Unsupervised Main Entity Extraction from News Articles using Latent Variables

Emory NLP

Yu-Hsin(Henry) Chen and Jinho D. Choi
Department of Mathematics and Computer Science, Emory University

INTRODUCTION

Entity Extraction

- Entity extraction adds to the semantic knowledge of documents.
- Entities consist of mainly proper nouns and pre-defined named entities.
- Additional entity information benefits other Natural Language Processing (NLP) tasks, such as relation extraction and coreference resolution.

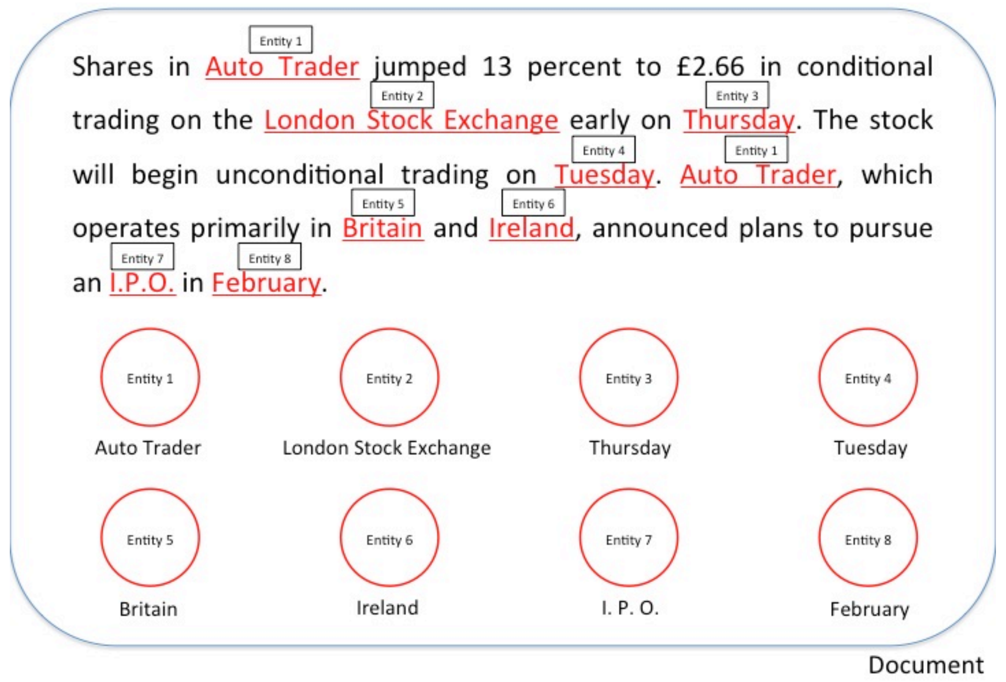

Figure 1. Example of entity extraction on a document

Main Entity Extraction

- Main entity extraction is the preliminary step for relation extraction.
- Main entities are subjects that the context of a document centers around.
- Extraction helps to filter singleton entities and reduce complexity of relation extraction.

