基于上市公司公告的事件要素抽取研究

鲁娇[1]and 秦文[2]

[1] 北京科技大学，北京，100083
[2]洞见时代（北京）信息技术有限公司，北京
lujiaoff@163.com
qinwen@insight-time.com

摘要. 事件抽取作为信息抽取的一个重要的子任务，可以从非结构化的自然语言文本中发现特定的事件，事件的事件要素中往往包含更加重要等信息，因此研究事件特别是其包含的事件要素具有十分重要的意义。本文使用上市公司的“人事变动”类型公告作为数据集，选择姓名实体作为事件的触发词来识别事件，使用添加注意力机制的BI-LSTM-CRF模型，将序列-触发词作为共同输入来进行触发词对应事件的事件要素的抽取，从而获取每一个文本对应的一个或多个事件信息，存储为该无结构化文本的结构化信息，为事件及事件要素提取提供新的思路。

关键词: 事件要素，注意力机制，BI-LSTM-CRF

1 引言

随着互联网的快速发展以及电子产品的普及，越来越多的信息以电子文本的形式呈现在人们眼前，人们也越来越倾向于从网上获取和发布大量的信息。但是网上的海量信息都是以半结构化的格式组织，和结构化数据有所不同，半结构化的数据形式在数据查询和获取时具有一定的难度。面对网上的海量数据，人们如何高效快速的从这些海量数据中得到有用的信息就成为了一个难题，也成为了众多领域学者的研究热点。正是在这种社会背景下，信息抽取成为了当前研究的一个重要方向。信息抽取是当前自然语言处理中的一个重要分支，事件抽取是信息抽取的重要组成部分。事件抽取就是从非结构化文档中抽取出用户感兴趣的事件,同时用结构化的形式描述,供用户查询及进一步分析。事件抽取在自动文摘、信息检索、问题题回答系统等方面有着广泛的应用。

事件抽取作为信息抽取的重要组成部分，主要包括两个步骤，第一步是事件的识别，第二步是对识别出来的事件进行分析，进而抽取其中的事件要素，这些要素包括事件发生的时间、地点、事件的参与者等。事件识别作为事件抽取的基础，事件识别的效果直接影响了事件抽取的结果。近年来，事件识别为面向事件的一些自然语言处理应用提供支持，比如在自动文摘、问答系统、信息检索、机器翻译等方面都有着广泛的应用。

事件抽取属于信息抽取领域，在事件抽取技术的发展过程中，MUC（ message understanding conference）会议和ACE（automatic content extraction）会议起了很大的推动作用。根据ACE中的定义，事件由事件触发词(trigger) 和描述事件结构的元素(argument) 构成。事件抽取的很多相关研究也就是围绕着触发词和事件元素来进行的。相应地，事件抽取的任务可分解为两步进行：a）从一篇文本的句子集中抽取出事件句： b）再从事件句中抽取出事件元素。ACE（ Automatic Content Extraction ）2005 事件抽取任务定义成法律制裁（Justice）、冲突（Conflict）、商业（Business）等8 个大类32 种子类型任务，但是ACE 2005 所定义的事件类型存在着类型过于宽泛、针对性不强的问题，例如Business 中的Start-Org（组织成立）、Movement 中的Transport（中转站）在使用中并无实际价值，不能真正满足现实社会对事件抽取的需求，因此还必须针对特定专业领域重新进行事件模型和类型的定义。

随着国内市场经济不断发展，特别是股市经济，对金融事件越来越敏感。然而在当下，面对海量的互联网金融信息，单纯依靠人工的分析很难达到实际的要求，所以研究面向金融领域的事件抽取对于深入分析金融领域的文本信息、为投资决策提供支持具有重要意义。

对于金融领域来说，仅仅识别公告的事件类型是远远不够的，大量的信息还包含在事件要素中，如在“人事变动”类型的公告中，仅仅提取辞职事件及辞职者姓名并无法获取完整的信息用于公司现况的判断或股价的预测，当我们提取完整的事件要素信息，包括辞职原因、职务、继任者的信息时，就可以得到更多该公司的信息，可以让我们高效的利用各种不同层次的信息，从而为支持投资决策提供重要的依据。

