbaatar, J. Weston, R. Fergus et al., “End-to-end memory networks,” in Advances in Neural Information Processing Systems, 2015, pp. 2431–2439.

[15] Yun-Nung Chen, Dilek Hakkani-Tur, Gokhan Tur, Jianfeng Gao, and Li Deng. 2016c. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In Proceedings of The 17th Annual Meeting of the International Speech Communication Association. pages 3245–3249.