A Binary Tagging Framework for Overlapping Relation Extraction

INTRODUCTION

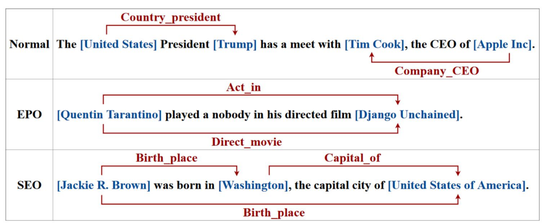
With the advent of the internet, a large amount of unstructured digital text data is generated every day in multiple forms such as news, publications, blogs, and social media. In the realm of information overload, it is easy for people to access and outspread information. At the same time, however, it is hard for us to filter and learn from such a vast amount of information efficiently. In this contest, information extraction technology emerges as the times require.

By utilizing information extraction technology, people are aiming to extract three types of information: 1) named entities, 2) relations, and 3) events. After comprehensive domain research within the information extraction field, we realized that there are still many challenges preventing us from extracting relations among name entities. Thus, **we decide to take a further step from the NER to the relation extraction(RE) task.**

CHALLENGE

The research on the normal relation triplets extraction is mature with the help of the development of RE methods. However, in the real-world case, a large proportion of the text data is not in the format of normal relations. For example, 'Jackie was born in Washington, the capital city of the U.S.' In this scenario, the entity Washington is both the object of (Jackie, born_place, Washington) and the subject of (Washington, capital_of, U.S.). We call it **relation overlapping**.

OVERLAPPING TRIPLE PROBLEM



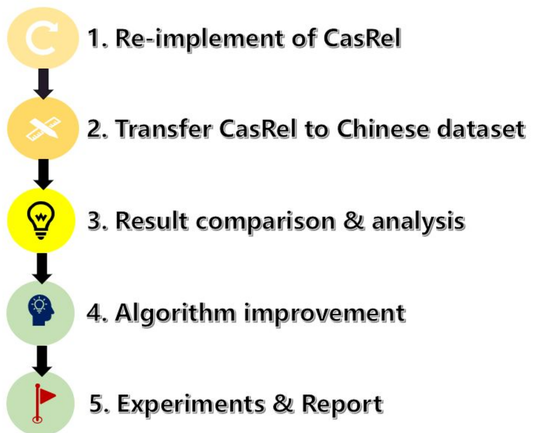
RESEARCH DIRECTION

Joint entity and relation extraction on text with relation overlapping issues. It is one of the fundamental steps on our way to explore the world of information and extract structured knowledge from it. Also, from the literature survey, we noticed that datasets researchers utilized extensively for the research for RE task are mainly in English

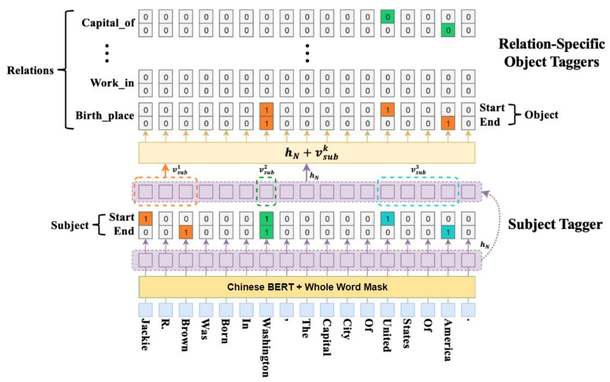
The NLP in English and Chinese are totally different. And the research in the Chinese RE task is not as extensive and deep as it in English. Thus, We want to divide it into such tasks and transfer state-of-the-art RE techniques to Chinese.

For our final project, we **transfered and improved** a novel RE technique called CasRel to deal with RE tasks on Chinese dataset. This experimental development of the CasRel would facilitate the research on RE tasks in Chinese language.