因此，对于事件要素的识别尤为重要。目前，对于事件要素的研究，国内还处在初始阶段，也还没有成熟的技术来自动识别事件中的事件要素，较为流行的方法有两种：一种方法是基于规则和模板的方法，一种是基于机器学习的方法。本文尝试通过采用神经网络模型的方法，将事件要素识别与抽取问题转变为序列标注任务，使用改进的序列标注模型进行不同事件要素的标注，实现对事件要素的识别与提取，以帮助人们获取有用的事件信息。事件要素识别在实际生活中具有重要的研究意义和实际价值。

针对以上问题，本文以公众公司“人事变动”类型公告文本为研究对象，使用通用命名实体识别模型抽取事件的触发词，根据抽取的触发词，提出了添加Attention机制的BI-LSTM-CRF模型进行事件要素的抽取，从而可以避免复杂的规则定制和特征工程，利用神经网络自动提取文本特征的优点，更好地实现事件要素的抽取。

2 数据来源

本文的课题来源于CCKS2019公众公司公告信息抽取评测任务，其中任务二为文本段落中的信息点提取，使用上市公司“人事变动”类型公告，提取其中的高管离职与继任信息。作为知识图谱构建的基础，结构化数据是必不可少的，因此，通过自动化的技术从公告中提取关键信息，将非结构化文本转化为

结构化数据是本课题的目标。本文的主要内容是基于给出的训练文本，建立模型用于事件及事件要素抽取的抽取。

本文的数据来源于评测任务给出的617个“人事变动”类型公告文本，文本中所涉及的事件为高管离职与继任信息，文本内容可大致分为三类：一类仅涉及一个或多个高管离任信息，一类仅涉及一个或多个高管的聘任信息，第三类则既包含一个或多个高管的离任信息，有包含继任其职位的聘任信息。除了文本之外，评测方还给出了人工提取的存储为JSON格式的每一个文本包含的事件信息，该事件信息包括：公告名称，证券代码、证券简称、公告类型（均为“人事变动”）、离职高管姓名、离职高管性别、离职高管职务、离职原因、继任者姓名、继任者性别、继任者职务，对于文本中未提及的的字段信息，结果为空。

3 事件要素抽取方法

近年来，随着互联网的广泛发展，人们想要获取想要的信息越来越困难，尤其是在金融领域，大量的金融信息充斥在视线中，导致人们对事件类信息的抽取提出了迫切的要求，研究面向金融领域的事件抽取对于深入分析金融领域的文本信息，为投资决策提供支持具有重要意义。对于事件类信息，仅仅检测事件的发生并不能够满足人们的需求，人们更想知道事件发生的具体的细节类信息，因此，对于事件要素的识别尤其重要。

本文拟将神经网络运用到事件要素的提取中来，面向上市公司公告“人事变动”类型事件要素抽取，将事件要素抽取问题转化为序列标注问题，使用标注了事件要素的训练语料进行事件要素抽取模型的训练。

下图表示本文的事件要素抽取流程框架，框架主要包含三部分：数据预处理，触发词提取，训练用于事件要素抽取的包含注意力机制的Bi-LSTM-CRF序列标注模型，使用测试数据在训练完成的模型上进行事件要素抽取；

本文提出的事件要素抽取流程为：根据已有的训练数据进行序列的标注，作为事件抽取模型 的训练语料，建立包含注意力机制的Bi-LSTM-CRF模型作为事件要素的抽取训练模型，将训练好的模型保存。对于输入的待抽取文本，对文本进行简单预处理后首先进行触发词的提取，之后使用训练好的事件要素抽取模型模型进行事件要素的抽取，最后得到事件抽取结果。下图展示了事件要素抽取流程的框架。

