Appendix III:

Recommended Resources：Text Understanding System TexSmart

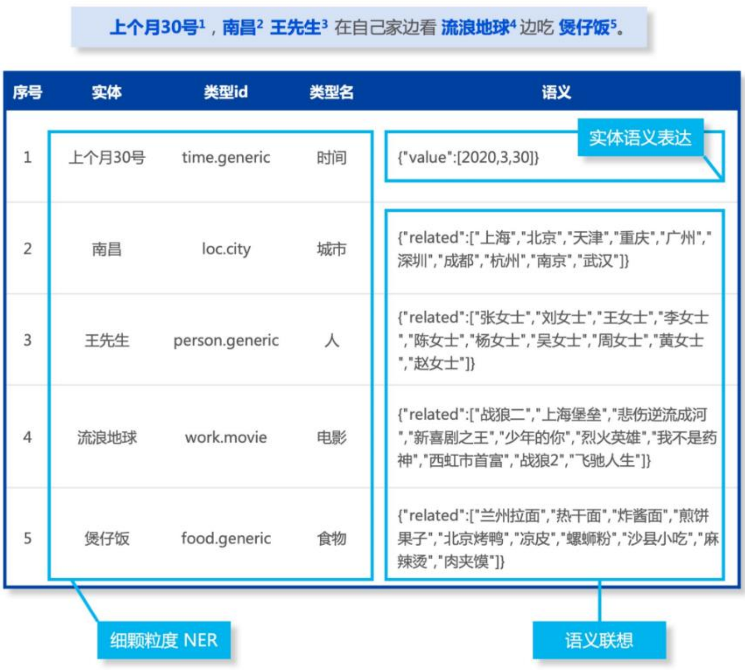
In view of the deficiencies of existing NLP tools in deep text understanding in the industry, Tencent AI Lab releases natural language understanding system TexSmart, for lexical, syntactic and semantic analysis of both Chinese and English. Besides supporting common features such as word segmentation, part-of-speech tagging, named entity recognition, syntactic analysis, semantic role labeling, TexSmart also provides special features such as fine-grained named entity recognition, semantic expansion, deep semantic expression and so on. The open-source of Tencent AI Lab is a demonstration of basic AI abilities itself based on the advantages of the company's data resource, which can provide more structural analysis and processing on semantic level for natural language text, and promote the improvement of NLP tasks effects in academic research and industrial application environment.

Why use Tencent AI Lab open-source TexSmart text understanding system?

1) **More Fine-grained Named Entity Recognition:** Support thousands of entity types with hierarchical structure and provide richer semantic information for downstream NLP applications. At present, most of the public text understanding tools only support a few or a dozen of coarse-grained entity types such as person, location, organization, etc.; while TexSmart can recognize thousands of entity types, including person, location, organization, product, trademark, work, time, numerical value, living creature, food, medicine, disease, subject, language, celestial body, organ, event, activity and so on. In the common coarse-grained entity types of people, location, organization, etc., TexSmart can further recognize many fine-grained sub-types such as actor, politician, athlete, country, city, company, university, financial institution and so on. As shown in the figure below:

2) **Enhanced Semantic Understanding**

Semantic Expansion: The function of semantic expansion is to provide a list of



上个月30号^{1}，南昌^{2} 王先生^{3} 在自己家边看 流浪地球^{4} 边吃 煲仔饭^{5}。

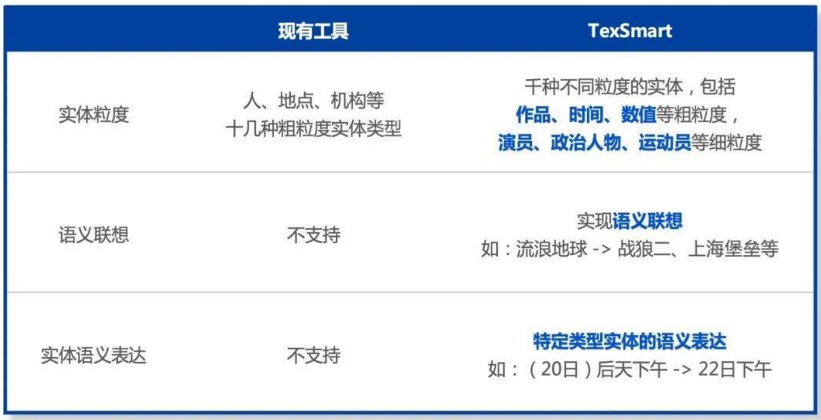
序号	实体	类型id	类型名	语义
1	上个月30号	time.generic	时间	{"value":[2020,3,30]}
2	南昌	loc.city	城市	{"related":["上海","北京","天津","重庆","广州","深圳","成都","杭州","南京","武汉"]}
3	王先生	person.generic	人	{"related":["张女士","刘女士","王女士","李女士","陈女士","杨女士","吴女士","周女士","黄女士","赵女士"]}
4	流浪地球	work.movie	电影	{"related":["战狼二","上海堡垒","悲伤逆流成河","新喜剧之王","少年的你","烈火英雄","我不是药神","西虹市首富","战狼2","飞驰人生"]}
5	煲仔饭	food.generic	食物	{"related":["兰州拉面","热干面","炸酱面","煎饼果子","北京烤鸭","凉皮","螺蛳粉","沙县小吃","麻辣烫","肉夹馍"]}

related entities for the entities in the input sentence. Semantic expansion is a way to enhance the understanding of each entity’s semantics. It has wide applications in industry, for instance in search engines and recommendation systems.

