Detecting Agent Mentions in U.S. Court Decisions

Jaromír Šavelka
Kevin D. Ashley

Intelligent Systems Program
University of Pittsburgh

jas438@pitt.edu

April 6, 2018

This work was supported in part by the National Institute of Justice Graduate Student Fellowship (Fellow: Jaromir Savelka) Award # 2016-R2-CX-0010, "Recommendation System for Statutory Interpretation in Cybercrime."

This work is about automatic detection of agent mentions in case law.

It is important because the capability is foundational for many applications.

We show that it is possible to detect the mentions automatically using simple sequence labeling model.

We explore relatedness of the task when performed on different domains (areas of legal regulation) showing that ...

- there are differences between distinct domains;
- but there are also similarities which enable utilization of knowledge across domains.

Overview

Motivation

Task Definition

Hypotheses

Data Set

Experiments

Discussion

Motivation

Case law analysis comprises (i) identification of relevant decisions and (ii) extraction of valuable information.

It has been argued that directly retrieving argument-related information (AR) would be extremely valuable.[1,2]

There is still a considerable gap between the state-of-the-art and a full-blown AR system.[3]

Even the most foundational NLP technology performs poorly when applied to legal texts . . .

. . . and detection of agent mentions is one such technology.

[1] Ashley and Walker 2013; [2] Grabmair et al. 2015; [3] Ashley 2017

Overview

Motivation

Task Definition

Hypotheses

Data Set

Experiments

Discussion

Task Definition

Detecting agent mentions amounts to recognizing when a word or a phrase denotes an agent.

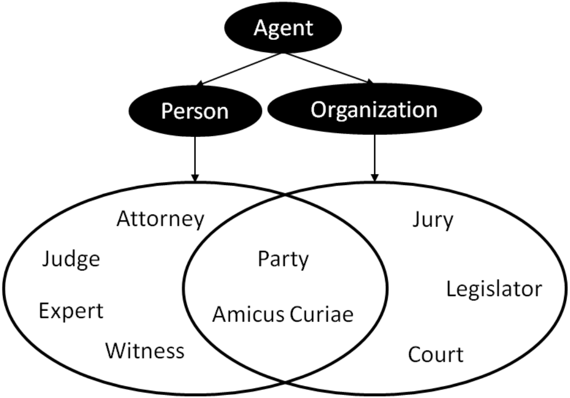
A typical court decision contains many mentions of agents.

> The magistrate judge denied the second motion to compel because Mavrix failed to notify the anonymous parties of the pending motion. Mavrix moved the district court for review of the magistrate judge's order, which the district court denied on the basis of the moderators' First Amendment right to anonymous internet speech.

Even words such as possessive adjectives (e.g., his, their) could be considered agent mentions.

Task Definition

We defined a light-weight and easilly extensible hierarchy of agents:



Task of detecting agent mentions: (1) find all text spans denoting agents; and (2) classify each span with most appropriate type.

Overview

Motivation

Task Definition

Hypotheses

Data Set

Experiments

Discussion

Hypotheses

hypothesis i A simple sequence labeling model that uses low-level textual features could learn to detect agent mentions automatically.

hypothesiis ii When a model is trained on decisions from one area of legal regulation and applied to texts from other area the performance decreases.

hypothesiiis iii Using texts from other domains to support the model trained on the decisions from the target domain may lead to better performance.

Overview

Motivation

Task Definition

Hypotheses

Data Set

Experiments

Discussion

Data Set: Decisions

We downloaded 10 court decisions from publicly available online sources.[1]

(a) 5 cyber crime (cyber bullying, credit card frauds, possession of electronic child pornography)

(b) 5 intellectual property (copyright, trade marks, patents)

		sum	longest	average	shortest
cyber-crime	# of chars	199980	61703	39996.0	28306
	# of tokens	71100	20881	14220.0	10414
	# of sentences	1772	513	354.4	250
intellectual-property	# of chars	247042	75625	49408.4	36823
	# of tokens	90286	27915	18057.2	13144
	# of sentences	2084	729	416.8	291

[1] www.courtlistener.com; scholar.google.com.

Data Set: Agent Mentions

We created annotation guidelines for manual annotation for the 12 types.[1]

The two guiding principles were full-coverage and maximum specificity.

