A Dataset for Open Event Extraction in English

Kiem-Hieu Nguyen[1], Xavier Tannier[2], Olivier Ferret[3], Romaric Besançon[3]

1. Hanoi Univ. of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam (hieunk@soict.hust.edu.vn)

2. LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France (xavier.tannier@limsi.fr)

3. CEA, LIST, Vision and Content Engineering Laboratory, F-91191, Gif-sur-Yvette, France (firstname.surname@cea.fr)
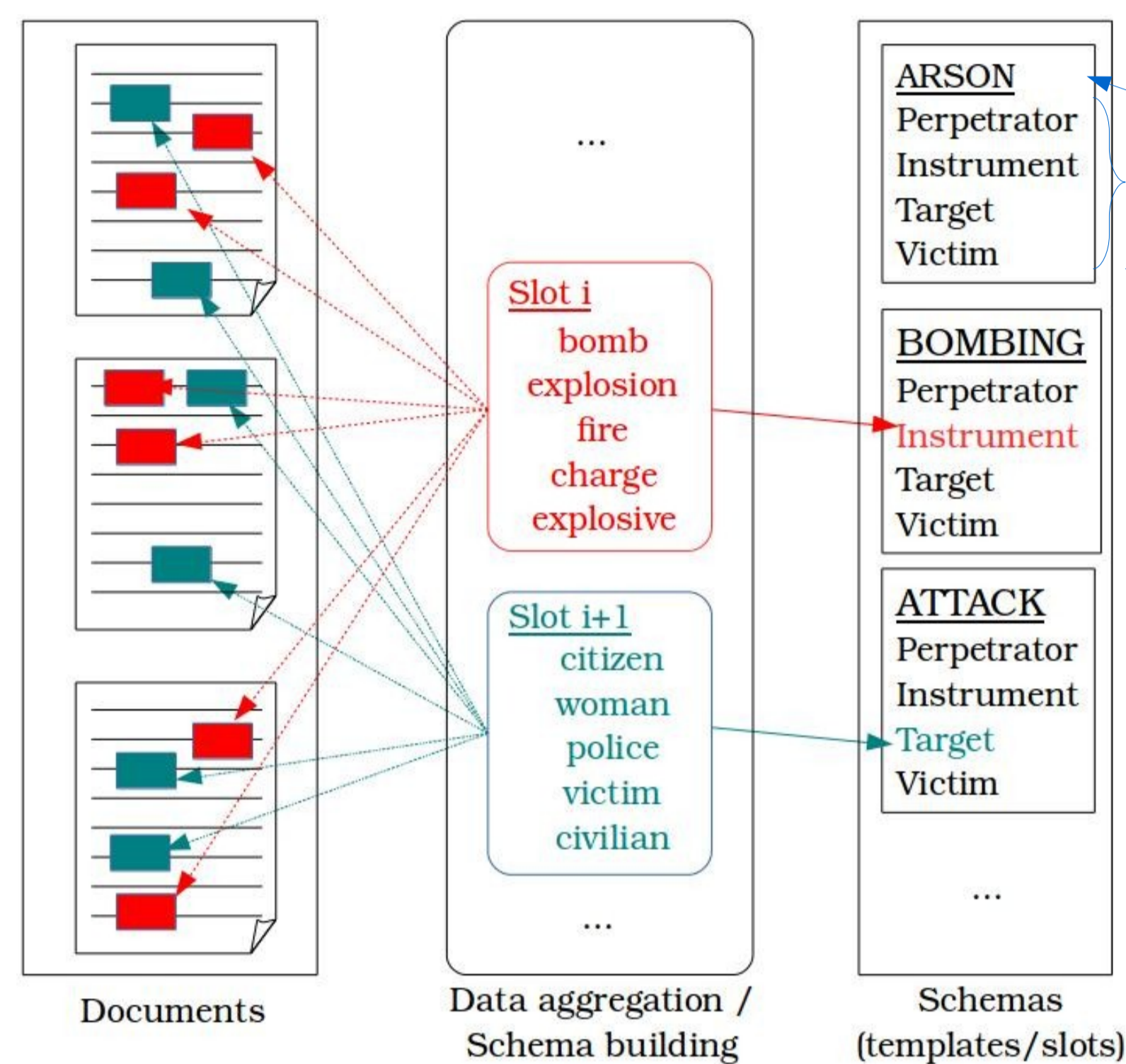
Open Event Extraction

