Word Similarity Datasets for Indian Languages: Annotation and Baseline Systems

Syed S. Akhtar Arihant Gupta Avijit Vajpayee Arjit Srivastava M. Shrivastava
International Institute of Information Technology
Hyderabad, Telangana, India
{syed.akhtar, arihant.gupta, arjit.srivastava}@research.iiit.ac.in,
avijit.vajpayee@students.iiit.ac.in,
manish.shrivastava@iiit.ac.in

Abstract

With the advent of word representations, word similarity tasks are becoming increasing popular as an evaluation metric for the quality of the representations. In this paper, we present manually annotated monolingual word similarity datasets of six Indian languages – Urdu, Telugu, Marathi, Punjabi, Tamil and Gujarati. These languages are most spoken Indian languages worldwide after Hindi and Bengali. For the construction of these datasets, our approach relies on translation and re-annotation of word similarity datasets of English. We also present baseline scores for word representation models using state-of-the-art techniques for Urdu, Telugu and Marathi by evaluating them on newly created word similarity datasets.

1 Introduction

Word representations are being increasingly popular in various areas of natural language processing like dependency parsing (Bansal et al., 2014), named entity recognition (Miller et al., 2004) and parsing (Socher et al., 2013). Word similarity task is one of the most popular benchmark for the evaluation of word representations. Applications of word similarity range from Word Sense Disambiguation (Patwardhan et al., 2005), Machine Translation Evaluation (Lavie and Denkowski, 2009), Question Answering (Mohler et al., 2011), and Lexical Substitution (Diana and Navigli, 2009).

Word Similarity task is a computationally efficient method to evaluate the quality of word vectors. It relies on finding correlation between human assigned semantic similarity (between words) and corresponding word vectors. We have used Spearman's Rho for calculating correlation. Unfortunately, most of the word similarity tasks have been majorly limited to English language because of availability of well annotated different word similarity test datasets and large corpora for learning good word representations, where as for Indian languages like Marathi, Punjabi, Telugu etc. – which even though are widely spoken by significant number of people, are still computationally resource poor languages. Even if there are models trained for these languages, word similarity datasets to test reliability of corresponding learned word representations do not exist.

Hence, primary motivation for creation of these six word similarity datasets has been to provide necessary evaluation resources for all the current and future work in field of word representations on these six Indian languages – all ranked in top 25 most spoken languages in the world, since no prior word similarity datasets have been publicly made available.

The main contribution of this paper is the set of newly created word similarity datasets which would allow for fast and efficient comparison between. Word similarity is one of the most important evaluation metric for word representations and hence as an evaluation metric, these datasets would promote development of better techniques that employ word representations for these languages. We also present baseline scores using state-of-the-art techniques which were evaluated using these datasets.

The paper is structured as follows. We first discuss the corpus and techniques used for training our models in section 2 which are later used for evaluation. We then talk about relevant related work that has been done with respect to word similarity datasets in section 3. We then move on to explain how these datasets have been created in section 4 followed by our evaluation criteria and

Proceedings of the 11th Linguistic Annotation Workshop, pages 91–94,
Valencia, Spain, April 3, 2017.

experimental results of various models evaluated on these datasets in section 5. Finally, we analyze and explain the results in section 6 and finish this paper with how we plan to extend our work in section 7.

2 Datasets

For all the models trained in this paper, we have used the Skip-gram, CBOW (Mikolov et al., 2013a) and FastText (Bojanowski et al., 2016) algorithms. The dimensionality has been fixed at 300 with a minimum count of 5 along with negative sampling.

As training set of Marathi, we use the monolingual corpus created by IIT-Bombay. This data contains 27 million tokens. For Urdu, we use the untagged corpus released by Jawaid et. al. (2014) containing 95 million tokens. For Telugu, we use Telugu wikidump available at `https://archive.org/details/tewiki-20150305` having 11 million tokens.

For testing, we use the newly created datasets. The word similarity datatsets for Urdu, Marathi, Telugu, Punjabi, Gujarati and Tamil contain 100, 104, 111, 143, 163 and 97 word pairs respectively.

For rest of the paper, we have calculated the Spearman ρ (multiplied by 100) between human assigned similarity and cosine similarity of our word embeddings for the word-pairs. For any word which was is not found, we assign it a zero vector.

In order to learn initial representations of the words, we train word embeddings (word2vec) using the parameters described above on the training set.

