Semantic Textual Similarity

TIAN Junfeng

September 30, 2017

Task Definition

Definition (Semantic Textual Similarity)

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

Example

$$\begin{cases} \text{The bird is bathing in the sink.} \\ \text{Birdie is washing itself in the water basin.} \end{cases}$$ (sys: ? / gs: 5.0)

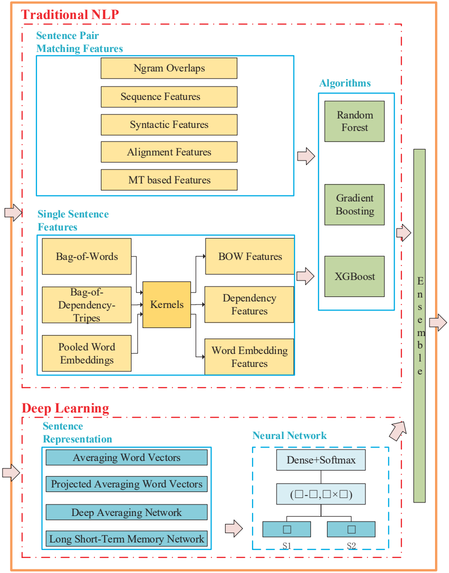
$$\begin{cases} \text{The woman is playing the violin.} \\ \text{The young lady enjoys listening to the guitar.} \end{cases}$$ (sys: ? / gs: 1.0)

Examples

5	*The two sentences are completely equivalent, as they mean the same thing.*
	The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. He is not a suspect anymore. John said.
2	*The two sentences are not equivalent, but share some details.*
	They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*
	The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

Outline

- Task Definition
- Our Systems
  - Traditional NLP
  - Deep Learning
  - Ensemble
- Experiments
- Results
- Conclusion



Sentence Matching Features (I / V)

N-grams Overlap

$$\text{ngo}(S_1, S_2) = 2 \cdot (\frac{|S_1|}{|S_1 \cap S_2|} + \frac{|S_2|}{|S_1 \cap S_2|})^{-1}$$

- word level (original and lemmatized) / character level.
- $n = \{1, 2, 3\}$ are used for the word level.
- $n = \{2, 3, 4, 5\}$ are used for the character level.

Remark (Other Coefficient)

dice coefficient: $2\frac{|A \cap B|}{|A|+|B|}$ *jaccard coefficient:* $\frac{|A \cap B|}{|A \cup B|}$
etc.

Sentence Matching Features (II / V)

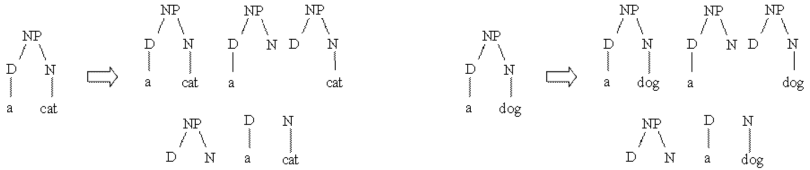
Sequence Features

- longest common substring / subsequence
- edit distance
- longest common prefix / suffix

Sentence Matching Features (III / V)

Syntactic Parse Features

Tree Kernels to calculate the similarity between two syntactic parse trees. (i.e., subtree (ST), subset tree (SST), partial tree (PT).)



Sentence Matching Features (IV / V)

Alignment Features

$$\left\{\begin{array}{l} \textit{12 killed in bus accident in Pakistan.} \\ \textit{10 killed in road accident in NW Pakistan.} \end{array}\right. \qquad \text{(sys: 3.3 / gs: 3.2)}$$

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

Remark

Not all alignments equal the same!

Sentence Matching Features (V / V)

MT based Features

1. Viewed as one input and one output of a MT system.
2. MT measures (e.g., BLEU, NIST, ROUGE-L, WER)

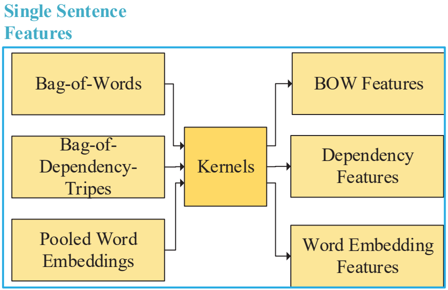
Single Sentence Features

- **Bag-of-Words**
  each word (i.e., dimension) is weighted by its IDF value.
- **Bag-of-Dependency-Triples**
- **Pooled Word Embeddings**
  concatenate min/max/average pooling of word vectors.

Remark

the dimensionality of single sentence features is huge, it would suppress the discriminating power of sentence pair matching features.

