Let Sense Bags Do Talking: Cross Lingual Word Semantic Similarity for English and Hindi

Apurva Nagvenkar
DCST, Goa University
apurv.nagvenkar@gmail.com

Jyoti Pawar
DCST, Goa University
jyotidpawar@gmail.com

Pushpak Bhattacharyya
CSE, IIT Bombay
pb@cse.iitb.ac.in

Abstract

Cross Lingual Word Semantic (CLWS) similarity is defined as a task to find the semantic similarity between two words across languages. Semantic similarity has been very popular in computing the similarity between two words in same language. CLWS similarity will prove to be very effective in the area of Cross Lingual Information Retrieval, Machine Translation, Cross Lingual Word Sense Disambiguation, etc.

In this paper, we discuss a system that is developed to compute CLWS similarity of words between two languages, where one language is treated as resourceful and other is resource scarce. The system is developed using WordNet. The intuition behind this system is that, two words are semantically similar if their senses are similar to each other. The system is tested for English and Hindi with the accuracy 60.5% precision@1 and 72.91% precision@3.

1 Introduction

Word Semantic Similarity between two words is represented by the similarity between concepts associated with it. It plays a vital role in Natural Language Processing (NLP) and Information Retrieval (IR). In NLP (Sinha and Mihalcea, 2007), it is widely used in Word Sense Disambiguation, Question Answering system, Machine Translation (MT) etc. In IR (Hliaoutakis et al., 2006) it can be used in Image Retrieval, Multimodal Document Retrieval, Query Expansions etc.

The goal of CLWS Similarity is to measure the semantic similarity between the two words across languages. In this paper, we have proposed a system that computes CLWS similarity between two languages i.e. Language L_s and L_t, where L_s is treated as resourceful language and L_t is treated as resource scarce language. Given two words in language L_s and L_t the CLWS Similarity engine will analyze which two concepts of the word from L_s and L_t are similar. Let m & n be the number of concepts for the word in L_s & L_t respectively then, the output will generate a sorted list of all the possible combination of concepts (i.e. $m * n$ sorted list) ordered by their similarity score and the topmost combination of the concepts from L_s and L_t are similar to each other. The system is developed and tested for the English and Hindi language where English is L_s and Hindi is L_t.

1.1 Semantic Similarity

Lot of research effort has been devoted to design semantic similarity measures having monolingual as parameter. WordNet has been widely adopted in semantic similarity measures for English due its large hierarchical organization of synsets. Monolingual semantic similarity can be computed using Edge Counting and Information Content (IC). Edge counting is path based approach which makes use of knowledge bases. It measures the similarity by computing shortest distance from two concepts and the distance is nothing but a IS-A hierarchy. There are different path based measures such as Path Length Measure, Leacock and Chodorow Measure (Leacock and Chodorow, 1998), Wu & Palmer Measure (Wu and Palmer, 1994). IC is probabilistic based approach which makes use of corpus. It computes the negative logarithm of the probability of the occurrence of the concept in a large corpus. The different IC approaches are Resnik Measure (Resnik, 1995), Jiang Conrath Measure (Jiang and Conrath, 1997), Lin Measure (Lin, 1998), etc.

2 The Proposed Idea

The aim of our work is to design a CLWS similarity engine that computes similarity between two words across languages. So, given a word in English and Hindi the system will analyze which two synsets of the words are similar by making use of WordNet.

To obtain CLWS similarity we follow the following steps.

1. Given a word W_{EN} and W_{HN} and its senses must be present in their respective WordNets i.e. English and Hindi WordNet.

2. Let $S_{EN} = \{{s_{EN}}^1, {s_{EN}}^2 ... {s_{EN}}^m\}$ & $S_{HN} = \{{s_{HN}}^1, {s_{HN}}^2 ... {s_{HN}}^n\}$ be the set sense bags. The sense bags are obtained from its synset constituents i.e. content words from concept, examples, synonyms and hypernym (depth=1). We make sure that the words in sense bag must be present in the WordNet.

3. S_{HN} Hindi sense bags must be translated to resourceful language i.e. English. The translations are obtained by making use of bilingual dictionary or GIZA++ (Och and Ney, 2003) aligner that identifies the word alignment considering a parallel corpus.

