Leveraging Word Embedding from Macro and Micro View to Boost Performance for Semantic Textual Similarity

October 7, 2016

Outline

- Task Definition
- Our Systems
  - Preprocess
  - Traditional NLP Feature Engineering
  - Word Embedding Feature Engineering
- Experiments
- Results
- Conclusion

Task Definition

Definition (Semantic Textual Similarity)

Input: given two sentences
Output: similarity score([0,5])
Gold Standard: human judgements
Evaluation: Pearson correlation

Example

{ The bird is bathing in the sink.
Birdie is washing itself in the water basin. (sys: ? / gs: 5.0)

{ The woman is playing the violin.
The young lady enjoys listening to the guitar. (sys: ? / gs: 1.0)

Our Systems

Traditional NLP Feature Engineering

1. String-Based Similarity
2. Machine Translation Similarity
3. Corpus-based Features
4. Alignment Measures

Word Embedding Feature Engineering

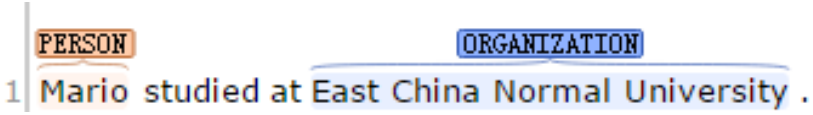
1. Word Centroid Distance
2. Word Mover's Distance

Preprocess

1. To Formal Writing (Search and Replace)
   e.g., doesn't → does not
2. Lemmatize (NLTK and Stanford CoreNLP)
   e.g., was → be
3. Parse (Stanford CoreNLP)

POS tag

NER tag



Traditional NLP Feature Engineering

1. String-Based Similarity
2. Machine Translation Similarity
3. Corpus-based Features
4. Alignment Measures

String-Based Similarity

- **Length Features (len):**

$$|A|, |B|, |A - B|, |B - A|, |A \cup B|, |A \cap B|, \frac{|A - B|}{|B|}, \frac{|B - A|}{|A|}$$

- **Syntactic Features (pos):**

$$|A_{pos}|, |B_{pos}|, |A_{pos} - B_{pos}|, |B_{pos} - A_{pos}|$$

$$|A_{pos} \cup B_{pos}|, |A_{pos} \cap B_{pos}|, \frac{|A_{pos} - B_{pos}|}{|B_{pos}|}, \frac{|B_{pos} - A_{pos}|}{|A_{pos}|}$$

- **Longest Common Sequence (lcs):**

$$\frac{|lcs(A, B)|}{min(|A|, |B|)}$$

String-Based Similarity

- **Ngrams Overlap Features (ngram):**
  1. word level (original and lemmatized) / character level.
  2. $n = \{1, 2, 3\}$ are used for the word level.
  3. $n = \{2, 3, 4\}$ are used for the character level.

- **Named Entities Features (ner):**

  location, organization, data, money, person, time, percent

Machine Translation Similarity

Machine Translation Similarity

1. Viewed as one input and one output of a MT system.
2. MT measures (i.e., WER, TER, PER, NIST, ROUGE-L, GTM-1)
3. Two strategies (i.e., average and concatenate)

Corpus-based Features

- **WordNet Rank Features (wordnet):**
  1. normalized ranking (sentence) vector.
  2. sentence vector distance: cosine, manhattan, Euclidean, Jaccard.

- **Vector Space Sentence Similarity (lsa):**
  1. Latent Semantic Analysis(LSA)
  2. New York Times Annotated Corpus(NYT) / Wikipedia
  3. convert to sentence level: sum up / use idf to weigh each word vector.

Alignment Measures

$$\begin{cases} \underline{\text{12 killed in bus accident in Pakistan.}} \\ \underline{\text{10 killed in road accident in NW Pakistan.}} \end{cases}$$

(sys: (2/3)*5 / gs: 3.2)

- **Global Alignment Features:**

$$sim(S_1, S_2) = \frac{n_a(S_1) + n_a(S_2)}{n(S_1) + n(S_2)}$$

- **POS-Specific Alignment Features:**

  calculate the aligned words proportion specifically according to POS tag(i.e., noun, verb, adjective, adverb).

