Désambiguïsation d'Entités pour l'Induction Non Supervisée de Schémas Événementiels

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

(1) LIMSI-CNRS
(2) CEA LIST
(3) Univ. Paris-Sud

TALN-RÉCITAL 2015 – Juin, Caen, France

What we do?

- Unsupervised event schema induction

- Unsupervised event schema induction
- Slot filling

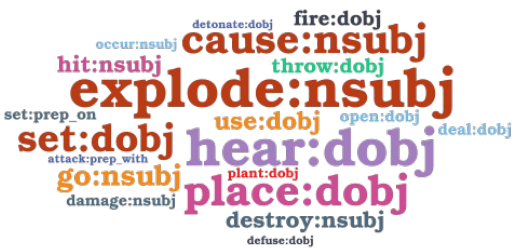
The group was traveling in a-4-wheel drive vehicle between Cucuta and the rural area known as Campanario when their vehicle was **blown up** by four **explosive charges** that **exploded** on the highway.

A UCR district headquarters in Buenos Aires province was "completely **destroyed** by a **bomb explosion**." The destroyed UCR headquaterss is in the Moreno district of Buenos Aires.

Alleged guerrilla urban commandos **launched** two **highpower bombs** against a car dealership in downtown San Salvador this morning. A police report said that the attack set the building of fire, but did not result in any casualties although economic losses are heavy.

The terrorist attack was launched from a moving vehicle, from which unidentified individuals **threw** three **dynamite charges** that **injured** a passerby and **damaged** the front of the building, according to the police.

Slot 1



BOMBING_instrument

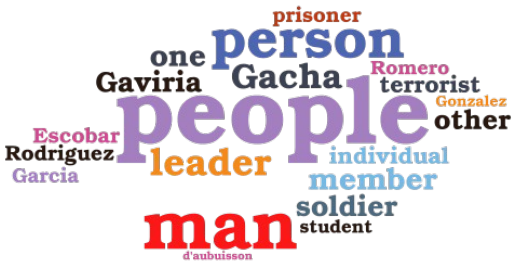
About 50 **peasants** of various ages have been **kidnapped** by FMLN in San Miguel department. According to that garrison, the mass kidnapping took place on 30 December in San Luis de la Reina.

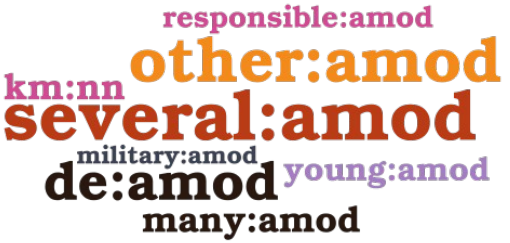
Ricardo Alfonso Castellar, mayor of Achi, who was **kidnapped** on 5 January, apparently by ELN guerrillas, was found dead today.

The MNR reports the **disappearance** and **kidnapping** of MNR assistant secretary general **Hector Oqueli Colindres** in Guatemala city today, 12 January.

The Salvadoran government today deplored the **disappearance** of Social Democratic leader **Hector Oqueli Colindres** this morning in Guatemala.