Event extraction ≡ template filling = assigning event roles to individual textual mentions

Schema induction = learning templates with no supervision from unlabeled texts

We focus here more specifically on event schema induction

Documents — Data aggregation / Schema building — Schemas (templates/slots)

State-of-the art evaluation: MUC-4 Corpus

A significant part of the work in the field of event schema induction from texts relies on the MUC-4 corpus for its evaluation

Templates
Slots

	Bombing	Kidnap	Attack	Arson
Perpetrator	X	X	X	X
Victim	X	X	X	X
Target	X		X	X
Instrument	X		X	

Slots in the hand-crafted MUC-4 templates (from Chambers & Jurafsky, 2011)

1,700 news articles about terrorist incidents happening in Latin America

Limits of MUC-4 corpus

1. Same roles for all templates (overcome by ACE 2005, TAC KBP)

2. Small corpus without information redundancy

ASTRE Corpus

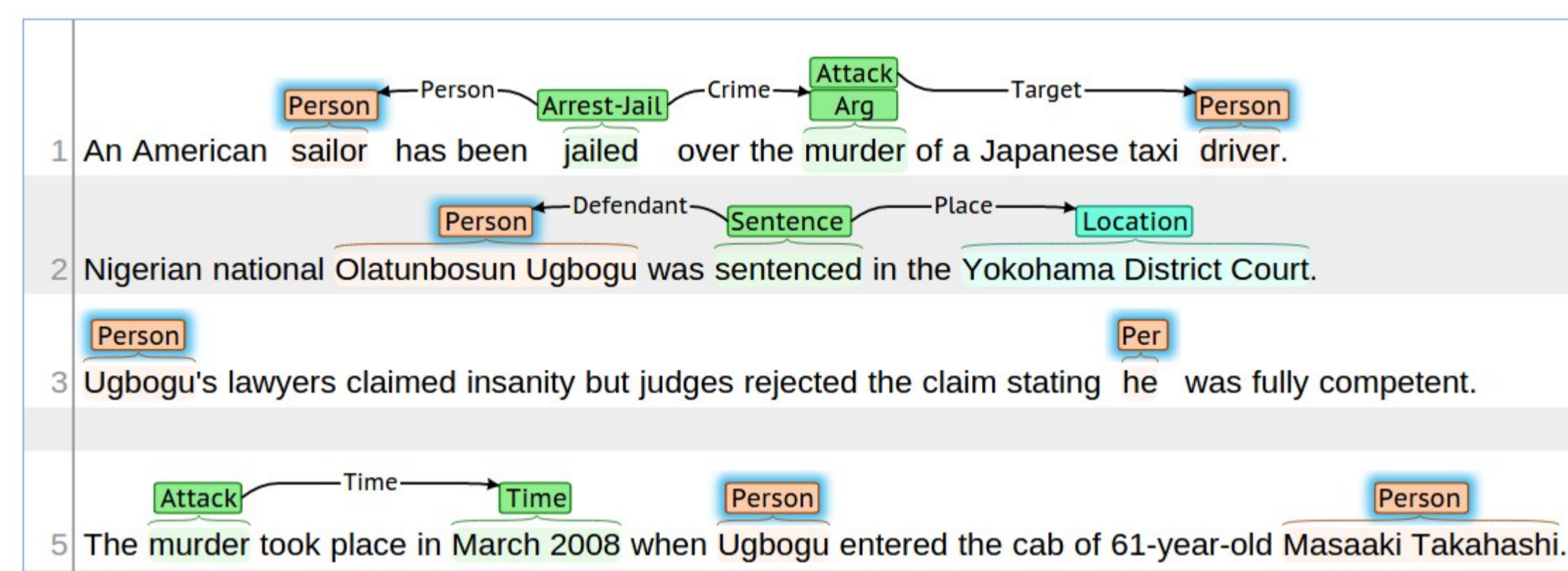ASTRE Corpus has the following characteristics:

