Named Entity Recognition for Science and Technology Policy Dynamics

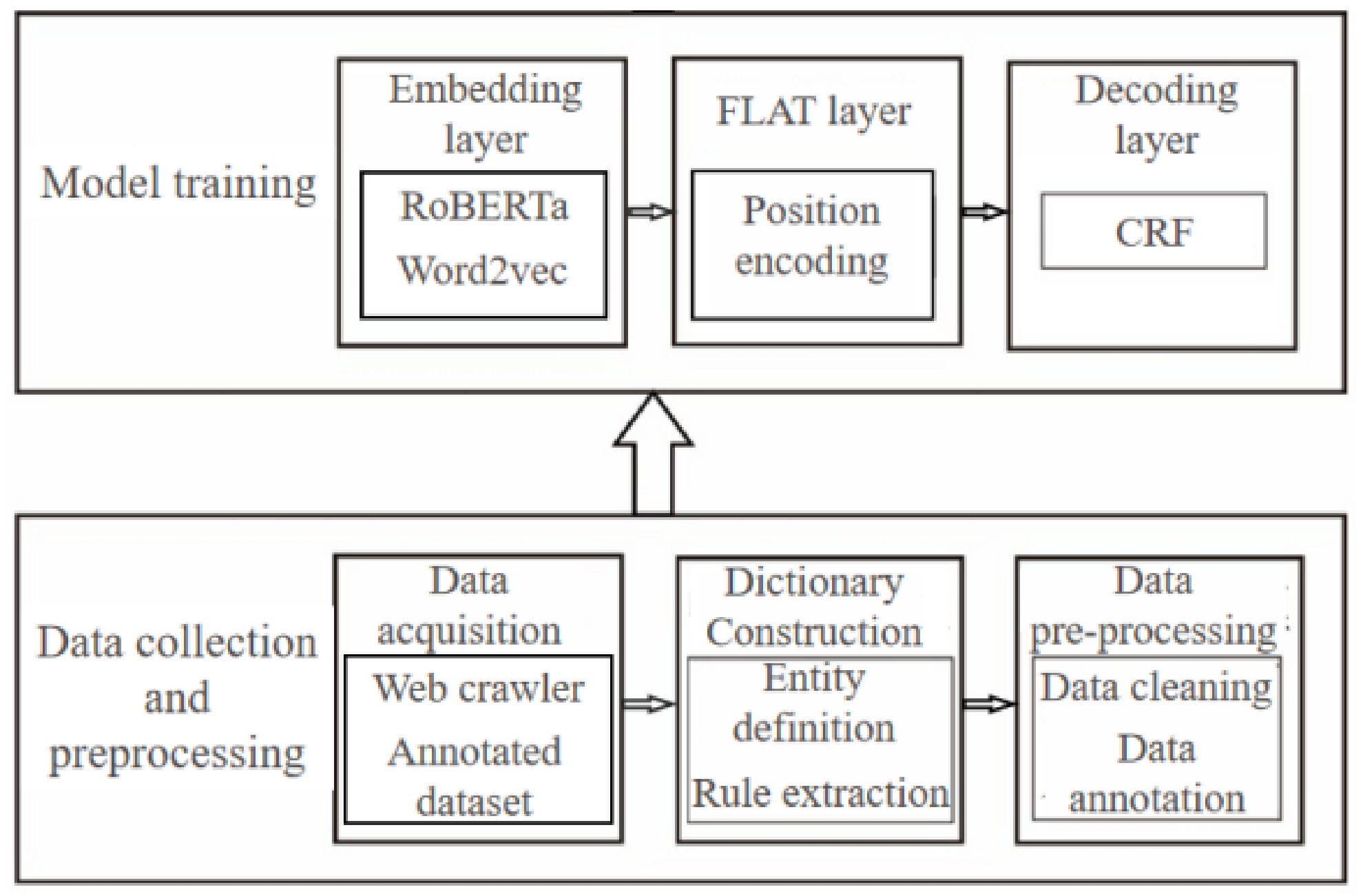
Wenjiao Zheng Bolin Hua

Introduction

Science and technology policy dynamics refers to media reports on the behavior of policy subjects in the formulation, implementation and supervision of national science and technology policy, including holding meetings, issuing proposals and revising legislation. There are a large number of entities related to science and technology policy in the dynamic text of science and technology policy, including all kinds of organizations, names, policy names, etc. Entity extraction of these entities can provide data basis for further downstream tasks such as entity relationship extraction and knowledge graph construction.

