Semi-automatic WordNet Linking using Word Embeddings

Kevin Patel, Diptesh Kanojia and Pushpak Bhattacharyya
Presented by: Ritesh Panjwani

January 11, 2018

Outline

Introduction

- Wordnet
  - Lexical resource
  - Groups words into sets of synonyms called Synsets
  - Records relations among these synsets
- Linked Wordnet
  - Synsets with same meaning, but belonging to wordnets of different languages are linked
  - EuroWordNet Vossen and Letteren (1997) and IndoWordNet Bhattacharyya (2010)
  - Used for Machine Translation Hovy (1998), Cross Lingual Information Retrieval Gonzalo et al. (1998), *etc.*
- Challenge in linking Wordnets
  - Linking done manually
  - Tools such as Joshi et al. (2012b) to assist humans

Background

- Princeton WordNet (Miller et al., 1990) or the English WordNet was the first wordnet.
- EuroWordNet (Vossen and Letteren, 1997) : linked wordnet comprising of wordnets for European languages.
  - Each wordnet separately captures a language-specific information.
  - Wordnets uses Princeton WordNet as an Inter-Lingual-Index.
  - Enables one to go from concepts in one language to similar concepts in any other language.
- IndoWordNet Bhattacharyya (2010) is a linked wordnet comprising of wordnets for 18 Indian languages.
  - Created using the expansion approach using Hindi WordNet as a pivot.
  - Partially linked to English WordNet.

Related Work

- Joshi et al. (2012a) developed a heuristic based measure where they use bilingual dictionaries to link two wordnets.
  - Combine scores using various heuristics and generate a list of potential candidates for linked synsets.
- Singh et al. (2016) discuss a method to improve the current status of Hindi-English linkage and present a generic methodology
  - Their method is beneficial for culture-specific synsets, or for non-existing concepts
  - Cost and time inefficient; requires a lot of manual effort on the part of a lexicographer.
- Our intention: reduce effort on the part of lexicographers

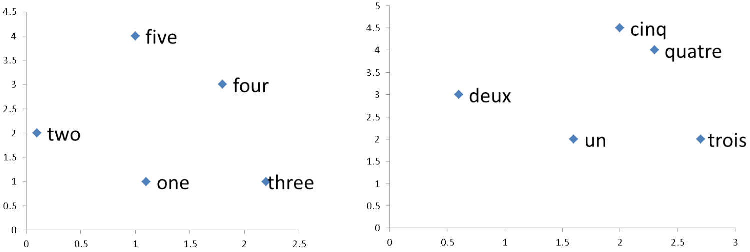
Problem Statement

- Given wordnets of two different languages E and F with sets of synsets $\{s_E^1, s_E^2, \ldots, s_E^m\}$ and $\{s_F^1, s_F^2, \ldots, s_F^n\}$ respectively, find mappings of the form $< s_E^i, s_F^j >$ which are semantically correct.

Hindi Synsets	**English Synsets**
1: {हाथ, हस्त, कर}	1: {hand, paw}
	2: {tax, revenue enhancement}

Motivation

- Adapted from Mikolov et al. (2013a)



English $\leftrightarrow$ French

$$\vec{y} = W\vec{x}$$

Algorithm: Notations

- Let E and F be two languages
- Let $|E|$ and $|F|$ be the number of synsets in wordnets of E and F respectively
- Let s_E^i and s_F^j be the i^{th} and j^{th} synsets of E and F respectively,
  - $s_E^i = \{e_\alpha^1, e_\alpha^2, \ldots, e_\alpha^{m_i}\}$
  - $s_F^j = \{f_\beta^1, f_\beta^2, \ldots, f_\beta^{n_j}\}$
    - e_α^p and f_β^q are words in vocabulary of E and F respectively for $1 \leq p \leq m_i$ and $1 \leq q \leq n_j$, and $1 \leq i \leq |E|$ and $1 \leq j \leq |F|$
- Let $v_{e_\alpha^p}$ be the word vector corresponding to e_α^p

Algorithm: Training

- Estimate $v_{s_E^i}$ as

$$v_{s_E^i} = \frac{1}{m_i} \sum_{p=0}^{m_i} v_{e_\alpha^p} \tag{1}$$

- Similarly,

$$v_{s_F^j} = \frac{1}{n_j} \sum_{q=0}^{n_j} v_{f_\beta^q} \tag{2}$$

- Given links of the form $\left\langle s_E^i, s_F^j \right\rangle$, we learn W such that the error Err

$$Err = \| W.v_{s_E^i} - v_{s_F^j} \|^2 \tag{3}$$

is minimized.