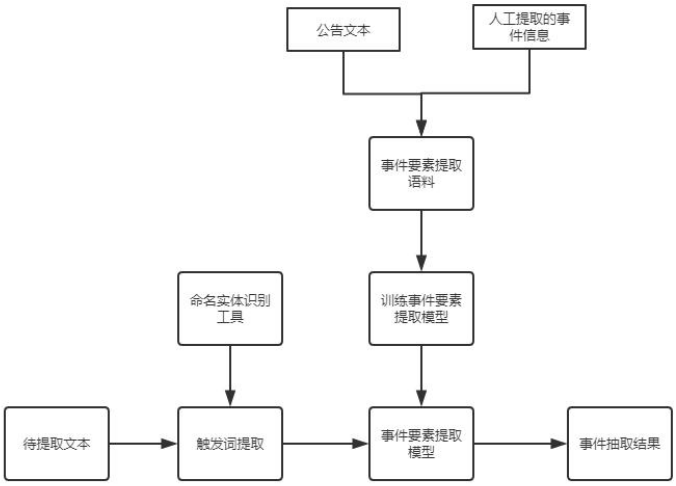

Fig. 1. 事件要素抽取框架

在生成事件要素抽取模型的部分，包含两块主要结构：数据预处理和事件要素抽取模型的构建。

数据预处理：将原始文本语料进行预处理，生成可以用于模型训练的长文本序列，划分为训练集和验证集，同时从人工提取的事件信息中找到触发词；

训练事件要素抽取模型：构建包含注意力机制的Bi-LSTM-CRF模型，首先对输入进行字嵌入操作，得到输入矩阵的高维矩阵表示，然后使用注意力机制融合文本序列与触发词的信息，之后通过双向LSTM层进行训练，最后经过CRF层进行序列标签的预测，完成了事件要素的抽取，输出的序列标注结果即为抽取结果。使用训练集进行训练，验证集来验证模型效果，实现最好效果时保存模型用于之后事件要素的抽取。

3.1 数据预处理

本文的训练语料使用CCKS官方给出的617条上市公司“人事变动”类型公告文本为生语料，根据已给出的每一篇文本的事件提取结果（包含离职高管信息与继任者信息），对生语料进行序列的标注。

首先，将原始的PDF文本进行处理，提取成长文本序列，预处理长文本序列，去掉大量的无信息句，如“本公司及监事会全体成员保证公告内容真实、准确和完整，没有虚假记载、误导性陈述或者重大遗漏。”根据每一篇文本对应的事件信息，在原序列中进行标注。对于包含单一事件的文本，在对应字符位置标注相应的标签即可。

其中：

离职高管姓名：LN

离职高管性别：LS

离职原因：LR

离职高管职务：LW
继任者姓名：RN
继任者性别：RS
继任者职务：RW
无标签：O

例1：
廖绮云女士表示因工作精力和工作安排等原因，申请辞去公司非职工代表监事职务。
对应的标注序列：
廖LN绮LN云LN女LS士LS表O示O因O工LR作LR精LR力LR和LR工LR作LR安LR排LR等O原O因O，O申O请O辞O去O公O司O非LW职LW工LW代LW表LW监LW事LW职O务O。O

对于包含多个事件的文本，如果将多个事件同时表示，在取出文本的过程中会出现事件元素错误匹配的问题，因此，对于这种情况，我们采用多次标注方法。

例2：
陈彦君先生、庄兴先生因个人原因申请辞去公司总经理、副总经理职务。
对应的标注序列：
陈LN彦LN君LN先LS生LS、O庄O兴O先O生O因O个LR人LR原LR因LR申O请O辞O去O公O司O总LW经LW理LW、O副O总O经O理O职O务O。O
陈O彦O君O先O生O、O庄LN兴LN先LS生LS因O个LR人LR原LR因LR申O请O辞O去O公O司O总O经O理O、O副LW总LW经LW理LW职O务O。O

因为我们采用了姓名作为事件的触发词，每个触发词可以唯一对应一个事件，所以在标注时尽管两个文本序列相同，根据其中含有的多个事件信息，会生成不同的标注序列，在用标注好的语料进行训练时，为了避免出现同一序列中相同的字对应不同标签的问题，我们需要使用该序列的触发词作为额外的输入，这样，使我们的训练数据从原来的纯文本及人工提取的json格式数据转换成了触发词-文本序列，这样就可以保证每一条训练数据在事件提取时只能够提取出一个事件及其事件元素。

