Chinese Open Relation Extraction and Knowledge Base Establishment

SHENGBIN JIA and SHIJIA E, MAOZHEN LI, YANG XIANG
ACM Trans
Content

➢ 1. Introduction
➢ 2. Related work
➢ 3. Unique Chinese linguistic phenomena
➢ 4. Dependency semantic normal forms
➢ 5. COER: Chinese open entity and relation knowledge base
➢ 6. Experiments
➢ 7. Conclusion

Abstract

➢ Research object
  ● Named entity relation extraction

➢ Research status
  ● Many English extractors have achieved reasonable performance, an effective system for Chinese relation remains undeveloped.

➢ Problems
  ● The lack of Chinese annotation corpora.
  ● The specificity of Chinese linguistics.

➢ Solution
  ● Summarize three kinds of unique but common phenomena in Chinese linguistics.
  ● Propose a novel unsupervised Chinese open relation extraction (ORE) model based on Dependency Semantic Normal Forms (DSNFs)

➢ Result
  ● The model achieve better performance than other methods.
  ● Establish a large-scale knowledge base of entity and relation, called COER.
1. Introduction

➢ **Entity relation extraction**
  - It is a core task of Information Extraction (IE) to extract semantic relations between named entities from unstructured text.
  - The extractions are retained in the form of three tuples: \((\text{Entity1}, \text{RelationWords}, \text{Entity2})\)
  - NLP applications: knowledge graph, intelligent search engines, and automatic question and answering systems.

➢ **ORE VS TRE**
  - Traditional Relation Extraction (TRE) Identify target relations by training extractors from pre-defined and manual annotating corpus (eg. Neural network and distant supervision).
  - ORE overcome the abovementioned obstacles.
1. Introduction

➢ **English relation extraction**
  - **Distance supervised** relation extraction, due to the availability of large-scale structured knowledge base. [Freebase(2008), DBpedia(2007), YAGO(2007)]
  - **Deep learning**, such as: LSTMs, CNNs.

➢ **Chinese relation extraction**
  - Extractors are primarily based on traditional supervised machine-learning algorithm and still undeveloped.

➢ **Reasons of the gap**
  - The lack of large-scale labeled Chinese knowledge corpora to train models.
  - Chinese language is more sophisticated and flexible than English in the inspects of morphology, syntax, and grammar.
1. Introduction

Example

For the sentence “德国总统高克访问中国，并在同济大学发表演讲”, the result of segments, POS tagging, named entity recognition, and syntactic parsing.

The Dependency Labels from LTP

<table>
<thead>
<tr>
<th>Types</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>主谓关系</td>
<td>SBV</td>
<td>subject-verb</td>
</tr>
<tr>
<td>核心关系</td>
<td>HED</td>
<td>head</td>
</tr>
<tr>
<td>动宾关系</td>
<td>VOB</td>
<td>直接宾语, verb-object</td>
</tr>
<tr>
<td>动补结构</td>
<td>CMP</td>
<td>complement</td>
</tr>
<tr>
<td>间宾关系</td>
<td>IOB</td>
<td>间接宾语, indirect-object</td>
</tr>
<tr>
<td>并列关系</td>
<td>COO</td>
<td>coordinate</td>
</tr>
<tr>
<td>前置宾语</td>
<td>FOB</td>
<td>前置宾语, fronting-object</td>
</tr>
<tr>
<td>介宾关系</td>
<td>POB</td>
<td>preposition-object</td>
</tr>
<tr>
<td>定中关系</td>
<td>ATT</td>
<td>attribute</td>
</tr>
<tr>
<td>状中结构</td>
<td>ADV</td>
<td>adverbia</td>
</tr>
<tr>
<td>左附加关系</td>
<td>LAD</td>
<td>left adjunct</td>
</tr>
<tr>
<td>右附加关系</td>
<td>RAD</td>
<td>right adjunct</td>
</tr>
<tr>
<td>标点</td>
<td>WP</td>
<td>punctuation</td>
</tr>
</tbody>
</table>
1. Introduction

➢ Major Contributions

• Generalize three kinds of unique but ubiquitous Chinese linguistic phenomena.
• Propose a novel unsupervised ORE model based on Dependency Semantic Normal Forms (DSNFs).
• Establish a large-scale Chinese entity and relation knowledge base. It is available on Github (https://github.com/TJUNLP/COER)
2. Related work

➢ **Chinese traditional relation extraction**
  - Supervised machine-learning algorithms
    - Feature-based
    - Kernel-based
    Rely heavily on massive manually labeled suitable corpora.
  - Unsupervised relation extraction
    Unreasonable clustering results or inaccurate relation extractions may be uncontrolled and unmeasured.