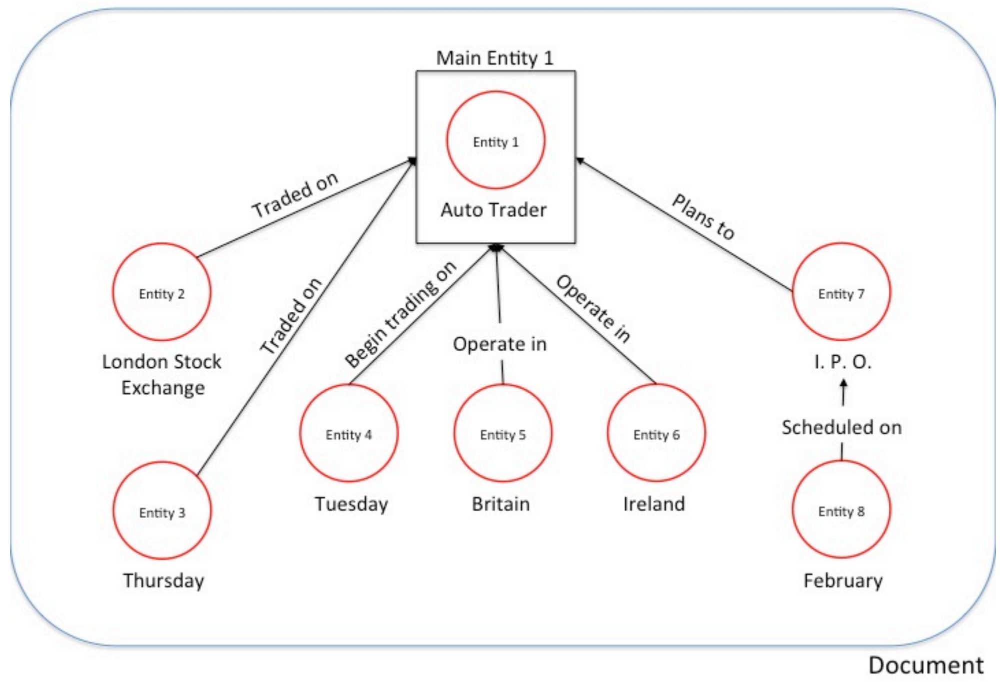

Figure 2. Example of extracted main entity and associated relations with other entities

METHODOLOGY

Natural Language Processing

- ClearNLP, an integrated NLP library developed at the Emory NLP Lab, parses raw text and gives semantic and lexical information.
- The information is used as features to train models for extraction.

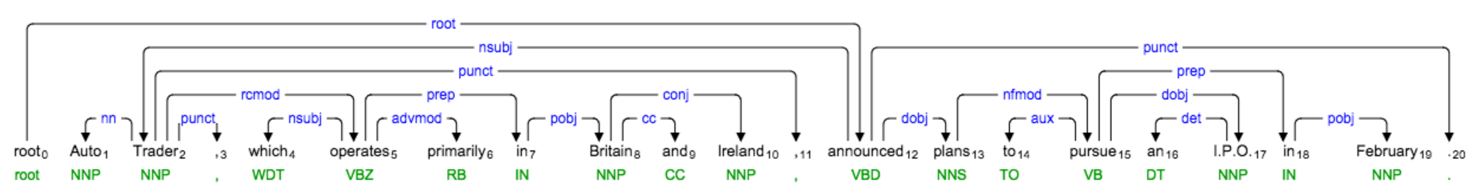

Figure 3. Example of dependency parsing

METHODOLOGY

Semi-Unsupervised Learning

- For initial supervised learning, we avoid the complication of obtaining annotated data by generating high-precision seed documents with human-defined features.
- Continuous model training with the output of previous training/decoding iterations.

