Argumentation-Driven Evidence Association in Criminal Cases

Yefei Teng and **Wenhan Chao**
School of Computer Science and Engineering, Beihang University, Beijing, China
flyxdoimp@gmail.com
chaowenhan@buaa.edu.cn

Abstract

Evidence association in criminal cases is dividing a set of judicial evidence into several non-overlapping subsets, improving the interpretability and legality of conviction. Observably, evidence divided into the same subset usually supports the same claim. Therefore, we propose an argumentation-driven supervised learning method to calculate the distance between evidence pairs for the following evidence association step in this paper. Experimental results on a real-world dataset demonstrate the effectiveness of our method.

1 Introduction

Previous work has put forward multiple legal assistant systems with various functions, such as searching relevant cases given the query (Chen et al., 2013), predicting the legal judgement (Ye et al., 2018), etc. Despite promising results in this area, the research of judicial evidence in criminal cases has been omitted in recent years. The role of judicial evidence is to support several sub-claims in favour of conviction and the evidence description is an essential part of criminal judgement documents. However, the organization of evidence varies in different legal documents. The form of evidence association mainly includes collection form and argumentation-driven form as shown in Figure 1. In most current criminal judgement documents, the evidence is only listed in the form of a collection without giving explicit claims, which is regarded as collection form. However, evidence collection is divided into several subsets according to related claim only in around 5% criminal judgement documents, which is regarded as argumentation-driven form.

As shown in Figure 1, evidence divided into the same subset could support the same claim and such kind of legal documents have better readability. Inspired by this observation, we propose to study the problem of evidence association in this paper. Evidence association is dividing a set of judicial evidence into several non-overlapping subsets according to their corresponding claims, improving the interpretability and legality of conviction. To our knowledge, there has been very limited research about evidence association in the legal field.

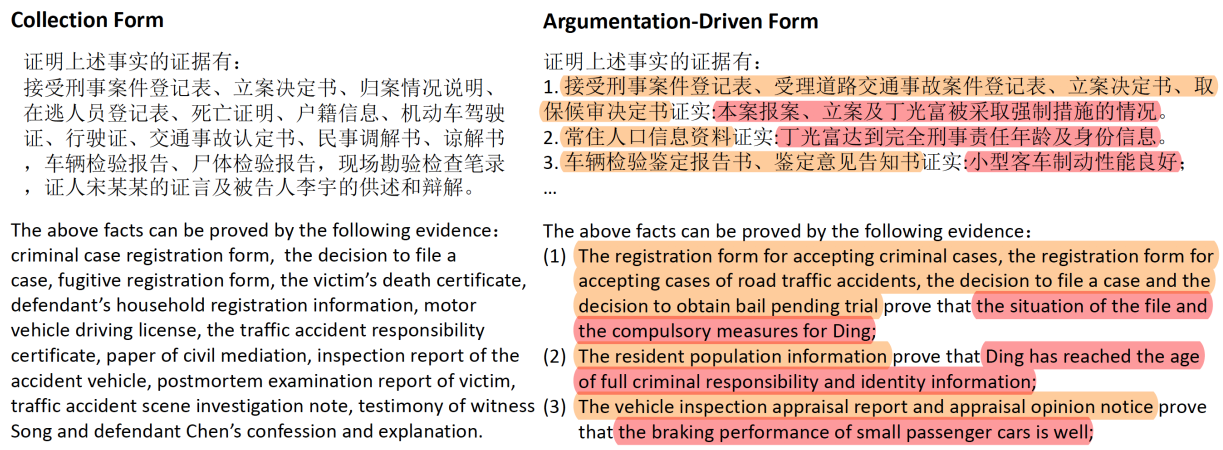
Evidence association could be treated as a clustering problem. Existing short text clustering methods broadly fall into two categories: representation-based methods and semantic textual similarity methods(Xu et al., 2017; Reimers et al., 2019). The representation-based methods concentrate on extracting rich semantic representation and then calculate cosine distance between text representations. The semantic textual similarity methods predict the distance between texts directly through supervised learning. However, the former methods perform poorly on the very short text and the latter methods require manually labelled data in the same field for supervised learning. We learn distance metric based on the probability supporting the same claim between evidence pairs directly on account of the short length of judicial evidence, which is regarded as an argumentation-driven method. Another challenge is that the number of clusters in each case is various. In this paper, we use agglomerative hierarchical clustering to learn the stopping threshold to avoid specifying the number of clusters. Our contributions of this paper are presented as follows:

1. We propose a task of evidence association in criminal cases which is significative but has not been well studied before and release a real-world dataset for this task.