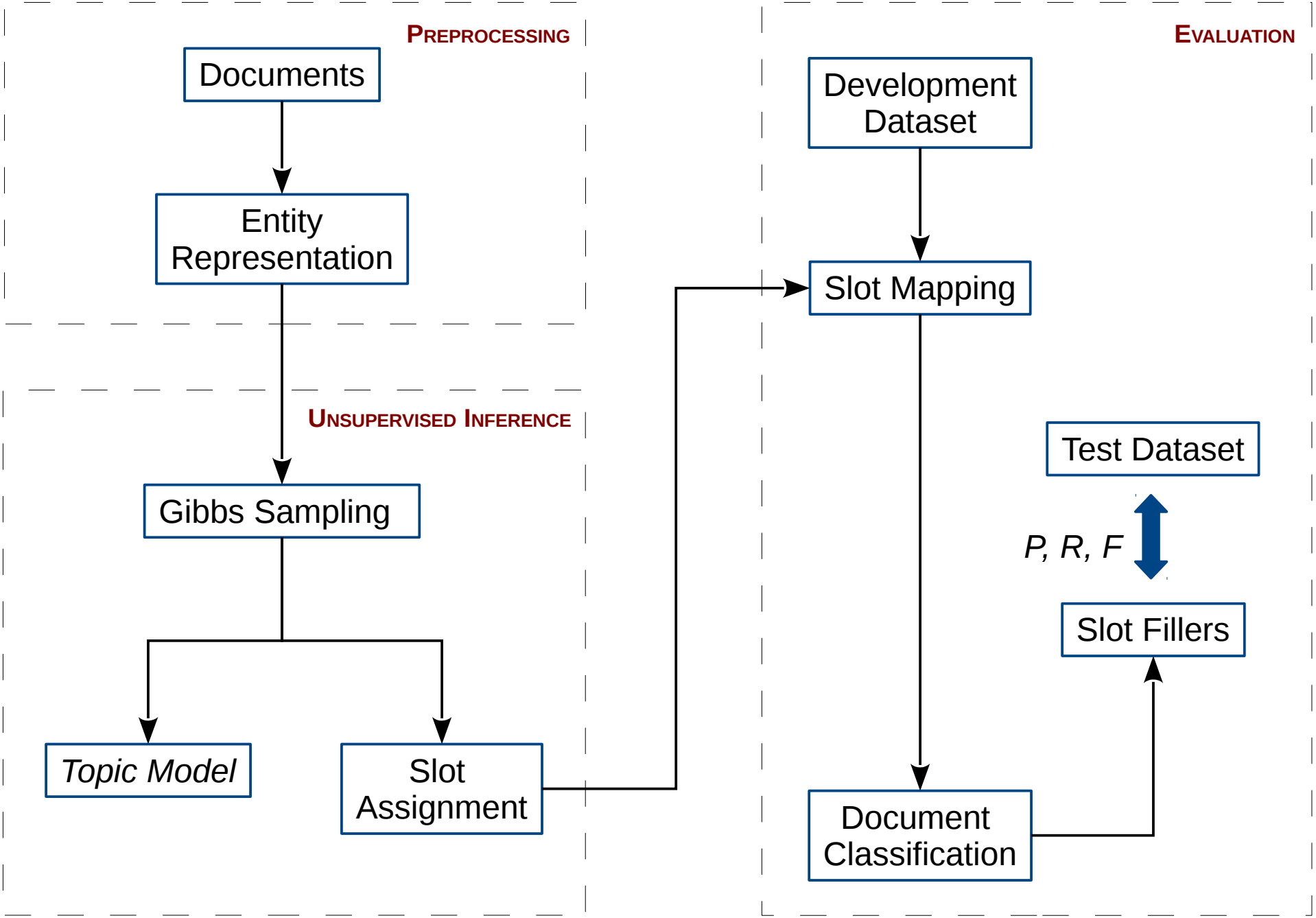
Slot 2

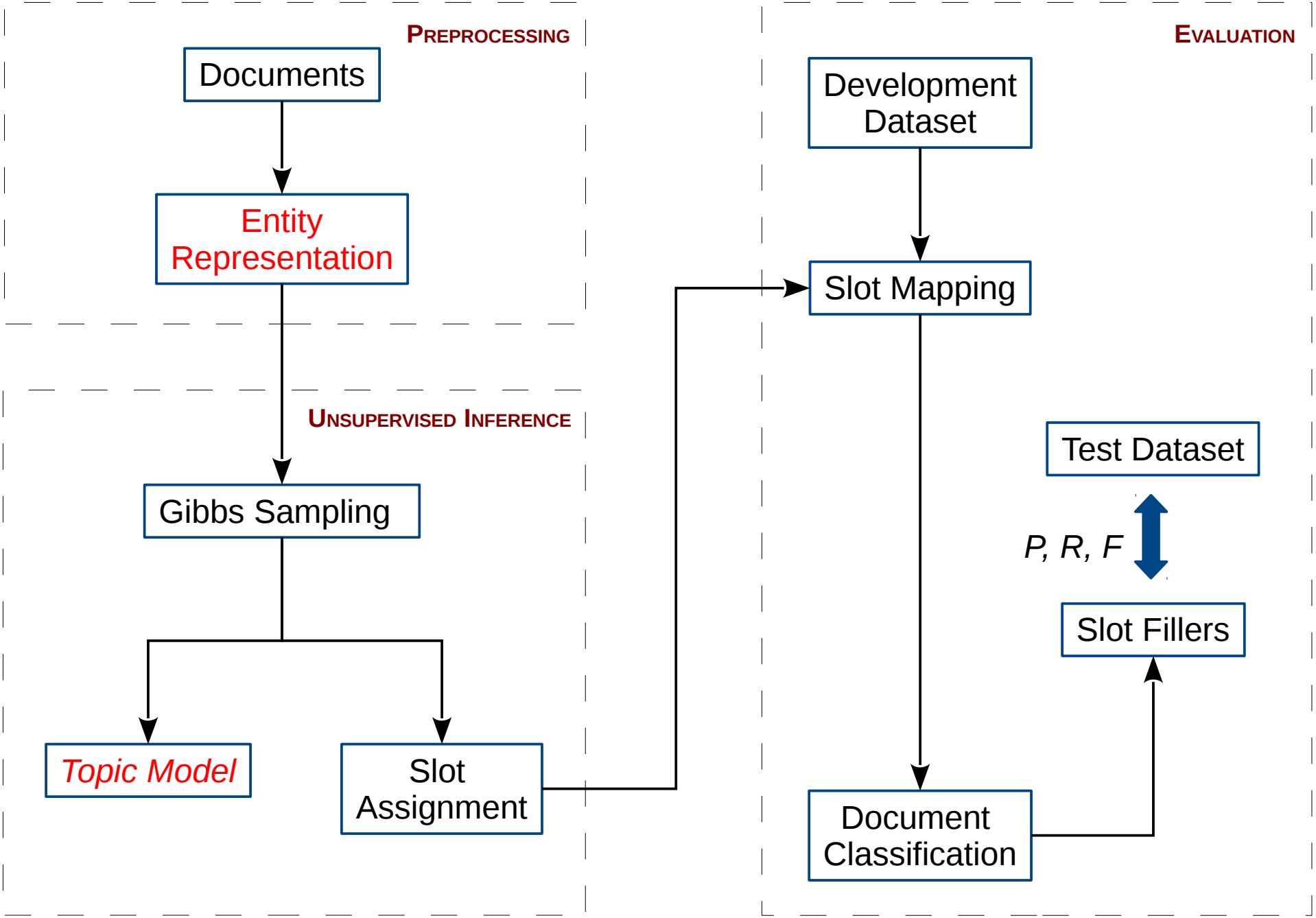


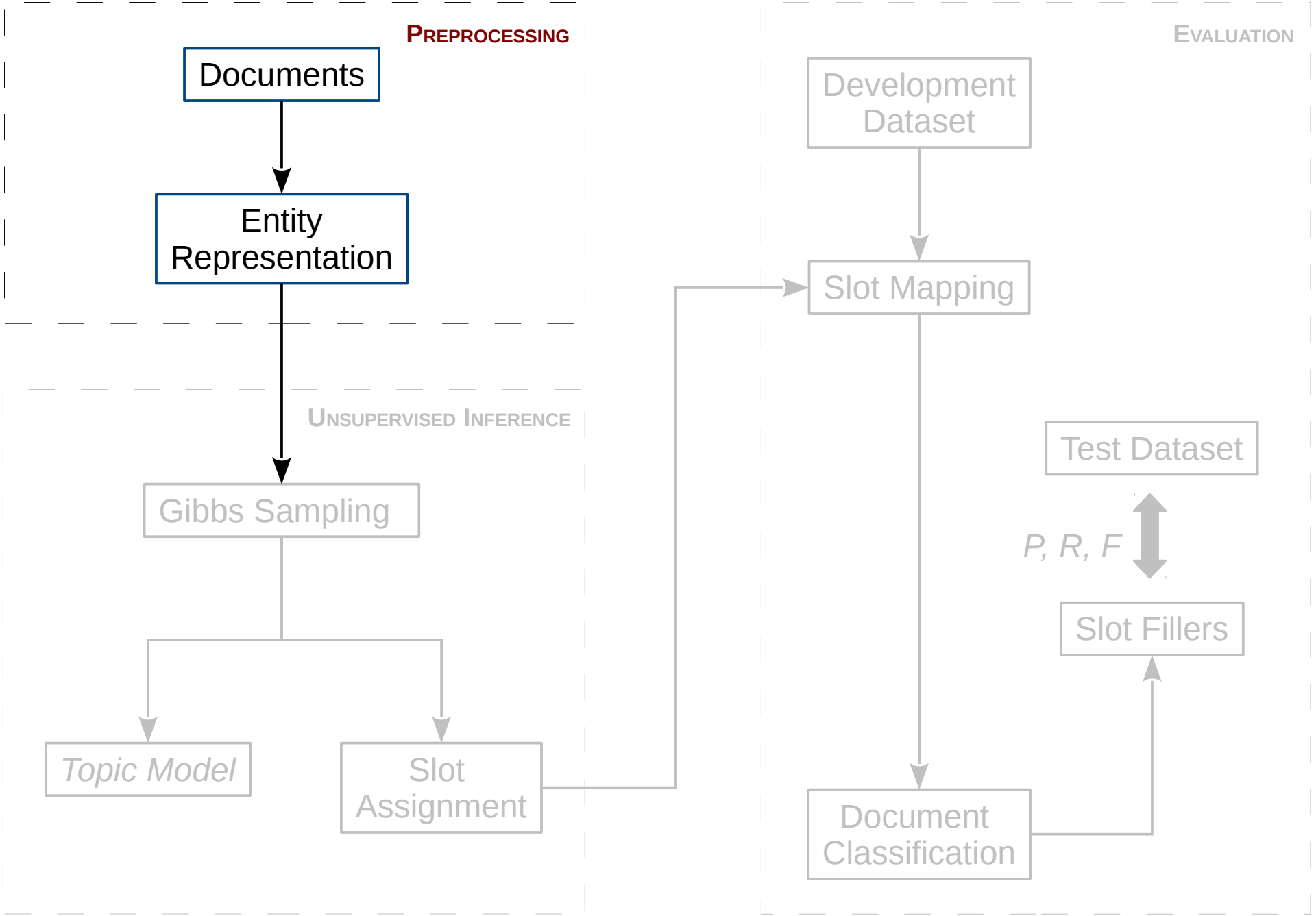


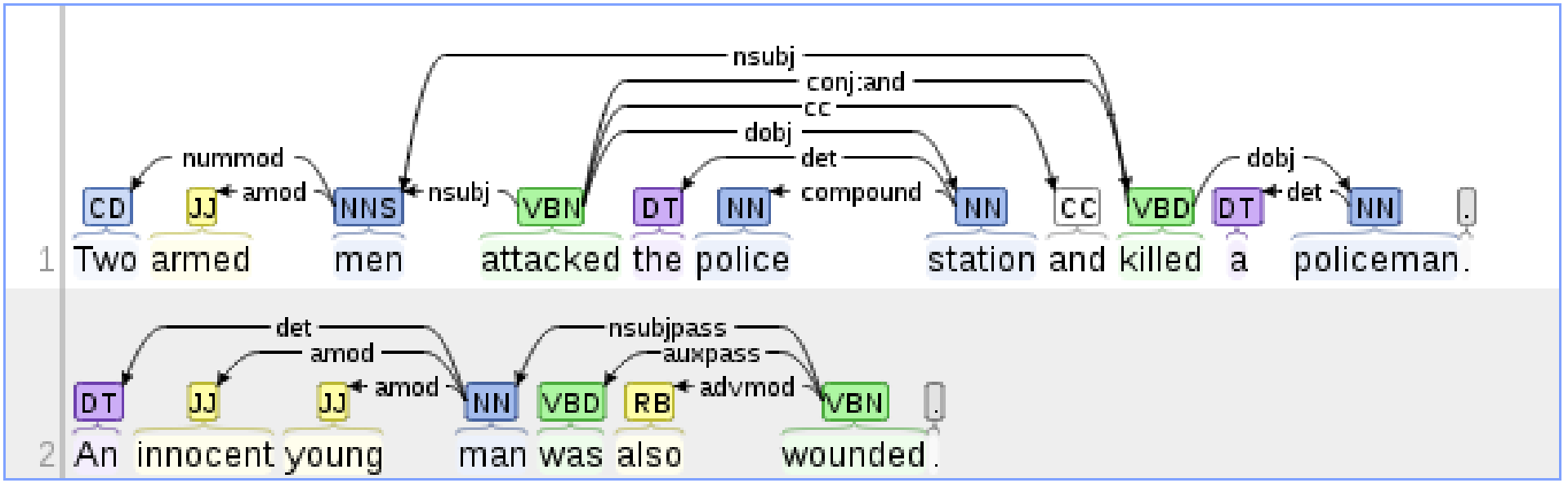
KIDNAPPING_victim

How we do it ?
(what new ?)

