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

Carmen Banea[♭*], Yoonjung Choi[‡], Lingjia Deng[‡], Samer Hassan[§], Michael Mohler[◇]
Bishan Yang[✓], Claire Cardie[✓], Rada Mihalcea[♭†], Janyce Wiebe[‡]

[♭]University of North Texas
Denton, TX

[‡]University of Pittsburgh
Pittsburgh, PA

[§]Google Inc.
Mountain View, CA

[◇]Language Computer Corp.
Richardson, TX

[✓]Cornell University
Ithaca, NY

Abstract

This article provides a detailed overview of the CPN text-to-text similarity system that we participated with in the Semantic Textual Similarity task evaluations hosted at *SEM 2013. In addition to more traditional components, such as knowledge-based and corpus-based metrics leveraged in a machine learning framework, we also use opinion analysis features to achieve a stronger semantic representation of textual units. While the evaluation datasets are not designed to test the similarity of opinions, as a component of textual similarity, nonetheless, our system variations ranked number 38, 39 and 45 among the 88 participating systems.

1 Introduction

Measures of text similarity have been used for a long time in applications in natural language processing and related areas. One of the earliest applications of text similarity is perhaps the vector-space model used in information retrieval, where the document most relevant to an input query is determined by ranking documents in a collection in reversed order of their angular distance with the given query (Salton and Lesk, 1971). Text similarity has also been used for relevance feedback and text classification (Rocchio, 1971), word sense disambiguation (Lesk, 1986; Schutze, 1998), and extractive summarization (Salton et al., 1997), in the automatic evaluation of machine translation (Papineni et al., 2002), text summarization (Lin and Hovy, 2003), text coherence (Lapata and Barzilay, 2005) and in plagiarism detection (Nawab et al., 2011).

Earlier work on this task has primarily focused on simple lexical matching methods, which produce a similarity score based on the number of lexical units that occur in both input segments. Improvements to this simple method have considered stemming, stopword removal, part-of-speech tagging, longest subsequence matching, as well as various weighting and normalization factors (Salton and Buckley, 1997). While successful to a certain degree, these lexical similarity methods cannot always identify the *semantic* similarity of texts. For instance, there is an obvious similarity between the text segments "she owns a dog" and "she has an animal," yet these methods will mostly fail to identify it.

More recently, researchers have started to consider the possibility of combining the large number of word-to-word semantic similarity measures (e.g., (Jiang and Conrath, 1997; Leacock and Chodorow, 1998; Lin, 1998; Resnik, 1995)) within a semantic similarity method that works for entire texts. The methods proposed to date in this direction mainly consist of either bipartite-graph matching strategies that aggregate word-to-word similarity into a text similarity score (Mihalcea et al., 2006; Islam and Inkpen, 2009; Hassan and Mihalcea, 2011; Mohler et al., 2011), or data-driven methods that perform component-wise additions of semantic vector representations as obtained with corpus measures such as latent semantic analysis (Landauer et al., 1997), explicit semantic analysis (Gabrilovich and Markovitch, 2007), or salient semantic analysis

*carmen.banea@gmail.com
†rada@cs.unt.edu

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 221–228, Atlanta, Georgia, June 13-14, 2013.

(Hassan and Mihalcea, 2011).

In this paper, we describe the system variations with which we participated in the *SEM 2013 task on semantic textual similarity (Agirre et al., 2013). The system builds upon our earlier work on corpus-based and knowledge-based methods of text semantic similarity (Mihalcea et al., 2006; Hassan and Mihalcea, 2011; Mohler et al., 2011; Banea et al., 2012), while also incorporating opinion aware features. Our observation is that text is not only similar on a semantic level, but also with respect to opinions. Let us consider the following text segments: "she owns a dog" and "I believe she owns a dog." The question then becomes how similar these text fragments truly are. Current systems will consider the two sentences semantically equivalent, yet to a human, they are not. A belief is not equivalent to a fact (and for the case in point, the person may very well have a cat or some other pet), and this should consequently lower the relatedness score. For this reason, we advocate that STS systems should also consider the opinions expressed and their equivalence. While the *SEM STS task is not formulated to evaluate this type of similarity, we complement more traditional corpus and knowledge-based methods with opinion aware features, and use them in a meta-learning framework in an arguably first attempt at incorporating this type of information to infer text-to-text similarity.