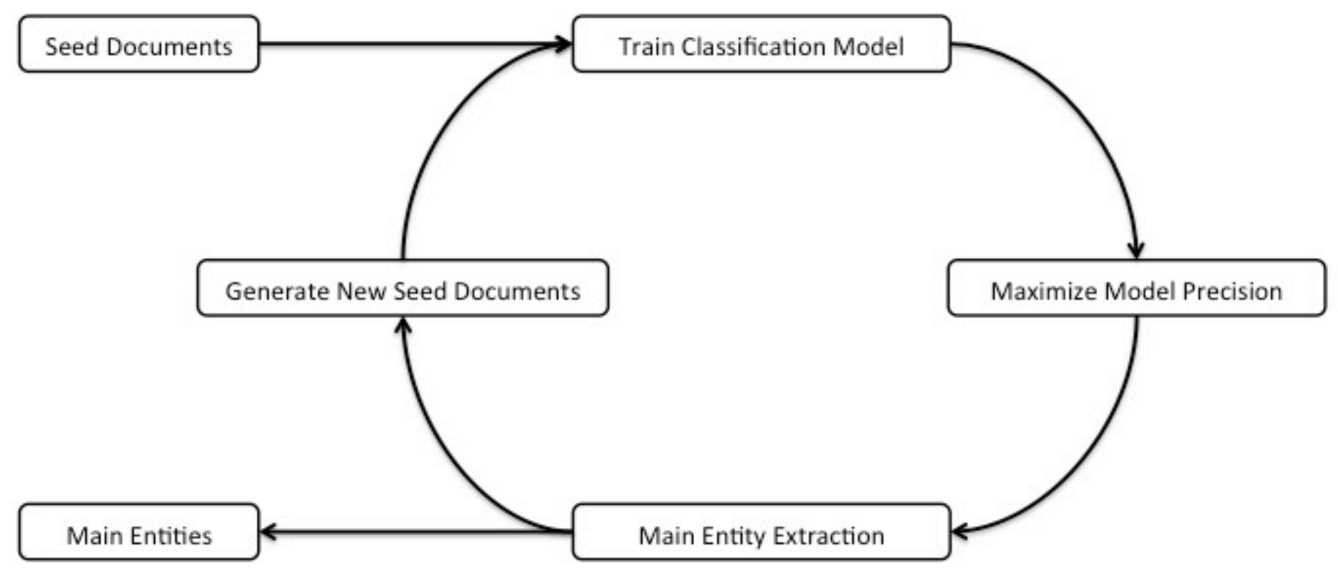

Figure 4. Illustration of semi-supervised learning implemented in this project

Model Training

- Entities are extracted using a proper noun chunker based on the dependency relations between words in sentences.
- Mentions of the same entity are connected based on either exact or relaxed string matches.
- Entities are given confidence scores based on the following human-defined features:
  1. Frequency count of an entity within a document
  2. Sentence where an entity is first mentioned
  3. Confidence of the mentions of the same entity
- Positive and negative samples are selected in the seed documents based on a pre-defined high cutoff and a calculated low cutoff of the entity confidence.

$$\mu_{mec} = \mu_{Main\ EntityConfidence}$$
$$\mu_{ec} = \mu_{EntityConfidence}$$
$$MD_{ec} = \mu_{mec} - \mu_{ec}$$
$$std_{ec} = standard\ deviation\ of\ enitity\ confidence$$
$$ME\% = Count(Main\ Entity)/(Count(Entity))$$

$$LowCutOff = \alpha * \frac{\mu_{mec}}{\mu_{ec}} * \left(\frac{MD_{ec}}{std_{ec}} - 1\right) * \left[\left(\frac{ME\%}{1-ME\%}\right) * \mu_{mec}\right]$$

Figure 5. Equation for calculating low cutoff of entity confidence

- Semantic and lexical features of the entities found with their surrounding contexts are extracted and converted into vectors.
- Entity feature vectors serve as instances in Adaptive Subgradient Support Vector Machine to train a binary classification model.
- The trained model is then used to decode the entire corpus in order to generate the next set of seed documents.

EXPERIMENT

Evaluation

- Precisions of extracted main entities are evaluated based on the word sequence matches between the entities and the titles of the news articles (in total of 3484 documents).

		Seed Document Statistics			
		(+) Sample Count	(-) Sample Count	Total Sample Count	Document Count
(Initial Seed)	Instance #0	387	1237	8685	383
	Instance #1	10970	159612	170582	3124
	Instance #2	7941	158650	166591	3053
	Instance #3	6861	156909	163770	3001
	Instance #4	6333	154656	160989	2954
	Instance #5	6031	153442	159473	2917
	Instance #6	5847	152283	158130	2889
	Instance #7	5821	151935	157756	2887
	Instance #8	5773	151680	157453	2878
	Instance #9	5664	150969	156633	2865
	Instance #10	5619	150499	156118	2853

Figure 6. Seed documents statistics of each learning instances

		Training Statisitcs			Evaluation
		Precision	Recall	F1 Score	Precision
(Initial Seed)	Instance #0	70.95%	99.22%	82.74%	33.81%
	Instance #1	42.06%	78.35%	54.74%	37.76%
	Instance #2	45.15%	87.21%	59.50%	39.70%
	Instance #3	46.73%	92.46%	62.09%	**40.54%**
	Instance #4	47.61%	94.63%	63.35%	40.31%
	Instance #5	47.93%	96.28%	64.00%	40.11%
	Instance #6	48.37%	**99.06%**	65.00%	40.27%
	Instance #7	**48.48%**	99.02%	**65.09%**	40.19%
	Instance #8	48.23%	97.76%	64.59%	40.01%
	Instance #9	48.30%	98.91%	64.91%	39.73%
	Instance #10	48.33%	98.33%	64.81%	39.65%

Figure 7. Evaluation results of each learning instances

Error Analysis

- The evaluation metric for the extracted entity is limited. It yields a lower precision since news article titles do not always include the main entities discussed in the articles.

CONCLUSION

- We define a feature template that generates initial seed documents from unlabeled data.
- We train a semi-supervised model with only semantic and lexical information from raw text to extract main entities from articles automatically.
- We need a better evaluation metric for this task.

REFERENCE

Agichtein, Eugene, and Luis Gravano. "Snowball: Extracting relations from large plain-text collections." Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000.

Choi, Jinho D., and Andrew McCallum. "Transition-based Dependency Parsing with Selectional Branching." ACL (1). 2013.

ACKNOWLEDGEMENT

This material is based upon work supported by the Howard Hughes Medical Institute Science Education Program award #52006923 to Emory University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Howard Hughes Medical Institute or Emory University.