2. We learn the distance metric by supervised argumentation-driven method for subsequent clustering without extra manual annotation.

3. Extensive experiments conducted on a real-world dataset show the efficiency of our methods and provide a simple baseline for future research.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2997–3001
November 7–11, 2021. ©2021 Association for Computational Linguistics



Collection Form

证明上述事实的证据有：
接受刑事案件登记表、立案决定书、归案情况说明、在逃人员登记表、死亡证明、户籍信息、机动车驾驶证、行驶证、交通事故认定书、民事调解书、谅解书，车辆检验报告、尸体检验报告，现场勘验检查笔录，证人宋某某的证言及被告人李宇的供述和辩解。

The above facts can be proved by the following evidence：
criminal case registration form, the decision to file a case, fugitive registration form, the victim's death certificate, defendant's household registration information, motor vehicle driving license, the traffic accident responsibility certificate, paper of civil mediation, inspection report of the accident vehicle, postmortem examination report of victim, traffic accident scene investigation note, testimony of witness Song and defendant Chen's confession and explanation.

Argumentation-Driven Form

证明上述事实的证据有：
1. 接受刑事案件登记表、受理道路交通事故案件登记表、立案决定书、取保候审决定书证实:本案报案、立案及丁光富被采取强制措施的情况。
2. 常住人口信息资料证实:丁光富达到完全刑事责任年龄及身份信息。
3. 车辆检验鉴定报告书、鉴定意见告知书证实:小型客车制动性能良好;
...

The above facts can be proved by the following evidence：
(1) The registration form for accepting criminal cases, the registration form for accepting cases of road traffic accidents, the decision to file a case and the decision to obtain bail pending trial prove that the situation of the file and the compulsory measures for Ding;
(2) The resident population information prove that Ding has reached the age of full criminal responsibility and identity information;
(3) The vehicle inspection appraisal report and appraisal opinion notice prove that the braking performance of small passenger cars is well;

Figure 1: A real-word example of evidence descriptions in argumentation-driven form and collection form. The part before "prove that" is the evidence subset and the part after "prove that" is the corresponding claim.

2 Related works

The evidence association task is motivated by previous research on the legal assistant system, especially by the work of improving the interpretability of charge prediction (Ye et al., 2018). To our knowledge, there has been very limited research about evidence associations in the legal field. One of the most related research work was done by Poudyal et al. (2018). They use clustering techniques to identify argumentative sentences in legal documents, whereas it is a sentence-level task.

As a part of argument mining, argument clustering aims to identify similar arguments. Boltužić and Šnajder (2015) identifies similar arguments in online debates using semantic textual similarity. Ajjour et al. (2019) groups arguments that emphasize a specific aspect of a controversial topic. Contextualized word embeddings methods are introduced in the classification and clustering of arguments in recent years (Reimers et al., 2019). In this paper, we mainly used the BERT(Devlin et al., 2019) and ESIM(Chen et al., 2017) model to learn the distance metric between evidence pairs.

3 Methodology

Given a set of evidence $E = \{e_1, e_2, ..., e_n\}$ involved with a criminal case, we expect to split the E into k non-overlapping subset $\{E_1, E_2, ..., E_k\}$ iff. $\bigcup_t^k E_t = E, E_i \cap E_j = \emptyset, 1 \leq i < j \leq k$. Each non-overlapping subset of evidence E_k proves the same claim c_k. We firstly study the latent argumentation-driven evidence association in the case of lacking explicit claims. We also explored how to associate evidence more accurately in the case of giving the explicit claim set $C = \{c_1, c_2, ..., c_k\}$ involved in the criminal case. Similarly, we define it as an explicit argumentation-driven evidence association. A suitable clustering method and a meaningful distance between evidence pairs are crucial for evidence association.

