The Poisson multi-media knowledge construction System for TAC KBP 2018

Zhicheng Sheng[1], Danping Wu[1], Zelong Li[1], Wei Zhang[1], Jingchen Lu[1], Zeming Xu[1], Dong Liu[1], Shuo Guo[1], Yantao Jia[1], Zhepei Wei[1], Erxing Yu[1], Xiongfeng Xiao[1], Zhuo Wang[2], Jianyong Wang[2]

[1]Poisson Lab, Distributed and Parallel Software Lab,
Huawei Technologies Co., Ltd, China
[2]Department of computer Science and Technology,
Tsinghua University, China
jiayantao@huawei.com

Abstract

This paper presents the Open knowledge System developed for the Streaming Multimedia Knowledge Base Population (SM-KBP) task in TAC KBP 2018. We participated in task 1a and task 2 of this track. Eight modules were developed to complete this task, including entity discovery module, the relation extraction module, the event nugget detection module, the image detection module, the video processing module, the Intra-document co-reference resolution module the Cross-document co-reference resolution module and the RDF graph construction module.

1 introduction

The goal of TAC KBP 2018 is to encourage research in Natural Language Processing and related applications. It contains several tracks, and we participated in Streaming Multimedia Knowledge Base Population (SM-KBP) this year, which contains three sub-tasks. We participated in task 1a and task 2 of this track.

The SM-KBP track aims to develop a multi-hypothesis semantic engine that generates explicit alternative interpretations of events, situations, and trends from a variety of multilingual multimedia sources which include text, speech, images, videos, and pdf files. For the pilot, the scenario is the Russian/Ukrainian conflict (2014-2015) and the scenario languages are English, Russian, and Ukrainian. This track has three main evaluation tasks and we participated in two of this tasks. Task 1 aims to extract knowledge entity (KE) and KE mentions from a stream of multi-media documents and produce a document-level knowledge graph for each document. Task 1a only considers generic background context. Task 2 will construct knowledge base (KB) by aggregating and linking document-level knowledge graph (KG) by Task 1. The knowledge graph format in defined by AIDA and we use Apache jena to perform graph construction and sparql query.

The SM-KBP track involves all the steps to construct the knowledge graph. In the entity discovery module the system must find all Person, Organization, Geopolitical Entity, Location, Organization, Vehicle and Weapon in the document, which are mentioned by named mentions or nominal mentions. For each entity, a KB node will be created, as well as the relation between each node. After construction for document level knowledge graph, co-reference resolution based on each KG will be performed and a corpus-level knowledge graph will be constructed.

The paper is organized as follows. Section 2 describes the architecture of the developed system. Section 3 will explain entity discovery process and relation extraction is elaborated in Section 4. Section 5 presents the event nugget extractor and image detection method is explained in Section 6. We introduce the video processing module in Section 7 and Section 8 introduce the Intra-document co-reference resolution. In Section 9, we presents our strategy for task2, namely cross document co-reference resolution. In Section 10 and Section 11 we introduce how to handle multi-language sources and the approach used to build RDF graph. Finally, we conclude the paper in Section 12.

2 The System Architecture

Our proposed system contains six steps, ie, entity discovery, relation extraction, event nugget detection, image detection Intra-document co-reference resolution and cross-document co-reference resolution, as illustrated in Figure 1.