	AGT	PER	ORG	ATT	JDG	EXP	WTN	PTY	AMC	JUR	LEG	CRT
cyber-crime												
seq	146	612	236	72	96	14	195	1352	0	82	17	334
seq/doc	29	122	47	14	19	3	39	270	0	16	3	67
intel-prop												
seq	241	661	433	76	115	37	34	1668	35	81	16	451
seq/doc	48	132	87	15	23	7	7	334	7	16	3	90
total												
seq	387	1273	669	148	211	51	229	3020	35	163	33	785
seq/doc	39	127	67	15	21	5	23	302	5	16	3	78

[1] luima.org; jas438@pitt.edu

Data Set: Inter-annotator Agreement

Three IP decisions were processed by two annotators to measure inter-annotator agreement (IA).

Two variants of IA measure:

(a) full agreement – ratio of the annotations created by the both users over the all annotations (agree in type and span)

(b) partial agreement – two annotations agree if they are of the same type and if they overlap by at least one character

	AGT	PER	ORG	ATT	JDG	EXP	WTN	PTY	AMC	JUR	LEG	CRT
full	.74	.53	.59	.63	.80	.00	.00	.81	.63	.00	.48	.71
partial	.87	.64	.74	.67	.84	.00	.00	.90	.71	.89	.48	.81

Overview

Motivation

Task Definition

Hypotheses

Data Set

Experiments

Discussion

Experiments: Design

We performed three experiments:

i. same domain – we assess the possibility of detecting agent mentions automatically for each of the domains separately.
ii. different domain – we applied models trained on one area of law to the texts from other domain.
iii. combined domains – we used data from both of the domains.

In all experiments we train a separate CRF model for each type.

Features – e.g. a token in lowercase, token's length, position within document, is-digit, is-whitespace

Labels – BILOU scheme.

Experiments: Evaluation

We use traditional IR metrics:

$$P = \frac{|Pred \cap Gold|}{|Pred|} \quad R = \frac{|Pred \cap Gold|}{|Gold|} \quad F = \frac{2*P*R}{P+R}$$

In the same domain experiment we use the leave one out cross-validation.

In the different domain experiment we evaluate the models on the documents from the different area of law.

In the combined domains experiment the data from both domains were pulled together (leave one out).

Experiments: Results

		AGT	PER	ORG	ATT	JDG	EXP	WTN	PTY	AMC	JUR	LEG	CRT
Same domain													
exact	P	.74	.65	.79	.67	.47	.00	.56	.73	.17	.87	.50	.81
	R	.36	.17	.39	.25	.16	.00	.04	.36	.03	.56	.06	.69
	F_1	.48	.27	.52	.37	.23	.00	.08	.48	.05	.68	.11	.75
overlap	P	.83	.73	.85	.73	.72	.00	.61	.84	.50	.91	1.0	.87
	R	.40	.19	.42	.27	.24	.00	.05	.41	.09	.59	.12	.74
	F_1	.54	.31	.57	.39	.36	.00	.09	.55	.15	.72	.22	.80
Different domain													
exact	P	.67	.48	.70	.59	.46	.00	.00	.63	.00	.85	.27	.80
	R	.28	.09	.39	.18	.20	.00	.00	.23	.00	.63	.09	.68
	F_1	.39	.16	.49	.27	.28	.00	.00	.33	.00	.73	.14	.73
overlap	P	.76	.58	.75	.64	.64	.00	.00	.74	.00	.90	.55	.85
	R	.31	.11	.42	.19	.28	.00	.00	.27	.00	.66	.18	.72
	F_1	.44	.19	.54	.29	.39	.00	.00	.39	.00	.77	.27	.78
Combined domains													
exact	P	.70	.66	.73	.68	.52	.00	.52	.69	.22	.88	.45	.79
	R	.37	.23	.43	.35	.26	.00	.06	.34	.06	.69	.15	.72
	F_1	.48	.34	.54	.46	.34	.00	.11	.46	.09	.77	.23	.76
overlap	P	.79	.74	.78	.72	.73	.00	.52	.80	.44	.92	.64	.85
	R	.41	.25	.46	.37	.36	.00	.06	.39	.11	.72	.21	.78
	F_1	.54	.38	.58	.49	.48	.00	.11	.53	.18	.81	.32	.81

Overview

Motivation

Task Definition

Hypotheses

Data Set

Experiments

Discussion

Discussion

The results clearly show that simple CRF models using low-level textual features are capable of detecting the agent mentions.

For some types (Jury, Court) the performance appears to be sufficient for actual use.