➢ **Chinese open relation extraction**

➢ **Open relation extraction for English**

➢ **Knowledge bases and corpora**
  - The Automatic Content Extraction (ACE) Program, research the detection and characterization of Entities, Relations, and Events. (830 documents for Chinese)
  - Freebase (more than 125,000,000 tuples)
  - DBpedia (approximately 103 million RDF triples)
  - YAGO (more than 1 million entities and 5 million facts)
3. Unique Chinese linguistic phenomena

Linguistics hold unique characteristics of generality, stability, and nationality.

➢ **The differences between English and Chinese in morphology**

- Dividing mark among words.
- English words vary in their tense, but Chinese words do not.
  - E.g. “托马斯(Tomas) 在(at) 肯德基(KFC) 吃(eat) 早餐(breakfast).”
- There is only one predicate head in an English sentence and many conjunctions among clause. While predicate heads might not be unique in a Chinese sentence, multiple subordinate clauses can be connected without any conjunctions.
  - Comparison
    “乔丹是美国职业篮球运动员，出生在纽约.”
    “Jordan is an American professional basketball player who was born in New York.”

➢ **Influence**

- Many methods using morphology features that usually perform well in English corpus, may achieve poor results in Chinese.
3. Unique Chinese linguistic phenomena

➢ Three main Chinese linguistic phenomena

● **Definition1: Nominal Modification-Center (NMC) phenomenon**
  ■ NMC consists of the modifiers and the head word.
  ■ The head word is a common noun and may play the role of the subject or object in a sentence.
  ■ **Comparison**
    “奥巴马 总统”, “华盛顿 警方”
    “president Obama”, “the police of Washington”
  ■ **Example**
    “奥巴马 总统 访问 中国” ➞ (奥巴马, 访问, 中国)
  ■ **Strategy**
    Instead of the entity directly being the subject and object, the head word usually becomes the main component of the phrase.
    Define the head word as Pseudo-entity to do the conversion.
3. Unique Chinese linguistic phenomena

➢ Three main Chinese linguistic phenomena

● Definition 2: Chinese Light Verb Construction (CLVC) phenomenon
  - CLVC is a co-occurrence that a verb must be with nouns, the patient of the verb appears in the form of a preposition object, and the position of the preposition is flexible and changeable.
  - A light verb carries little semantic content and typically forms a predicate with a noun.

➢ Problem

The English extractor solved the problem by syntactic constrains: the light verbal phrase must be a contiguous word sequence of “verb-noun-preposition”. However, this will no longer be maintained in Chinese because of the flexibility of the position of the preposition in CLVC.
3. Unique Chinese linguistic phenomena

Three main Chinese linguistic phenomena

- Definition 2: Chinese Light Verb Construction (CLVC) phenomenon
  - Solution
    Employ the dependency structure to solve the trouble of CLVC.
  - Construction and relation
    The preposition depends on the verb by **ADV**, and there is an entity object that depends on the preposition by **POB**. Meanwhile, a noun closely follows the light verb by **VOB** as the direct object. Define the combination of the verb and its direct object as a relation expression.
  - Example
    (习近平, 看望师生, 北京八一学校)
3. Unique Chinese linguistic phenomena

➢ Three main Chinese linguistic phenomena

- **Definition 3: Intransitive Verb (IV) phenomenon**

  - IV means that intransitive verb must be linked to its patient by **preposition**, and the preposition may be on the left or right side of the verb.

- **Example**

  “Hudson born in Hampstead, which is a suburb of London”

  “哈德森 出生在 伦敦的 郊区 汉普斯特德” (verb – [CMP] – preposition)

  “哈德森 在 伦敦的 郊区 汉普斯特德 出生” (preposition <- [ADV] – verb)

    (哈德森, 出生在, 汉普斯特德)
3. Unique Chinese linguistic phenomena

➢ Three main Chinese linguistic phenomena distribution

- Randomly collect 500 sentences from web news text, in which they manually extract 118 correct relations.
- Statics show that relation triples that agree with the unique phenomena account for over 60% of the total.
- Among them, NMC(40.37%), CLVC(14.64%), IVC(8.78%)
Two kinds of definitions of DSNFs and the triples.

**Base:** Since entity pairs must appear in the dependency trees as noun phrases, the dependency paths between entities will reflect the corresponding relationships of pairs.

DSNFs can be divided into four categories:
- **Modified construction (MOD)**
- **Verb construction (VERB)**
- **Coordination construction (COOR)**
- **Formulaic construction (FORM)**
4. Dependency semantic normal forms

➢ Modified Construction

- Construction [DSNF1]
  - Relation words are the modifier components in the context of entity pairs.
  - The head word is an entity and modifiers are called the modifying attributives.
- The types of composition of the modifier
  - Proper nouns (entity), common nouns, 的 (De) phrases, numbers, ...
- Example
  - “德国 总统 高克”
  - 德国 <- [ATT] - 总统 <- [ATT] - 高克
  - (德国, 总统, 高克)
- Note
  - There may be two or three attributives and a combination of several modifiers is a relation word.
4. Dependency semantic normal forms

❖ Verb Construction

● Transitive verb relations
  ■ Construction [DSNF2]

  Both entities, as the subject or object of the sentence, depend on the head predicate word by labels \textit{SBV} or \textit{VOB}.

