Improving Event Coreference using Knowledge Bases

Chris Ying and Bishan Yang

Machine Learning Department, Carnegie Mellon University

Introduction

Event extraction is the task of identifying discrete events from free text. It is generally divided into four steps [1]:

1. Identify event anchors
2. Match related entities
3. Assign attributes
4. Coreference event mentions

> A powerful bomb **tore** through a waiting shed at the Davao City international airport.
>
> Bomb **explodes** in the airport of the fourth largest city in the Philippines last Tuesday.

Figure 1: Two coreferring events. The event anchors are bolded and the entities are underlined. Attributes not shown.

The motivation for this project is to utilize **prior world knowledge** to construct entity relations which provide evidence for event coreference.

Objectives

- Develop a model for representing events, entities, and prior world knowledge
- Extract salient features from the model and train a pairwise classifier for coreference
- Improve the performance of event coreference by utilizing rich features

Resources used in this project:

- ECB+ corpus: 982 annotated news documents with 90 topics
- YAGO ontology: semantic knowledge base created using Wikipedia and WordNet
- DBpedia ontology: semantic knowledge graph with over 4.5 million entities and their relations

Text-KB Graph

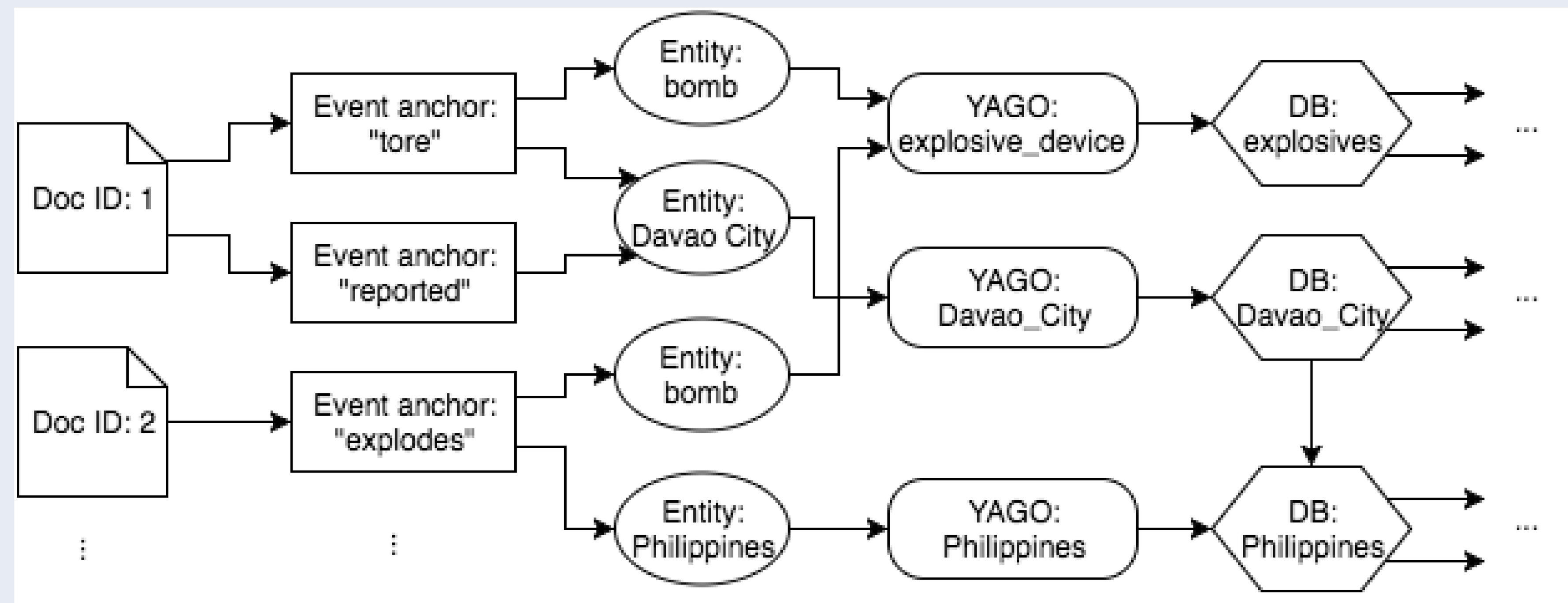

Figure 2: Text-KB graph. Layers of the graph from left to right: documents, events, entities, YAGO entities, DBpedia entities. The "rich" features are YAGO and DBpedia entities. We can choose to traverse any number of layers into the DBpedia graph.

