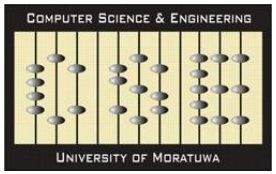
Semi-Supervised Instance Population of an Ontology using Word Vector Embeddings



Computer Science and Engineering Department
University of Moratuwa, Sri Lanka

University of London International Programmes
University of London

International Conference on Advances in ICT for Emerging Regions (ICTer 2017)

September 7th-8th, 2017 | Colombo, Sri Lanka

OUTLINE

1. Introduction
2. Background Work
3. Proposed Methodology
4. Results
5. Conclusion and Future Work

INTRODUCTION

1. What is an ontology
2. What are word embeddings
3. Word embeddings and ontology population

INTRODUCTION

- An ontology is a “formal and explicit specification of a shared conceptualization”.
- Ontology population has become a problematic process due to its nature of heavy coupling with manual human intervention.
- In word embedding, words or phrases from the vocabulary are mapped to vectors of real numbers.
- Proposed here is a novel way of semi-supervised ontology population through word embeddings as the basis.

BACKGROUND WORK

1. Ontologies
2. Word Vector Embeddings
3. Word Set Expansion
4. Ontology Population
5. Semi Supervised Ontology Population

BACKGROUND

1. Distributed representations of words and phrases and their compositionality

T. Mikolov, I. Sutskever, K. Chen et al., “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing
systems, pp. 3111–3119, 2013.

2. Deriving a representative vector for ontology classes with instance word vector embeddings

V. Jayawardana, D. Lakmal, N. de Silva, A.S. Perera, K. Sugathadasa, and B. Ayesha, “Deriving a representative vector for ontology classes with instance word vector embeddings,” arXiv preprint arXiv:1706.02909, 2017.

3. Semi-supervised algorithm for concept ontology based word set expansion

N. H. N. D. de Silva, A. S. Perera, and M. K. D. T. Maldeniya, “Semi-supervised algorithm for concept ontology based word set expansion,” Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on, pp. 125–131, 2013.

Methodology

1. Ontology Creation
2. Training word Embeddings
3. Deriving Representative Class Vectors
4. Creating Instance Corpus for Ontology Population
5. Candidate Model Building
6. Ensemble Model

