社交媒体中法律法条引用信息抽取算法研究

苏展[1]　窦志成[2]　文继荣[3,4]

（1.中国人民大学 信息学院，北京 100872；2.中国人民大学 高瓴人工智能学院，北京 100872；
3. 大数据管理与分析方法研究北京市重点实验室；4. 数据工程与知识工程教育部重点实验室）

论文摘要

摘要：该文重点研究互联网文本中法律法条引用信息的抽取问题，主要任务是从互联网文本中抽取出内容中引用的法律以及法条信息。由于文本引用法律内容时存在省略、跳跃以及转述等不规范方式，难以直接从文本中获取引用的法律法条。针对这一难点，该文提出了基于DFA、特征词序列以及深度匹配的混合层次抽取模型DS-LSTM，将法律法条的抽取问题转换为文本检索问题。基于确定有限自动机(DFA)，初步筛选出文本可能对应的法律和法条集合；结合TF-IDF选取特征词来表示文本，能够进一步区分同一部法律不同历史版本的法条；通过计算文本和法条的特征词序列相似度，生成候选的匹配法条，再将文本与候选法条通过MV-LSTM模型计算语义匹配度，根据语义匹配度判断文本与法条是否匹配。实验结果表明，该匹配算法在法律、法条级别的F_1分别达到0.97和0.92。

关键词：确定有限自动机；特征词序列；法条匹配；历史版本

Abstract : This article focuses on the problem of extracting laws and clauses mentioned in the web text. The main task is to match the web text with the corresponding laws and clauses. Due to the irregular expression such as omitting, skipping, and retelling, it is difficult to obtain laws and clauses directly from the text. For this difficulty, we proposed DS-LSTM, a hybrid hierarchical extraction model based on DFA, feature word sequence and deep text matching model, this article converts the extraction of laws and clauses into text retrieval problems, and based on the Deterministic Finite Automaton(DFA), we preliminarily filtered out the set of laws and clauses corresponding to the text. Then we selected the feature words to represent the clauses according to the TF-IDF. The match of the feature words can distinguish different history versions of the law. By calculating the similarity between the feature word sequences of the text and the clause, candidate clauses can be given, and then the text and the candidate clauses can be put into the MV-LSTM to obtain the semantic similarity. We used the semantic similarity to match the text and the clause. The results show that the matching algorithm's F_1 score at the laws and clauses level reaches 0.97 and 0.92, respectively.

Key words: DFA; feature word sequences; clause match; history version

论文简介

法律法条引用信息的抽取是研究法律相关文本的基础，与该任务相关的工作有命名实体识别与消岐、法律文本检索和深度文本匹配等。然而法律法条引用信息的抽取具有其特殊性，一个是法律多版本的问题，另一个是隐式引用的问题。目前的命名实体识别算法面向的是一般实体，由于法律法条抽取的特殊性难以适用；此外法律法条匹配的稀疏性造成数据倾斜问题，无法直接训练出有效的深度匹配模型。大部分现有的方法无法解决法律法条引用信息的抽取问题，尤其是属于大陆法系的中国法律。

为了解决上述问题，我们提出了一种融合确定有限自动机（DFA）、特征词序列和深度匹配的混合层次抽取模型DS-LSTM，而不是单纯使用深度模型或者DFA进行简单的匹配。

系统模型

由于DFA只能对明确指出法律名和第几条的文本识别成功，对于多版本法律的情形，DFA的判断只能作为一个初步筛选的结果，因此为了进一步提升法律法条匹配的准确度，在DFA识别模块后添加了历史版本区分、特征词序列匹配和深度文本匹配模块。

