

# Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions

Siddharth Patwardhan and Ellen Riloff

## Introduction

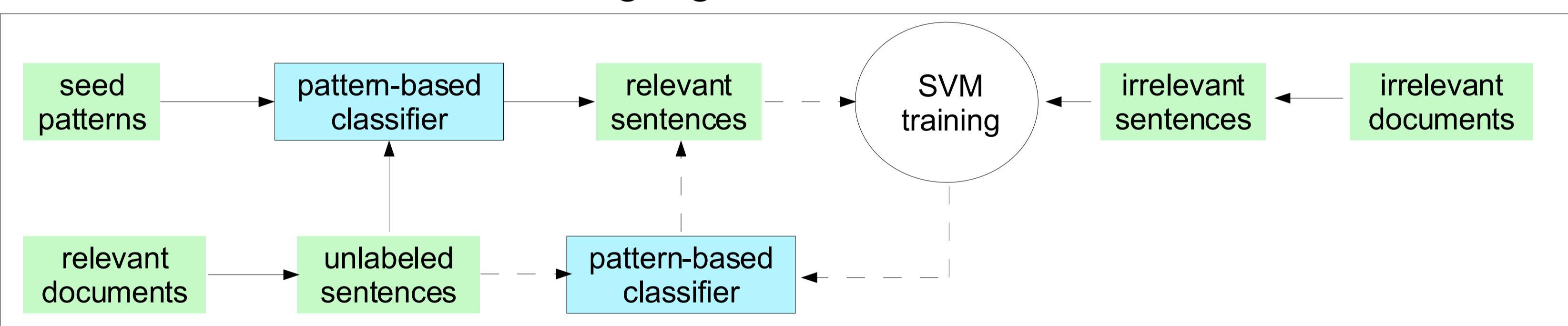
- Information Extraction (IE) is the task of extracting information about specific events from free text.
- For example, an ideal IE system could assist in locating terrorist event in news articles, and output records of the form:
 

```
{Victim="3 travelers", Weapon="bomb", PerpOrg="Al Qaeda", Location...}
```
- Typically, IE systems try to learn rules or patterns for extracting information relying on local context surrounding the relevant information.
 

```
<subj> was assassinated, <subj> exploded, was kidnapped by <np>
```
- In this work, we decouple the task of relevant region identification and locating potential extractions. This has the following benefits:
  - allows the use of more general semantically appropriate patterns.
  - prevents the use of good patterns in obviously irrelevant contexts.
  - simplifies the learning process.
- We use a self-trained sentence classifier along with a *semantic affinity* pattern learner, to generate a completely automated IE system.
- Our automated system matches the performance of an existing system that benefits from human oversight.

## Self-Trained Sentence Classifier

- For training, takes as input seed extraction patterns, relevant documents and irrelevant documents.
- Uses an iterative self-training algorithm to train an SVM.



**Iteration 0:** Initiates the iterative process with a simple seed pattern based classifier. Sentences in relevant documents with seeds in them are relevant.

**Iteration n:** Uses the relevant and irrelevant sentences generated at iteration  $n-1$  to train an SVM. Sentences from the relevant documents classified as relevant by the SVM are added to the relevant sentence set.

Equal number of irrelevant sentences used for training at each iteration.

## Automatic Pattern Selection

- In general, for patterns to work reliably they must meet two constraints:
  - pattern matches text only in event descriptions.
  - text extracted by the pattern pertains to the event.
- By using the relevant region classifier, the first constraint is met, and the patterns need only ensure that the extracted text is potentially relevant.
- This can be done by using *semantic affinity* to make sure that the extractions are semantically acceptable.

**Semantic Affinity:** Is a measure of the tendency of a pattern to extract phrases of a particular semantic category ( $\text{sem\_aff}(p, c_n)$ ):

$$\text{sem\_aff}(p, c_n) = \frac{f(p, c_n)}{\sum_{i=1}^{|C|} f(p, c_i)} - \log_2 f(p, c_n)$$

where  $C$  is the set of semantic categories  $\{c_1, c_2, \dots, c_{|C|}\}$ .

### Automated IE System:

- Our automated IE system uses AutoSlog-TS to generate an exhaustive set of extraction patterns from relevant and irrelevant documents.
- The semantic affinity of each of the patterns to each of the semantic categories is computed using the training data and a semantic dictionary.
- The top patterns ranked by semantic affinity with each category are picked.
- Relevant sentence classifier applied to test data, and the selected patterns are applied only to the relevant-classified sentences.

## Primary vs. Secondary Patterns

- So far, we are applying all patterns only to relevant sentences.
- In doing so we ignore patterns that can in and of themselves indicate the relevance of text regions (for e.g. `<subj> was assassinated`).
- Primary patterns:** Patterns with a conditional probability of occurring in a relevant document is greater than or equal to 0.8.
- Secondary patterns:** Patterns with probability between 0.5 and 0.8.
- Some of the top ranked patterns by *semantic affinity*:

Target	Victim	Weapon	PerpOrg	Disease
destroyed <dobj>	murder of <np>	<subj> exploded	<subj> claimed	cases of <np>
barrels of <np>	assassination of <np>	planted <dobj>	command of <np>	spread of <np>
shattered <dobj>	killing of <np>	fired <dobj>	panama from <np>	outbreak of <np>
<subj> was damaged	<subj> question	<subj> was planted	wing of <np>	$n^{\text{th}}$ outbreak
blew up <dobj>	murdered <dobj>	explosion of <np>	kidnapped by <np>	$n$ outbreaks

## Data Sets

- This work was evaluated on **two** domains.
- The terrorism data set consists of 1700 news stories about Latin American terrorism. We used 1300 for training, 200 for tuning, 200 for testing.
- Focused on 5 string slots: *victim*, *target*, *weapon*, *perpOrg*, and *perpInd*.
- The disease outbreak data consists of ProMed messages about disease outbreaks. We used 6000 training, 125 tuning, and 120 testing documents.
- We focused on the *disease* and *victim* slots.