  ■ Example

  “高克 访问 中国”

  高克 <-[SBV]– 访问 –[VOB]– 中国

  \((\text{高克, 访问, 中国})\)
4. Dependency semantic normal forms

➢ Verb Construction

- CLVC or extend CLVC relations

There is no verbal object following the transitive verb. Instead, a prepositional phrase modifies the verb.

- Construction [DSNF3]

Both entities, as the subject or object of the sentence, depend on the head predicate word by labels \textit{SBV} or \textit{VOB}.

- Example

```
“习近平 在 上海 视察”
李克强 <[SBV]– 视察 –[ADV]– 在 –[POB]–> 上海
(习近平, 视察, 上海) (E1, Pred-[Dobj]?+, E2)
```

- Extend example

```
“习近平 对 埃及 进行 国事访问”
习近平 <[SBV]– 进行 –[ADV]– 对 –[POB]–> 埃及, 进行 –[VOB]–> 国事访问
(习近平, 进行国事访问, 埃及)
```
4. Dependency semantic normal forms

➢ Verb Construction

- **IVC Relations**
  - **Construction [DSNF4]**
    DSNF3 is applicable for the condition that the preposition lies on the left of the verb, but when the preposition lies on the right of the verb,...

- **Example**

  “奥巴马 毕业 于 哈佛大学”

  奥巴马 ←[SBV]– 毕业 ←[CMP]– 于 ←[POB]– 哈佛大学

  (奥巴马，毕业于，哈佛大学) (E1, Pred-Prep, E2)
4. Dependency semantic normal forms

➢ Coordination Construction

● Coordinate entities

■ Construction[DSNF5]

As for coordinate entities E1 and E2, all of triples that involve E1 are suitable for E2. Commas and conjunctions are commonly used to connect coordinate entities.

■ Example

“拉里·佩奇 和 谢尔盖·布林 在 1996年 创建 Google”

和 <-[LAD]– 谢尔盖·布林 –[COO]– 拉里·佩奇 <-[SBV]– 创建 –[VOB]– 谷歌

(拉里·佩奇, 创建, 谷歌) (E2, Pred-Prep, E3)

(谢尔盖·布林, 创建, 谷歌) (E2, Pred-Prep, E1)
4. Dependency semantic normal forms

➢ Coordination Construction

- Coordinate verbs
  - Construction [DSNF7]

  The coordinate verbs that are mainly described several different actions are taken by same entity.

- Example

  “高克 访问 中国, 并 在 同济大学 发表 演讲”

  高克 ←[SBV]– 访问 –[COO]– 发表 –[VOB]– 演讲
  发表 –[ADV]– 在 –[POB]– 同济大学

  (高克, 发表演讲, 同济大学) (E1, Pred2, E2)
4. Dependency semantic normal forms

➢ Formulaic Construction

Following the writing styles of Chinese articles, there are a number of fixed expression formats that are difficult to dependency parse.

● Example

“拉里·佩奇, Google 首席执行官, 发表 …”

There is no dependency path to use for extracting!
But just build templates to perform hard matching can get satisfying result.

(拉里·佩奇, 首席执行官, Google)
4. Dependency semantic normal forms

➢ Constraints for DSNFs

To avoid cascading effects caused by dependency parsing errors as much as possible, they adopted many constraints for the DSNFs.

● To maintain the semantic integrity, the extraction is done in sentences, and the length of a sentence is smaller than threshold $M$.

● The number of other entities between the entity pair must not exceed $X_1$, and the number of any other words between the entity pair must be less than $X_2$.

● If $E_1$ is dependent on predicate by $SBV$, there should be no other words between $E_1$ and the predicate.

● For DSNF1, the distance between $E_1$ and $E_2$ should no more than 4.

● For DSNF3, when a word is a direct object that depends on the predicate by $VOB$, the word must be a noun.

● For DSNF3, when there are multiple adverbial phrases in the front of the predicate, only match the one that is nearest to the predicate.

● Divide a noun into two types: the occupation-related noun and naive noun. Extend the entity by joining the naive nouns, E.g. “纽约警方”
5. COER: Chinese open entity and relation knowledge base

➢ About COER

COER is a scalable entity and relation knowledge base.

● Data sources and scale
  ■ Approximately 15G text data.

● Knowledge base
  ■ Approximately 1 million relation triples, 556,012 named entities and 282,347 open relationship phrases.