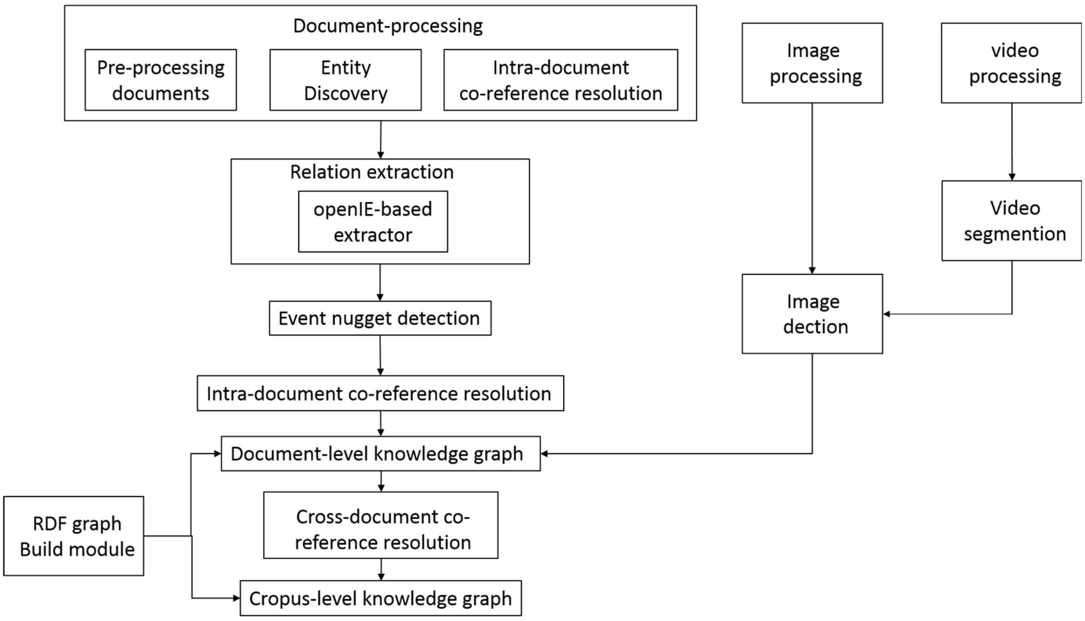

Figure 1

In entity discovery module, we use Bi-LSTM Conditional Random Field (CRF) model to extract entity from the text. Then the relation of entities are extracted from document from openIE-based method. Event detection is carried out by RPI Joint Information Extraction System (Qi Li, et al ,2014) and image detection is performed by the popular Single shot multibox detector(SSD) framework (W Liu, et al, 2017). Cross document co-reference resolution is based on hierarchical clustering.

3 Entity discovery

Our entity discovery module is based on Bi-LSTM Conditional Random field (Huang Z, et al, 2015). We also use Stanford NER (Manning et al., 2014) to extract person, organization, geo-political entity, facility and location. The LSTM tagger above is typically sufficient for part-of-speech tagging, but a sequence model like the CRF is really essential for strong performance on NER. The CRF computes a conditional probability. Let y be a tag sequence and x an input sequence of words. Then we compute:

$$P(y \mid x) = \frac{\exp(Score(x,y))}{\sum_{y'} \exp(Score(x,y'))}$$

Where the score is determined by defining some log potentials such that

$$Score(x,y) = \sum_i \log \psi_i(x,y)$$

In the Bi-LSTM CRF, we define two kinds of potentials: emission and transition. The emission potential for the word at index i comes from the hidden state of the Bi-LSTM at timestep i. The transition scores are stored in a matrix P. $P_{j,k}$ is the score of transitioning to tag j from tag k, so

$$\begin{aligned} Score(x,y) &= \sum_i \log \psi_{EMIT}(y_i \to x_i) + \log \psi_{TRANS}(y_{i-1} \to y_i) \\ &= \sum_i h_i \left| y_i \right| + \mathbf{P}_{y_i, y_{i-1}} \end{aligned} .$$

We trained our model on corpus from ERE and ACE. To get the segmentation of location entity, latest Geonames data was employed. With regard to weapon and vehicle entity, because this two entity rarely show up in the training or evaluation data, we introduce wiki information to annotate the document and build the pre-training model.

4 Relation extractor

The relation extraction module is based on OpenIE relation extractor module and Implicit Relation extractor module.

4.1 OpenIE-based Relation Extraction

Open information extraction is a useful tool to get the relation tuples. Although the extracted result may not match the relation type defined in the TAC and the result may also contain many noise, this method can provide a preliminary result for relation extraction.

Open Information Extraction (OpenIE) system (Soderland S, et al, 2013) is from the university of Washington(UW). An Open IE system runs over sentences and creates extractions that represent relations in text, and it produces tuples of the form (arg1, relation,arg2) for the given sentence. However, there are conflicts direct result of and TAC demands. For example, openIE may extract tuples like (Barack Obama, is the president of, the U.S.) of the sentence “The U.S. president Barack Obama gave his speech on Tuesday to thousands of people.” The output is expressed as a triple (A, B, C) where A and B are arguments, C is the relation between those arguments. This tuple define a relation of Barack Obama and the U.S. but this relation has not been defined clearly in the TAC. We design several rules to help us to filter the result of openIE. Firstly, only arguments that contains entities which have been extracted from the entity discovery module can be kept. Secondly, we build a map between TAC relation and the words and expressions to describe the relationship and it helps us to filter the relation that fits the demands of TAC. In the meantime, our proposed strategy also provide a method to assign the extracted tuples to the exact relation.

This year we initially got 196978 extraction results from the English ltf text. After a series of screening , we retained 2356 results.

5 Event nugget detection

Our event extractor is based on RPI joint information extraction(Qi Li, et al ,2014). This is a joint framework based on structured prediction which extracts triggers and arguments together. We perform the extraction on each document and filter the result by the entity name using the entity discovery result. We employed a simple heuristic method for event co-reference resolution. We just merged the event nuggets with the same trigger mention in a document.

6 Image detection module

In the SM-KBP track, a large number of image and video were used to extract information. We use the single shot multibox detector(SSD) based on tensorflow framework to perform image detection. SSD approach use small convolutional filter to predict object categories and offsets in bounding box locations. It use separate predictors for different aspect ration detections, and apply these filters to multiple feature maps from the later stages of a network in order to perform detection at multiple scales. Five steps were involved in training the SSD model:

1). Matching strategy: match the ground-truth box with the default box with biggest jaccard overlap to guarantee that every ground-truth can have the corresponding default box.