Word Embedding Feature Engineering

1. Word Centroid Distance
2. Word Mover's Distance

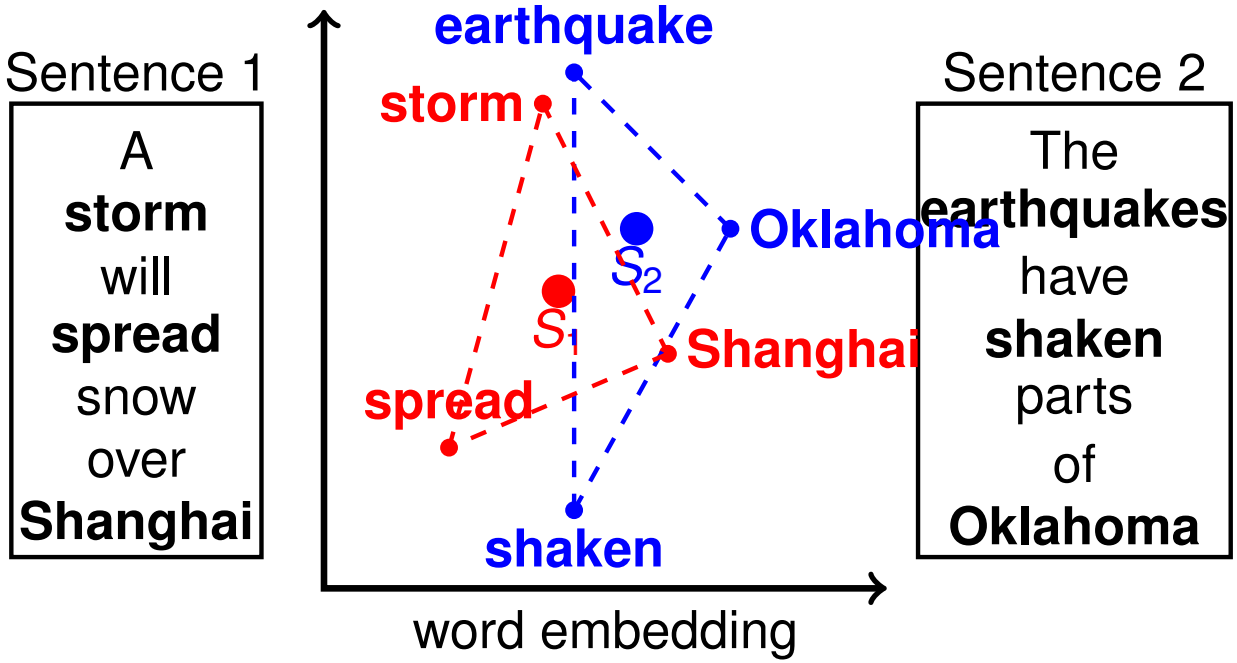


Figure: An illustration of the word centroid distance.

Word Embedding Feature Engineering

1. Word Centroid Distance
2. Word Mover's Distance

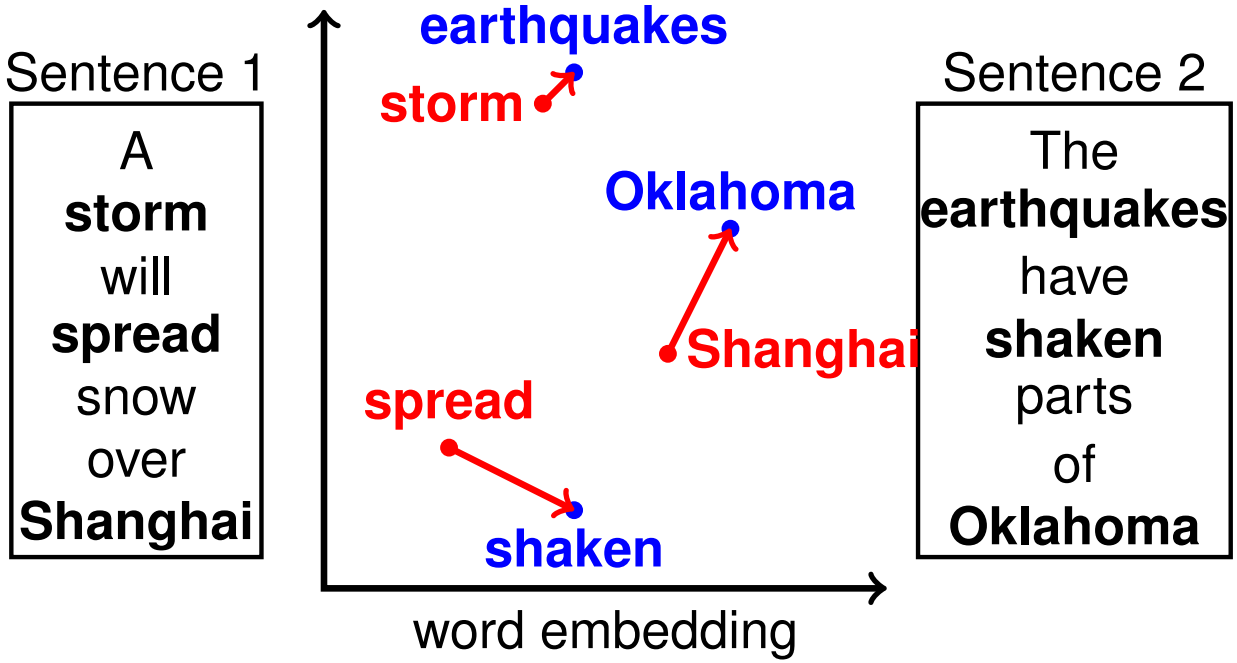


Figure: An illustration of the word mover's distance.