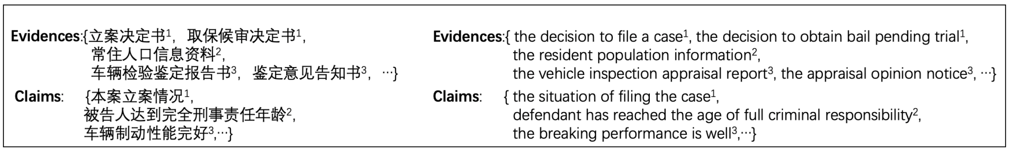
3.1 Clustering Method

It is a prior that the number of clusters in each case is various so that we can not set a specific cluster number like the K-Means method. We try to cluster evidence via agglomerative hierarchical clustering (Day and Edelsbrunner, 1984), which learns a stopping threshold that determines when to stop merging two clusters without giving the specific number of clusters.

3.2 Distance Metric

Latent Distance

Without giving the explicit claims, we can only use the information of the evidence pairs to calculate the distance between them. Nogueira and Cho (2019) define the correlation between relevant query-passage pairs as 0 and irrelevant query-passage pairs as 1 on account of the lack of labeled dataset. Similarly, we assume a smaller distance between two pieces of evidence that support the same claim. For simplification, the distance between evidence pairs that supports the same claim is labeled to 0. And the distance between evidence pairs involved in the same criminal case that prove different claims is labeled to 1. If p is the possibility that the distance between evidence pairs is 0 predicted by the model, then we simply regard the latent distance between evidence pairs as $1 - p$.



Evidences:{立案决定书^1，取保候审决定书^1,
常住人口信息资料^2,
车辆检验鉴定报告书^3，鉴定意见告知书^3，…}
Claims: {本案立案情况^1,
被告人达到完全刑事责任年龄^2,
车辆制动性能完好^3,…}

Evidences:{ the decision to file a case^1, the decision to obtain bail pending trial^1,
the resident population information^2,
the vehicle inspection appraisal report^3, the appraisal opinion notice^3, …}
Claims: { the situation of filing the case^1,
defendant has reached the age of full criminal responsibility^2,
the breaking performance is well^3,…}

Figure 2: An example of pre-processed samples. The superscript of the claim text represents the ID number of the claim. The superscript of the evidence text represents which claim the evidence can support.

Explicit Distance

There is strong relevance between evidence and the corresponding claim. For example, the traffic accident responsibility certificate can support the division of responsibility for traffic accidents. Therefore, we assume a higher relevance score between evidence and the corresponding claim. Similar to the sampling method mentioned above, the relevance score between evidence and the corresponding claim is 1 and the relevance score between evidence and any other claim is 0.

For a given criminal case, there is a evidence set denoted as $E = \{e_1, e_2, ..., e_n\}$ and a claim set denoted as $C = \{c_1, c_2, ..., c_m\}$. Models predict a relevance score matrix denoted as $A \in \mathbb{R}^{n \times m}$. Each element a_{ij} in matrix A means the relevance score between the evidence e_i and the claim c_j. We assume that evidence belonging to the same cluster have a similar relevance score distribution. More specifically, suppose the relevance score distribution of evidence e_1 is $P \in \mathbb{R}^{1 \times m}$, where each element P_j is the relevance score between evidence e_1 and claim c_j. Similarly, $Q \in \mathbb{R}^{1 \times m}$ is the relevance score distribution of evidence e_2. We view Jensen–Shannon divergence (Endres and Schindelin, 2003) between these two distributions as the explicit distance between e_1 and e_2.

Ensemble Distance

The latent distance only uses the semantic information between the evidence texts to calculate the similarity. The explicit distance only uses the inference relationship between evidence and claim to calculate the distance between evidence. We try to use the semantic information between the evidence and the inference information between the evidence and the claim at the same time by fusing these two methods. We define the ensemble distance as the weighted sum of these two distances.