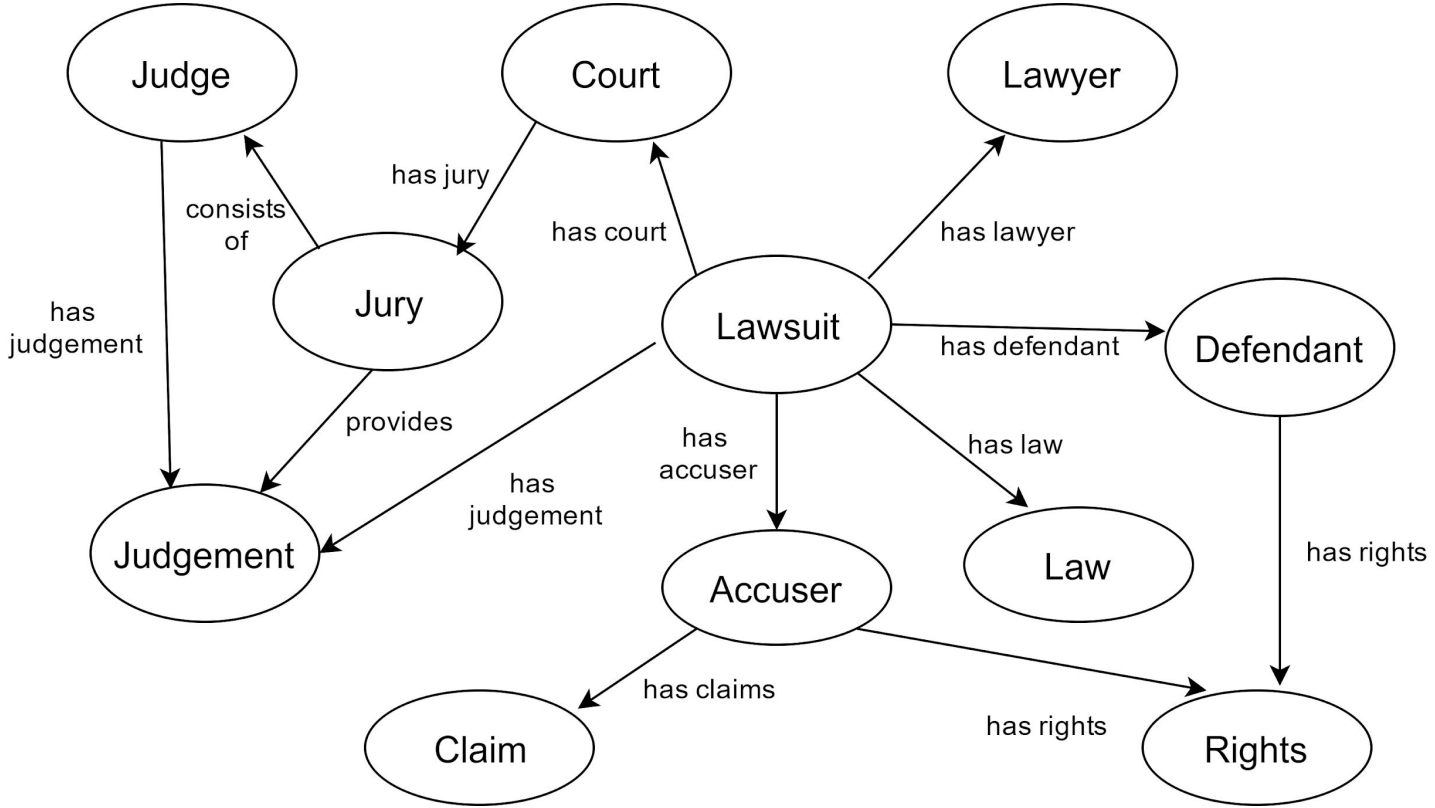
METHODOLOGY

1. **Ontology Creation**

- Legal ontology based on the consumer protection law.
- Findlaw as the reference.
- Part of the ontology was selected mainly focusing on more sophisticated relationships and taxonomic presences.
- Manually added seed instances for all the classes in the selected ontology part.

METHODOLOGY

Selected Ontology Part



METHODOLOGY

2. Training word Embeddings

- Training of the word embeddings was the process of building a word2vec model.
- The text corpus consisted of legal cases under 78 law categories from FindLaw.
- Stanford CoreNLP for preprocessing the text with tokenizing, sentence splitting, Part of Speech (PoS) tagging and lemmatizing.

METHODOLOGY

3. Deriving Representative Class Vectors

- A methodology to derive a representative vector for ontology classes, whose instances were mapped to a vector space by Jayawardana at el. was used.

METHODOLOGY

4. Creating Instance Corpus for Ontology Population

- Legal cases from Findlaw to create an instance corpus.
- Stanford CoreNLP based preprocessing on the raw text with tokenization and sentence splitting was used.

METHODOLOGY

5. Candidate Model Building

- Membership by distance model (M1)
- Membership by dissimilar exclusion model (M2)
- Set expansion based model (M3)
- Semi-supervised K-Means clustering based model (M4)
- Semi-supervised hierarchical clustering based model (M5)

METHODOLOGY

Membership by distance model (M1)

- Candidate vectors for the ontology classes were generated from the instance corpus based on the minimum distance to the representative class vector.

Membership by dissimilar exclusion model (M2)

- word2vec based dissimilar exclusion method in identifying the membership of a particular instance to a given class.

METHODOLOGY

Set expansion based model (M3)

- Set expansion algorithm to expand the seed instances of an ontology class.

Semi-supervised K-Means clustering based model (M4)

- A semi-supervised model.
- By mixing up the labeled and unlabeled data and clustering with k-means clustering where k is the number of classes in the selected ontology.