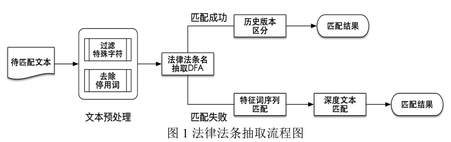

图 1 法律法条抽取流程图

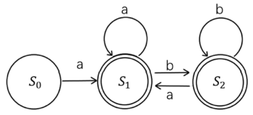

图 2 法律法条名识别DFA的状态转换图

定义 DFA M=(S, Σ, δ, S_0, F),法律名集合A={‘宪法’,’刑法’,…}，法条名集合B={‘第(.*)条’,’第(.*)条到第(.*)条’,’第(.*)条至第(.*)条’}，其中状态集S={S_0,S_1,S_2}，字母表Σ=A∪B，终态集F={S_1,S_2}，a∈A，b∈B。S_1,S_2分别表示识别末尾是法律名和法条名的终态，从S_0→S_1对应的是法律的识别过程，从S_1→S_2对应的是法条的识别过程。

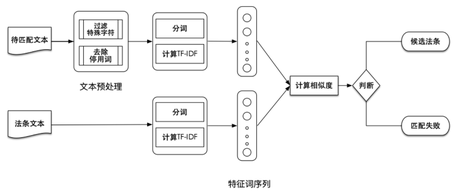

图 3 特征词序列匹配流程图

在匹配矩阵$\boldsymbol{M}$上使用卷积核*kernel*进行卷积运算，将计算的结果使用最大池化得到两个序列之间的匹配值。

$$\text{seqscore}(s_1,s_2)=\frac{p}{T}$$

算法原理

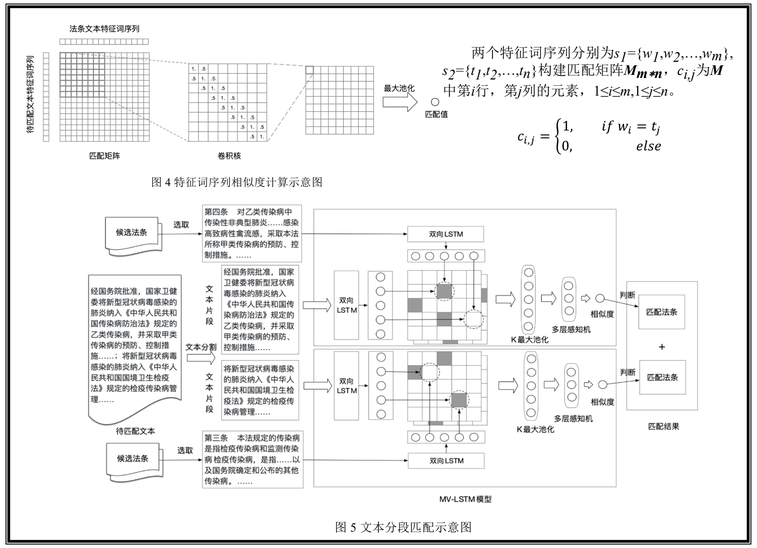

图 4 特征词序列相似度计算示意图

两个特征词序列分别为$s_1=\{w_1,w_2,\ldots,w_m\}$，$s_2=\{t_1,t_2,\ldots,t_n\}$构建匹配矩阵$\boldsymbol{M_{m*n}}$，$c_{i,j}$为$\boldsymbol{M}$中第$i$行，第$j$列的元素，$1\le i\le m, 1\le j\le n$。

$$c_{i,j}=\begin{cases}1, & if\ w_i=t_j\\ 0, & else\end{cases}$$

图 5 文本分段匹配示意图

实验结果

本实验采用了法律相关的微博文本数据约为370万条，数据库中含有匹配对应的法律274部，司法解释124部，包含历史版本共有596部，其中法条3万9千多条，在法律级别匹配上约330万次，在法条级别匹配上约32万次。

表 1 测试集上DFA模型抽取效果

匹配级别	P (%)	R (%)	F_1(%)
法律级别	96.14	99.67	97.87
法条级别	97.4	88.24	92.59

表 2 深度文本匹配模型法条级别抽取效果

模型	F_1	P	R
DSSM	0.912	0.940	0.886
ARC-I	0.921	0.932	0.910
ARC-II	0.928	0.964	0.895
KNRM	0.719	0.621	0.853
Cov-KNRM	0.748	0.659	0.863
DUET[27]	0.908	0.991	0.838
ANMM	0.866	0.927	0.812
MVLSTM	0.926	0.951	0.902