Method design

The overall research design of the research on dynamic named entity recognition of science and technology policy includes two main modules, namely data collection and pre-processing, model training.



Data collection and preprocessing

- Words segmentation and entity extraction

The research selected the weekly science and technology policy dynamic reports published on the official website of American Physical Society as the data source, and obtained a total of 3012 reports from 2020-2022 by using the method of web crawlers.

The original English corpus is automatically translated, and the original corpus is converted into Chinese by calling baidu translation interface for subsequent training tasks. As there is no publicly annotated data set in the field of science and technology policy dynamics, Four types of entities from the CLUENER2020 Chinese fine-grained named entity recognition data set, including location, company, name and position, are selected for this study.

Data collection and preprocessing

- Dictionary Construction

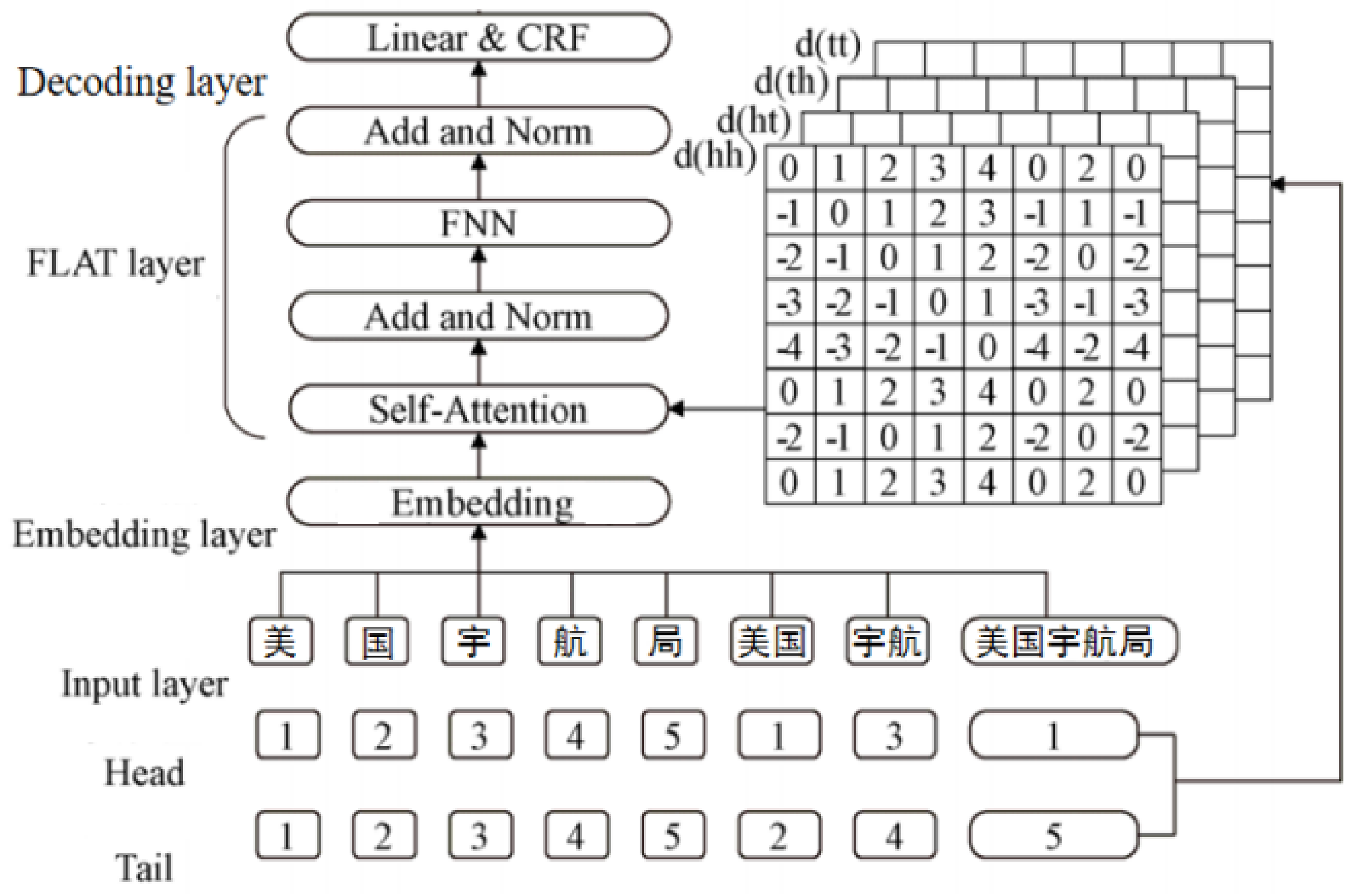
Nine types of dynamically related entities of science and technology policy are defined, including Government, Company, Research institution, Name, Position, Policy, Conference, Location and Time. Since there is no open domain dictionary in the field of science and technology policy, the study chooses to construct a domain dictionary by adding domain words to the general domain lexicon.