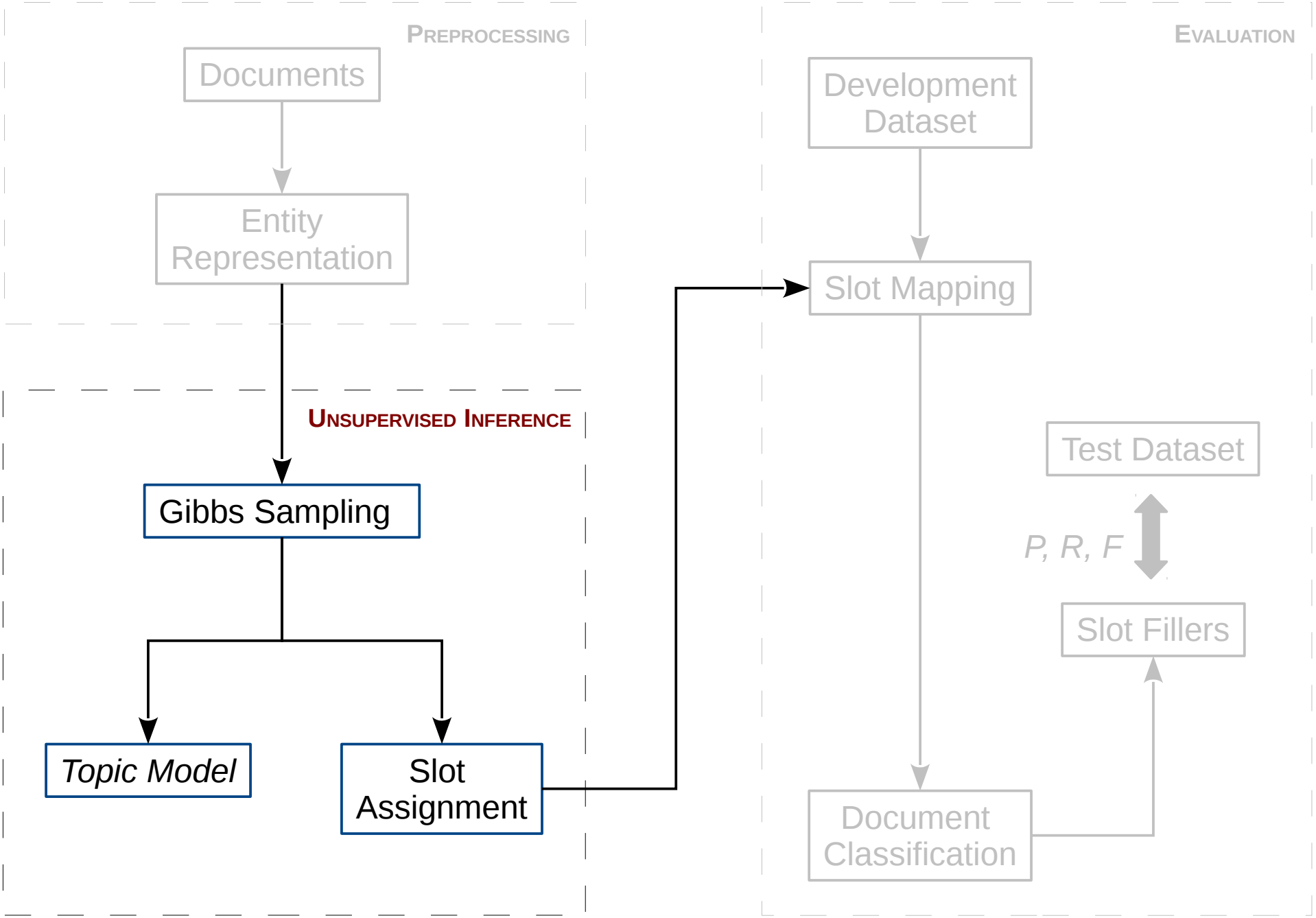

PREPROCESSING
Documents
Entity Representation
UNSUPERVISED INFERENCE
Gibbs Sampling
Topic Model
Slot Assignment
EVALUATION
Development Dataset
Slot Mapping
Test Dataset
P, R, F
Slot Fillers
Document Classification

