IHLT
STS Project

Ferran Agulló, Adrián Tormos

Índex

1. Preprocessing
2. Features
3. Regression models
4. Explored runs
5. Results

Índice

1. Preprocessing
2. Features
3. Regression models
4. Explored runs
5. Results

Preprocessing

1. Lowercasing
2. Tokenizing
3. Removing punctuation
4. Tagging parts of speech
5. Lemmatizing
6. Word sense disambiguation (Lesk algorithm)
7. Compute Tf-idf

Preprocessing

1. Lowercasing
2. Tokenizing
3. Removing punctuation
4. Tagging parts of speech
5. Lemmatizing
6. Word sense disambiguation (Lesk algorithm)
7. Compute Tf-idf

The original and the results of steps 1, 4, 5, 6 and 7 are stored for each sentence.

Índice

1. Preprocessing
2. Features
3. Regression models
4. Explored runs
5. Results

Explored features

Explored lexical features:

- Word, stopword, stem and character n-gram overlaps
- Word n-gram weighted overlap
- Length of longest common substring and subsequence
- Tf-idf cosine similarity
- Machine translation similarity*

Explored features

Explored lexical features:

- Word, stopword, stem and character n-gram overlaps
- Word n-gram weighted overlap
- Length of longest common substring and subsequence
- Tf-idf cosine similarity
- Machine translation similarity*

Other explored features:

- Sentence length difference
- Number overlap

Explored features

Explored lexical features:

- Word, stopword, stem and character n-gram overlaps
- Word n-gram weighted overlap
- Length of longest common substring and subsequence
- Tf-idf cosine similarity
- Machine translation similarity*

Other explored features:

- Sentence length difference
- Number overlap

Explored syntactic features:

- Part of speech n-gram overlap
- Dependency overlap
- Difference in amount of phrases*
- Syntactic role similarity*

Explored features

Explored lexical features:

- Word, stopword, stem and character n-gram overlaps
- Word n-gram weighted overlap
- Length of longest common substring and subsequence
- Tf-idf cosine similarity
- Machine translation similarity*

Other explored features:

- Sentence length difference
- Number overlap

Explored syntactic features:

- Part of speech n-gram overlap
- Dependency overlap
- Difference in amount of phrases*
- Syntactic role similarity*

Explored semantic features:

- Content n-gram overlap
- Wordnet-based pairwise word similarity
- Wordnet-based weighted pairwise word similarity

Individiual correlations

Correlations between the golden standard and each feature:

Word 1-gram	0.52	Stopw. 4-gram	0.24	**Tf-idf cos.**	**0.73**
Word 2-gram	0.39	Stopw. 5-gram	0.22	W-Net L-Ch	0.28
Word 3-gram	0.32	PoS 1-gram	0.02	W-Net 5s L-Ch	0.20
Wgt. W. 1-gram	**0.66**	PoS 2-gram	0.09	W-Net Path	0.38
Stem 1-gram	0.53	**Char. 2-gram**	**0.69**	W-Net 5s Path	0.55
Cont. 1-gram	0.50	**Char. 3-gram**	**0.62**	W-Net W. L-Ch	0.45
Cont. 2-gram	0.31	Char. 4-gram	0.58	W-Net W. Path	0.52
Cont. 3-gram	0.32	Char. 5-gram	0.55	Number overlap	0.19
Cont. 4-gram	0.27	Char. 6-gram	0.52	S. length	0.07
Stopw. 1-gram	0.01	Char. 7-gram	0.40	LC Subseq.	0.52
Stopw. 2-gram	0.09	Char. 8-gram	0.49	LC Substr.	0.44
Stopw. 3-gram	0.11	Char. 9-gram	0.47	Dependency	0.43