We say that two words are semantically similar if

1. W_{EN} is compared with ${s_{EN}}^i$ and ${s_{HN}}^j$ then their score must be similar. i.e. $CLWS_w(W_{EN}, {s_{EN}}^i) \approx {CLWS_w}^{tr}(W_{EN}, {s_{HN}}^j)$ this is further explained in section 3.1.

2. If they have similar sense bags. i.e. $CLWS_s({s_{EN}}^i, {s_{HN}}^j) \approx 1$ & ${CLWS_s}^{weight}(w_{EN}, {s_{EN}}^i, {s_{HN}}^j) \approx 1$ this is further explained in section 3.2.1 and section 3.2.2.

3 CLWS Similarity Measures

Following are the different measures that are used to compute CLWS similarity score.

3.1 Measure 1: Word to Sense Bag Similarity Measure

In this measure, the source word W_{EN} is compared with the words present in the sense bag using monolingual similarity function i.e. equation (1). Where the parameters w_1 & w_2 are the words from the source (English) language and the parameter τ is the approach viz. obtained from either edge counting or information content.

$$sim(w_1, w_2, \tau) \quad (1)$$

3.1.1 Source Word to Source Sense Bag Similarity Measure

The source word i.e. W_{EN} is compared with the words present in the source sense bag ${s_{EN}}^i$ i.e. ${s_{EN}}^i$ = (ew_1,ew_2,..,ew_p). So, W_{EN} is compared with the first word ew_1 from ${s_{EN}}^i$ using monolingual similarity function i.e. $sim(W_{EN}, ew_1, \tau)$ the same procedure is applied for the other words from the source sense bag and we obtain the following measure.

$$CLWS_w(W_{EN}, {s_{EN}}^i) = \frac{1}{p}\sum_{k}^{p} sim(W_{EN}, ew_k, \tau) \quad (2)$$

3.1.2 Source Word to Target Sense Bag Similarity Measure

Here, the source word i.e. W_{EN} is compared with the words present in the target (Hindi) sense bag ${s_{HN}}^j$ i.e. ${s_{HN}}^j$ = (hw_1,hw_2,..,hw_q). So, W_{EN} is compared with the first word hw_1 from ${s_{HN}}^j$. In this case, hw_1 is replaced with its translation. If a word has more than one translation then maximum score between the translations is considered to be the winner candidate (equation (3)). The same procedure is applied for the other words from the target sense bag and we obtain the equation (4) for this measure.

$$sim_{tr}(W_{EN}, hw_j, \tau) = \max_{k=1}^{l} sim\left(W_{EN}, {ew_k}^{tr}, \tau\right)$$
$$where\ hw_j = ({ew_1}^{tr}, {ew_2}^{tr}, .., {ew_l}^{tr}) \quad (3)$$

$${CLWS_w}^{tr}(W_{EN}, {s_{HN}}^j) = \frac{1}{q}\sum_{k=1}^{q} sim_{tr}(W_{EN}, hw_k, \tau) \quad (4)$$

We say that, two (source and target) sense bags are similar to each other if they are related to the source word.

$$score_1 = 1 - \mid CLWS_w(W_{EN}, {s_{EN}}^i) - {CLWS_w}^{tr}(W_{EN}, {s_{HN}}^j) \mid \quad (5)$$

3.2 Sense Bag to Sense Bag Similarity Measure

In this measure, the source sense bag ${s_{EN}}^i$ is compared with target sense bag ${s_{HN}}^j$.

3.2.1 Measure 2: Sense Bag to Sense Bag Similarity without weight

In this measure, every word from ${s_{EN}}^i$ is compared with all the words from ${s_{HN}}^j$. For example, the first word ew_1 is compared with all the words from ${S_{HN}}^j$ using equation (3) and the maximum score is retrieved. The same process is continued for the remaining words present in the source sense bag to get equation (6).

$$CLWS_s({s_{EN}}^i, {s_{HN}}^j) = \frac{1}{p}\sum_{k=1}^{p}\left(\max_{l=1}^{q} sim_{tr}(ew_k, hw_l, \tau)\right) \quad (6)$$

$$score_2 = CLWS_s({s_{EN}}^i, {s_{HN}}^j) \quad (7)$$

3.2.2 Measure 3: Sense Bag to Sense Bag Similarity with weight

This is similar to the above measure (Section 3.2.1) but at every iteration the weights are assigned to the words that are compared. Every word from both the sense bags is assigned with a weight viz. obtained by comparing it with the source word i.e. W_{EN}, by using equation(1) & (3). This weight depicts how closely the words from sense bags are related to W_{EN}.

