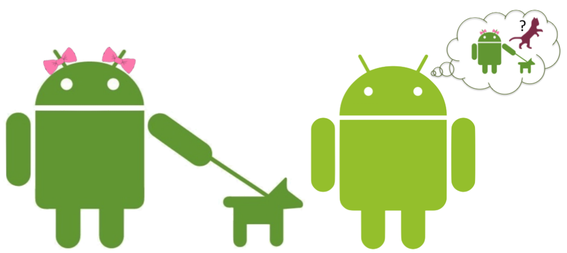
CPN-CORE: A Text Semantic Similarity System Infused with Opinion Knowledge

[♭]C. Banea, [‡]Y. Choi, [‡]L. Deng, [§]S. Hassan, [◇]M. Mohler, [√]B. Yang, [√]C. Cardie, [‡]J. Wiebe, and [♭]R. Mihalcea

[♭]University of North Texas [‡]University of Pittsburgh [§]Google Inc. [◇]Language Computer Corp. [√]Cornell University

Contribution

STS has traditionally focused on corpus and knowledge-based methods to compute similarity. We posit that textual similarity also has an opinion component which should be taken into account for a correct interpretation.



She owns a dog.

I believe she owns a dog.

Semantic Text Similarity

- Allows a computer to establish a measure of similarity or relatedness between two text fragments, even when they do not share common words.
- Central task in Natural Language Processing; used in information retrieval, relevance feedback and text classification, word sense disambiguation, extractive summarization, automatic evaluation of machine translation, text summarization, text coherence, and in plagiarism detection.
- **Training data**: SemEval 2012 data.
- **Test data**: text pairs extracted from headlines (*headlines*, 750 pairs), sense definitions from WordNet and OntoNotes (*OnWN*, 561 pairs), sense definitions from WordNet and FrameNet (*FNWN*, 189 pairs), and data used in the evaluation of machine translation systems (*SMT*, 750 pairs).

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340 and IIS awards #1018613, #0208798 and #0916046, as well as by DARPA-BAA-12-47 DEFT grant #12475008. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

Machine Learning Features

Knowledge-based features

Similarity scores obtained based on WordNet.

PATH shortest path
LCH Leacock & Chodorow (1998)
Lesk Lesk (1986)
WUP Wu & Palmer (1994)
RES Resnik (1995)
JCN Jiang and Conrath (1997)

Corpus-based features

Similarity scores obtained based on building word-concept vectorial models on Wikipedia.

LSA Latent Semantic Analysis (Landauer et al., 1997); implicit concepts obtained after a dimensionality reduction operation (SVD).
RP Random Projection (Dasgupta, 1999); concepts obtained by projecting to a random lower-dimension space.
ESA Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007); expresses the semantic profile of a word using explicit concepts leveraged from Wikipedia's structure.
SSA Salient Semantic Analysis (Hassan and Mihalcea, 2011); uses *salient* concepts, where a "concept" is an expression which affords an encyclopedic definition. *Saliency* is determined based on the word being hyperlinked in context, implying that it is highly relevant to the given text.

Opinion aware features

Features obtained using OpinionFinder (Wilson et al., 2005).

SUBJSL (per sentence pair) whether both sentences are classified as subjective or objective
NUMEX1 number of subjective expressions in the first sentence
NUMEX2 number of subjective expression in the second sentence
EXPR (per sentence pair) the number of tokens the subjective expressions in each sentence share

Features obtained using a logistic regression classifier trained on the MPQA corpus.

SUBJDIFF the difference in probability between the two sentences being subjective

Features obtained using the opinion extraction model by Yang and Cardie (2012).

SUBJCNT binary feature which equals 1 if both sentences contain a subjective expression
DSEALGN number of shared words in the subjective expressions in two sentences
DSESIM similarity of subjective expressions in two sentences
AGENT for all subjective expressions in a sentence pair, the number of tokens shared by their agents

Results

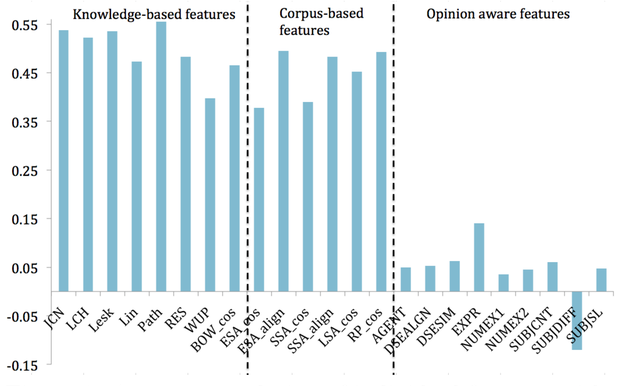


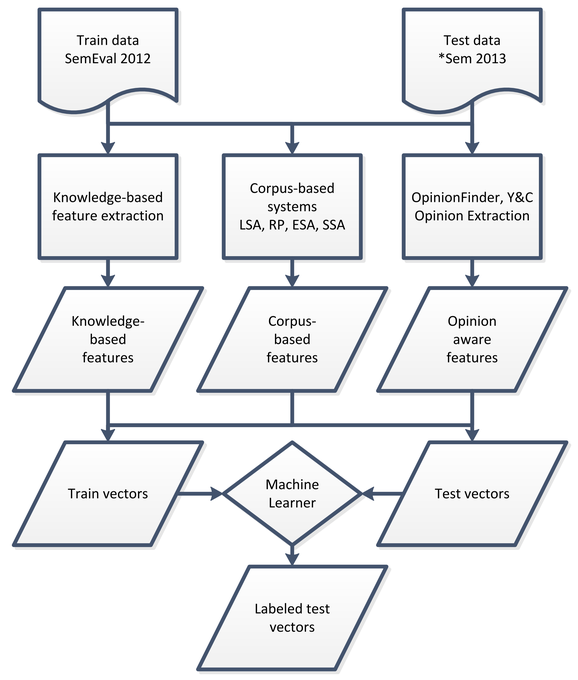
Figure 1: Average correlation of individual features with the gold standard across the test datasets.

System	FNWN	Headlines	OnWN	SMT	Mean
comb.RandSubSpace	0.331	**0.677**	0.514	0.337	**0.494**
comb.SVR	**0.362**	0.669	0.51	**0.341**	**0.494**
indv.RandSubspace	0.331	**0.677**	**0.548**	0.277	0.483
baseline-tokencos	0.215	0.54	0.283	0.286	0.364

Table 1: Evaluation and ranking as published by the task organizers.

- RP achieves competitive results with LSA, while also being computationally efficient.
- Lower correlations for the subjectivity features seem to be associated with shorter spans of text (*MSRvid*, *OnWN*, *headlines*).
- Training and testing on the same type of data achieves the best results (*OnWN*).

Meta-learning Framework



Conclusion

- Our system ranked 38, 39 and 45 among the 88 participating systems.
- Corpus based measures have a similar average performance with knowledge-based methods.
- Despite being the simplest knowledge-based metric we computed, *PATH* has the highest average correlation across the datasets.
- Among corpus-based metrics, *ESA*, *SSA* and *RP* are the top contenders.
- Among opinion aware features, *EXPR* reaches the highest average correlation at 0.15. It computes the overlap across subjective expressions.
- Opinion-based measures exhibit a low performance on the test datasets. However, these datasets do not display a consistent opinion content, nor were they annotated with this aspect in mind.

Contact information

Carmen Banea
carmen.banea@gmail.com