另外，由于公告文本较长，在标注语料时较容易出现问题导致影响训练效果：
1、原人工事件提取数据中的职务信息为多个并列，标注出错导致提取结果职务出现错误；
例3：张三辞去董事会董事职务，同时辞去总经理职务
在已经由人工提取的结果中，职务信息为“董事会董事、总经理”，由于在语料标注时的方式为文字匹配，这就导致这句话在语料标注时会出现无法匹配的情况，即该句没有标注职务，导致的错误率的提升。

2、对于某些职务的情况，由于是对字进行标注，容易出现歧义，如将董事会中的董事标注为职务名称，将监事会的监事标注为职务。

因此，需要对自动标注的结果进行二次校验，才能保证训练语料的准确性，从而保证后续模型训练的效果。

3.2 触发词提取

在传统的事件抽取任务中，触发词提取是事件抽取的一个子任务，也是非常重要的任务，它主要从句子中抽取事件的触发词，一般为句子的谓语动词，抽取触发词后，判断事件类型。

例：廖绮云女士表示因工作精力和工作安排等原因，申请辞去公司非职工代表监事职务。

其中“辞去”为本句中的触发词，事件类型为离职，可以提取的离职事件为{廖绮云-辞去-非职工代表监事}。

但是在本文所使用的数据集中，经常会包含多个离职信息，出现多个主语对应同一个谓语动词的情况，此时，事件发生的谓语动词无法唯一代表一个事件，可能对应多个事件，此时，提取“辞去”为触发词对于后续提取事件要素产生了一定的阻碍。

例：陈彦君先生、庄兴先生因个人原因申请辞去公司总经理、副总经理职务。

在上例中，在事件提取时，传统方法会将“辞去”作为事件的触发词，但此时，“辞去”这一触发词对应了两个离职事件{陈彦君-辞去-总经理}，{庄兴-辞去-副总经理}，此时“辞去”无法唯一对应文本中的一个事件，对于后续提取事件要素时会产生错误的累积。因此，在本文中，我们使用“姓名”作为事件的的触发词。在上市公司“人事变动”类型的公告中，每一个公告的内容都是离职或聘任事件，其中必然会包含姓名实体，该实体将唯一代表一个离职或聘任事件。所以在使用标注好的语料进行训练时我们需要使用离职高管姓名或继任者姓名（仅在只有聘任者信息时）作为触发词，使我们的训练数据从原来的纯文本及人工提取的json格式数据转换成了语句标注序列和唯一代表事件的姓名信息，这样就可以保证每一条训练文本在事件提取时只能够提取出该姓名实体对应 的事件及其事件要素。

在自然语言处理的各项任务中，命名实体识别是一项比较基础的任务，已有的工具在这项任务上可以达到非常不错的效果，本文在姓名触发词的提取上使用哈工大的LTP语言技术平台进行姓名触发词的提取。

3.3 基于注意力机制的事件要素抽取模型

注意力机制源于对人类视觉的研究，在认知科学中，由于信息处理的瓶颈，人类会选择性地关注所有信息的一部分，同时忽略其他可见的信息，上述机制通常被称为注意力机制。例如，人们在阅读时，关注和处理关键词汇即可帮助完成对全文的理解。与阅读过程相似，在演化关系抽取中，关注语料中重要词汇对实体要素抽取会有重要帮助，所以本文在模型中引入注意力机制。

本文使用添加注意力机制的Bi-LSTM-CRF模型进行事件元素抽取。将触发词抽取模型得到的姓名触发词对和长文本作为输入，预测各个事件要素。由于我们把事件元素抽取作为序列标注任务来做，使用字嵌入方式进行向量的输入，添加注意力层可以使得文本中与该触发词相关的文字可以获得更高的注意力权重，之后经过双向LSTM层进行序列特征的建模，最后经过CRF层实现序列标签的标注，得到事件要素抽取结果。