表 3 实验方法效果对比

编号	方法	F_1	P	R
①	DFA	0.636	0.721	0.570
②	DFA+版本区分	0.731	0.792	0.679
③	DFA+特征词序列匹配	0.813	0.910	0.735
④	DFA+版本区分+特征词序列	0.913	0.973	0.859
⑤	DFA+深度文本匹配(全文)	0.070	0.039	0.375
⑥	DFA+深度文本匹配(片段)	0.081	0.045	0.388
⑦	DFA+版本区分+深度文本匹配(全文)	0.099	0.056	0.437
⑧	DFA+版本区分+深度文本匹配(片段)	0.102	0.057	0.472
⑨	DFA+特征词序列匹配+深度文本匹配(全文)	0.844	0.878	0.813
⑩	DFA+特征词序列匹配+深度文本匹配(片段)	0.865	0.918	0.817
⑪	DFA+版本区分+特征词序列匹配+深度文本匹配(全文)	0.883	0.935	0.837
⑫	DFA+版本区分+特征词序列匹配+深度文本匹配(片段)	0.926	0.974	0.882

表 4 法律法条抽取示例

编号	待匹配文本	匹配法律	匹配法条	对应法条文本	匹配结果
①	电影好看，但拒绝盗摄！尊重电影主创来之不易的劳动成果。《电影产业促进法》第31条指出：未经权利人许可，任何人不得对正在放映的电影进行录音录像。……	《中华人民共和国电影产业促进法》	第三十一条	未经权利人许可，任何人不得对正在放映的电影进行录音录像。发现进行录音录像的，电影院工作人员有权予以制止，并要求其删除；对拒不听从的，有权要求其离场。	✓
②	报考公务员，除应当具备公务员法第十三条规定的条件以外，还应当具备省级以上公务员主管部门规定的拟任职位所要求的资格条件。……	《中华人民共和国公务员法》(2018修订版)	第十三条	公务员应当具备下列条件 （一）具有中华人民共和国国籍 （二）年满十八周岁 （三）拥护中华人民共和国宪法，拥护中国共产党领导和社会主义制度 ……	✓
	国家对行政机关中初次从事行政处罚决定审核、行政复议、行政裁决、法律顾问的公务员实行统一法律职业资格考试制度，由国务院司法行政部……	《中华人民共和国公务员法》(2017修订版)	第十三条	第十三条　公务员享有下列权利 （一）获得履行职责应当具有的工作条件； （二）非因法定事由、非经法定程序，不被免职、降职、辞退或者处分； ……	×
③	经国务院批准，国家卫健委将新型冠状病毒感染的肺炎纳入《中华人民共和国传染病防治法》第四条规定的乙类传染病，并采取甲类传染病的预防、控制措施……；将新型冠状病毒感染的肺炎纳入《中华人民共和国国境卫生检疫法》规定的检疫传染病管理。……	《中华人民共和国传染病防治法》	第四条	对乙类传染病中传染性非典型肺炎、炭疽中的肺炭疽和人感染高致病性禽流感，采取本法所称甲类传染病的预防、控制措施。……	✓
		《中华人民共和国国境卫生检疫法》	第三条	本法规定的传染病是指检疫传染病和监测传染病检疫传染病，是指鼠疫、霍乱、黄热病以及国务院确定和公布的其他传染病。……	✓

论文结论

本文深入研究了互联网文本中法律法条的抽取问题，利用了DFA模型抽取文本中引用的法律和法条，使用TF-IDF选取特征词来区分同一部法律中不同历史版本的法条，通过计算特征词序列之间的相似度来筛选可能匹配上的候选法条，消除了数据倾斜的问题，将候选法条输入MV-LSTM[1]计算得到文本与法条的语义相似度，根据语义相似度来判断文本与法条是否匹配，在法律以及法条级别匹配的F_1分别达到了0.97和0.92。总的来说，本文采用的基于DFA、特征词序列以及深度匹配的混合层次抽取模型DS-LSTM在法律法条引用信息抽取任务中有较好的效果。本文的混合层次抽取算法DS-LSTM还有需要改进的地方，比如区分中国和外国的同名法律等。

西安电子科技大学 XIDIAN UNIVERSITY