$$CLWS_s{}^{weight}(w_{EN}, {s_{EN}}^i, {s_{HN}}^j) = \frac{1}{p}\sum_{k=1}^{p} sim(w_{EN}, ew_k, \tau) \times \max_{l=1}^{q}\left(sim_{tr}\left(w_{EN}, hw_l, \tau\right) \times sim_{tr}\left(ew_k, hw_l, \tau\right)\right) \quad (8)$$

$$score_3 = CLWS_s{}^{weight}(w_{EN}, {s_{EN}}^i, {s_{HN}}^j) \quad (9)$$

3.3 Measure 4: Incorporating Monolingual Corpus

To check the frequency of a sense we compare the sense bag to a large monolingual corpora. A context bag (CB_{EN}), is obtained for a word W_{EN} by using word2vec[1] toolkit (Mikolov et al., 2013). CB_{EN} is compared with all the sense bags of S_{EN} (using Sense Bag to Sense Bag Similarity with weight since, it gives higher accuracy than others, refer section 5.3). This method will assign a similarity score to all the sense bags. The sense bag with most frequent usage will be assigned with high value than the sense bag with low frequent usage. This score is further multiplied with the similarity score obtained from ${s_{EN}}^i$ and ${s_{HN}}^j$.

$$score_4 = CLWS_s{}^{weight}(w_{EN}, {s_{EN}}^i, CB_{EN}) \times CLWS_s{}^{weight}(w_{EN}, {s_{EN}}^i, {s_{HN}}^j) \quad (10)$$

4 Resources and Dataset Used

Princeton WordNet (Miller et al., 1990) has 117791 synsets and 147478 words whereas, Hindi WordNet (Bhattacharyya, 2010) has 38782 synsets and 99435 words in its inventory. These wordNets are used to derive sense bags for each language. We have used publicly available pre-trained word embeddings for English which are trained on Google News dataset[2] (about 100 billion words). These word embeddings are available for around 3 million words and phrases. We also make use of ILCI parallel corpus[3] to obtain the word alignment of Hindi with respect to English for translating Hindi word to English. The size of ILCI corpus contains 50,000 parallel sentences.

To check the performance of our system we need to evaluate it against human judgment. Currently, synset linking task is carried out at CFILT[4], this task is carried out manually where the word-pairs (Hindi-English) across the languages having the similar senses are linked together. We take 2000 word pairs for development. Figure 1 shows the number of occurrences of degree of polysemy for English and Hindi words as well as the word pairs (i.e. the product of degree of polysemy for English and Hindi word) that are used in development of the system.

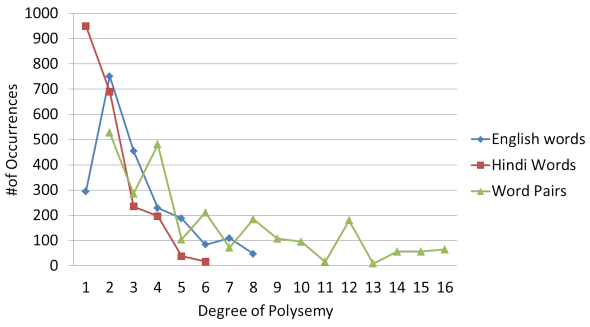

Figure 1: degree of polysemy v/s #of occurrences

[1] https://code.google.com/p/word2vec/

[2] Downloaded from https://code.google.com/p/word2vec/

[3] Downloaded from http://tdil-dc.in/

[4] http://www.cfilt.iitb.ac.in/

5 Experimental Results

For system evaluation we take 80% of word pairs for training and 20% of word pairs for testing. The system is evaluated for all the measures described above. To get the best possible accuracy we have to set following parameters at training:

- α - threshold value that will decide the size of sense bag
- τ - the monolingual similarity measures like PATH, JCN, LIN, WUP, etc.

5.1 The Baseline system

To date, there is no work carried out in computing the CLWS similarity for English and Hindi word pairs. So we define our own baseline system. The synsets in English and Hindi WordNet are organized based on their most frequent sense usage. As a baseline we say that given a word pair their most frequent sense is similar to each other. The baseline system makes an assumption by considering most frequent sense for both the word pairs. So, given a word in English and Hindi their most frequent sense are similar to each other.