METHODOLOGY

Semi-supervised hierarchical clustering based model (M5)

- A semi-supervised model.
- A model which creates hierarchy of clusters using the word embeddings taken from the word2vec model of the entire corpus.
- Then extracted the slice of hierarchical clusters such that the number of clusters in the slice is equal to the number of classes in the sub-ontology

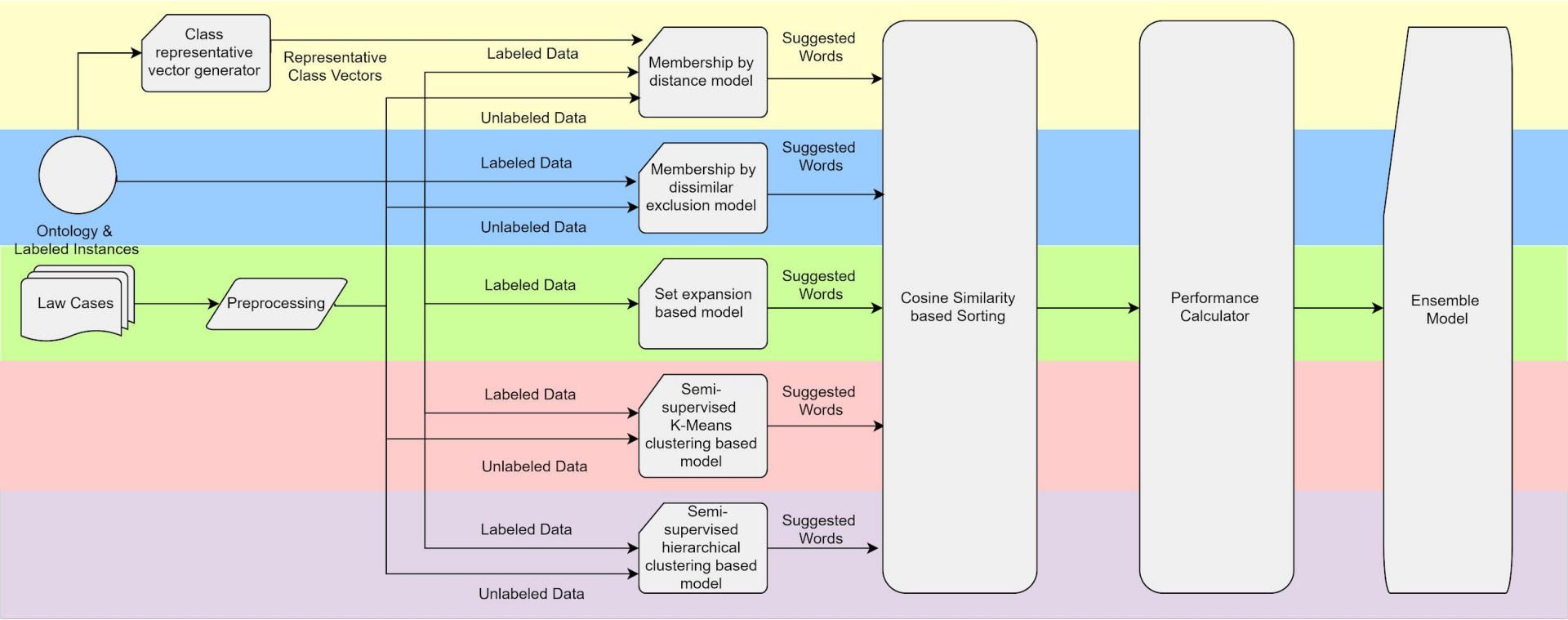
METHODOLOGY

6. Ensemble Model

- Ensemble model based on the models identified earlier.
- In creating the ensemble model, allocated a candidate weight for each model based on each model's F1 measure.