事件要素抽取模型如下图所示：

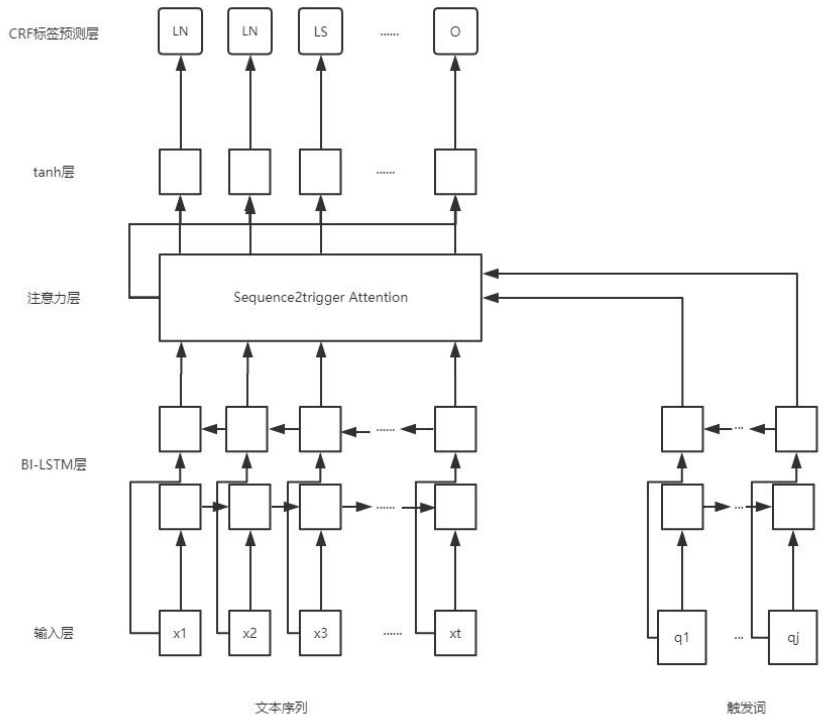

Fig. 2. 添加Attention机制的Bi-LSTM-CRF模型

模型的网络结构包含文本输入层，双向LSTM层，注意力层CRF层，tanh层和CRF标签预测层。

输入层：网络结构的第一层为输入层，主要步骤包含文本的字嵌入与触发词对的字嵌入。字嵌入是将每个字转化向量的过程。输入N个文本序列，对于输入的每一个文本序列S={s1,s2,s3,…st}，将其进行文本的向量化，得到整个句子的字嵌入矩阵X={x1,x2,x3,…xt}，最后得到N*T*DIM的三维字嵌入矩阵。

Bi-LSTM层：网络结构的第二层双向LSTM层，输入对象为字嵌入操作后的三维矩阵，对于每一条句子文本数据，输入到LSTM模型中进行特征的提取，双向LSTM能够充分利用整个文本序列的信息已经各个字之间的信息来进行特征提取，在处理时序性问题上有更好的效果。

注意力层：网络结构第三层为注意力层。注意力层用于融合文本序列和触发词的信息，注意力层对双向 LSTM 提取的状态信息序列进行加权变换，使得文本中与该触发词相关的文字可以获得更高的注意力权重，突出重要状态信息的贡献，有效提高模型演化关系抽取准确性。

注意力层的主要作用是获取长文本序列和姓名触发词之间的交互信息。由于同一文本序列可能在所有的训练数据中多次出现，根据触发词的信息才能唯一标识文本序列中的某一事件。Attention机制其实就是一个相似性的度量，当输入的文本与目标触发词越详细，那么当前的输入的权重就会越大，说明当前的包含触发词事件信息越多。

假设输入到注意力层的文本序列维度为R^{n*T}，触发词的维度为R^{n*J}，那么可以得到一个相似度矩阵S，维度为R^{T*J}，然后将每个序列中的字对应的向量归一化，即得到触发词中的每个字关于序列中所有字的权重分布，再将每个系数和触发词的每个字相乘，得到序列中每个字包含触发词信息的表示。