2 Related Work

Over the past years, the research community has focused on computing semantic relatedness using methods that are either knowledge-based or corpus-based. Knowledge-based methods derive a measure of relatedness by utilizing lexical resources and ontologies such as WordNet (Miller, 1995) to measure definitional overlap, term distance within a graphical taxonomy, or term depth in the taxonomy as a measure of specificity. We explore several of these measures in depth in Section 3.3.1. On the other side, corpus-based measures such as Latent Semantic Analysis (LSA) (Landauer et al., 1997), Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007), Salient Semantic Analysis (SSA) (Hassan and Mihalcea, 2011), Pointwise Mutual Information (PMI) (Church and Hanks, 1990), PMI-IR (Turney, 2001), Second Order PMI (Islam and Inkpen, 2006), Hyperspace Analogues to Language (Burgess et al., 1998) and distributional similarity (Lin, 1998) employ probabilistic approaches to decode the semantics of words. They consist of unsupervised methods that utilize the contextual information and patterns observed in raw text to build semantic profiles of words. Unlike knowledge-based methods, which suffer from limited coverage, corpus-based measures are able to induce a similarity between any given two words, as long as they appear in the very large corpus used as training.

3 Semantic Textual Similarity System

3.1 Task Setup

The STS task consists of labeling one sentence pair at a time, based on the semantic similarity existent between its two component sentences. Human assigned similarity scores range from 0 (no relation) to 5 (semantivally equivalent). The *SEM 2013 STS task did not provide additional labeled data to the training and testing sets released as part of the STS task hosted at SEMEVAL 2012 (Agirre et al., 2012); our system variations were trained on SEMEVAL 2012 data.

The test sets (Agirre et al., 2013) consist of 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).

3.2 Resources

Various subparts of our framework use several resources that are described in more detail below.

Wikipedia[1] is the most comprehensive encyclopedia to date, and it is an open collaborative effort hosted on-line. Its basic entry is an *article* which in addition to describing an entity or an event also contains hyperlinks to other pages within or outside of Wikipedia. This structure (articles and hyperlinks) is directly exploited by semantic similarity methods such as ESA (Gabrilovich and Markovitch, 2007), or SSA (Hassan and Mihalcea, 2011)[2].

[1] www.wikipedia.org

[2] In the experiments reported in this paper, all the corpus-based methods are trained on the English Wikipedia download from October 2008.

WordNet (Miller, 1995) is a manually crafted lexical resource that maintains semantic relationships such as synonymy, antonymy, hypernymy, etc., between basic units of meaning, or *synsets*. These relationships are employed by various knowledge-based methods to derive semantic similarity.

The MPQA corpus (Wiebe and Riloff, 2005) is a newswire data set that was manually annotated at the expression level for opinion-related content. Some of the features derived by our opinion extraction models were based on training on this corpus.

3.3 Features

Our system variations derive the similarity score of a given sentence-pair by integrating information from knowledge, corpus, and opinion-based sources[3].

3.3.1 Knowledge-Based Features

Following prior work from our group (Mihalcea et al., 2006; Mohler and Mihalcea, 2009), we employ several WordNet-based similarity metrics for the task of sentence-level similarity. Briefly, for each open-class word in one of the input texts, we compute the maximum semantic similarity[4] that can be obtained by pairing it with any open-class word in the other input text. All the word-to-word similarity scores obtained in this way are summed and normalized to the length of the two input texts. We provide below a short description for each of the similarity metrics employed by this system.

The **shortest path** ($Path$) similarity is equal to:

$$Sim_{path} = \frac{1}{length} \quad (1)$$

where $length$ is the length of the shortest path between two concepts using node-counting.