Índice

1. Preprocessing
2. Features
3. Regression models
4. Explored runs
5. Results

Regression models

- Support Vector Regression (RBF kernel, grid search on C and γ)
- Kernel Ridge Regression (RBF kernel, grid search on α and γ)
- AdaBoost (grid search on estimators and learning rate)
- Random Forest (grid search on minimum amount of samples to keep splitting)
- Voting Method (RBF SVR, RBF KRR, AdaBoost, Random Forest and Extra-Trees)

Índice

1. Preprocessing
2. Features
3. Regression models
4. Explored runs
5. Results

Runs

Lexical run:

- Word n-gram overlap ($n \in \{1, ..., 3\}$)
- Word n-gram weighted overlap ($n = 1$)
- Stopword n-gram overlap ($n \in \{1, ..., 5\}$)
- Stem n-gram overlap ($n = 1$)
- Character n-gram overlap ($n \in \{2, ..., 9\}$)
- Length of longest common substring and subsequence
- Tf-idf cosine similarity

Runs

Lexical run:

- Word n-gram overlap ($n \in \{1, ..., 3\}$)
- Word n-gram weighted overlap ($n = 1$)
- Stopword n-gram overlap ($n \in \{1, ..., 5\}$)
- Stem n-gram overlap ($n = 1$)
- Character n-gram overlap ($n \in \{2, ..., 9\}$)
- Length of longest common substring and subsequence
- Tf-idf cosine similarity

Semantic run:

- Content n-gram overlap ($n \in \{1, ..., 4\}$)
- Wordnet-based pairwise word similarity (Leacock-Chodrow, Wu-Palmer and path sims.; 1 and 5 best synsets)
- Wordnet-based weighted pairwise word similarity (Leacock-Chodrow, Wu-Palmer and path sims.)

Runs

Syntactic run:

- Part of speech n-gram overlap ($n \in \{1, 2\}$)
- Dependency overlap

Runs

Syntactic run:

- Part of speech n-gram overlap ($n \in \{1, 2\}$)
- Dependency overlap

Best features run (with and without PCA):

- Word n-gram overlap ($n \in \{1, ..., 4\}$)
- Content n-gram overlap ($n \in \{1, ..., 3\}$)
- Stopword n-gram overlap ($n \in \{1, ..., 5\}$)
- Part of speech n-gram overlap ($n \in \{1, 2\}$)
- Stem n-gram overlap ($n = 1$)
- Character n-gram overlap ($n \in \{2, ..., 9\}$)
- Tf-idf cosine similarity
- Sentence length difference
- Wordnet-based pairwise word similarity (unweighted for 1 and 5 best synsets, weighted for best synset; L-Ch. and path sims.)
- Length of longest common substring
- Number overlap

Índice

1. Preprocessing
2. Features
3. Regression models
4. Explored runs
5. Results

Results

	d	SVR	KRR	AdaB	RF	Voting
Lexical	21	.7423	.7193	.7256	.7539	.7486
Semantic	10	.6585	.6622	.6640	.7105	.7013
Syntactic	3	.4493	.4200	.4576	.4744	.5075
Best (PCA)	$33 \rightarrow 16$	.7320	.6708	.7135	.7347	.7482
Best	33	.7444	.7381	.7369	**.7685**	**.7672**

Comparison with original STS

	d	Score
baer/task6-UKP-run2_plus_postprocessing_smt_twsi	21	.8239
jan_snajder/task6-takelab-syntax	31	.8138
jan_snajder/task6-takelab-simple	14	.8133
baer/task6-UKP-run1	19	.8117
rada/task6-UNT-IndividualRegression	17	.7846
mheilman/task6-ETS-PERPphrases	36	.7834
mheilman/task6-ETS-PERP	36	.7808
baer/task6-UKP-run3_plus_random	22	.7790
ferranagullolopez/ihlt-sts_best-RF	**33**	**.7685**
rada/task6-UNT-IndividualDecTree	17	.7677
ferranagullolopez/ihlt-sts_best-VS	**33**	**.7672**
yeh/task6-SRIUBC-SYSTEM2	> 18	.7562