Methods

Features extracted for each pair of events:

1. Event anchor match (baseline)
2. Distance between bag-of-words-of-entities
3. Distance between YAGO entities
4. Distance between DBpedia entities

To give more weight to more salient entities, features 2 - 4 use TF-IDF weighting (treat topics as documents). We represent each event as a vector v.

$$v_i = \mathtt{tf}_i * \log \frac{N}{\mathtt{df}_i} \tag{1}$$

Since the vector is very sparse, we use cosine distance to measure event similarity.

$$dist_{u,v} = 1 - \frac{u \cdot v}{\|u\| \|v\|} \tag{2}$$

Using these extracted features, we train a **logistic regression** classifier to output whether the event pair is coreferencing or not.

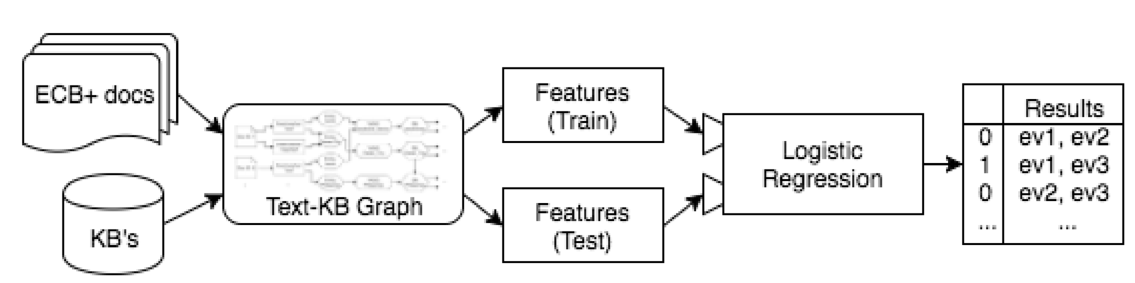

Figure 3: Coreference pipeline.