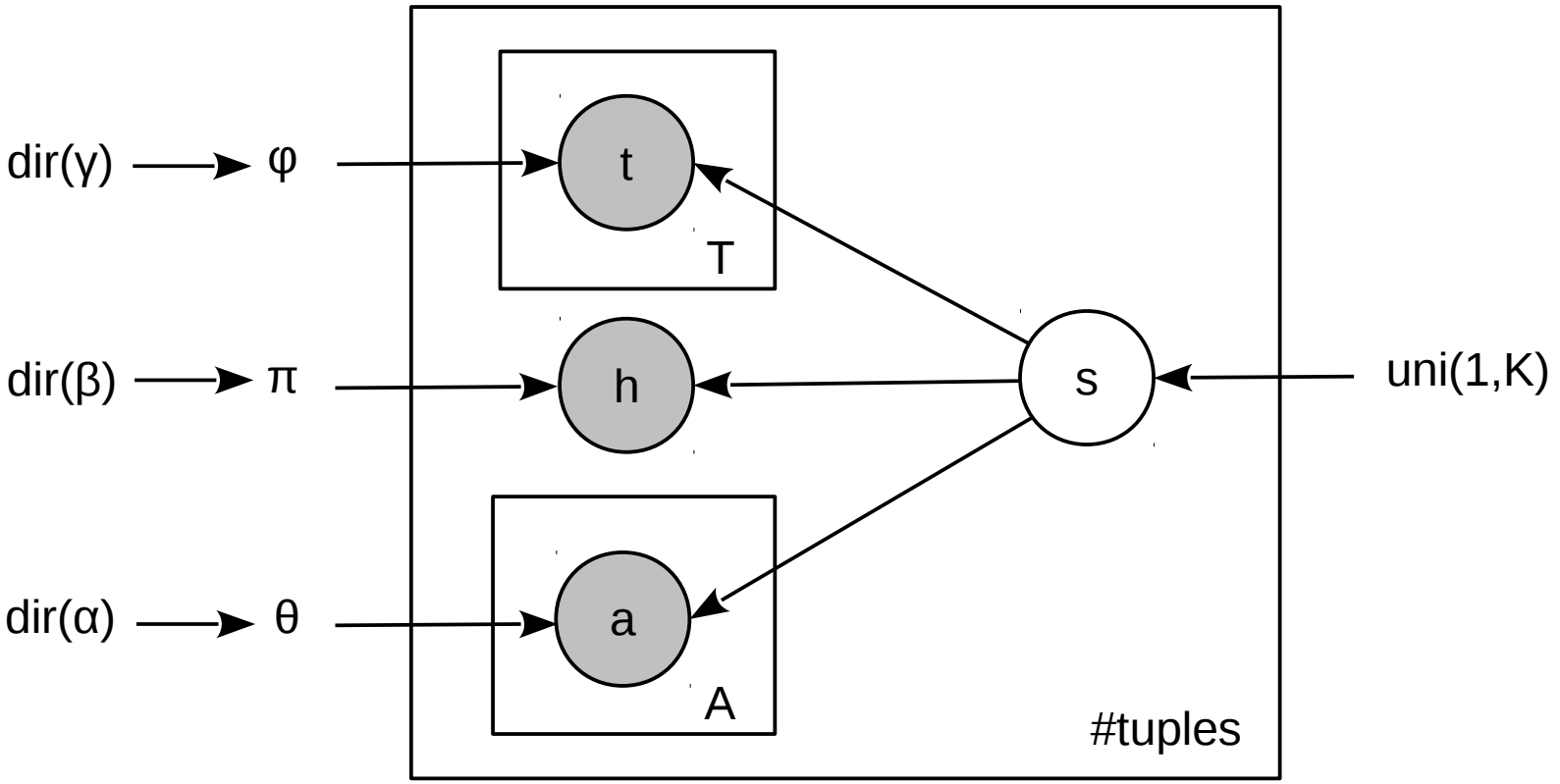

PREPROCESSING
Documents
Entity Representation
UNSUPERVISED INFERENCE
Gibbs Sampling
Topic Model
Slot Assignment
EVALUATION
Development Dataset
Slot Mapping
Test Dataset
P, R, F
Slot Fillers
Document Classification


PREPROCESSING
Documents
Entity Representation
UNSUPERVISED INFERENCE
Gibbs Sampling
Topic Model
Slot Assignment
EVALUATION
Development Dataset
Slot Mapping
Test Dataset
P, R, F
Slot Fillers
Document Classification



	Attributes	Head	Triggers
#1	[armed:amod]	man	[attack:nsubj, kill:nsubj]
#2	[police:nn]	station	[attack:dobj]
#3	[]	policeman	[kill:dobj]
#4	[innocent:amod, young:amod]	man	[wound:dobj]


Preprocessing
Documents
Entity Representation
Unsupervised Inference
Gibbs Sampling
Topic Model
Slot Assignment
Evaluation
Development Dataset
Slot Mapping
Test Dataset
P, R, F
Slot Fillers
Document Classification


dir(γ)
φ
t
T
dir(β)
π
h
s
uni(1,K)
dir(α)
θ
a
A
#tuples

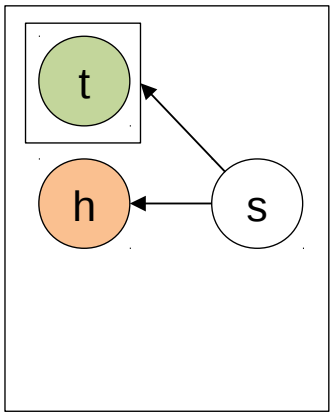
Corpus probability:

$$P_{\pi,\phi,\theta}(E) = \prod_{e \in E} P_{\pi,\phi,\theta}(e)$$

Entity probability:

$$\begin{aligned} P_{\pi,\phi,\theta}(e) &= P(s) \\ &\times P(h|s) \\ &\times \prod_{t \in T_e} P(t|s) \\ &\times \prod_{a \in A_e} P(a|s) \end{aligned}$$

Why ambiguity matters?
& Our resolution



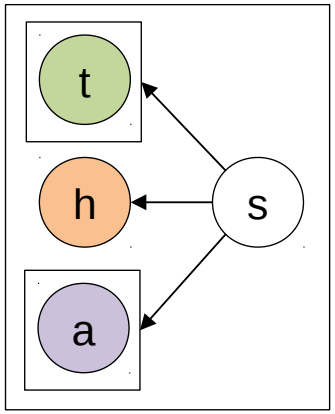
Slot: **ATTACK_victim**

Head *pr(h|s)*	Triggers *pr(t|s)*
citizen	*kill:dobj*
woman	*murder:dobj*
man	*die:nsubj*
police	*die:prep_of*
terrorist	*wound:dobj*

El Comercio reported that alleged Shining Path members also attacked public facilities in Huarpacha yesterday.

Municipal official **Sergio Horna** was seriously **wounded**.

Two extremist **terrorists** were reported **killed** by national officers.