3 Related Work

Multitude of word similarity datasets have been created for English, like WordSim-353 (Finkelstein et al., 2002), MC-30 (Miller and Charles, 1991), Simlex-999 (Hill et al., 2016), RG-65 (Rubenstein and Goodenough, 2006) etc. RG-65 is one of the oldest and most popular datasets, being used as a standard benchmark for measuring reliability of word representations.

RG-65 has also acted as base for various other word similarity datasets created in different languages: French (Joubarne and Inkpen, 2011), German (Zesch and Gurevyc, 2006), Portuguese (Granada et al., 2014), Spanish and Farsi (Camacho-Collados et al., 2015). While German and Portuguese reported IAA (Inter Annotator Agreement) of 0.81 and 0.71 respectively, no IAA was calculated for French. For Spanish and Farsi, inter annotator agreement of 0.83 and 0.88 respectively was reported. Our datasets were created using RG-65 and WordSim-353 as base, and their respective IAA(s) are mentioned later in the paper.

4 Construction of Monolingual Word Similarity datasets

4.1 Translation

English RG-65 and WordSim-353 were used as base for creating all of our six different word similarity datasets. Translation of English data set to target language (one of the six languages) was manually done by a set of three annotators who are native speakers of the target language and are fluent in English. Initially, translations are provided by two of them, and in case of disparity, third annotator was used as a tie breaker.

Finally, all three annotators reached a final set of translated word pairs in target language, ensuring that there were no repeated word pairs. This approach was followed by Camacho-Callados et al. (2015) where they created word similarity datasets for Spanish and Farsi in a similar manner.

4.2 Scoring

For each of the six languages, 8 native speakers were asked to manually evaluate each word similarity data set individually. They were instructed to indicate, for each pair, their opinion of how similar in meaning the two words are on a scale of 0-10, with 10 for words that mean the same thing, and 0 for words that mean completely different things. The guidelines provided to the annotators were based on the SemEval task on Cross-Level Semantic Similarity (Jurgens et al., 2014), which provides clear indications in order to distinguish similarity and relatedness.

The results were averaged over the 8 responses for each word similarity data set, and each data set saw good agreement amongst the evaluators, except for Tamil, which saw relatively weaker agreement with respect to other languages (see table 1).

5 Evaluation

5.1 Inter Annotator Agreement (IAA)

The meaning of a sentence and its words can be interpreted in different ways by different read-

ers. This subjectivity can also reflect in annotation of sentences of a language despite the annotation guidelines being well defined. Therefore, inter-annotator agreement is calculated to give a measure of how well the annotators can make the same annotation decision for a certain category.

Language	Inter Annotator Agreement
Urdu	0.887
Punjabi	0.821
Marathi	0.808
Tamil	0.756
Telugu	0.866
Gujarati	0.867

Table 1: Inter Annotator Agreement (Fleiss Kappa) scores for word similarity datasets created for six languages.

5.1.1 Fleiss' Kappa

Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. This contrasts with other kappas such as Cohen's kappa, which only work when assessing the agreement between not more than two raters or the interrater reliability for one appraiser versus himself. The measure calculates the degree of agreement in classification over that which would be expected by chance (Wikipedia contributors, 2017).

We have calculated Fleiss' Kappa for all our word similarity datasets (see table 1).

6 Result and Analysis

System	Score	OOV	Vocab
CBOW	28.30	19	130K
SG	34.40	19	130K
FastText	34.61	19	130K
FastText w/ OOV	45.47	14	-

Table 2: Results for **Urdu**

We present baseline scores using state of the art techniques – CBOW and Skipgram (Mikolov et al., 2013a) and FastText-SG (Bojanowski et al., 2016), evaluated using our word similarity datasets in tables 2, 3 and 4. As we can see the models trained encountered unseen word pairs when evaluated on their corresponding word similarity datasets. This goes on to show that all word pairs in our word similarity sets are not too common, and contain word pairs with some rarity.

System	Score	OOV	Vocab
CBOW	36.16	3	194K
SG	41.22	3	194K
FastText	33.68	3	194K
FastText w/ OOV	38.66	0	-

Table 3: Results for **Marathi**

System	Score	OOV	Vocab
CBOW	26.01	14	174K
SG	27.04	14	174K
FastText	34.29	14	174K
FastText w/ OOV	46.02	0	-

Table 4: Results for **Telugu**

We see that FastText w/ OOV (Out of Vocabulary) performed better than FastText in all the experiments, because character based models perform better than rest of the models since they are able to handle unseen words by generating word embeddings for missing words via character model.