WORKFLOW



OUR ARCHITECTURE



1. **BERT-Chinese Encoder**
   The encoder module extracts feature information from sentence, which will feed into subsequent tagging modules 2. We employ a pre-trained BERT model (Devlin et al., 2019) to encode the context information.

2. **Whole Word Mask (WWM)**
   In the original BERT, a WordPiece tokenizer is used to split the text into WordPiece tokens, where some words are split into several small fragments. The whole word masking (wwm) mitigates the drawback of masking only a part of the whole word, which is easier for the model to predict. In Chinese condition, WordPiece tokenizer no longer splits the word into small fragments, as Chinese characters are not formed by alphabet-like symbols.

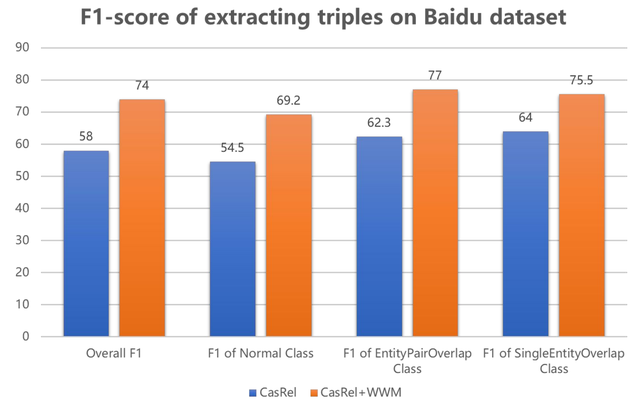
	English
Original	Use a language model to predict the probability of the next word
Normal Masking	Use a language [M] to [M] ##di ##ct the pro [M] ##bility of the next word
Whole Word Masking	Use a language [M] to [M] [M] [M] the [M] [M] [M] of the next word

3. **Cascade Decoder**
   The basic idea is to extract triples in two cascade steps.
   - Subject Tagger: we detect subjects from the input sentence.
   - Relation-specific object taggers: for each candidate subject, we check all possible relations to see if a relation can associate objects in the sentence with that subject.

RESULTS

Below table (Wei et al. 2020) above shows the performance on the NYT and WebNLG dataset of current SOTA RE methods. We can tell that there is a significant improvement regarding metrics such as Precision, Recall, and F1 score for the CasRel, which is a novel RE technique proposed in 2020.

Method	NYT			WebNLG		
	Prec.	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*
NovelTagging (Zheng et al., 2017)	62.4	31.7	42.0	52.5	19.3	28.3
CopyR$_{OneDecoder}$ (Zeng et al., 2018)	59.4	53.1	56.0	32.2	28.9	30.5
CopyR$_{MultiDecoder}$ (Zeng et al., 2018)	61.0	56.6	58.7	37.7	36.4	37.1
GraphRel$_{1p}$ (Fu et al., 2019)	62.9	57.3	60.0	42.3	39.2	40.7
GraphRel$_{2p}$ (Fu et al., 2019)	63.9	60.0	61.9	44.7	41.1	42.9
CopyR$_{RL}$ (Zeng et al., 2019)	77.9	67.2	72.1	63.3	59.9	61.6
CopyR$^{*}_{RL}$	72.8	69.4	71.1	60.9	61.1	61.0
CASREL$_{random}$	81.5	75.7	78.5	84.7	79.5	82.0
CASREL$_{LSTM}$	84.2	83.0	83.6	86.9	80.6	83.7
CASREL	**89.7**	**89.5**	**89.6**	**93.4**	**90.1**	**91.8**



CONTRIBUTIONS

1. We identified the main gap between NLP in English and NLP in Chinese regarding the Relation Extraction tasks.
2. We deployed the SOTA CasRel algorithm to a new Chinese dataset, and improved the technique by adding Whole Word Masking mechanism.
3. Our experimental development of the CasRel would facilitate the research on RE tasks in Chinese language, as well as form part of the foundation of constructing the Knowledge Graph systems in Chinese.