## Evaluation

Self-training iterations of classifier evaluated on tuning set:

	Relevant				Irrelevant		
	Acc	Rec	Pr	F	Rec	Pr	F
<b>Terrorism Domain</b>							
It #0	0.87	0.26	0.52	0.35	0.96	0.90	0.93
It #1	0.86	0.48	0.48	0.48	0.92	0.92	0.92
It #2	0.84	0.63	0.43	0.51	0.88	0.94	0.91
It #3	0.82	0.71	0.42	0.51	0.88	0.94	0.91
<b>Disease Outbreak Domain</b>							
It #0	0.88	0.04	0.25	0.07	0.98	0.89	0.93
It #1	0.86	0.26	0.33	0.29	0.93	0.91	0.92
It #2	0.86	0.26	0.34	0.29	0.93	0.91	0.92
It #3	0.86	0.26	0.34	0.29	0.93	0.91	0.92

Relevant Sentence-based IE system on the terrorism domain:

Patterns	App	PerpInd			PerpOrg			Target			Victim			Weapon		
		Rec	Pr	F	Rec	Pr	F	Rec	Pr	F	Rec	Pr	F	Rec	Pr	F
ASlogTS	All	0.49	0.35	0.41	0.33	0.49	0.40	0.64	0.42	0.51	0.52	0.48	0.50	0.45	0.39	0.42
ASlogTS	Rel	0.42	0.52	0.46	0.26	0.59	0.36	0.59	0.50	0.54	0.50	0.56	0.53	0.40	0.50	0.44
SA-50	All	0.24	0.29	0.26	0.20	0.42	0.27	0.42	0.43	0.42	0.41	0.43	0.42	0.53	0.46	0.40
SA-50	Rel	0.20	0.34	0.25	0.18	0.60	0.28	0.39	0.48	0.43	0.39	0.55	0.46	0.41	0.55	0.47
SA-50	Sel	0.21	0.35	0.36	0.18	0.52	0.27	0.42	0.50	0.45	0.40	0.54	0.46	0.43	0.53	0.48
SA-100	All	0.40	0.30	0.34	0.30	0.43	0.35	0.56	0.38	0.45	0.45	0.37	0.41	0.55	0.43	0.48
SA-100	Rel	0.35	0.39	0.37	0.24	0.55	0.33	0.52	0.45	0.48	0.42	0.49	0.45	0.45	0.50	0.47
SA-100	Sel	0.38	0.40	0.39	0.24	0.50	0.32	0.56	0.45	0.50	0.43	0.49	0.45	0.47	0.49	0.48
SA-150	All	0.50	0.27	0.35	0.34	0.39	0.37	0.62	0.30	0.40	0.50	0.33	0.40	0.55	0.39	0.45
SA-150	Rel	0.45	0.38	0.42	0.27	0.55	0.36	0.56	0.37	0.45	0.46	0.46	0.46	0.45	0.48	0.46
SA-150	Sel	0.48	0.39	0.43	0.28	0.51	0.37	0.60	0.37	0.45	0.47	0.45	0.46	0.47	0.47	0.47
SA-200	All	0.73	0.08	0.15	0.42	0.43	0.42	0.64	0.29	0.40	0.54	0.32	0.40	0.64	0.17	0.27
SA-200	Rel	0.68	0.15	0.24	0.32	0.58	0.41	0.58	0.36	0.44	0.49	0.44	0.47	0.52	0.28	0.36
SA-200	Sel	0.71	0.12	0.21	0.33	0.55	0.41	0.61	0.35	0.45	0.50	0.44	0.47	0.53	0.22	0.31

Relevant Sentence-based IE system on the disease outbreak domain:

Patterns	App	Disease			Victim		
		Rec	Pre	F	Rec	Pre	F
ASlogTS	All	0.51	0.27	0.36	0.48	0.35	0.41
ASlogTS	Rel	0.46	0.31	0.37	0.44	0.38	0.41
SA-50	All	0.51	0.25	0.34	0.47	0.41	0.44
SA-50	Rel	0.49	0.31	0.38	0.44	0.43	0.43
SA-50	Sel	0.50	0.29	0.36	0.46	0.41	0.44
SA-100	All	0.57	0.22	0.32	0.52	0.33	0.40
SA-100	Rel	0.55	0.28	0.37	0.49	0.36	0.41
SA-100	Sel	0.56	0.26	0.35	0.51	0.34	0.41
SA-150	All	0.66	0.20	0.31	0.55	0.27	0.37
SA-150	Rel	0.61	0.26	0.36	0.51	0.31	0.38
SA-150	Sel	0.63	0.24	0.35	0.53	0.29	0.37
SA-200	All	0.68	0.19	0.30	0.56	0.26	0.36
SA-200	Rel	0.63	0.25	0.35	0.52	0.30	0.38
SA-200	Sel	0.65	0.23	0.34	0.54	0.28	0.37