Results

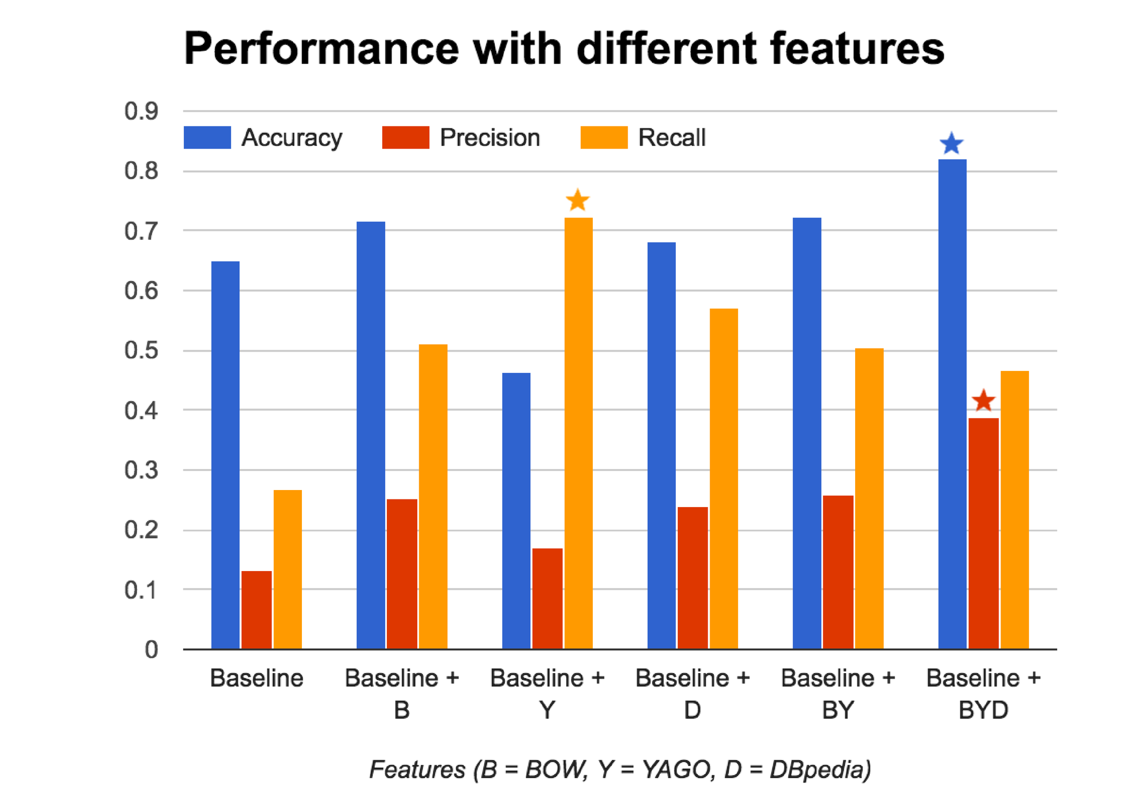

Figure 4: Pairwise coreference performance on a test set.

California earthquake today has erupted outside Santa Rosa.	An earthquake measuring 4.6 rattled Sonoma and Lake Counties.
Explaining its attack on al-Fahora school, the Israeli military claimed that a mortar was fired from the playground.	Three shells hit Fakhura, a girls' elementary school in the Jabalya refugee camp in northern Gaza.
Quarterback Peyton Manning was nearly perfect as the Colts beat the host Jaguars, 31-24, on Thursday night.	Indianapolis (11-4) made an NFL-best fourth double-digit comeback this this season to lock up the five seed in the AFC.

Figure 5: Events coreferenced by the rich model.

AMD to acquire microserver vendor SeaMicro.	In an unprecedented move that suggests there might be life in the old dog yet, AMD has acquired SeaMicro, a company that builds low-power servers using Intel CPU's.
T-Mobile is the only major U.S. wireless company to offer the Q10 with no annual service contract plus unlimited talk, text, and Web on a fast, nationwide 4G network.	T-Mobile has announced that it will be carrying the keyboard touting Q10, with pre-registration starting April 29th.
One of the key suspected Mafia bosses arrested yesterday in one of Sicily's biggest police operations has hanged himself in his prison cell.	Police have launched an investigation into how Gaetano Lo Presti was able to commit suicide.

Figure 6: Events NOT coreferenced by the rich model.

Conclusion

As seen from the results in Figure 4, the model utilizing all **rich features beats the baseline and shallow models** in nearly all metrics. The Text-KB graph allows us to utilize real-world knowledge to better match events in free-text.

From manually inspecting the coreferenced outputs, we know that the system:

Performs well with:

- Similar event mention lengths
- Closely related entities (e.g. geographic)
- Well-known entities, esp. from Wikipedia

Performs poorly with:

- Significantly different mention lengths
- Multiple unrelated events/entities incl.
- Unrecognized named entities

Future Work

- Extract features from the structure of the graph (e.g. edges, connectivity)
- Link the Text-KB graph to additional knowledge bases including NELL
- Use dependency parsing and event frames to better represent event-entity relations

References

[1] David Ahn.
The Stages of Event Extraction.
Proceedings of the Workshop on Annotating and Reasoning about Time and Events, 2006.

Acknowledgements

Thank you to Bishan Yang, Tom Mitchell, and the entire Read the Web team for feedback on improving my project.