The **Leacock & Chodorow** (Leacock and Chodorow, 1998) (LCH) metric is equal to:

$$Sim_{lch} = -\log \frac{length}{2 * D} \quad (2)$$

where $length$ is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the taxonomy.

The **Lesk** ($Lesk$) similarity of two concepts is defined as a function of the overlap between the corresponding definitions, as provided by a dictionary. It is based on an algorithm proposed by Lesk (1986) as a solution for word sense disambiguation.

The **Wu & Palmer** (Wu and Palmer, 1994) (WUP) similarity metric measures the depth of two given concepts in the WordNet taxonomy, and the depth of the least common subsumer (LCS), and combines these figures into a similarity score:

$$Sim_{wup} = \frac{2 * depth(LCS)}{depth(concept_1) + depth(concept_2)} \quad (3)$$

The measure introduced by **Resnik** (Resnik, 1995) (RES) returns the information content (IC) of the LCS of two concepts:

$$Sim_{res} = IC(LCS) \quad (4)$$

where IC is defined as:

$$IC(c) = -\log P(c) \quad (5)$$

and $P(c)$ is the probability of encountering an instance of concept c in a large corpus.

The measure introduced by **Lin** (Lin, 1998) (Lin) builds on Resnik's measure of similarity, and adds a normalization factor consisting of the information content of the two input concepts:

$$Sim_{lin} = \frac{2 * IC(LCS)}{IC(concept_1) + IC(concept_2)} \quad (6)$$

We also consider the **Jiang & Conrath** (Jiang and Conrath, 1997) (JCN) measure of similarity:

$$Sim_{jnc} = \frac{1}{IC(concept_1) + IC(concept_2) - 2 * IC(LCS)} \quad (7)$$

3.3.2 Corpus Based Features

While most of the corpus-based methods induce semantic profiles in a word-space, where the semantic profile of a word is expressed in terms of its co-occurrence with other words, LSA, ESA and SSA rely on a concept-space representation, thus expressing a word's semantic profile in terms of the implicit (LSA), explicit (ESA), or salient (SSA) concepts. This departure from the sparse word-space to a denser, richer, and unambiguous concept-space resolves one of the fundamental problems in semantic relatedness, namely the vocabulary mismatch.

Latent Semantic Analysis (LSA) (Landauer et al., 1997). In LSA, term-context associations are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD)

[3]The abbreviation in italics accompanying each method allows for cross-referencing with the results listed in Table 2.

[4]We use the WordNet::Similarity package (Pedersen et al., 2004).

on the term-by-context matrix $\mathbf{T}$, where the matrix is induced from a large corpus. This reduction entails the abstraction of meaning by collapsing similar contexts and discounting noisy and irrelevant ones, hence transforming the real world term-context space into a word-latent-concept space which achieves a much deeper and concrete semantic representation of words[5].

Random Projection (RP) (Dasgupta, 1999). In RP, a high dimensional space is projected onto a lower dimensional one, using a randomly generated matrix. (Bingham and Mannila, 2001) show that unlike LSA or principal component analysis (PCA), RP is computationally efficient for large corpora, while also retaining accurate vector similarity and yielding comparable results.

Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007). ESA uses encyclopedic knowledge in an information retrieval framework to generate a semantic interpretation of words. It relies on the distribution of words inside Wikipedia articles, thus building a semantic representation for a given word using a word-document association.

Salient Semantic Analysis (SSA) (Hassan and Mihalcea, 2011). SSA incorporates a similar semantic abstraction as ESA, yet it uses salient concepts gathered from encyclopedic knowledge, where a "concept" represents an unambiguous expression which affords an encyclopedic definition. Saliency in this case is determined based on the word being hyperlinked in context, implying that it is highly relevant to the given text.

In order to determine the similarity of two text fragments, we employ two variations: the typical cosine similarity (cos) and a best alignment strategy ($align$), which we explain in more detail in the paragraph below. Both variations were paired with the ESA, and SSA systems resulting in four similarity scores that were used as features by our meta-system, namely ESA_{cos}, ESA_{align}, SSA_{cos}, and SSA_{align}; in addition, we also used BOW_{cos}, LSA_{cos}, and RP_{cos}.