● Knowledge form
  ■ The extracted triple are stored in a series of XML files.

```xml
<relation id="7">
  <origin_text>新华网北京5月6日电5月2日，从事服装设计的胡明明第四次飞往泰国普吉岛度假</origin_text>
  <entity_pair type="PER-LOC">
    <argument id="1" POS="nh" position="13">胡明明</argument>
    <argument id="2" POS="ns" position="17">泰国普吉岛</argument>
  </entity_pair>
  <relation_phrase phrase_text="飞往">
    <mention id="1" POS="v" position="16" isHead="true">飞往</mention>
  </relation_phrase>
  <dependency_path>A(1)-SBV-M(1); A(2)-VOB-M(1);</dependency_path>
</relation>
```
5. COER: Chinese open entity and relation knowledge base

➢ About COER

COER is a scalable entity and relation knowledge base.

- Knowledge form
6. Experiments

- Experimental setup
  - Step
    - Step 1. Pre-processing input text
      Sentence split ($M=40$), POS tagging, and dependency parsing from the LTP.
    - Step 2. Selecting candidate entities
      Named entity recognizer from the LTP, according POS list \{\textit{ni}, \textit{nh}, \textit{ns}, \textit{nz}, \textit{j}\}
      \textbf{Iterated heuristic algorithm}, is a supplement to the former.
    - Step 3. Matching DSNFs
      Pairing entities separately with the thresholds X1 and X2 are set as 3 and 12.
    - Step 4. Output relation triples according to the match result
  - Data preparation
    - Another subset $\text{Set\_Sample}$: randomly sampled 500 examples from the above four subsets.
    - The sentence that can’t get two or more entities are filtered (high quality).
    - Follow the method in REVERB to label and evaluate the extraction results.
6. Experiments

➢ Experimental setup

● Metrics

■ Precision

\[ P = \frac{C_1}{C_2} \times 100\% \]

■ Recall

\[ R = \frac{C_1}{C_3} \times 100\% \]

■ F-measure

\[ F = \frac{2 \times P \times R}{(P + R)} \times 100\% \]

Where,

\( C_1 \) is the number of correct extractions
\( C_2 \) is the number of results gained by an extractor
\( C_3 \) is the existed instances in corpus
6. Experiments

➢ Experimental results

- Confidence function

ZORE use a logic regression classifier to evaluate each extraction by assigning a confidence score, imitate the process to assign a confidence score to each extraction.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>The $RW_s$ between $E_1$ and $E_2$</td>
<td>0.36</td>
</tr>
<tr>
<td>The $RW_s$ on the right of $E_2$</td>
<td>0.16</td>
</tr>
<tr>
<td>Entities distribute in different clauses</td>
<td>−0.49</td>
</tr>
<tr>
<td>The length of sentence $\leq 18$</td>
<td>0.65</td>
</tr>
<tr>
<td>The $RW_s$ are noun</td>
<td>0.44</td>
</tr>
<tr>
<td>The $RW_s$ are verb</td>
<td>0.40</td>
</tr>
<tr>
<td>The $RW_s$ are compound words</td>
<td>−0.35</td>
</tr>
<tr>
<td>The dependency label of $E_1$ is $SBV$</td>
<td>−0.18</td>
</tr>
<tr>
<td>The dependency label of $E_1$ is $ATT$</td>
<td>−0.10</td>
</tr>
<tr>
<td>The dependency label of $E_1$ is $COO$</td>
<td>0.09</td>
</tr>
<tr>
<td>The dependency label of $E_2$ is $VOB$</td>
<td>0.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>The dependency label of $E_2$ is $POB$</td>
<td>−0.56</td>
</tr>
<tr>
<td>The dependency label of $E_2$ is $COO$</td>
<td>0.64</td>
</tr>
<tr>
<td>The dependency label of $RW_h$ is $HED$</td>
<td>−0.13</td>
</tr>
<tr>
<td>The dependency label of $RW_h$ is $COO$</td>
<td>−0.24</td>
</tr>
<tr>
<td>The dependency label of $RW_h$ is $ATT$</td>
<td>0.58</td>
</tr>
<tr>
<td>The number of entities $= 2$</td>
<td>0.47</td>
</tr>
<tr>
<td>The number of words between entities $\leq 5$</td>
<td>0.71</td>
</tr>
<tr>
<td>First word on left of $E_1$ is noun</td>
<td>−1.14</td>
</tr>
<tr>
<td>First word on right of $E_2$ is noun</td>
<td>−0.58</td>
</tr>
</tbody>
</table>

*Note: This table shows the features and weights of the logistics regression classifier to mark for an extraction ($E_1$, $RW_s$, $E_2$). Here, $RW_h$ indicates the head of relationship words.*
6. Experiments

➢ Experimental results

● Confidence function

Compute precision-recall curves of DSNFs and ZORE by varying the confidence threshold and then compare the area under the curve (AUC) on dataset Set_Sample.