Algorithm: Prediction

- To find a mapping for a new synset s_E^k, one needs to
  - Calculate $v' = W.v_{s_E^k}$
  - Find $v_{s_F^l}$ such that $v_{s_F^l}.v'$ is maximized
  - Create link $\langle s_E^k, s_F^l \rangle$

Datasets

- Linking Hindi WordNet to English WordNet
- **English Vectors**: Pretrained vectors from Google's word2vec tool Mikolov et al. (2013b), trained on News dataset (around 100 billion tokens)
- **Hindi Vectors**: Trained using word2vec on Bojar corpus Bojar et al. (2014) (around 365 million tokens)
- Linked data: Created at CFILT, IITB
  - Of the form $\langle \textit{hindi\_synset\_id}, \textit{english\_synset\_id}, \textit{link\_type} \rangle$, where $\textit{link\_type} \in \{\textit{DIRECT}, \textit{HYPERNYMY}, \textit{etc.}\}$
  - Focus on only DIRECT links
  - 6863 such links available

Distribution of links

Class	Count
Noun	4757
Adjective	1283
Verb	680
Adverb	143

Distribution of available links among various classes

Evaluation metric

- Accuracy@n: One of the top *n* predictions can be correct

	Predicted Label	Accuracy @1	Accuracy @3	Accuracy @5
True label	Prediction1			
	Prediction2			
	Prediction3			
	Prediction4			
	Prediction5			

Results: Overall

	Acc@1	**Acc@3**	**Acc@5**	**Acc@8**	**Acc@10**
Overall	0.29	0.45	0.52	0.58	0.60

Results for the overall setting: Dimension of English embeddings=300, Dimensions of Hindi embeddings=300

Results: Per word class I

Word Class	Acc@1	Acc@3	Acc@5	Acc@8	Acc@10
Noun	0.35	0.53	0.60	0.65	0.67
Adjective	0.26	0.44	0.50	0.57	0.60
Verb	0.15	0.25	0.29	0.33	0.37
Adverb	0.28	0.51	0.59	0.70	0.73

Results for the setting: Dimension of English Vectors=300, Dimensions of Hindi Vectors=300

Results: Per word class II

Word Class	Acc@1	Acc@3	Acc@5	Acc@8	Acc@10
Noun	0.35	0.51	0.58	0.64	0.66
Adjective	0.12	0.20	0.24	0.30	0.32
Verb	0.17	0.27	0.32	0.36	0.39
Adverb	0.38	0.52	0.65	0.76	0.80

Results for the setting: Dimension of English Vectors=300, Dimensions of Hindi Vectors=1200

Discussion

- Possible reasons for poor performance

Discussion

- Possible reasons for poor performance
  - Something is fundamentally missing in word vectors. Probably presence of only co-occurence information, and lack of other information such as word ordering, argument frames(for verbs), etc.

Discussion

- Possible reasons for poor performance
  - Something is fundamentally missing in word vectors. Probably presence of only co-occurence information, and lack of other information such as word ordering, argument frames(for verbs), etc.
  - The approach to create synset vectors is not optimal.

Discussion

- Possible reasons for poor performance
  - Something is fundamentally missing in word vectors. Probably presence of only co-occurence information, and lack of other information such as word ordering, argument frames(for verbs), etc.
  - The approach to create synset vectors is not optimal.
  - The linear transformation approach is not optimal.

Discussion

- Possible reasons for poor performance
  - Something is fundamentally missing in word vectors. Probably presence of only co-occurence information, and lack of other information such as word ordering, argument frames(for verbs), etc.
  - The approach to create synset vectors is not optimal.
  - The linear transformation approach is not optimal.
  - Synset members are often phrases instead of words. How to create phrase vectors?

Discussion

- Possible reasons for poor performance
  - Something is fundamentally missing in word vectors. Probably presence of only co-occurence information, and lack of other information such as word ordering, argument frames(for verbs), etc.
  - The approach to create synset vectors is not optimal.
  - The linear transformation approach is not optimal.
  - Synset members are often phrases instead of words. How to create phrase vectors?
  - Currently, a word has only one vector. That is a one of the reason for ambiguity. Perhaps for each word, multiple vectors (one vector per sense) is the way to go.

Conclusion and Future Work

- Described an approach to link wordnets
- Creates synset embeddings using word embeddings, followed by learning transformation from source to target language synsets
- Our approach achieves accuracy@10 of approximately 60% and 70% of all synsets and noun synsets, respectively
- Discussed reasons for poor performance on classes such as verbs
- Plan to integrate it in tools such as Joshi et al. (2012a)

References

Bhattacharyya, P. (2010). Indowordnet. In *Lexical Resources Engineering Conference 2010 (LREC 2010)*.

Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014). HindMonoCorp 0.5.

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*.

Hovy, E. (1998). Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, pages 535–542.

Joshi, S., Chatterjee, A., Karra, A. K., and Bhattacharyya, P. U. (2012a). Eating your own cooking: automatically linking wordnet synsets of two languages.

References

Joshi, S., Chatterjee, A., Karra, K. A., and Bhattacharyya, P. (2012b). Eating your own cooking: Automatically linking wordnet synsets of two languages. In *Proceedings of COLING 2012: Demonstration Papers*, pages 239–246. The COLING 2012 Organizing Committee.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

References

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Singh, M., Shukla, R., Jha, J., Kashyap, L., Kanojia, D., and Bhattacharyya, P. (2016). Mapping it differently: A solution to the linking challenges. In *Eighth Global Wordnet Conference*. GWC 2016.

Vossen, P. and Letteren, C. C. (1997). Eurowordnet: a multilingual database for information retrieval. In *In: Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.

Thank You

Questions?

For more details, write to: kevin.patel@cse.iitb.ac.in