5.2 The size of sense bag

The IC measures the specificity of the concept (Pedersen, 2010). The general concepts are assigned low value and specific concepts are assigned high value. In this scenario, we measure the specificity of the word instead of concept i.e. $IC(w) = -log(P(w))$. The IC value is computed for every word present in the sense bag. The size of sense bag depends on the IC threshold i.e. α. The sense bag will contain only those words with $IC(w) \geq \alpha$. Figure 2 shows how the size of the sense bag affects the performance of the system. In this, figure the α value is iterated from 0 to 13. When $\alpha = 0$ the sense bag will contain all the words but as α value increases to t it will contain only those word whose $IC(w) \geq t$. The system reported best performance when the size of sense bag was 6.

5.3 Results

The system is also evaluated by making use of bilingual dictionary and GIZA++ for word alignment. The reason behind this is that many of the under resource languages may not have well defined bilingual dictionary and therefore we used unsupervised approach i.e. word alignment.

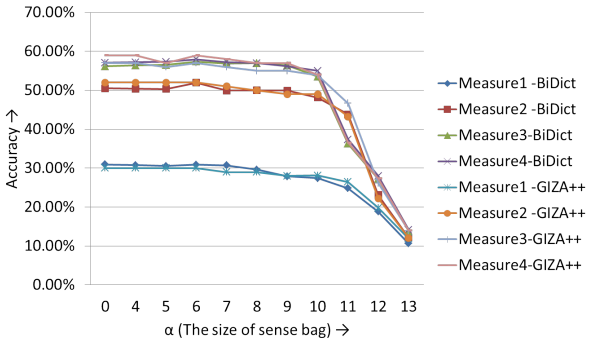

Figure 2: Accuracy of the CLWS Similarity when α value iterated from 0 to 13.

After training the CLWS similarity system, parameters were assigned with values α=6.0 and τ=RES.

	Precision					
Baseline	42%					
	Bilingual Dictionary			GIZA++		
	P@1	P@2	P@3	P@1	P@2	P@3
Measure1	25.5%	34.38%	45.0%	22.7%	28.21%	36.97%
Measure2	51.0%	64.91%	72.91%	42.2%	56.78%	63.44%
Measure3	58.5%	65.26%	73.3%	47.34%	56.78%	64.70%
Measure4	60.5%	65.26%	72.91%	52.65%	59.64%	69.74%

Table 1: Accuracy of CLWS Similarity measures computed from test dataset of 400 word pairs.

Table 1 contains the accuracy in terms of precision of the system computed from test dataset i.e. 400 word pairs. Form the table it is very clear that the system outperforms baseline system in most of the cases. The system performs best for Measure 4 with precision 60.5% .

6 Conclusion and Future Work

In this paper we presented several measures that are used to achieve the CLWS similarity. The main objective is to compute CLWS similarity for settings in which one language has many resources and the other is resource scarce. The system is tested for English-Hindi language pair and demonstrates which approach is better over the other. The accuracy of the system is further enhanced by making use of large monolingual corpora and Word2vec toolkit. We achieve 60.5% precision@1 and 72.91% precision@3 for English and Hindi word pairs.

In future we will like to implement the CLWS similarity for other resource scarce languages. CLWS similarity models are very much required for resource scarce languages but we need to think about the ways to reduce dependency of existing

resources by making use of mathematical modeling.

References

Pushpak Bhattacharyya. 2010. Indowordnet. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

Angelos Hliaoutakis, Giannis Varelas, Epimeneidis Voutsakis, Euripides G. M. Petrakis, and Evangelos Milios. 2006. Information retrieval by semantic similarity. In *Intern. Journal on Semantic Web and Information Systems (IJSWIS), 3(3):5573, July/Sept. 2006. Special Issue of Multimedia Semantics*.

J.J. Jiang and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.

C. Leacock and M. Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In Christiane Fellfaum, editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Ted Pedersen. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 329–332, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ravi Sinha and Rada Mihalcea. 2007. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *Proceedings of the International Conference on Semantic Computing*, ICSC '07, pages 363–369, Washington, DC, USA. IEEE Computer Society.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.