Slot: **ATTACK_victim**

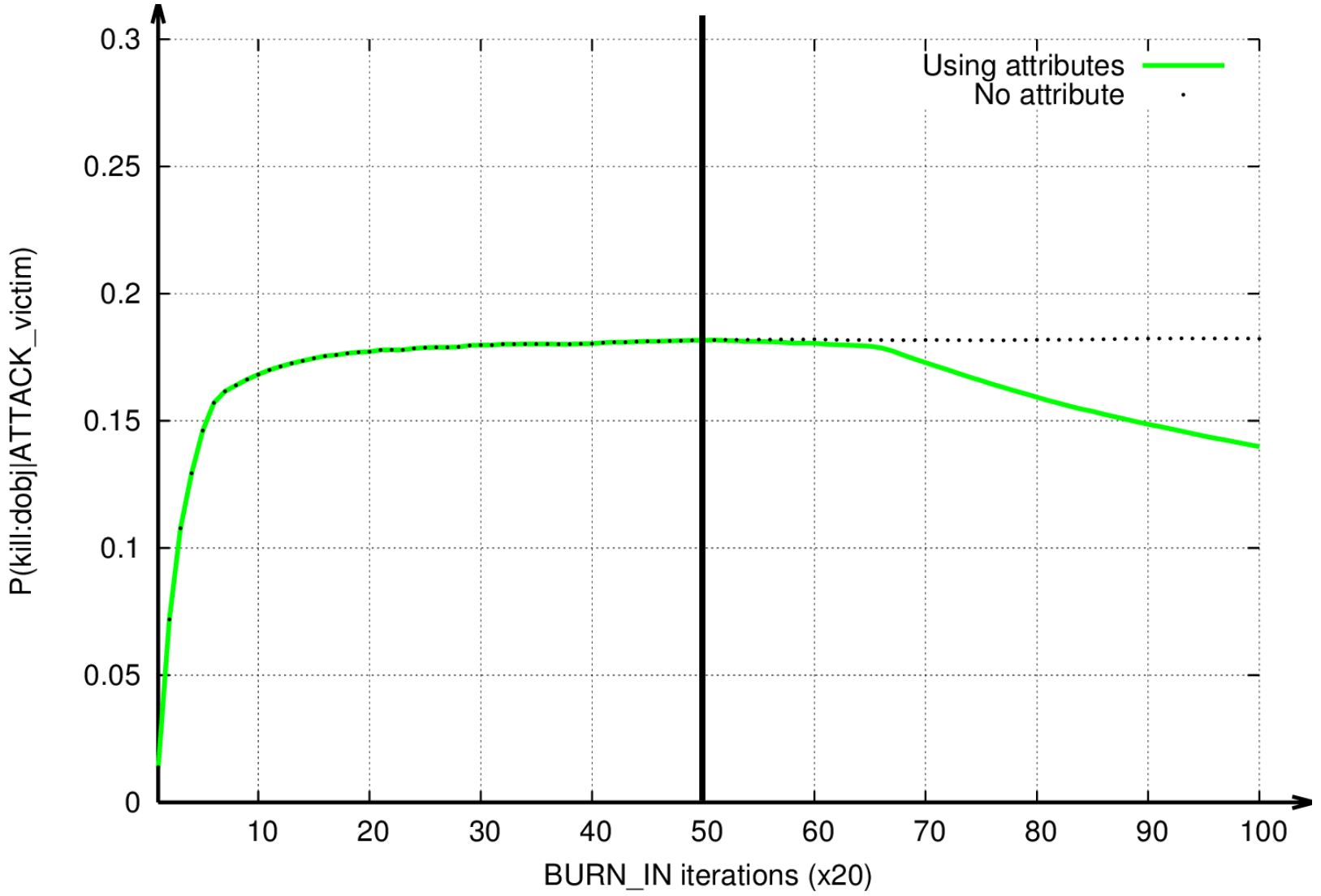
Attributes pr(a|s)	Head *pr(h|s)*	Triggers *pr(t|s)*
innocent:amod	*citizen*	*murder:dobj*
wounded:amod	*woman*	*kill:dobj*
U.N:nn	*police*	*die:nsubj*
young:amod	*victim*	*die:prep_of*
official:nn	*civilian*	*assassinate:dobj*

El Comercio reported that alleged Shining Path members also attacked public facilities in Huarpacha yesterday.

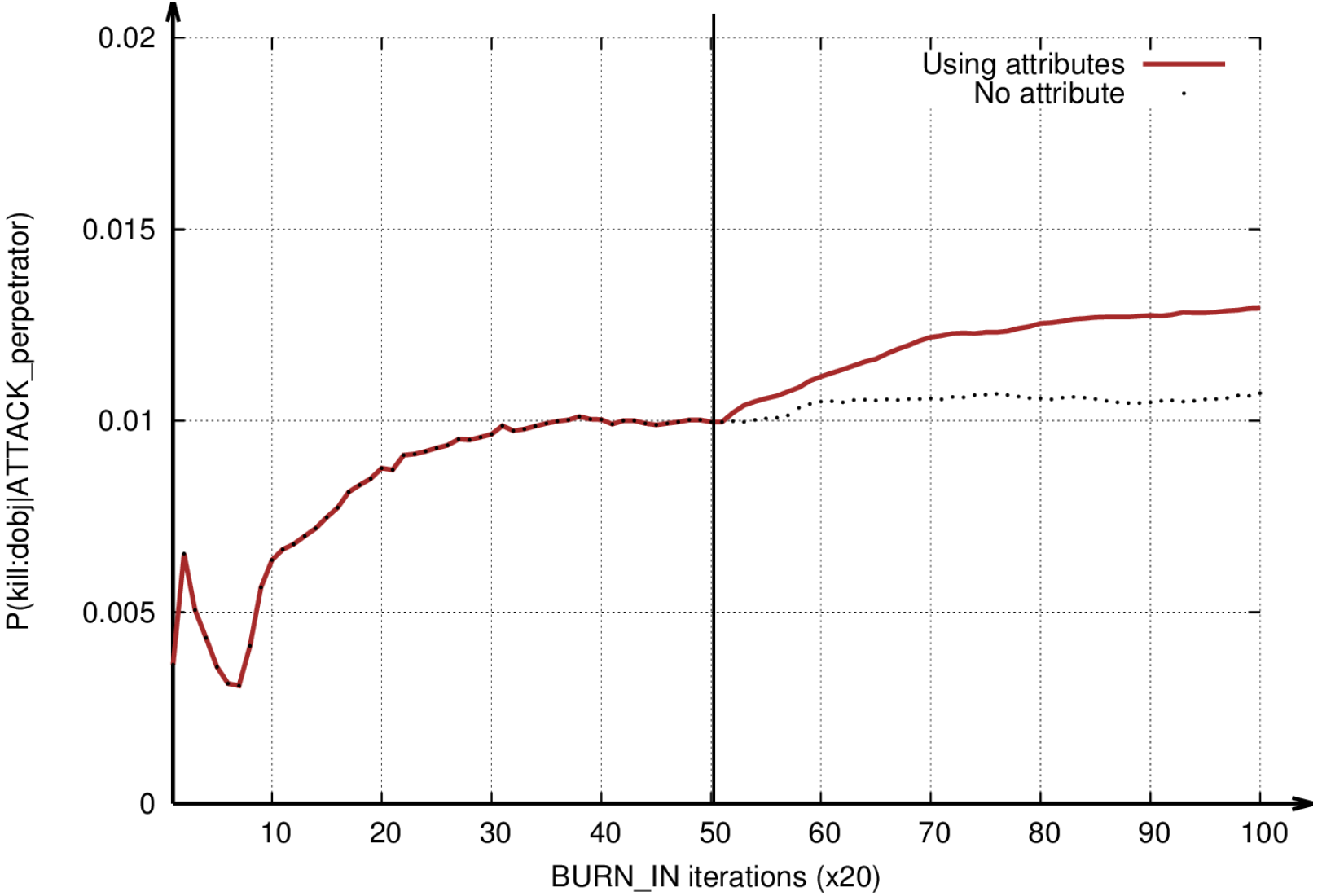
Municipal **official Sergio Horna** was seriously **wounded**.

Two **extremist terrorists** were reported **killed** by national officers.