Best Alignment Strategy ($align$). Let T_a and T_b be two text fragments of size a and b respectively. After removing all stopwords, we first determine the number of shared terms (ω) between T_a and T_b. Second, we calculate the semantic relatedness of all possible pairings between non-shared terms in T_a and T_b. We further filter these possible combinations by creating a list φ which holds the strongest semantic pairings between the fragments' terms, such that each term can only belong to one and only one pair.

$$Sim(T_a, T_b) = \frac{(\omega + \sum_{i=1}^{|\varphi|} \varphi_i) \times (2ab)}{a + b} \quad (8)$$

where φ_i is the similarity score for the ith pairing.

3.3.3 Opinion Aware Features

We design opinion-aware features to capture sentence similarity on the subjectivity level based on the output of three subjectivity analysis systems. Intuitively, two sentences are similar in terms of subjectivity if there exists similar opinion expressions which also share similar opinion holders.

OpinionFinder (Wilson et al., 2005) is a publicly available opinion extraction model that annotates the subjectivity of new text based on the presence (or absence) of words or phrases in a large lexicon. The system consists of a two step process, by feeding the sentences identified as subjective or objective by a rule-based high-precision classifier to a high-recall classifier that iteratively learns from the remaining corpus. For each sentence in a STS pair, the two classifiers provide two predictions; a subjectivity similarity score (*SUBJSL*) is computed as follows. If both sentences are classified as subjective or objective, the score is 1; if one is subjective and the other one is objective, the score is -1; otherwise it is 0. We also make use of the output of the subjective expression identifier in OpinionFinder. We first record how many expressions the two sentences have: feature *NUMEX1* and *NUMEX2*. Then we compare how many tokens these expressions share and we normalize by the total number of expressions (feature *EXPR*).

We compute the difference between the probabilities of the two sentences being subjective (*SUBJDIFF*), by employing a logistic regression classifier using LIBLINEAR (Fan et al., 2008) trained on the MPQA corpus. The smaller the difference, the more similar the sentences are in terms of subjectivity.

We also employ features produced by the opinion-extraction model of Yang and Cardie (Yang and Cardie, 2012), which is better suited to process ex-

[5]We use the LSA implementation available at `code.google.com/p/semanticvectors/`.

pressions of arbitrary length. Specifically, for each sentence, we extract subjective expressions and generate the following features. *SUBJCNT* is a binary feature which is equal to 1 if both sentences contain a subjective expression. *DSEALGN* marks the number of shared words between subjective expressions in two sentences, while *DSESIM* represents their similarity beyond the word level. We represent the subjective expressions in each sentence as a feature vector, containing unigrams extracted from the expressions, their part-of-speech, their WordNet hypernyms and their subjectivity label[6], and compute the cosine similarity between the feature vectors. The holder of the opinion expressions is extracted with the aid of a dependency parser[7]. In most cases, the opinion holder and the opinion expression are related by the dependency relation *subj*. This relation is used to expand the verb dependents in the opinion expression and identify the opinion holder or *AGENT*.

3.4 Meta-learning

Each metric described above provides one individual score for every sentence-pair in both the training and test set. These scores then serve as input to a meta-learner, which adjusts their importance, and thus their bearing on the overall similarity score predicted by the system. We experimented with regression and decision tree based algorithms by performing 10-fold cross validation on the 2012 training data; these types of learners are particularly well suited to maintain the ordinality of the semantic similarity scores (i.e. a score of 4.5 is closer to either 4 or 5, implying that the two sentences are mostly or fully equivalent, while also being far further away from 0, implying no semantic relatedness between the two sentences). We obtained consistent results when using support vector regression with polynomial kernel (Drucker et al., 1997; Smola and Schoelkopf, 1998) (SVR) and random subspace meta-classification with tree learners (Ho, 1998) ($RandSubspace$)[8].