2) Training objective: define the loss function and perform optimization.

3) Choosing scales and aspect ratios for default boxes.

4) Hard negative mining: Sort the default box by its confidence coefficient and choose the top ones.

5) Data argumentation.

Open data sets were used to train the model. For example, to detect the weapon shown up in the evaluation data set, we trained our model on nine classes of weapon, including fighter, submarine, bullet, bomb, warship, gun, tank, guided missile and torpedo.

7 Video processing

We firstly preprocess the video data, which mainly to segment the video according to the format requirements. Then, we use the officially available tools to extract key-frame images, and perform target detection using SSD algorithm based on tensorflow frame.

8 Intra-document co-reference resolution

There are five steps in our intra-document co-reference resolution: a) the mentions whose name are the same are linked together; b) a co-reference chain is generated by Stanford CoreNLP (Manning et al., 2014); c) we combined the two chains as the final co-reference chain; d) the entity types of mentions in one chain should be unified into the entity type same as the majority; e) the canonical mention is selected by following standards:

1) The canonical mention should appear in the main body of the document.

2) The start offset of the canonical mention should be as small as possible.

9 Cross-document co-reference resolution

To address the tagging of entities, we employ the clustering method to cluster the cross-document entities across target documents. We cluster the entities in terms of the similarity between entity mentions.

In order to reduce the time complexity of the process, we use two steps to cluster the entities. We firstly use the Elastic Search to generate the candidates among the entity mentions in Coarse-grained calculation manner. Secondly, we use the other attributes perform fine-grained similarity calculations.

9.1 Candidate Entity Generation

We employ Elastic Search to create index with all entity mentions which include various attributes that we extract before, such as mention names, types, and so on. Then, we initially select mention names and types to generate corresponding entity candidate sets.

9.2 Entity Clustering

Based on the candidate entities generated by the Elastic Search, we then use rules and hierarchical clustering to perform final clustering. We mainly employ the mention features to measure the similarity for each pair of entity mention and candidate, and the similarity calculating methods mainly include edit distance similarity, vector distance similarity, nearest distance similarity. These list some features below:

1) Embedding similarity. The similarity between a mention embedding and a candidate mention embedding.

2) Name similarity. Namely, the string similarity between the reference entity mention and the candidate entity.

3) Context similarity. We select K words surrounding an entity mention as its context, and then compute the similarity between the entity mention and the candidate entity.

4) Acronym matching, which indicates whether the entity mention is an acronym of the candidate entity and whether the candidate entity appears in the document text.

On the basis of rules, and based on the features and the similarity calculating methods above, we weighted summation of the similarities, and employ hierarchically cluster method to cluster each candidate entity pair by a pre-specified threshold that is used to judge whether add an entity to a cluster or create a new cluster by itself.

In summary, we use Cross-document co-reference resolution to aggregate the entity mentions that are referred as the same entity. The method mainly employ features that we extract before to measure the similarity between the reference entity and the candidate entity mentions.

10 multilingual multimedia sources

The scenario is the Russian/Ukrainian conflict (2014-2015) and the scenario languages are English, Russian, and Ukrainian. So for the Russian and Ukrainian, we identity entity name mentions and classifies them into pre-defined types using a Cross-lingual Entity Extraction, Linking and Localization System(Xiaoman Pan, et al., 2018). The system consider name tagging as a sequence labeling problem, to tag each token in a sentence as the Beginning (B), Inside (I) or Outside (O) of an entity mention with a certain type. The model is based on a bi-directional long short-term memory (LSTM) networks with a Conditional Random Fields (CRFs) layer.

11 RDF Graph construction module

In this track, we use Apache jena to build RDF graph and perform sparql query according to the built graph. The SM-KBP track provide the AIF format to define the ontology needed to build the graph, and all the query can be performed by the docker environment provided by the TAC.

12 Conclusion

In this paper, we present our system for SM-KBP track of TAC 2018, and we participated in task 1a and task 2. The proposed system contains six modules, namely, entity discovery module, the relation extraction module, the event nugget detection module, the image detection module, the Intra-document co-reference resolution and the entity co-reference resolution.

References

[1] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector, European Conference on Computer Vision. Springer International Publishing, 2016:21-37.

[2] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-6

[3] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging, arXiv preprint arXiv:1508.01991, 2015.

[4] Soderland S, Gilmer J, Bart R, et al. Open Information Extraction to KBP Relations in 3 Hours, TAC. 2013.

[5] Qi Li and Heng Ji. 2014. Incremental Joint Extraction of Entity Mentions and Relations. In Proc. ACL.

[6] Lin H, Zhao Z, Jia Y, et.al. OpenKN at TAC KBP 2014. In Proceedings of TAC-KBP 2014.

[7] Soderland S., Gilmer J., Robert Bart, Oren Et- zioni, and Daniel S. Weld. 2013. Open information extraction to KBP relations in 3 hours. In Proceedings of TAC-KBP 2013.

[8] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In Proc. ACL 2017.