Single Sentence Features

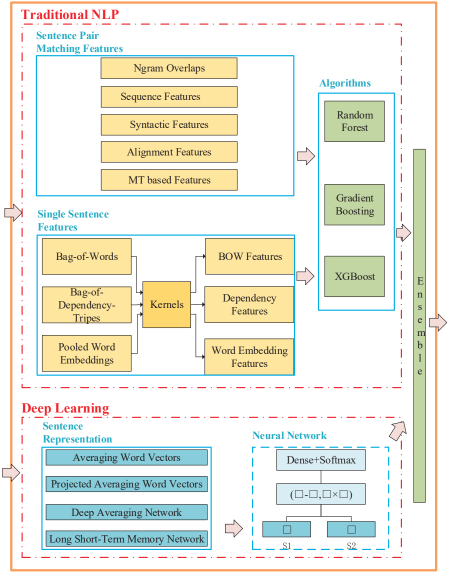


Type	Measures
linear kernel	Cosine distance, Manhanttan distance, Euclidean distance, Chebyshev distance
stat kernel	Pearson coefficient, Spearman coefficient, Kendall tau coefficient
non-linear kernel	polynomial, rbf, laplacian, sigmoid

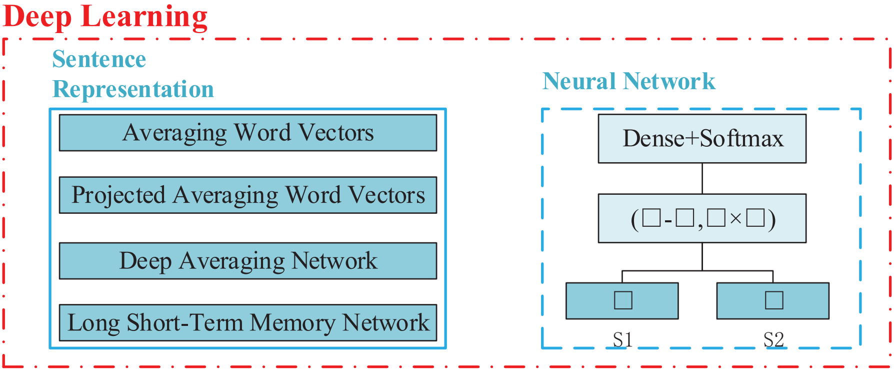
Table: List of 11 kernel functions

Outline

- Task Definition
- Our Systems
  - Traditional NLP
  - Deep Learning
  - Ensemble
- Experiments
- Results
- Conclusion

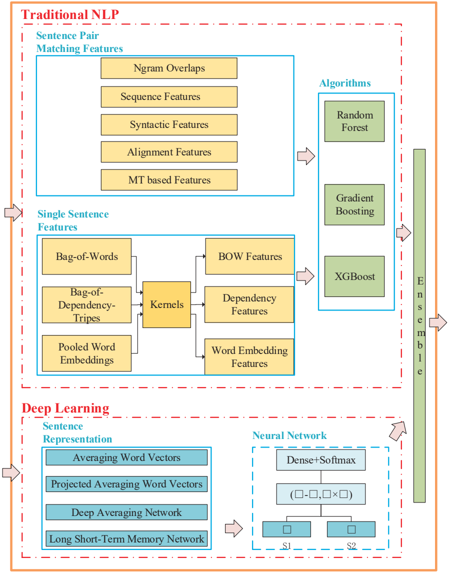


Deep Learning



Ensemble

The NLP-based scores and the deep learning based scores are **averaged** in the ensemble module to obtain the final score.



Datasets

Training set:

SemEval STS task (2012-2015): 13,592 sentence pairs.

Development and Test set:

Track	Language Pair	Development		Test
		Pairs	Dataset	Pairs
Track 1	Arabic-Arabic (AR-AR)	1088	MSRpar, MSRvid, SMTeuroparl (2017)	250
Track 2	Arabic-English (AR-EN)	2176	MSRpar, MSRvid, SMTeuroparl (2017)	250
Track 3	Spanish-Spanish (SP-SP)	1555	News, Wiki (2014, 2015)	250
Track 4a	Spanish-English (SP-EN)	595	News, Multi-source (2016)	250
Track 4b	Spanish-English WMT news data (SP-EN-WMT)	1000	WMT (2017)	250
Track 5	English-English (EN-EN)	1186	Plagiarsism, Postediting, Ans.-Ans., Quest.-Quest., HDL (2016)	250
Track 6	English-Turkish (EN-TR)	-	-	500

Table: The statistics of development and test set.

Experiments

Comparison of NLP Features

Table: Feature analysis on English STS 2016 datasets, the last three rows are the top three systems in that year.