For some other types (Expert, Witness, Legislator) the performance is clearly too low to produce useful results.

For the remaining types it is not clear if the results would have the potential to be useful in practice (may depend on an application).

Discussion II

For each domain there may be certain agent mentions that are rare or non-existent in other domains.

In cyber crime decisions one of the prosecuting parties was often mentioned as “the government.”

Certain patterns in mentioning agents transfer across domains.

The Court type mentions transfer well since even the models trained on the different domain retained good performance.

“We” is universally being used to mention the deciding majority.

Future Work

The models trained in our experiments are quite simplistic (especially in terms of features they use).

Simple textual features do not provide sufficient information to detect certain mentions and to distinguish among different types.

The models struggled to distinguish mentions of the Amicus Curiae type from mentions of the Party and the Organization types.

Most of the times Amici could be detected by a rule-based system in the header.

Using the detected tokens as contextual features could raise the performance of our models from very bad to excellent.

Future Work II

Corpus extension (there is a clear data sparsity problem for some of the types)

Extension of the agent hierarchy (we use only a handful of very basic types)

More powerful prediction model (e.g., long short-term memory networks)

Co-reference resolution

Use in an actual application (e.g., attribution resolution for improving IR)

Conclusion

We examined the possibility of automatically detecting agent mentions in case law analysis.

We have shown that:

i. it is possible to detect mentions of different agent types automatically.
ii. the task is domain dependent in a sense that models trained on one area of law do not perform as well for a different area.
iii. there is relatedness between domains allowing the use of data from different areas of law.

Thank you!

Questions, comments and suggestions are welcome now or any time at jas438@pitt.edu.

References I

Al-Kofahi, Khalid, Brian Grom, and Peter Jackson. "Anaphora resolution in the extraction of treatment history language from court opinions by partial parsing." *Proceedings of the 7th international conference on Artificial intelligence and law*. ACM, 1999.

Ashley, Kevin D. *Artificial Intelligence and Legal Analytics*. Cambridge University Press, 2017.

Ashley, Kevin D. *Modeling legal arguments: Reasoning with cases and hypotheticals*. MIT press, 1991.

Ashley, Kevin D., and Vern R. Walker. "From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study." *JURIX*. 2013.

Ashley, Kevin D., and Vern R. Walker. "Toward constructing evidence-based legal arguments using legal decision documents and machine learning." *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*. ACM, 2013.

Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. "Named entity recognition and resolution in legal text." Semantic Processing of Legal Texts. Springer Berlin Heidelberg, 2010. 27–43.

References II

Faiz, Syeed Ibn, and Robert Mercer. "Extracting higher order relations from biomedical text." *Proceedings of the First Workshop on Argumentation Mining*. 2014.

Francesconi, Enrico. "An Approach to Legal Rules Modelling and Automatic Learning." *JURIX*. 2009.

Grabmair, Matthias, Ashley, K. D., Chen, R., Sureshkumar, P., Wang, C., Nyberg, E., and Walker, V. R. "Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools." *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM, 2015.

Grabmair, M., Ashley, K. D., Hwa, R., and Sweeney, P. M. "Toward Extracting Information from Public Health Statutes using Text Classification Machine Learning." *JURIX*. 2011.

John Lafferty, Andrew McCallum, Fernando Pereira, and others. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning*, ICML, Vol. 1. 282–289.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields. (2007).

References III

- Peters, Wim, and Adam Z. Wyner. “Legal Text Interpretation: Identifying Hohfeldian Relations from Text.” *LREC*. 2016.
- Quaresma, Paulo, and Teresa Goncalves. “Using linguistic information and machine learning techniques to identify entities from juridical documents.” *Semantic Processing of Legal Texts*. Springer Berlin Heidelberg, 2010. 44–59.
- Savelka, Jaromir, Matthias Grabmair, and Kevin D. Ashley. “Mining Information from Statutory Texts in Multi-Jurisdictional Settings.” *JURIX*. 2014.
- Walker, Vern R. “The Need for Annotated Corpora from Legal Documents, and for (Human) Protocols for Creating Them: The Attribution Problem.” *Dagstuhl Seminar on Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments*, 2016.
- Walker, Vern R., Parisa Bagheri, and Andrew J. Lauria. “Argumentation Mining from Judicial Decisions: The Attribution Problem and the Need for Legal Discourse Models.” 2015.