- Data pre-processing

In this study, the corpus was cleaned and processed by clause processing, and 14,047 sentences were obtained. In this study, 3000 sentences after processing were annotated, and the annotation results were combined with CLUENER annotation dataset to obtain a total of 13, 748 data pieces.

Model training

The Roberta-Flat model combining character and lexical information is selected for data training. Firstly, the RoBERTa pre-trained language model and Word2vec word vector coding model are used to vectorize the character and vocabulary information respectively, and then the character vector and corresponding word vector are matched and spliced. The spliced vector will be used as the output of the embedding layer to enter the FLAT layer for position coding. The model builds a head position encoding and a tail position encoding for each character and word, respectively, and fully models the encoding results using Transformer. The output of the FLAT layer will be decoded into the CRF layer to obtain the predicted results. The specific structure is shown in Figure 2.



0	1	2	3	4	0	2	0
-1	0	1	2	3	-1	1	-1
-2	-1	0	1	2	-2	0	-2
-3	-2	-1	0	1	-3	-1	-3
-4	-3	-2	-1	0	-4	-2	-4
0	1	2	3	4	0	2	0
-2	-1	0	1	2	-2	0	-2
0	1	2	3	4	0	2	0

Input layer	美	国	宇	航	局	美国	宇航	美国宇航局
Head	1	2	3	4	5	1	3	1
Tail	1	2	3	4	5	2	4	5

Analysis and measurement of results

- Analysis of experimental results

In order to verify the effectiveness of relative position coding and the introduction of external word lists, comparative experiments are conducted on BiLSTM+CRF, Iterative expansive convolutional Neural Network (IDCNN) and FLAT.