English STS 2016						
Features	**Postediting**	**Ques.-Ques.**	**HDL**	**Plagiarism**	**Ans.-Ans.**	**Weighted mean**
BOW features	0.8388	0.6577	0.7338	0.7817	0.6302	0.7322
Alignment Features	0.8125	0.6243	0.7642	0.7883	**0.6432**	0.7312
Ngram Overlaps	0.8424	0.5864	0.7581	0.8070	0.5756	0.7203
Sequence Features	**0.8428**	0.6115	0.7337	0.7983	0.4838	0.7000
Word Embedding Features	0.8128	0.6378	0.7625	0.7955	0.4598	0.6992
MT based Features	0.8412	0.5558	0.7259	0.7617	0.5084	0.6851
Dependency Features	0.7264	0.5381	0.4634	0.5820	0.3431	0.5328
Syntactic Parse Features	0.5773	0.0846	0.4940	0.3976	0.0775	0.3376
All Features	0.8357	**0.6967**	**0.7964**	**0.8293**	0.6306	**0.7618**
Rychalska et al. (2016)	0.8352	0.6871	**0.8275**	**0.8414**	**0.6924**	**0.7781**
Brychcin and Svoboda (2016)	0.8209	0.7020	0.8189	0.8236	0.6215	0.7573
Afzal et al. (2016)	**0.8484**	**0.7471**	0.7726	0.8050	0.6143	0.7561

Experiments

Comparison of Learning Algorithms

Table: Algorithms comparison on English STS 2016 datasets

English STS 2016								
Algorithm		**Postediting**	**Ques.-Ques.**	**HDL**	**Plagiarism**	**Ans.-Ans.**	**Weighted mean**
Single Model	RF	0.8394	0.6858	0.7966	0.8259	0.5882	0.7518
	GB	0.8357	0.6967	0.7964	0.8293	0.6306	0.7618
	XGB	0.7917	0.6237	0.7879	0.8175	0.6190	0.7333
	DL-word	0.8097	0.6635	0.7839	0.8003	0.5614	0.7283
	DL-proj	0.7983	0.6584	0.7910	0.7892	0.5573	0.7234
	DL-dan	0.7695	0.4200	0.7411	0.6876	0.4756	0.6274
	DL-lstm	0.7864	0.5895	0.7584	0.7783	0.5182	0.6921
Ensemble	RF+GB+XGB	0.8298	0.6969	0.8086	0.8313	0.6234	0.7622
	DL-all	0.8308	0.6817	0.8160	0.8261	0.5854	0.7528
	EN-seven	**0.8513**	**0.7077**	**0.8288**	**0.8515**	**0.6647**	**0.7851**

Results on Test Data

Table: The results of our three runs on STS 2017 test datasets. Baseline is provided by the organizer, using cosine similarity of one-hot vector representations of sentence pairs.

Run	**Primary**	**Track 1** AR-AR	**Track 2** AR-EN	**Track 3** SP-SP	**Track 4a** SP-EN	**Track 4b** SP-EN-WMT	**Track 5** EN-EN	**Track 6** EN-TR
Run 1: RF	0.6940	0.7271	0.6975	0.8247	0.7649	0.2633	0.8387	0.7420
Run 2: GB	0.7044	0.7380	0.7126	0.8456	0.7495	0.3311	0.8181	0.7362
Run 3: EN-seven	**0.7316**	**0.7440**	**0.7493**	**0.8559**	**0.8131**	**0.3363**	**0.8518**	**0.7706**
Rank 2: BIT	0.6789	0.7417	0.6965	0.8499	0.7828	0.1107	0.8400	0.7305
Rank 3: HCTI	0.6598	0.7130	0.6836	0.8263	0.7621	0.1483	0.8113	0.6741
Baseline	0.5370	0.6045	0.5155	0.7117	0.6220	0.0320	0.7278	0.5456

Results on Test Data

Table: Case Study (DL V.S. NLP)

	Human	DL	NLP
A kid is talking in class. A girl is going to class.	1.80	2.55	3.35
There is a young girl. There is a young boy with the woman.	1.00	2.95	3.7
Friends walk into a building. A man walks along walkway to the s-tore.	0.40	2.80	1.25
A boy does a skateboard trick on the stairs downtown. A boy is running on the sidewalk.	1.20	2.60	1.50

Results on Test Data

Table: Difficult Sentence Pairs

Examples	Human	Ours
word sense disambiguation, making and preparing are very similar in the context of food		
There is a cook preparing food. A cook is making food.	5.0	4.1
attribute importance, outside and deserted are minor details		
The man is in a deserted field. The man is outside in the field.	4.0	3.1
compositional meaning		
A man is carrying a canoe with a dog. A dog is carrying a man in a canoe.	1.8	4.7
negation systems score		
A girl in water without goggles or a swimming cap. A girl in water, with goggles and swimming cap.	3.0	4.6
semantic blending		
There is a young girl. There is a young boy with the woman.	1.0	3.3

Conclusion

1. both the traditional NLP methods and the deep learning methods make contribution to performance improvement.
2. Current work:
    - Low-resource language STS without MT.
    - Cross-lingual STS