METHODOLOGY

Flow of Semi-Supervised Instance Population of an Ontology using Word Vector Embeddings



RESULTS

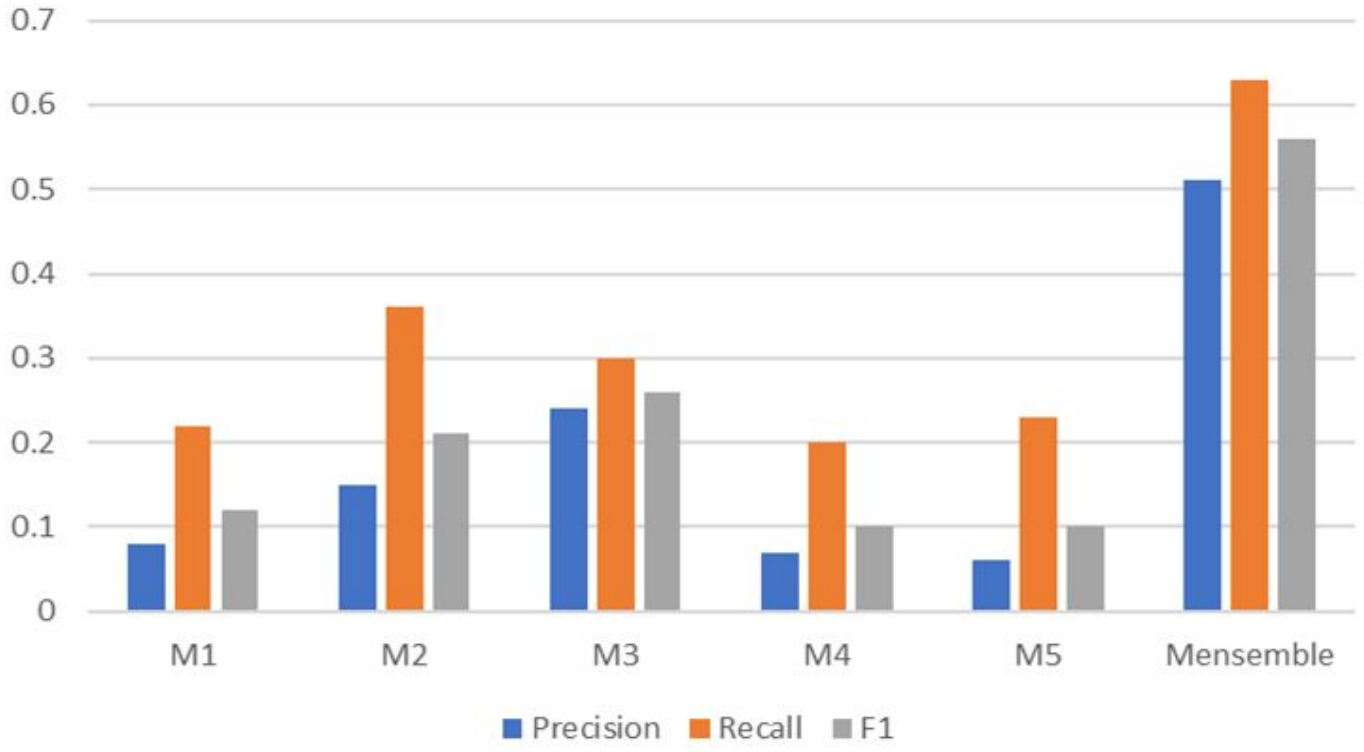
RESULTS

COMPARISON OF PERFORMANCE OF MODELS

	Precision	Recall	F1
M_1	0.08	0.22	0.12
M_2	0.15	0.36	0.21
M_3	0.24	0.30	0.26
M_4	0.07	0.20	0.10
M_5	0.06	0.23	0.10
$M_{ensemble}$	**0.51**	**0.63**	**0.56**

RESULTS

Comparison of precision, recall and F1 of the models



CONCLUSION & FUTURE WORK

CONCLUSION AND FUTURE WORK

- The work demonstrated the use of word embeddings on semi-supervised ontology population.
- As showed in the results, proposed ensemble model outperforms the five individual models in populating the selected legal ontology.
- This study is mainly important in two ways
  - To keep a handcrafted ontology up to date.
  - The novelty in the methodology proposed.
- Only considered the single word instances in populating the ontology using the defined models.
- Word Set Expansion algorithm used has the weakness of being dependent on the WordNet lexicon.

THANK YOU !!