- Redundancy, i.e. it contains several documents about the same event
- Partial annotation.
  - annotated data for evaluation purpose
  - larger amount of unannotated data for inducing event schemas
- A larger variety of templates / MUC-4
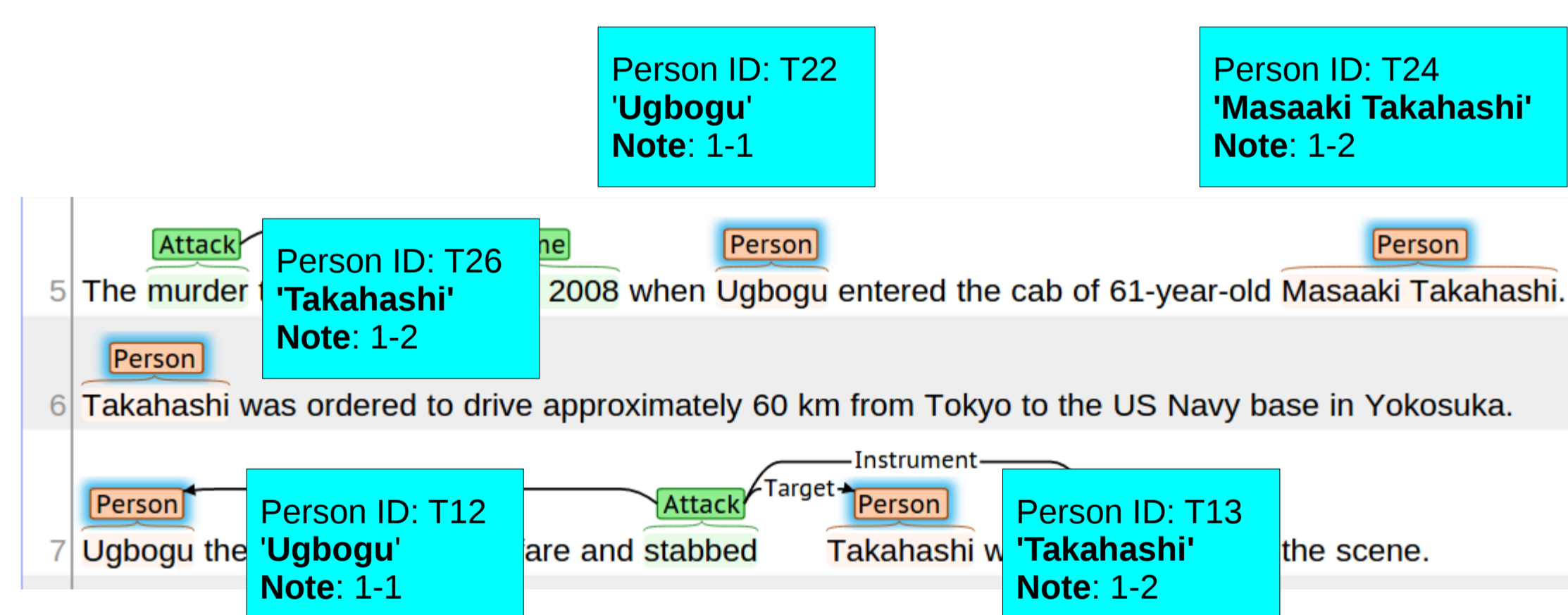
1. Document Annotation

Documents: 100 Wikinews articles from category "Law & Justice"

