Constructing Knowledge Graph from Unstructured Text

Kundan Kumar and Siddhant Manocha
Under the guidance of Dr. Amitabha Mukerjee
Indian Institute of Technology Kanpur

Problem Statement

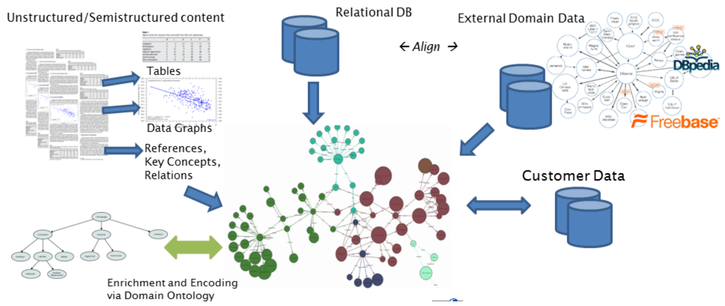

Figure 1: Construct a knowledge graph from unstructured text [1]

Existing Methods

Supervised Models:

- Learn classifiers from +/- examples, typical features: context words + POS, dependency path between entities, named entity tags

Semi-supervised Models:

- Bootstrap Algorithms: Use seed examples to learn initial set of relations utilizing pattern recognition
- Generate +ve/-ve examples to learn a classifier

Distant Supervision:

- Existing knowledge base + unlabeled text generate examples
- Generate +ve/-ve examples to learn a classifier

Approach

- Linear sum of input vectors are projected onto the projection layer
- The projection matrix from the projection layer to the hidden layer give the word vector embeddings [5]
- Word that occur in similar context lie close together in the word embedding space[6]
- Word Vectors is semantically consistent and capture many linguistic properties

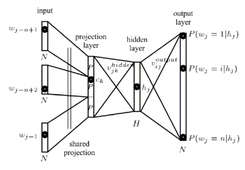

Figure 2: CBOW Model[5]

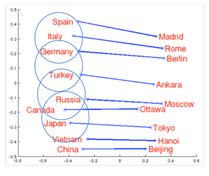

Figure 3: Semantic Regularities in Word Embedding[6]

Methodology

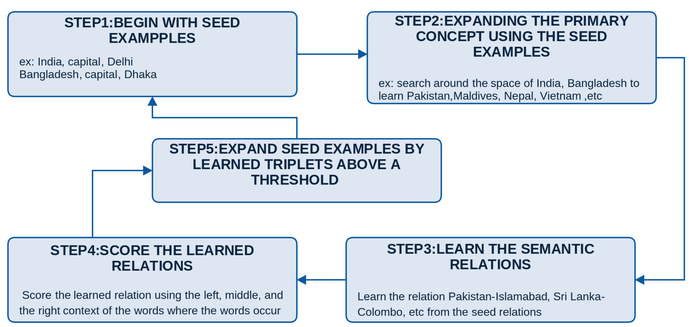

Figure 4: Bootstrapping Relations using Distributed Word Vector Embedding

Similarity Metric

- Context Words Approach: Weighted sum of similarity metric between left, right and middle contexts of two relations respectively
- Tree Kernel based on dependency parse tree
- Semantic Nets (Word Net,etc) based approach

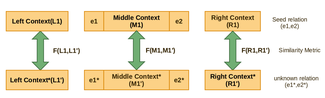

Figure 5: Context Words[3]

Sentence1: 'his actions in Brecko' : his -> actions <- in <- Brecko
Sentence2: 'hi arrival in Beijing': his -> arrival <- in <- Beijing

his PRP Person	→	actions NNS Noun	←	in IN	←	Brecko NNP Noun Location	
his PRP Person	→	arrival NN Noun	←	in IN	←	Beijing NNP Noun Location	
3	1	1	1	2	1	3	18

Figure 6: Dependency Tree Kernels[4]

Question Answering

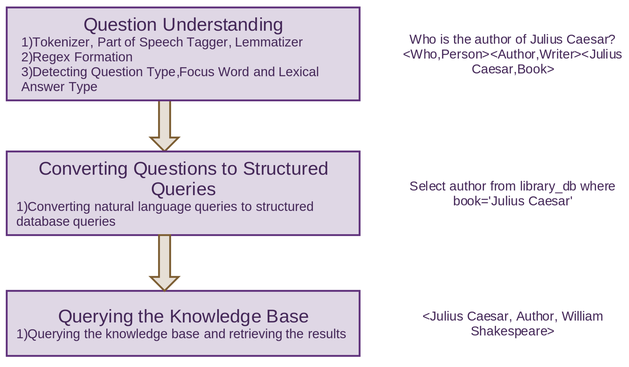

Figure 7: Work flow: Question Answering System

Results

Country	Language	Confidence
India	Hindi	1.0/S
France	French	1.0/S
Croatia	Croatian	0.81/P
Austria	German	0.75/P
Belgium	Dutch	0.78/P
Serbia	Polje	0.64/N
Poland	Polish	0.85/P
Moldova	Romanian	0.72/P
Slovakia	Czech	0.55/N
Belarus	Belarusian	0.57/P

Table: Relation Confidence

Relation(Type)	Correct	Incorrect
Demonym	19	9
Language	25	10
Capital	21	11

Table: Performance

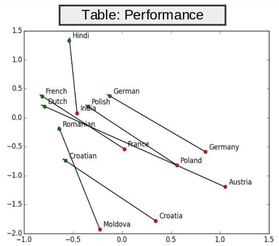

Fig: Word Embeddings

Figure 8: Results

Conclusion and Future Work

- Word embeddings are consistent and can be combined with existing approaches to extract semantic relations
- Regular expressions can be used to form structured queries from natural language questions
- In future, we will like to scale our system to general domains
- Include a relevant similarity metric for evaluation of learned relations

References

[1] Knowledge Graphs.
`http://www.sindicetech.com/overview.html`.
Accessed: 2015-03-10.

[2] E. Agichtein and L. Gravano.
Snowball: Extracting relations from large plain-text collections.
In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA, 2000. ACM.

[3] N. Bach and S. Badaskar.
A survey on relation extraction.

[4] A. Culotta and J. Sorensen.
Dependency tree kernels for relation extraction.
In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics, 2004.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean.
Efficient estimation of word representations in vector space.
CoRR, abs/1301.3781, 2013.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean.
Distributed representations of words and phrases and their compositionality.
In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.