Word Embedding Feature Engineering

Word Embedding

1. word2vec
2. GloVe
3. C&W
4. Wiki

Distance Measurements

1. Cosine Distance
2. Manhattan Distance
3. Euclidean Distance
4. Pearson coefficient
5. Spearman coefficient
6. Kendall tau coefficient

Datasets

Training Set			Test Set		
Dataset	**Input**	**Gold**	**Dataset**	**Input**	**Gold**
MSRpar	1500	1500	answers-answers	1572	254
SMTeuroparl	1193	1193	plagiarism	1271	230
headlines*	3000	2250	headlines*	1498	249
SMTnews	399	399	postediting	3287	244
MSRvid	1500	1500	question-question	1555	209
OnWN	2061	2061	-	-	-
FNWN	189	189	-	-	-
images	2250	1500	-	-	-
deft-forum	450	450	-	-	-
deft-news	300	300	-	-	-
tweet-news	750	750	-	-	-
answers-forums	1500	375	-	-	-
answers-students	1500	750	-	-	-
belief	2000	375	-	-	-
All	19092	13592	All	9183	1186

Table: The statistics of all datasets for STS task.

Questions

Q_1 Supervised Model?
A: Learning Algorithm: SVM? RF? GB?

Q_2 Difference between Training Data and Test Data?
A: Training Data: All? Selected Data?

Q_3 Efficient Feature Set?
A: Feature Set: ? ?

Q_1 : Learning Algorithm

Regression	belief	answers -students	headlines	images	answers -forums	Weighted Mean
SVR(c=1)	0.7413	0.7359	0.8168	0.8660	0.7400	0.7898
RF(n=40)	0.7466	0.7100	0.8200	0.8534	0.7398	0.7816
GB(n=140)	**0.7655**	0.7484	**0.8439**	**0.8791**	**0.7469**	**0.8080**
DLSCU-S1	0.7491	**0.7725**	0.8250	0.8644	0.7390	0.8015

Table: Results of different algorithm on STS 2015 test data.

Q_2 : Training Data

Measurements

1. source
2. average length of sentences
3. word mover's distance

Training Data for STS 2016 test data

1. **headlines**: headlines
2. **answers-answers, question-question**: belief, deft-forums, answers-students, answers-forums
3. **postediting**: SMTeuropar, MSRpar
4. **plagiarism**: onWN, FNWN

Q_3 : Feature Selection

Feature		belief	answers -students	headlines	image	answers -forums
String-based	len	-	-	√	√	-
	pos	-	-	√	-	√
	lcs	-	-	√	√	√
	ngram	-	√	√	√	√
	ner	-	√	√	-	-
Machine Translation	average	√	-	√	√	-
	concat	-	-	-	-	-
Corpus-based	wordnet	√	√	√	-	√
	lsa	-	√	√	√	√
Alignment	global	-	√	√	√	√
	specific	√	√	√	√	√
Word Centroid Distance	word2vec	√	√	√	√	√
	glove	-	-	-	-	-
	turian's	√	√	√	√	-
Word Mover's Distance	wmd	√	√	√	√	√
Our Results		**0.7835**	0.7713	**0.8455**	**0.8808**	**0.7636**
Best Scores		0.7717	**0.7879**	0.8417	0.8713	0.7390

Table: Results of feature selection experiments on STS 2015 test data.

Setups

- U-SEVEN:
  1. longest common sequence
  2. alignment feature
  3. corpus-based feature
  4. word centroid distance from from four word embedding. cosine distance, Pearson coefficient, Spearman coefficient.
- S1-All
  1. all the training datasets
  2. regression model: GB(n=140)
  3. feature selection: hill climbing
- S2
  1. selected training datasets
  2. regression model: GB(n=140)
  3. feature selection: hill climbing

Results

Dataset	Runs			Best Score
	U-SEVEN	S1-All	S2	
answers-answers	0.4774	0.5697	**0.5715**	0.6923
plagiarism	**0.8301**	0.8250	0.7733	0.8413
headlines	0.7668	**0.8121**	0.7903	0.8274
postediting	**0.8423**	0.8234	0.7496	0.8669
question-question	0.7191	**0.7311**	0.6763	0.7470
weighted mean	0.7242	**0.7507**	0.7116	0.7780

Table: The results of our three runs on STS 2016 test datasets.

Results

{ You should do it.
  You can do it, too. } (sys: 4.0 / gs: 1.0)

{ It's pretty much up to you.
  It's much better to ask. } (sys: 3.2 / gs: 0.0)

Conclusion

1. The difference between top system and our best system is about 2.8%
2. Future work: Deep Learning