Templates: a subset of TAC-KBP events

LIFE.{Injure, Die},

CONFLICT.Attack,

JUSTICE.{Charge-Indict, Arrest-Jail, Release-Parole,

Sentence, Convict, Appeal, Acquit, Execute, Extradite}

2. Relevant Document Retrieval

1. Submit the document title to the Google search engine

2. Keep only documents with creation time around the event date

3. Corpus Building

1. Clean documents with Boilerpipe

2. Remove duplicates with SpotSigs and mcl

3. Clean semi-automatically remaining texts

A sample of annotations

An example of entity coreference annotations

#docs	#sentences	#words	#tokens
1,038	42.6 K	969.5 K	1.19 M

Corpus statistics
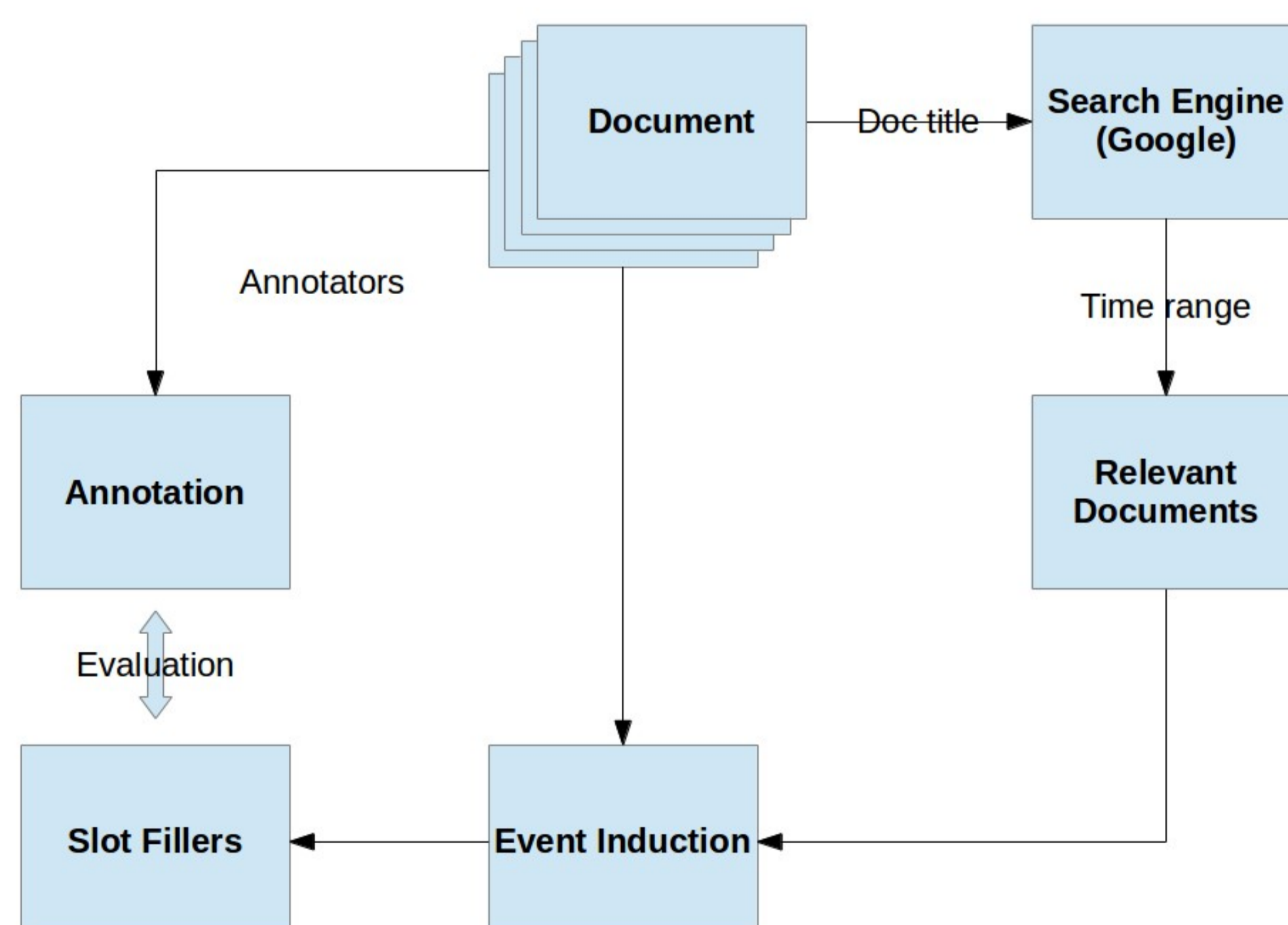
Evaluation

- Unannotated documents retrieved from the Web were used for model learning

- Manually annotated data were used as development and test datasets

System	MUC-4			ASTRE dev			ASTRE test		
	P	R	F	P	R	F	P	R	F
Chambers 2013	.41	.41	.41	.33	.34	.34	.15	.28	.19
Nguyen et al. 2015	.36	.54	.43	.41	.30	.35	.21	.26	.23