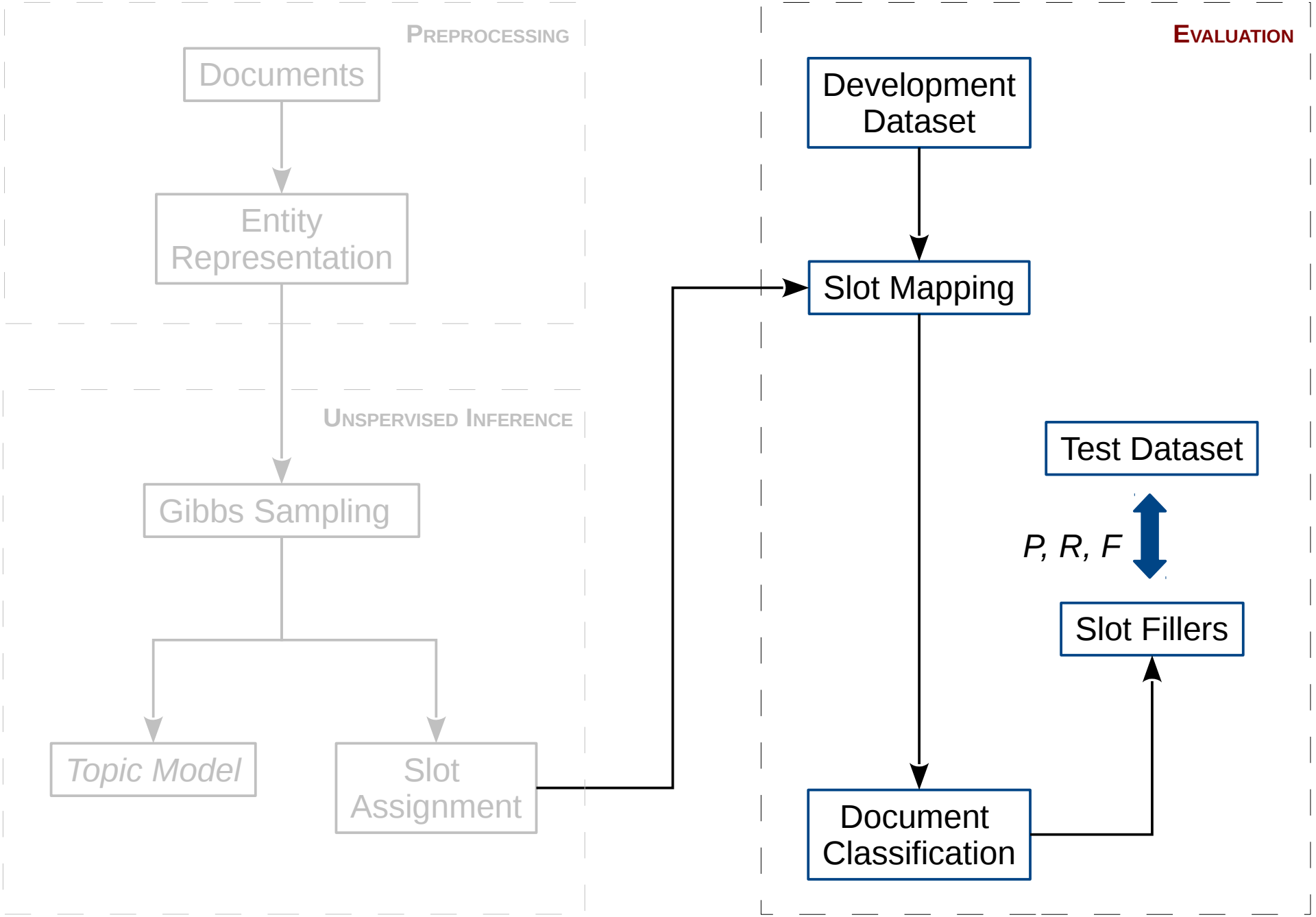
p(kill:dobj) in slot ATTACK_victim



p(kill:dobj) in slot ATTACK_perpetrator



Evaluation

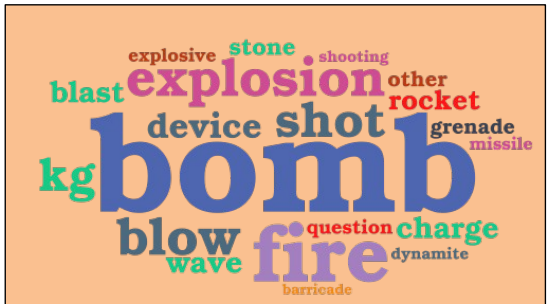
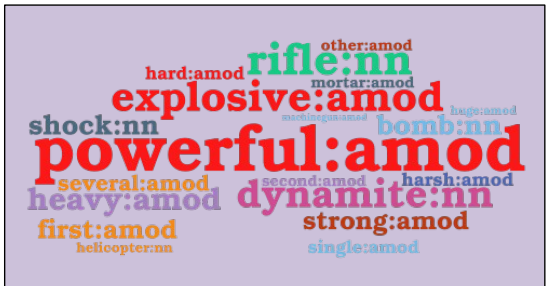
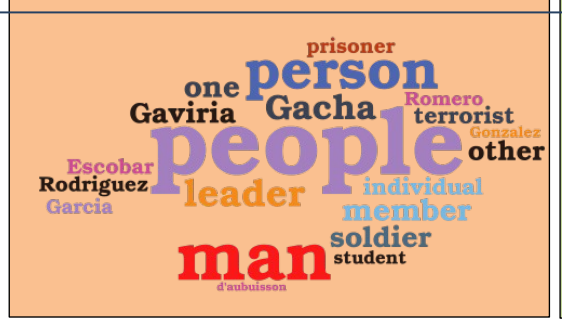
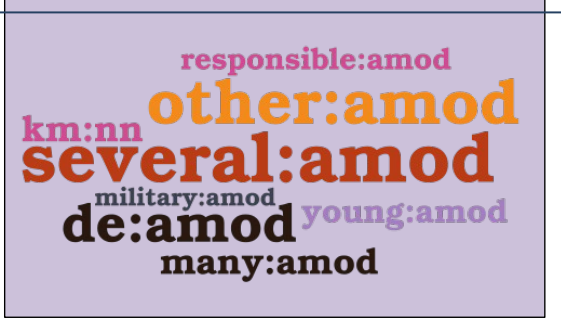

Preprocessing
Documents
Entity Representation
Unspervised Inference
Gibbs Sampling
Topic Model
Slot Assignment
Evaluation
Development Dataset
Slot Mapping
Test Dataset
P, R, F
Slot Fillers
Document Classification