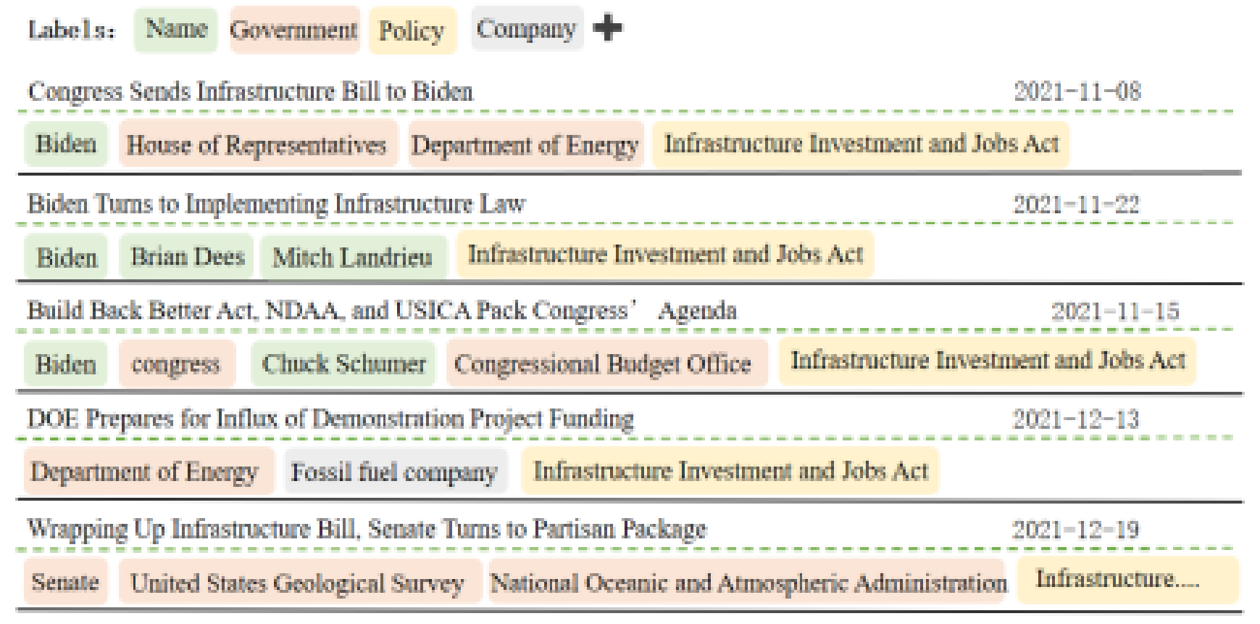
Training program	The F value
BiLSTM+CRF	78. 48
IDCNN-CRF	70. 52
FLAT	76. 15
RoBERTa+FLAT	78. 99

- Recognition details analysis

Entity extraction was carried out on 3012 dynamic texts of science and technology policies. After statistical analysis, a total of 21, 320 entities were extracted in the experiment, and the average number of entities extracted from each report was 7. 08.

- System Display

We designed entity label display system based on the entity extraction results. Take reports related to the Infrastructure Investment and Jobs Act as an example. Different types of entities are distinguished by different background colors. The specific search result interface is shown in Figure 2.



Labels: Name Government Policy Company ➕

Congress Sends Infrastructure Bill to Biden 2021-11-08
Biden House of Representatives Department of Energy Infrastructure Investment and Jobs Act

Biden Turns to Implementing Infrastructure Law 2021-11-22
Biden Brian Dees Mitch Landrieu Infrastructure Investment and Jobs Act

Build Back Better Act, NDAA, and USICA Pack Congress' Agenda 2021-11-15
Biden congress Chuck Schumer Congressional Budget Office Infrastructure Investment and Jobs Act

DOE Prepares for Influx of Demonstration Project Funding 2021-12-13
Department of Energy Fossil fuel company Infrastructure Investment and Jobs Act

Wrapping Up Infrastructure Bill, Senate Turns to Partisan Package 2021-12-19
Senate United States Geological Survey National Oceanic and Atmospheric Administration Infrastructure....

Conclusion and Discussion

Based on the research of domain named entity recognition, this paper adopts the method of RoBERTa+FLAT integrating lexical information to extract the entity from the dynamic text of science and technology policy. The experimental results show that the method we used compared with the traditional method has a better effect of entity recognition.

However, the research in this paper also has some limitations. Small-scale annotated datasets affect the training effect of the model. The subsequent research will try to overcome the obstacles of small-scale datasets by using related methods such as domain migration.