■ Analysis

◆ DSNFs has much higher precision than ZORE at all levels of recall.
◆ DSNFs achieve an AUC 18% higher than ZORE.
◆ DSNFs more than 40% of extractions at a precision of 0.9, ZORE’s results have a precision of greater than 0.8 only about 30%.
◆ DSNFs doesn’t achieve the symmetrical improvement of precision with sacrifice of recall.
6. Experiments

➢ Experimental results

● Self-Testing Performance

The performance of the model in different datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encyclopedia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set_Wiki</td>
<td>84.42</td>
<td>57.75</td>
<td>69.03</td>
</tr>
<tr>
<td>Set_BaiduBaike</td>
<td>83.93</td>
<td>58.50</td>
<td>68.94</td>
</tr>
<tr>
<td>News</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set_SinaNews</td>
<td>82.66</td>
<td>57.72</td>
<td>67.97</td>
</tr>
<tr>
<td>Set_SogouNews</td>
<td>81.87</td>
<td>58.98</td>
<td>68.57</td>
</tr>
</tbody>
</table>

■ Analysis

◆ The performance in the encyclopedia sources is better than that in the news sources.

◆ The are more errors when the NLP tools process news text.
6. Experiments

➢ Experimental results

● Performance comparing

<table>
<thead>
<tr>
<th>methods</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSNFs</td>
<td>83.76</td>
<td>58.68</td>
<td>69.01</td>
</tr>
<tr>
<td>UnCORE</td>
<td>80.57</td>
<td>47.62</td>
<td>59.86</td>
</tr>
<tr>
<td>ZORE</td>
<td>83.76</td>
<td>14.47</td>
<td>24.68</td>
</tr>
<tr>
<td>( \text{REVERB}^{-n} )</td>
<td>59.93</td>
<td>58.68</td>
<td>59.30</td>
</tr>
<tr>
<td>( \text{REVERB}^{-v} )</td>
<td>81.69</td>
<td>25.00</td>
<td>38.28</td>
</tr>
<tr>
<td>( \text{REVERB}^{-v} )</td>
<td>47.22</td>
<td>3.39</td>
<td>6.34</td>
</tr>
</tbody>
</table>

■ Analysis
The results prove that the English extractors are not suitable for processing the Chinese corpus; and the model’s performance is better than that of the other open Chinese relation extraction systems.

■ Extension of model
To extract the relations between entity mentions. Adopt a method to obtain base noun phrases (NPs) from dependency parsing as entity mentions.

<table>
<thead>
<tr>
<th>methods</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSNFs</td>
<td>83.33</td>
<td>57.14</td>
<td>67.79</td>
</tr>
<tr>
<td>ZORE</td>
<td>60.47</td>
<td>74.29</td>
<td>66.67</td>
</tr>
</tbody>
</table>
6. Experiments

➢ Error Analysis

● Precision loss

The major errors analysis of incorrect extractions

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Named Entity Recognition errors</td>
<td>31%</td>
</tr>
<tr>
<td>Dependency parsing errors</td>
<td>28%</td>
</tr>
<tr>
<td>Model defects</td>
<td>25%</td>
</tr>
<tr>
<td>Chinese Segmentation and POS tagging errors</td>
<td>16%</td>
</tr>
</tbody>
</table>

● Recall loss

The major errors analysis of missed correct extractions

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity pair is incomplete</td>
<td>24%</td>
</tr>
<tr>
<td>Model defects</td>
<td>21%</td>
</tr>
<tr>
<td>Entity Recognition omissions</td>
<td>19%</td>
</tr>
<tr>
<td>Other NLP processing errors</td>
<td>18%</td>
</tr>
<tr>
<td>The employ of pronouns</td>
<td>13%</td>
</tr>
<tr>
<td>Relation phrases are expatiatory and complicated</td>
<td>5%</td>
</tr>
</tbody>
</table>
6. Experiments

➢ Evaluation of DSNFs

The statistical amount distributions of triples extracting based on each DSNF in the main syntactic classes. This method can achieve presentable accuracy so that we can use it to label automatically to help to train those extractors that are established by appropriate machine-learning algorithms even deep learning.
7. Conclusion

➢ The main work
  ● Propose an unsupervised Chinese open relation extraction model that solves the Chinese linguistic troubles.
  ● Establish a knowledge base COER.

➢ The Limit
  ● The method is limited by the performance of NLP technologies including Chinese segmenting, POS tagging, dependency parsing and so on.