7 Future Work

There are a lot of Indian languages that are still computationally resource poor even though they are widely spoken by significant number of people. Our work is a small step towards generating resources to further the research involving word representations on Indian languages.

To further extend our work, we will create rare-word word similarity datasets for six languages we worked on in this paper, and creating word similarity datasets for other major Indian languages as well.

We will also work on improving word representations for the languages we worked on, hence improve the baseline scores that we present here. This will require us to build new corpus to train our models for three languages that we couldn't provide baseline scores for – Punjabi, Tamil and Gujarati and build more corpus for Urdu, Telugu and Marathi to train better word embeddings.

References

Alon Lavie, and Michael J. Denkowski. 2009. *The METEOR metric for automatic evaluation of ma-*

chine translation. Machine translation 23, no. 2-3 (2009): 105-115.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov 2016. *Enriching word vectors with subword information*. arXiv preprint arXiv:1607.04606 (2016).

Bushra Jawaid, Amir Kamran, and Ondrej Bojar. 2014. *A Tagged Corpus and a Tagger for Urdu*. LREC 2014.

Camacho-Collados, Jos, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. *A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets*. In ACL (2) (pp. 1-7).

Colette Joubarne and Diana Inkpen. 2011. *Comparison of semantic similarity for different languages using the Google n-gram corpus and second-order co-occurrence measures*. In Advances in Artificial Intelligence – 24th Canadian Conference on Artificial Intelligence, pages 216-221.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. *Semeval-2014 task 3: Cross-level semantic similarity*. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), in conjunction with COLING. 2014.

Diana McCarthy, and Roberto Navigli. 2009. *The English lexical substitution task*. Language resources and evaluation 43, no. 2 (2009): 139-159.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. *Simlex-999: Evaluating semantic models with (genuine) similarity estimation*. Computational Linguistics.

George A. Miller, and Walter G. Charles. 1991. *Contextual correlates of semantic similarity*. Language and cognitive processes 6.1 (1991): 1-28.

Herbert Rubenstein and John B. Goodenough. 2006. *Contextual correlates of synonym*. Communications of the ACM, volume 8, number 10, pages 627-633.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman and Eytan Ruppin. 2002. *Placing search in context: The concept revisited*. Proceedings of the 10th international conference on World Wide Web, pages 406-414.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea.. 2011. *Learning to grade short answer questions using semantic similarity measures and dependency graph alignments*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 752-762). Association for Computational Linguistics.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. *Tailoring continuous word representations for dependency parsing*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, Baltimore, MD, USA, Volume 2: Short Papers, pages 809815.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. *Parsing With Compositional Vector Grammars*. In ACL, pages 455-465.

Roger Granada, Cassia Trojahn, and Renata Vieira. 2014. *Comparing semantic relatedness between word pairs in Portuguese using Wikipedia*. International Conference on Computational Processing of the Portuguese Language. Springer International Publishing, 2014.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. *Name tagging with word clusters and discriminative training*. In Proceedings of HLT-NAACL, volume 4, pages 337-342.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen 2005. *SenseRelate:: TargetWord: a generalized framework for word sense disambiguation*. Proceedings of the ACL 2005 on Interactive poster and demonstration sessions (pp. 73-76). Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeff Dean. 2006. *Efficient estimation of word representations in vector space*. In In arXiv preprint arXiv:1301.3781

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. *Linguistic regularities in continuous space word representations*. In Proceedings of HLT-NAACL, volume 13, pages 746-751.

Tomas Mikolov, Quoc V. Le, and lya Sutskever. 2013. *Exploiting similarities among languages for machine translation*. arXiv preprint arXiv:1309.4168 (2013).

Torsten Zesch and Iryna Gurevyc. 2006. *Automatically creating datasets for measures of semantic relatedness*. In Proceedings of the Workshop on Linguistic Distances, pages 16-24.

Wikipedia contributors. 2017. *"Fleiss' kappa." Wikipedia, The Free Encyclopedia.*. Wikipedia, The Free Encyclopedia, 6 Feb. 2017. Web. 16 Feb. 2017.