Table 1: Statistics of our dataset

Avg. number of evidence	16.2
Avg. number of claim	11.9
Avg. length of evidence	10.7
Avg. length of claim	45.5

4 Experiments

4.1 Datasets

We construct a new dataset from the published legal documents in China Judgements Online[1]. We selected the legal documents where the evidence description is the argumentation-driven form as shown in Figure 1 for experiments. For those evidence descriptions of argument-driven form, we can extract the evidence and corresponding claims without manual annotation easily. A subset of evidence and the corresponding claim are always on the same line. The part before “prove that” is the evidence subset and the part after “prove that” is the corresponding claim. Evidence in the same subset is usually separated by punctuations. After pre-processing, each judicial evidence description sample can be composed of an evidence set and a claim set as the illustration of our data in Figure 2.

We select 500 cases of the Traffic Accident Crime, which is one of the most frequent criminal charges. We counted the average number of judicial evidence and claims per case. The average length of evidence and claims of Chinese characters are calculated. The detailed statistical results of the datasets are shown in Table 1.

4.2 Experimental Setup

We calculate the cosine distance between the average word GloVe embeddings of evidence pairs as a baseline. We mainly adopt ESIM and BERT to predict the distance via supervised learning.

ESIM. We tokenize the Chinese texts with the open-source tool of HanLP[2] and use the Glove

[1] http:/wenshu.court.gov.cn

[2] https://github.com/hankcs/HanLP

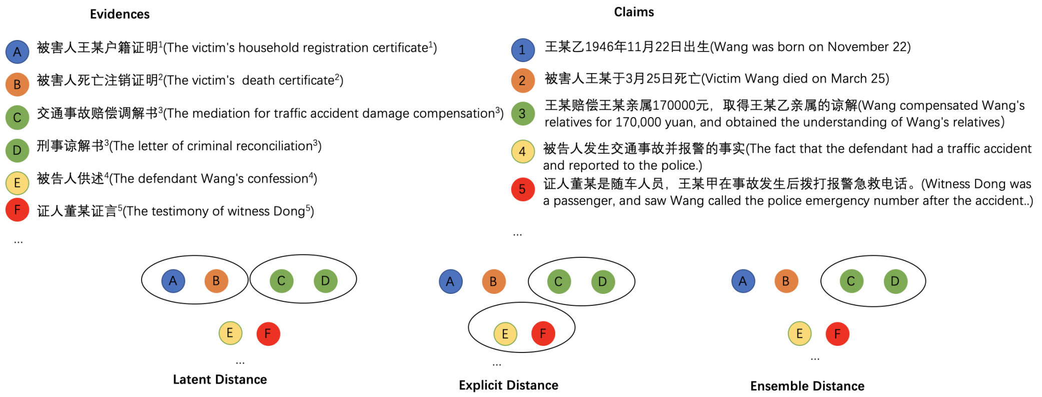

Evidences
A 被害人王某户籍证明¹(The victim's household registration certificate¹)
B 被害人死亡注销证明²(The victim's death certificate²)
C 交通事故赔偿调解书³(The mediation for traffic accident damage compensation³)
D 刑事谅解书³(The letter of criminal reconciliation³)
E 被告人供述⁴(The defendant Wang's confession⁴)
F 证人董某证言⁵(The testimony of witness Dong⁵)
...
Claims
1 王某乙1946年11月22日出生(Wang was born on November 22)
2 被害人王某于3月25日死亡(Victim Wang died on March 25)
3 王某赔偿王某亲属170000元，取得王某乙亲属的谅解(Wang compensated Wang's relatives for 170,000 yuan, and obtained the understanding of Wang's relatives)
4 被告人发生交通事故并报警的事实(The fact that the defendant had a traffic accident and reported to the police.)
5 证人董某是随车人员，王某甲在事故发生后拨打报警急救电话。(Witness Dong was a passenger, and saw Wang called the police emergency number after the accident..)
...
A B C D E F
Latent Distance
Explicit Distance
Ensemble Distance

Figure 3: An example of clustering results via different distance metrics. The superscript of the evidence text represents which claim the evidence can support and evidence with the same superscript should be grouped together.