➢ The next contemplation
  Utilize the corpus to do subsequent research, such as
  ● Chinese distant supervised relation extraction
  ● Chinese relation extraction based on deep learning
  ● Chinese entity linking and relation inferences
Knowledge Base Establishment on Legal Traffic Instruments

汇报人：胡志强
2018年5月30日
Content

➢ 1. Data introduction
➢ 2. Expert knowledge
➢ 3. The knowledge extraction framework
➢ 4. The extraction result
➢ 5. Knowledge visualization
1. Data introduction

➢ Data overview

<table>
<thead>
<tr>
<th>ID</th>
<th>Text file</th>
<th>Data scale (piece)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>机动车交通事故责任纠纷-案例.txt</td>
<td>668001</td>
</tr>
<tr>
<td>2</td>
<td>机动车交通事故责任纠纷-争议焦点.txt</td>
<td>180222</td>
</tr>
</tbody>
</table>
1. Data introduction

Data sample (Instrument)
1. Data introduction

Data sample (Instrument)

<table>
<thead>
<tr>
<th>Code</th>
<th>Value</th>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>37</td>
<td>2014年3月10日2014年5月4日，原告在中国人民解放军第二炮兵总医院住院55天，开支医疗费152376.81元，期间于2014年3月28日遵医嘱外购硫酸钠人工骨粉盒6237元，合计158613.81元；</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>38</td>
<td>2014年7月14日至2014年8月25日，原告在中国人民解放军第二炮兵总医院住院42天，开支医疗费64333.96元；</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>39</td>
<td>2014年10月13日至2014年11月8日，原告在中国人民解放军第二炮兵总医院住院26天，开支医疗费23810.92元；2016年4月1日至2016年5月3日，原告在中国人民解放军第二炮兵总医院住院32天，开支医疗费33650.28元。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>41</td>
<td>以上医疗费总计人民币301271.60元。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>42</td>
<td>2017年3月24日，中国人民解放军火箭军总医院医务部证明：由于误诊，患者葛继廷在第2-4次住院病历中出现时间差错，2013年3月18日因右胫骨骨折致残定为2014年03月10日。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>43</td>
<td>又查，车牌号冀D×××车架号LAKCA2ZVH9CY1519，2013年4月15日因购买转移登记为冀J×××××，该车在被告华安财产保险股份有限公司邯郸中心支公司投保了机动车交通事故强制保险122000元，保险期间自2012年8月18日至2013年8月17日。本次交通事故发生在保险期间内。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>44</td>
<td>冀J×××××/冀J×××××车辆事故实际所有权人为被告孙果城，冀J×××××/冀J×××××车辆在被告华安财产保险股份有限公司邯郸中心支公司投保了机动车交通事故强制保险122000元，保险期间自2012年9月15日至2013年9月14日。该车辆未投保商业险。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>45</td>
<td>又查，豫F×××××/豫F×××××车辆信息在唐山市公安局交通警察支队外环高速大队出具道路交通事故认定书中均记载为豫F×××××/豫F×××××车辆所有人吴海强，检验有效期至2013年10月31日。本次交通事故发生在保险期间内。2012年12月11日至2013年12月10日。该车辆在被告中国人民财产保险股份有限公司鹤壁分公司投保了机动车交通事故强制保险122000元及不计免赔率险各2份，保险期间自2012年12月11日至2013年12月10日。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>7</td>
<td>46</td>
<td>以上事故，有当事人陈述及有关证据予以证实。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>8</td>
<td>47</td>
<td>本院认为，原、被告对唐山市公安局交通警察支队外环高速大队的道路交通事故认定书均未提出异议，本院予以采纳。葛继廷、刘李胜、孙果城的主次、责任按70%、15%、15%分担。该事故有2人受伤，由于损失比例无法确定，本院确定伤者各自承担医疗费的1/2。原告误工的医疗费人民币301271.60元，本院予以确认，由被告华安财产保险股份有限公司邯郸中心支公司在机动车交通事故强制保险医疗费用赔偿限额内赔偿10000元；由于其他保险公司机动车交通事故强制保险医疗费用赔偿限额已用尽，超出交强险部分（人民币301271.60元-10000元）×38%=87381.64元，由被告吴海强投保的被告中国人民财产保险股份有限公司鹤壁分公司和孙果城各赔偿1/2即43690.74元。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>9</td>
<td>48</td>
<td>依据《中华人民共和国侵权责任法》第六条第一款、第十六条、第四十八条、《中华人民共和国道路交通安全法》第七十六条，以及《最高人民法院关于审理人身损害案件适用法律若干问题的解释》第十七条和有关法律规定。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>16</td>
<td>49</td>
<td>判决如下：</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>16</td>
<td>50</td>
<td>一、被告华安财产保险股份有限公司邯郸中心支公司赔偿原告医疗费人民币10000元；</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>16</td>
<td>51</td>
<td>二、被告中国人民财产保险股份有限公司鹤壁分公司赔偿原告医疗费人民币43690.74元；</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>16</td>
<td>52</td>
<td>三、被告孙果城赔偿原告医疗费人民币87381.64元；</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>16</td>
<td>53</td>
<td>以上条款均于本判决生效后10日内履行。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>16</td>
<td>54</td>
<td>四、驳回原告的其他诉讼请求。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>16</td>
<td>55</td>
<td>如被告未按本判决书指定的期间履行给付义务的，应当依照《中华人民共和国民事诉讼法》第二百五十三条之规定，加倍支付迟延履行期间的债务利息。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>10</td>
<td>56</td>
<td>案件受理费人民币770元，由被告吴海强、孙果城各自负担385元。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>11</td>
<td>57</td>
<td>如不服本判决，可在判决书送达之日起15日内，向本院递交上诉状，并按对方当事人的人数提出副本，按照河北省高级人民法院冀高法[2015]79号意见提交支持上诉理由和上诉诉讼请求的证据等材料，上诉于河北省唐山市中级人民法院。</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>12</td>
<td>58</td>
<td>审判长</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>12</td>
<td>59</td>
<td>审判长 王树宁</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>12</td>
<td>60</td>
<td>审判员 孟洋洋</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>12</td>
<td>61</td>
<td>审判员 赵立新</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>12</td>
<td>62</td>
<td>二〇一七年三月二十四日</td>
</tr>
<tr>
<td>009f3c7-119d-11e8-9a7d-17221d1a4ea3</td>
<td>12</td>
<td>63</td>
<td>书记员 杨有丽</td>
</tr>
</tbody>
</table>
| 009f3c7-119d-11e8-9a7d-17221d1a4ea3 | 99 | 65 | 特别提示：根据冀高法[2015]79号意见，当事人不服一审判决、裁定，在法律规定的期限内上诉的，通过原一审人民法院提出，并提交以下材料：
①上诉状是自然人的，提交身份证明复印件；上诉人是法人或其他组织的，提交营业执照复印件、组织机构代码证复印件、法定代表人或主要负责人身份证明书和身份证复印件，委托他人代理上诉的，还应当提交授权委托书和代理人身份证明。
②上诉状副本及本院按上诉状提交的上诉状副本。
③一审裁判文书及复印件，或者经核实无误的复印件。
④支持上诉理由和上诉诉讼请求的证据材料。
⑤上诉人送达地址确认书。
⑥材料清单。材料不齐全的，上级人民法院将退回上诉。 |
| 009f3c7-119d-11e8-9a7d-17221d1a4ea3 | 66 | 66 | 法院账号开户行：中国建设银行股份有限公司唐山丰南文化路支行（简称建行唐山丰南文化路支行）户名：唐山市丰南区人民法院 |
1. Data introduction