Deep Semantic Expression for Specific Type of Entities:

For time, quantity and other specific types of entities, TexSmart can analyze their potential structured expressions, so as to further derive the precise semantics of these entities. This kind of deep semantic understanding is essential for certain NLP applications. For example, in intelligence dialogue systems, a user sends a request to the bot on April 20, 2020, which is “Please help me book an air ticket to Beijing at 4 pm the day after tomorrow.”. The bot not only needs to know that “at 4 pm the day after tomorrow” is a time entity, but also needs to know the deep semantics of this entity refers to “4 pm, April 22, 2020”. At present, most public NLP tools do not provide the function of deep semantic expression like this, which is needed to be implemented by the application layer itself.

Designed for multi-dimensional application requirements: There are different requirements for speed, precision and timeliness in various applications scenarios in academia and industry; and it is often difficult to achieve both speed and precision. Therefore, we fully consider the needs of these three aspects in the design of TexSmart. First of all, for a certain function (such as named entity recognition), TexSmart implements a variety of algorithms and models with different speed and precision to customize the upper applications, so as to meet the diverse application needs in different scenarios. Secondly, TexSmart takes advantage of large-scale unstructured data and unsupervised or weakly-supervised methods. On the one hand, these unstructured data covers a large number of words and entities with strong timeliness (such as “Captain Marvel” above); on the other hand, using unsupervised or weakly-supervised methods can make the system update at a low cost, so as to ensure its strong timeliness.



	现有工具	TexSmart
实体粒度	人、地点、机构等 十几种粗粒度实体类型	千种不同粒度的实体，包括 **作品、时间、数值**等粗粒度， **演员、政治人物、运动员**等细粒度
语义联想	不支持	实现**语义联想** 如：流浪地球 -> 战狼二、上海堡垒等
实体语义表达	不支持	**特定类型实体的语义表达** 如：（20日）后天下午 -> 22日下午

Please refer to the following links for relevant home pages and articles:

https://texsmart.qq.com

https://mp.weixin.qq.com/s/pLdKTgXogtlTR2BvpwEvmg

Recommended Resources：Tencent AI Lab Pre-trained Graph Neural Network Model for Molecular Representation

Introduction

AI aided drug discovery has been suffering from two thorny problems:

1) Insufficient labeled drug molecular data;

2) Insufficient generalization capabilities of existing models: models trained on one molecular dataset are often difficult to be generalized to another one.

In pursuit of solving the above problems, we developed a new Transformer-based Graph Neural Network encoder, termed GX, to capture the rich implicit structural and semantic information from molecules. We design a self-supervised training strategy to pre-train GX with huge amounts of unlabeled molecular data (about 10M unlabeled molecules). We hope that the release of GX models can help with boosting the performance of drug discovery applications, such as molecular property prediction and virtual screening.

We will release several model instances with different sizes, to ease the usage of GX. Researchers and practitioners can flexibly choose the model instances with proper sizes according to the computing resource budget in practical projects. For the researchers without machine learning experience, we also provide the pre-computed fingerprints for 200M molecules, which come from the Chembl dataset and several benchmark datasets.

Highlights

GX has encoded rich structural information of molecules through the designing of self-supervision tasks. It can also produce feature vectors of atoms and molecule fingerprints, which can serve as inputs of downstream tasks.

GX is designed based on graph neural networks and all the parameters are fully differentiable. So it is easily to fine-tune GX in conjunction with specific drug discovery tasks, in order to achieve better performance.

How to Use

Researchers and practitioners can easily use GX models in two ways:

1) Without fine-tuning: Use the output of GX as the molecular fingerprints directly.

2) With fine-tuning: Use the GX models as building blocks in drug development projects that need to encode drug molecular data in an end-to-end fashion.

Release Plan

We are now working hard for the release of the pre-computed fingerprints and pre-trained models. The whole release pipeline contains two phases:

1) Before 30th May, we will release the pre-computed fingerprints.

2) About mid-June, we will release the source code and the pre-trained models.

Download

Please refer to https://ai.tencent.com/ailab/ml/gnnpretrain.html for downloads and the documents.

Disclaimer

The pre-computed fingerprints, the source code and the pre-trained models are for research purpose only.