CRF标签预测层：网络结构的最后一层是使用CRF对注意力层之后的序列进行序列标签的预测，考虑单词标签之间的制约关系，加入标签转移概率矩阵，给出全局最优标注序列。CRF 主要综合了隐马尔科夫模型和最大熵模型的优点。

4 实验验证

对于待处理的文本，进行事件及事件要素抽取的步骤：

（1）处理文本序列，去掉无信息句子，形成待预测语句；

（2）使用哈工大LTP工具进行姓名触发词的提取；

（3）对于只包含一个事件触发词 文本的序列，经过一次事件要素抽取模型从而获得完整的事件信息；

（4）对于包含多个姓名触发词的文本序列，根据不同的姓名触发词将文本序列多次输入事件要素抽取模型进行多个事件的事件要素抽取，并将结果合并，得到该文本的完整事件抽取结果。

在实验中，使用Word2vec对文本进行字向量的训练，上下文窗口大小为5，向量的维度为300；对于本文使用的添加Attention的Bi-LSTM-CRF模型，隐藏层的单元数为128，使用Adam优化器来训练我们的网络，设定初始学习率为0.001，由于训练数据相对较少，为了防止过拟合，Dropout大小设置为0.5。

使用处理过的标注序列进行模型的训练，使用验证集进行事件要素抽取结果的验证，总体的F1值达到了86%不再继续提升。使用从深圳证券交易所爬取的上市公司“人事变动”类型公告进行抽取结果的预测，同样可以达到期望的效果。

5 总结与展望

本文针对上市公司“人事变动”类型公告，建立了姓名触发词抽取模型和事件要素抽取模型，从公告文本中抽取一个或多个高管离职或继任信息的结构化数据。实验表明，该方法可以从该类型的任意公告中提取对应的信息，具有比较优良的效果，但由于标注语料的稀缺，无法使用大量数据进行训练与模型的评估。该方法的一个不足之处在只能针对这一种类型的公告文本，对于不同的公告类型，其触发词提取模型需要根据事件类型的不同做出相应的改变。下一步的工作是扩展该方法的应用范围，使其能够应用到更广泛的类型中去。

参考文献

1. NGUYEN T H ,GRISHMAN R． Event detection and domain adaptation with convolutional neural networks [C]. Proceedings of 53rd ACL. Beijing , China: Association for Computational Linguistics ,2015: 365－371．
2. NGUYEN T H ,GRISHMAN R． Modeling skip － grams for event detection with convolutional neural networks [C]. Proceedings of 2016 Conference on Empirical Method in Natural Language Processing． Austin ,Texas ,USA: Association for Computational Linguistics ,2016: 886－891．
3. FENG Xiaocheng ,HUANG Lifu ,TANG Duyu ,et al． A language independent neural network for event detection [C]. Proceedings of 54th Annual Meeting of ACL． Berlin ,Germany: Association for Computational Linguistics ,2016: 66－71．
4. LIU Shulin ,CHEN Yubo ,LIU Kang ,et al． Exploiting argument information to improve event detection via supervised attention mechanisms[C]. Proceedings of 55th ACL． Vancouver ,Canada: Association for Computational Linguistics ,2017: 1789－1798．
5. BENGIO Y ,DUCHARME R ,VINCENT P． A neural probabilistic language model[J]. Journal of Machine Learning Research ,2000 ,3(6) : 932－938．
6. MIKOLOV T ,SUTSKEVER I ,CHEN K ,et al． Distributed representations of words and phrases and their compositionality[J]. arXiv preprint arXiv: 1310．4546 ,2013．
7. COLLOBERT R ,WESTON J ,BOTTOU L , et al． Natural language processing (almost) from scratch[J]. arXiv preprint arXiv: 1103．0398 ,2.
8. HOCHREITER S ,SCHMIDHUBER J.Long short-term memory[J]. Neural Computing ,1997 ,9(8) : 1735－1780．