We submitted three system variations based on the training corpus (*first word* in the system name) or the learning methodology (*second word*) used: $comb.RandSubspace$, $comb.SVR$ and $indv.RandSubspace$. For $comb$, training was performed on the merged version of the entire 2012 SEMEVAL dataset. For $indv$, predictions for $OnWN$ and SMT test data were based on training on matching $OnWN$ and SMT[9] data from 2012, predictions for the other test sets were computed using the combined version ($comb$).

System	FNWN	headlines	OnWN	SMT	Mean
comb.RandSubSpace	0.331	**0.677**	0.514	0.337	**0.494**
comb.SVR	**0.362**	0.669	0.510	**0.341**	**0.494**
indv.RandSubspace	0.331	**0.677**	**0.548**	0.277	0.483
baseline-tokencos	0.215	0.540	0.283	0.286	0.364

Table 1: Evaluation results (Agirre et al., 2013).

4 Results and Discussion

Table 2 lists the correlations obtained between the scores assigned by each one of the features we used and the scores assigned by the human judges. It is interesting to note that overall, corpus-based measures are stronger performers compared to knowledge-based measures. The top contenders in the former group are ESA_{align}, SSA_{align}, LSA_{cos}, and RP_{cos}, indicating that these methods are able to leverage a significant amount of semantic information from text. While LSA_{cos} achieves high correlations on many of the datasets, replacing the singular value decomposition operation by random projection to a lower-dimension space (RP) achieves competitive results while also being computationally efficient. This observation is in line with prior literature (Bingham and Mannila, 2001). Among the knowledge-based methods, JCN and $Path$ achieve high performance on more than five of the datasets. In some cases, particularly on the 2013 test data, the shortest path method ($Path$) peforms better or on par with the performance attained by other knowledge-based measures, despite its computational simplicity. While opinion-based measures do not exhibit the same high correlation, we should remember that none of the datasets displays consistent opinion content, nor were they annotated with this aspect in mind, in order for this information to be properly leveraged and evaluated.

[6]Label is based on the OpinionFinder subjectivity lexicon (Wiebe et al., 2005).

[7]`nlp.stanford.edu/software/`

[8]Included with the Weka framework (Hall et al., 2009); we used the default values for both algorithms.

[9]The SMT training set is a combination of $SMTeuroparl$ (in this paper abbreviated as $SMTep$) and $SMTnews$ data.