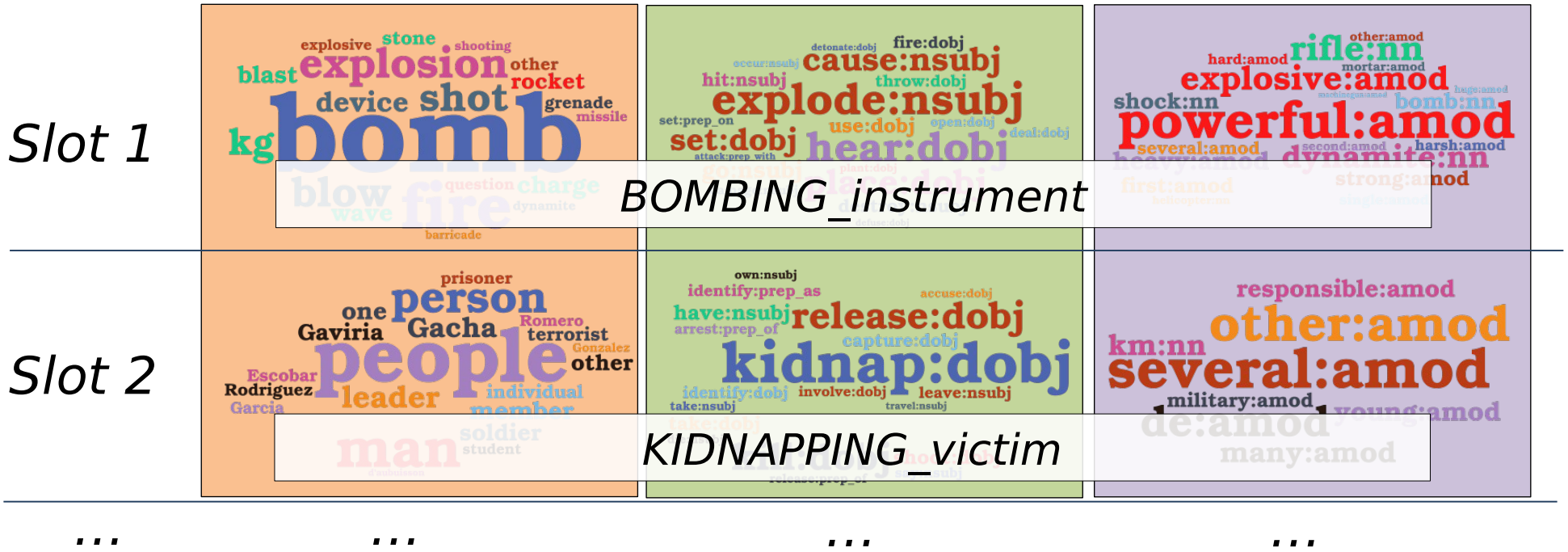
- MUC-4 corpus
  - Terrorist events
  - 1,700 docs: 1,300 train / 200 dev (tst1-2)/ 200 test (tst3-4)
  - Uppercase restoration, linguistic analysis using Stanford NLP toolkit

- Slot filling
  - Precision, recall, f-score
  - 'Right most word' phrase matching
  - Slots of different templates are exchangeable

Slot mapping

- To map learned slots to reference slots
- For each reference slot, map it to the best performed learned slots
- Set of reference slots + development reference dataset + scoring function

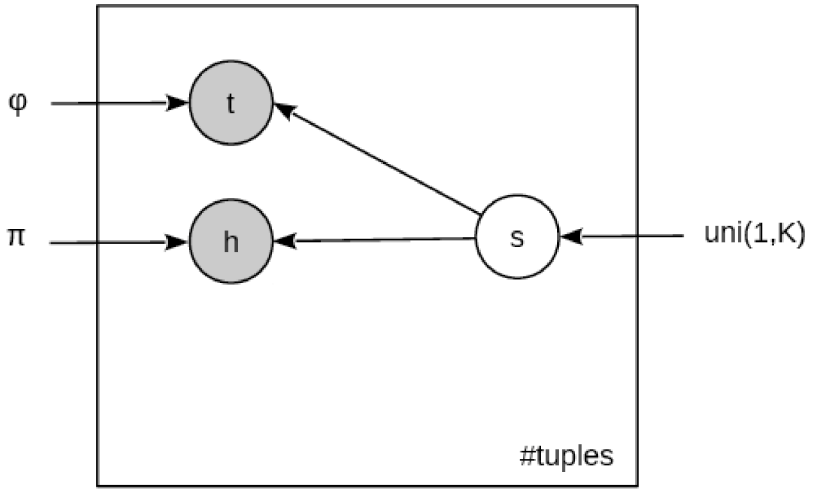
Slot 1	 explosive stone shooting blast explosion other rocket device shot grenade missile kg bomb blow fire question charge wave dynamite barricade	 detonate:dobj fire:dobj occur:nsubj cause:nsubj hit:nsubj throw:dobj explode:nsubj set:prep_on use:dobj open:dobj deal:dobj set:dobj hear:dobj attack:prep_with plant:dobj go:nsubj place:dobj damage:nsubj destroy:nsubj defuse:dobj	 other:amod rifle:nn hard:amod mortar:amod explosive:amod shock:nn bomb:nn powerful:amod several:amod second:amod harsh:amod heavy:amod dynamite:nn strong:amod first:amod helicopter:nn single:amod
Slot 2	 prisoner person one Romero Gaviria Gacha terrorist Gonzalez people other Escobar Rodriguez leader individual Garcia member soldier man student d'aubuisson	 own:nsubj identify:prep_as accuse:dobj have:nsubj release:dobj arrest:prep_of capture:dobj kidnap:dobj identify:dobj involve:dobj leave:nsubj take:nsubj travel:nsubj take:dobj die:nsubj kill:dobj shoot:dobj release:prep_of say:nsubj	 responsible:amod other:amod km:nn several:amod military:amod de:amod young:amod many:amod
...	...	...	...	


Slot 1
explosive
stone
shooting
blast
explosion
other
rocket
device
shot
grenade
missile
kg
bomb
blow
wave
fire
question
charge
dynamite
barricade
detonate:dobj
fire:dobj
occur:nsubj
cause:nsubj
hit:nsubj
throw:dobj
explode:nsubj
set:prep_on
use:dobj
open:dobj
set:dobj
attack:prep_with
hear:dobj
deal:dobj
other:amod
hard:amod
rifle:nn
mortar:amod
explosive:amod
shock:nn
bomb:nn
powerful:amod
several:amod
second:amod
harsh:amod
dynamite:nn
strong:amod
first:amod
BOMBING_instrument
Slot 2
prisoner
person
one
Gaviria
Gacha
Romero
terrorist
Gonzalez
other
Escobar
Rodriguez
people
leader
individual
Garcia
member
man
soldier
student
own:nsubj
identify:prep_as
accuse:dobj
have:nsubj
release:dobj
arrest:prep_of
capture:dobj
kidnap:dobj
identify:dobj
involve:dobj
leave:nsubj
take:nsubj
travel:nsubj
responsible:amod
other:amod
km:nn
several:amod
military:amod
young:amod
many:amod
KIDNAPPING_victim
...
...
...
...

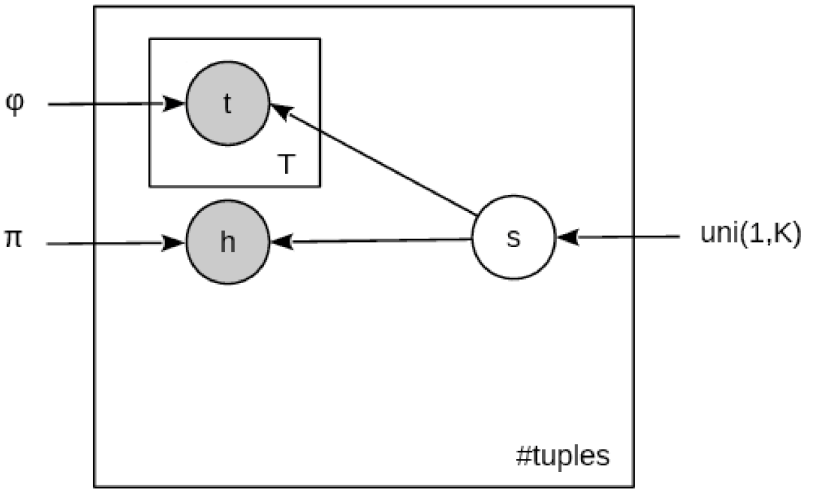
Document classification

- Irrelevant documents (in MUC-4 dataset)
  - Containing *terrorist* entities
  - Not containing actual *terrorist* events
- Classifying relevance *vs.* Irrelevance