Data sample

Instrument
1. Data introduction

- Data sample
  - Focus of disputation

00005279-296b-11e7-990f-17070a1d658f @hc关于本案争议的焦点：1、关于人寿保险公司辩称原告的起诉超过诉讼时效问题，本案交通事故的发生时间是2014年3月6日，因李如强涉嫌交通肇事，被淮北市杜集区人民检察院提起公诉，本院于2014年9月4日作出（2014）杜刑初字第00056号刑事判决。在刑事案件的审理过程中，死亡受害人的亲属张长根作为刑事附带民事诉讼的原告人曾提起过附带民事诉讼，后撤诉。现张祚环于2015年5月28日向本院提起民事赔偿诉讼，未超过诉讼时效，人寿保险公司的上述辩由依法不能成立。2、关于原告的主体资格问题。人寿保险公司辩称张长根应是死者财产的合法继承人，因原告提供的公证书涉及的事项与本案交通事故受害人权利的主张无直接因果关系。因死亡受害人张祚义未婚、无子女，无第一顺序继承人，张祚环作为张祚义的近亲属有权提起诉讼，故张祚环是本案的适格主体。3、关于李如强要求返还精神损害抚慰金问题。因实际侵权人李如强犯交通肇事罪被本院判刑，精神损害抚慰金依法不予支持。李如强给付受害人精神抚慰金系双方自愿给付，且给付后李如强取得了受害人亲属的谅解。李如强要求返还精神损害抚慰金，依法不予支持。@hc关于原告的损失认定问题。1、死亡赔偿金，受害人张祚义死亡时未满72周岁，原告要求按照9年计算，本院予以支持。原告要求按照城镇居民标准计算死亡赔偿金，所举证据不足，本院不予支持。张祚义系农村居民，死亡赔偿金为89244元（9916元*9年）。2、丧葬费，原告要求赔偿丧葬费22300元，符合法律规定，本院予以支持。3、关于精神抚慰金，因实际侵权人李如强犯交通肇事罪被本院判刑，原告要求赔偿精神损害抚慰金，本院不予支持。@hc综上，原告本起交通事故的赔偿项目为：死亡赔偿金89244元、丧葬费22300元，合计111544元。对于原告的损失，根据事故责任认定，由人寿保险公司在交强险范围内赔偿110000元，在商业三者险限额内赔偿1544元。

(Note: “@hc” indicates the break of natural paragraph)
2. Expert knowledge
3. The knowledge extraction framework