(Pennington et al., 2014) word embedding trained on the corpus crawled from China Judgements Online with the embedding size of 300. We trained the model for 20 epochs, with a learning rate of 1e-4, a hidden size of 300, and a batch-size of 32.

BERT. We concatenate evidence pairs (evidence-claim pairs while calculating explicit distance, both separated by a special [SEP] token) and add a sigmoid layer to the special [CLS] token. We only fine-tune the last two layers of the BERT model for 10 epochs with a learning rate of 5e-5 and a batch-size of 32.

We choose the weights between latent and explicit distance after testing the results of different proportions.

The agglomerative hierarchical clustering method has a stopping threshold parameter. We choose the best parameter on the validation dataset in the range of 0 to 0.2 with a step size of 0.001. To ensure the stability of the experimental results, we evaluate methods via 5-fold cross-validation.

4.3 Result and Analysis

As the constructed datasets include ground truth cluster labels, we adopt the Adjusted Rand Index(ARI)(Hubert and Arabie, 1985) and the Adjusted Mutual Information(AMI)(Vinh et al., 2009) to evaluate the clustering performance.

Table 2 presents the experiment results. Encouragingly, compared with unsupervised methods, the performance of any one of the supervised methods is much higher. Meanwhile, the BERT model outperforms the ESIM model. Firstly, the deeper neural network produces better performance. Another possible reason may be that the evidence pairs supporting the same claim have a co-occurrence tendency, which could be learned by the next sentence prediction task of the BERT model. The performance of latent distance is better than the explicit distance because it utilizes the semantic information between evidence pairs. The clustering result via the ensemble distance has a great improvement than any single distance owing to integrating the relationship between evidence pairs and evidence-claim pairs.

Table 2: The clustering results

Metrics		ARI	AMI
Unsupervised Methods(Average Embeddings)			
GloVe	cosine	0.169	0.204
Supervised Methods			
ESIM	$dist_{latent}$	0.582	0.599
	$dist_{explicit}$	0.519	0.540
	$dist_{ensemble}$	0.633	0.646
BERT	$dist_{latent}$	0.603	0.611
	$dist_{explicit}$	0.534	0.555
	$dist_{ensemble}$	**0.643**	**0.656**

As shown in Figure 3, claims 1 and 2 represent the victim's date of birth and death, respectively. Both the victim's household registration certificate and the victim's death certificate can partly support the victim's identification information, and they were clustered together by mistake while using latent distance because no explicit claims were given so that only the semantic relationship between evidence pairs are used. Claims 4 and 5 are similar and they are both descriptions of the scene of a traffic accident. The defendant Wang's confession and the testimony of witness Dong are clustered together by mistake because almost no semantic relationship between evidence pairs is considered while using explicit distance. The clustering result via the ensemble distance is correct via combining the semantic relationship between evidence pairs and the information introduced by explicit claims.

5 Conclusion

In this paper, we propose a novel task of evidence association. The experiment results show that supervised methods significantly improve the clustering results even with a few training data. The clustering results have been greatly improved by introducing the information from explicit claims. Since explicit claims are not given in most cases, we are now studying how to model the claims through the fact description of the case in order to take advantage of the improvement of explicit claims.

Ethics Statement

The dataset constructed in this paper is from China Judgements Online[3], which is an official legal documents website. The names of all participants in the dataset are anonymized before being published online. And there are already lots of datasets constructed from this website used in Chinese law-related research. We do not perform analysis at the user level rather than the evidence level, which is less intrusive for specific people. Finally, This technology mainly plays an auxiliary role to provide a reference for judges rather than play a decisive role.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *EMNLP/IJCNLP*.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Yen-Liang Chen, Yi-Hung Liu, and Wu-Liang Ho. 2013. A text mining approach to assist the general public in the retrieval of legal documents. *Journal of the American Society for Information Science and Technology*, 64(2):280–290.

William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dominik Maria Endres and Johannes E Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Prakash Poudyal, Teresa Gonçalves, and Paulo Quaresma. 2018. Using clustering techniques to identify arguments in legal documents.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. 2017. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 88:22–31.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864.

[3] http:/wenshu.court.gov.cn