	Train 2012			Test 2012					Test 2013			
Feature	SMTep	MSRpar	MSRvid	SMTep	MSRpar	MSRvid	OnWN	SMTnews	FNWN	headlines	OnWN	SMT
Knowledge-based measures												
JCN	**0.51**	**0.49**	**0.63**	0.48	0.48	0.64	0.62	0.28	0.38	**0.72**	0.71	**0.34**
LCH	0.45	0.48	0.49	0.47	**0.49**	0.54	0.54	0.3	0.39	0.69	0.69	0.32
$Lesk$	0.5	0.48	0.59	**0.5**	0.47	0.63	**0.64**	**0.4**	0.4	0.71	0.7	0.33
Lin	0.48	**0.49**	0.54	0.48	0.48	0.56	0.57	0.27	0.28	0.65	0.66	0.3
$Path$	0.5	**0.49**	0.62	0.48	**0.49**	**0.65**	0.62	0.35	**0.43**	**0.72**	**0.73**	**0.34**
RES	0.48	0.47	0.55	0.49	0.47	0.6	0.62	0.33	0.28	0.64	0.7	0.31
WUP	0.42	0.46	0.38	0.44	0.48	0.42	0.48	0.26	0.19	0.55	0.6	0.25
Corpus-based measures												
BOW_cos	0.51	0.47	0.69	0.32	0.44	0.71	0.66	0.37	0.34	0.68	0.52	0.32
ESA_cos	0.53	0.34	0.71	0.44	0.3	0.77	0.63	0.44	0.34	0.55	0.35	0.27
ESA_align	0.55	**0.56**	0.75	**0.49**	**0.52**	0.78	0.69	0.38	**0.46**	**0.71**	0.47	**0.34**
SSA_cos	0.4	0.34	0.63	0.4	0.22	0.71	0.6	0.42	0.35	0.48	0.47	0.26
SSA_align	0.54	**0.56**	0.74	**0.49**	0.51	0.77	0.68	0.38	0.44	0.69	0.46	**0.34**
LSA_cos	**0.65**	0.48	0.76	0.36	0.45	**0.79**	0.67	**0.45**	0.25	0.63	**0.61**	0.32
RP_cos	0.6	0.49	**0.78**	0.46	0.43	**0.79**	**0.7**	**0.45**	0.38	0.68	0.57	**0.34**
Opinion-aware measures												
$AGENT$	0.16	0.15	0.05	0.11	0.12	0.03	n/a	-0.01	n/a	**0.08**	-0.04	0.11
$DSEALGN$	**0.18**	0.2	**0.11**	0.05	0.11	**0.11**	0.07	0.06	-0.1	**0.08**	0.13	0.1
$DSESIM$	0.12	0.15	0.05	0.1	0.08	0.07	0.04	**0.08**	0.05	**0.08**	0.04	0.08
$EXPR$	0.17	0.19	0.06	0.18	**0.18**	0.02	0.07	0	0.13	**0.08**	**0.18**	**0.17**
$NUMEX1$	0.12	**0.22**	-0.03	0.07	0.16	-0.05	-0.01	-0.01	-0.01	-0.03	0.08	0.1
$NUMEX2$	-0.25	0.19	0.01	0.06	0.14	-0.03	0.01	0.06	0.09	-0.05	0.03	0.11
$SUBJCNT$	0.14	0.19	0.01	0.09	0.07	0.03	0.02	**0.08**	0.05	0.05	0.05	0.09
$SUBJDIFF$	-0.07	-0.07	-0.17	-0.27	-0.13	-0.22	-0.17	-0.12	-0.04	-0.12	-0.2	-0.12
$SUBJSL$	0.15	-0.11	0.07	**0.23**	0.01	0.07	**0.11**	-0.08	**0.15**	0.07	-0.03	0

Table 2: Correlation of individual features for the training and test sets with the gold standard.

Nonetheless, we notice several promising features, such as $DSEALIGN$ and $EXPR$. Lower correlations seem to be associated with shorter spans of text, since when averaging all opinion-based correlations per dataset, $MSRvid$ (x2), $OnWN$ (x2), and $headlines$ display the lowest average correlation, ranging from 0 to 0.03. This matches the expectation that opinionated content can be easier identified in longer contexts, as additional subjective elements amount to a stronger prediction. The other seven datasets consist of longer spans of text; they display an average opinion-based correlation between 0.07 and 0.12, with the exception of $FNWN$ and $SMTnews$ at 0.04 and 0.01, respectively.

Our systems performed well, ranking 38, 39 and 45 among the 88 competing systems in *SEM 2013 (see Table 1), with the best being *comb.SVR* and *comb.RandSubspace*, both with a mean correlation of 0.494. We noticed from our participation in SEMEVAL 2012 (Banea et al., 2012), that training and testing on the same type of data achieves the best results; this receives further support when considering the performance of the *indv.RandSubspace* variation on the OnWN data[10], which exhibits a 0.034 correlation increase over our next best system (*comb.RandSubspace*). While we do surpass the bag-of-words cosine baseline (*baseline-tokencos*) computed by the task organizers by a 0.13 difference in correlation, we fall short by 0.124 from the performance of the best system in the STS task.

5 Conclusions

To participate in the STS *SEM 2013 task, we constructed a meta-learner framework that combines traditional knowledge and corpus-based methods, while also introducing novel opinion analysis based metrics. While the *SEM data is not particularly suited for evaluating the performance of opinion features, this is nonetheless a first step toward conducting text similarity research while also considering the subjective dimension of text. Our system variations ranked 38, 39 and 45 among the 88 participating systems.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340 and IIS awards #1018613,

[10]The SMT test data is not part of the same corpus as either $SMTep$ or $SMTnews$.