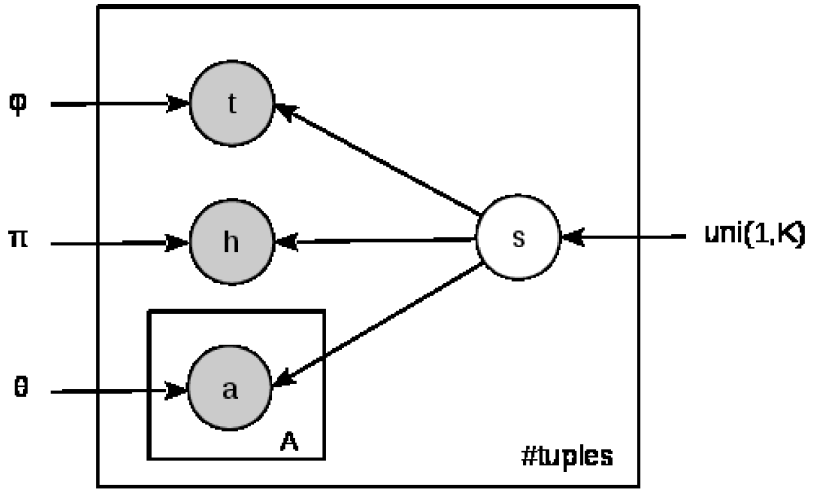
$$relevance(d) = \frac{\sum_{e \in d : s_e \in S_m} \sum_{t \in T_e} P(t|s_e)}{\sum_{e \in d} \sum_{t \in T_e} P(t|s_e)}$$



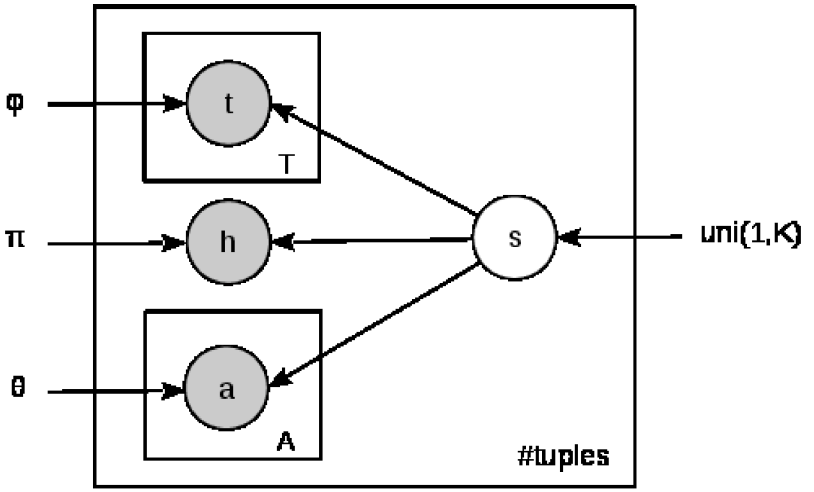
a) HT&Single



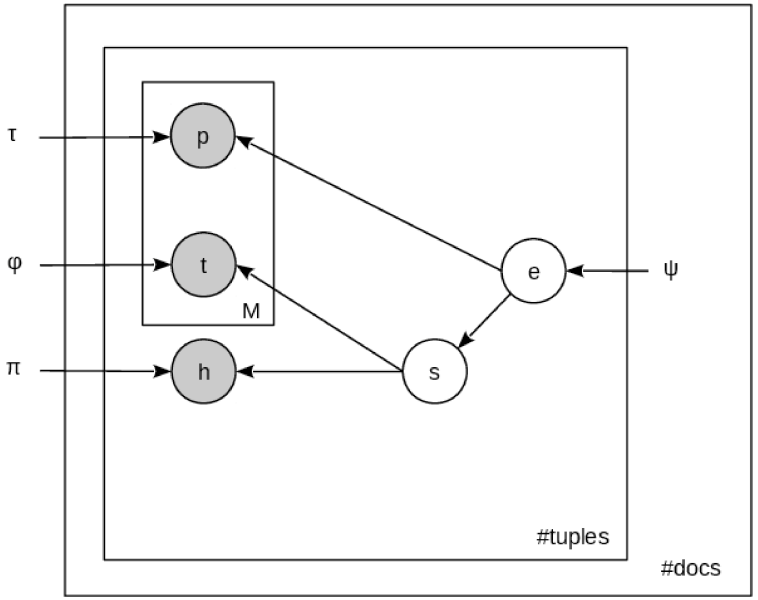
b) HT&Multi/Coref



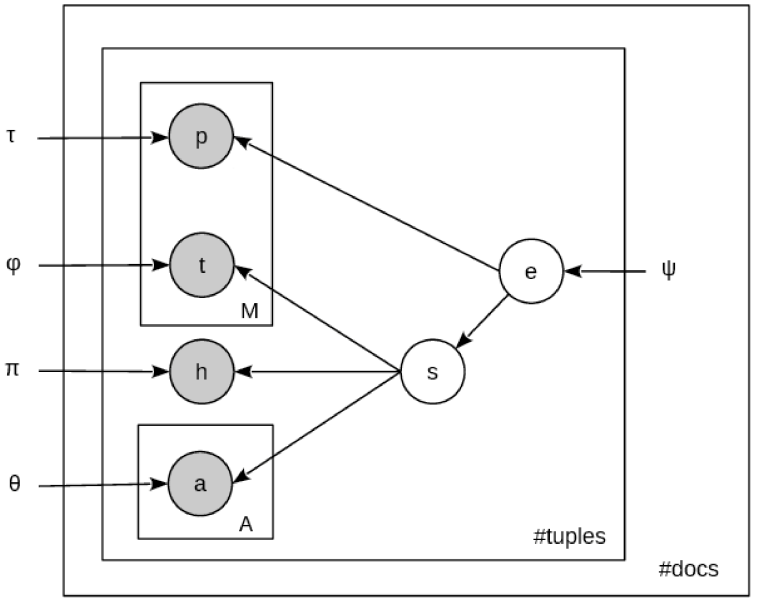
c) HT+A&Single



d) HT+A&Multi/Coref



a) Chambers'13 model



b) Chambers'13+ Attributes

	Precision	Recall	F-score
HT & Single	29.59	51.17	37.48
HT & Multi	29.32	52.21	37.52
HT & Coref	29.99	53.53	40.01
HT + A & Single	30.22	52.41	38.33
HT + A & Multi	30.82	51.68	38.55
HT + A & Coref	32.42	**54.59**	40.62
HT + A + Doc classification	35.57	53.89	**42.79**
HT + A + Oracle doc classification	44.58	54.59	49.08
Chambers'13(reimpl)	38.65	42.68	40.56
Chambers'13(reimpl) +Attr	**39.25**	43.68	41.31

	Precision	Recall	F-score
Cheung et al	32	37	34
Chambers'11	**48**	25	33
Chambers'13 (paper values)	41	41	41
HT+A+Doc classification	36	**54**	**43**

Conclusions & Discussions

- We proposed
  - Joint model for role induction & filling
  - Entity disambiguation using syntactic information
- Future work
  - Evaluation on other domains, corpora
  - Integration of temporal-spatial information about events

THANK YOU !