➢ Overall framework
3. The knowledge extraction framework

➢ Detail workflow

The involved NLP step for unstructured extraction

- Sentence split (*)
- Word segment (NLPIR[ICTCLAS] + THUOCL_law.txt)
- POS tagging (LTP)
- Name entity recognition (LTP)
- Dependency parsing (LTP)
3. The knowledge extraction framework

➢ The model

- **DSNFs** (Basic implementation)
- **DSNF for instrument** (An example, need expert knowledge)
4. The extraction result

<table>
<thead>
<tr>
<th>文书名称</th>
<th>基于制定规则的方法</th>
</tr>
</thead>
<tbody>
<tr>
<td>文书类型</td>
<td>处理第二列分段标号及说明：</td>
</tr>
<tr>
<td>案号</td>
<td>1---判决书名称</td>
</tr>
<tr>
<td>原告</td>
<td>14---法院名称</td>
</tr>
<tr>
<td>被告</td>
<td>15---文书类型</td>
</tr>
<tr>
<td>华安财产保险有限公司邯郸中心支公司</td>
<td>2---案号</td>
</tr>
<tr>
<td>一</td>
<td>3---当事人信息</td>
</tr>
</tbody>
</table>

基于依存句法分析的方法

处理第二列分段标号及说明：

7---法院认定事实

依据的几种句法：

- 主语谓语宾语关系
- 定语后置动宾关系
- 介宾关系主谓动补关系
4. The extraction result

{ "文书编号": "600603c7-119d-11e8-9a7d-172214d14ea3",
"文书标题": "葛继民与孙某城、青县凯达汽车运输队机动车交通事故责任纠纷一审民事判决书",
"受理法院": "河北省唐山市丰南区人民法院",
"文某判决": "民事判决书",
"案件编号": "(2017)冀0207民初512号",
"原告诉求": {
"编号": 1,
"姓名": "葛继民",
"委托代理人": "葛印楼"
}
},
"被告": {
"编号": 1,
"姓名": "孙某城",
"委托代理人": "崔志学"
},
"编号": 2,
"姓名": "青县凯达汽车运输队",
"法定代表人": "李海伦"
},
"编号": 3,
"姓名": "南皮县某大汽车运输队"
},
"编号": 4,
"姓名": "刘太胜"
},
"编号": 5,
"姓名": "吴海强"
},
"编号": 6,
"姓名": "李建平"
},
"编号": 7,
"姓名": "肖海永"
},
"编号": 8,
"姓名": "中国人民财产保险股份有限公司鹤壁分公司",
"负责人": "陈良",
"委托代理人": "李现亮"
},
"编号": 9,
"姓名": "中国人民财产保险股份有限公司青县支公司",
"负责人": "郑建广",
"委托代理人": "李春辉"
},
"编号": 10,
"姓名": "华安财产保险有限公司邯郸中心支公司",
"负责人": "李立刚",
"委托代理人": "解河海"}

"案情事实": {
"编号": 1,
"句子": "葛继民驾驶冀B××××号重型半挂牵引车由北向南行驶至长××(×)下××980KM+200M处时",
"知识": "[葛继民", "重型半挂车", "冀B××××号"
},
"编号": 2,
"句子": "葛继民驾驶冀B××××号重型半挂车由北向南行驶至长××(×)下××980KM+200M处时",
"知识": "[冀B××××号"
},
"编号": 3,
"句子": "行至此处刘太胜驾驶的豫F××××/豫F××××号重型半挂车右前部与冀B××××/冀B××××号重型半挂车的左侧相撞",
"知识": "[豫F××××/豫F××××号"
},
"编号": 4,
"句子": "该车在被告华安财产保险股份有限公司邯郸中心支公司投保了机动车交通事故强制保险122000元",
"知识": "[该车", "投保元", "华安财产保险股份有限公司邯郸中心"
},
"编号": 5,
"句子": "冀J×××××牵引车挂靠在被告青县凯达汽车运输队",
"知识": "[冀J×××××", "挂靠在", "青县"
},
"编号": 6,
"句子": "冀J×××××车挂靠在被告南皮县某大汽车运输队",
"知识": "[冀J×××××", "挂靠在", "南皮县"
},
"编号": 7,
"句子": "其中冀J×××××在被告中国人民财产保险股份有限公司青县支公司投保了机动车交通事故强制保险122000元",
"知识": "[冀J×××××", "投保元", "中国人民财产保险股份有限公司青县支公司"
},
"编号": 8,
"句子": "该车在被告中国人民财产保险股份有限公司鹤壁分公司投保了机动车交通事故强制保险122000元，第三者责任保险300000元及不计免赔率等险种各2份",
"知识": "[该车", "投保份", "中国人民财产保险股份有限公司鹤壁分公司"]
}
5. Knowledge visualization

➢ Import Neo4j
Thanks