#0208798 and #0916046. This work was supported in part by DARPA-BAA-12-47 DEFT grant #12475008. Any opinions, findings, and conclusions or recommendations 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.

References

E. Agirre, D. Cer, M. Diab, and A. Gonzalez. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*.

E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo. 2013. *SEM 2013 Shared Task: Semantic Textual Similarity, including a Pilot on Typed-Similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, Atlanta, GA, USA.

C. Banea, S. Hassan, M. Mohler, and R. Mihalcea. 2012. UNT: A supervised synergistic approach to semantic text similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 635–642, Montreal, Canada.

E. Bingham and H. Mannila. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2001)*, pages 245–250, San Francisco, CA, USA.

C. Burgess, K. Livesay, and K. Lund. 1998. Explorations in context space: words, sentences, discourse. *Discourse Processes*, 25(2):211–257.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

S. Dasgupta. 1999. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (FOCS 1999)*, pages 634–644, New York, NY, USA.

H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and Vladimir Vapnik. 1997. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9:155–161.

R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th AAAI International Conference on Artificial Intelligence (AAAI'07)*, pages 1606–1611, Hyderabad, India.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

S. Hassan and R. Mihalcea. 2011. Measuring semantic relatedness using salient encyclopedic concepts. *Artificial Intelligence, Special Issue*.

T. K. Ho. 1998. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844.

A. Islam and D. Inkpen. 2006. Second order co-occurrence PMI for determining the semantic similarity of words. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06)*, volume 2, pages 1033–1038, Genoa, Italy, July.

A. Islam and D. Inkpen. 2009. Semantic Similarity of Short Texts. In Nicolas Nicolov, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing V*, volume 309 of *Current Issues in Linguistic Theory*, pages 227–236. John Benjamins, Amsterdam & Philadelphia.

J. J. Jiang and D. W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 9008+, September.

T. K. Landauer, T. K. L, D. Laham, B. Rehder, and M. E. Schreiner. 1997. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans.

M. Lapata and R. Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh.

C. Leacock and M. Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*, pages 305–332.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA. ACM.

C. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth Interna-*

tional Conference on Machine Learning, pages 296–304, Madison, Wisconsin.

R. Mihalcea, C. Corley, and C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*, pages 775–780, Boston, MA, US.

G. A. Miller. 1995. WordNet: a Lexical database for English. *Communications of the Association for Computing Machinery*, 38(11):39–41.

M. Mohler and R. Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the European Association for Computational Linguistics (EACL 2009)*, Athens, Greece.

M. Mohler, R. Bunescu, and R. Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the Association for Computational Linguistics – Human Language Technologies (ACL-HLT 2011)*, Portland, Oregon, USA.

R. M. A. Nawab, M. Stevenson, and P. Clough. 2011. External plagiarism detection using information retrieval and sequence alignment: Notebook for PAN at CLEF 2011. In *Proceedings of the 5th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2011)*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet:: Similarity-Measuring the Relatedness of Concepts. *Proceedings of the National Conference on Artificial Intelligence*, pages 1024–1025.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

J. Rocchio, 1971. *Relevance feedback in information retrieval*. Prentice Hall, Ing. Englewood Cliffs, New Jersey.

G. Salton and C. Buckley. 1997. Term weighting approaches in automatic text retrieval. In *Readings in Information Retrieval*. Morgan Kaufmann Publishers, San Francisco, CA.

G. Salton and M. Lesk, 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Computer evaluation of indexing and text processing. Prentice Hall, Ing. Englewood Cliffs, New Jersey.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 2(32).

H. Schutze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

A. Smola and B. Schoelkopf. 1998. A tutorial on support vector regression. NeuroCOLT2 Technical Report NC2-TR-1998-030.

P. D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML'01)*, pages 491–502, Freiburg, Germany.

J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing (CICLing 2005)*, pages 486–497, Mexico City, Mexico.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, Vancouver, BC, Canada.

Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133—-138, Las Cruces, New Mexico.

B. Yang and C. Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.