SimCompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity

Carmen Banea, Di Chen, Rada Mihalcea*
University of Michigan
Ann Arbor, MI

Claire Cardie
Cornell University
Ithaca, NY

Janyce Wiebe
University of Pittsburgh
Pittsburgh, PA

Abstract

This article presents our team's participating system at SemEval-2014 Task 3. Using a meta-learning framework, we experiment with traditional knowledge-based metrics, as well as novel corpus-based measures based on deep learning paradigms, paired with varying degrees of context expansion. The framework enabled us to reach the highest overall performance among all competing systems.

1 Introduction

Semantic textual similarity is one of the key components behind a multitude of natural language processing applications, such as information retrieval (Salton and Lesk, 1971), relevance feedback and text classification (Rocchio, 1971), word sense disambiguation (Lesk, 1986; Schutze, 1998), summarization (Salton et al., 1997; Lin and Hovy, 2003), automatic evaluation of machine translation (Papineni et al., 2002), plagiarism detection (Nawab et al., 2011), and more.

To date, semantic similarity research has primarily focused on comparing text snippets of similar length (see the semantic textual similarity tasks organized during **Sem 2013* (Agirre et al., 2013) and *SemEval 2012* (Agirre et al., 2012)). Yet, as new challenges emerge, such as augmenting a knowledge-base with textual evidence, assessing similarity across different context granularities is gaining traction. The SemEval Cross-level semantic similarity task is aimed at this latter scenario, and is described in more details in the task paper (Jurgens et al., 2014).

* {carmennb,chenditc,mihalcea}@umich.edu

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) or Roget (Rog, 1995) to measure definitional overlap, term distance within a graphical taxonomy, or term depth in the taxonomy as a measure of specificity. There are many knowledge-based measures that were proposed in the past, e.g., (Leacock and Chodorow, 1998; Lesk, 1986; Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998; Jarmasz and Szpakowicz, 2003; Hughes and Ramage, 2007).

On the other side, corpus-based measures such as Latent Semantic Analysis (LSA) (Landauer and Dumais, 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 the similarity between any two words, as long as they appear in the corpus used for training.

3 System Description

3.1 Generic Features

Our system employs both knowledge and corpus-based measures as detailed below.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 560–565,
Dublin, Ireland, August 23-24, 2014.

Knowledge-based features
Knowledge-based metrics were shown to provide high correlation scores with the goldstandard in text similarity tasks (Agirre et al., 2012; Agirre et al., 2013). We used three WordNet-based similarity measures that employ information content. We chose these metrics because they are able to incorporate external information derived from a large corpus: Resnik (Resnik, 1995) (RES), Lin (Lin, 1998) (LIN), and Jiang & Conrath (Jiang and Conrath, 1997) (JCN).

Corpus based features
Our corpus based features are derived from a deep learning vector space model that is able to "understand" word meaning without human input. Distributed word embeddings are learned using a skip-gram recurrent neural net architecture running over a large raw corpus (Mikolov et al., 2013b; Mikolov et al., 2013a). A primary advantage of such a model is that, by breaking away from the typical n-gram model that sees individual units with no relationship to each other, it is able to generalize and produce word vectors that are similar for related words, thus encoding linguistic regularities and patterns (Mikolov et al., 2013b). For example, vec(Madrid)-vec(Spain)+vec(France) is closer to vec(Paris) than any other word vector (Mikolov et al., 2013a). We used the pretrained Google News word2vec model (WTV) built over a 100 billion words corpus, and containing 3 million 300-dimension vectors for words and phrases. The model is distributed with the word2vec toolkit. [1]

Since the methods outlined above provide similarity scores at the sense or word level, we derive text level metrics by employing two methods.

VectorSum. We add the vectors corresponding to the non-stopwords tokens in bag of words (BOW) A and B, resulting in vectors V_A and V_B, respectively. The assumption is that these vectors are able to capture the semantic meaning associated with the contexts, enabling us to gauge their relatedness using cosine similarity.

Align. Given two BOW A and B as input, we compare them using a word-alignment-based similarity measure (Mihalcea et al., 2006). We calculate the pairwise similarity between the words in A and B, and match each word in A with its most similar counterpart in B. For corpus-based features, the similarity measure represents the average over these scores, while for knowledge-based measures, we consider the top 40% ranking pairs.

We use the DKPro Similarity package (Bär et al., 2013) to compute knowledge-based metrics, and the word2vec implementation from the Gensim toolkit (Rehurek and Sojka, 2010).

3.2 Feature Variations

Since our system participated in all four lexical levels evaluations, we describe below the modifications pertaining to each.

word2sense. At the word2sense level, we employ both knowledge and corpus-based features. Since the information available in each pair is extremely limited (only a word and a sense key) we infuse contextual information by drawing on WordNet (Miller, 1995). In WordNet, the sense of each word is encapsulated in a uniquely identifiable synset, consisting of the definition (gloss), usage examples and its synonyms. We can derive three variations (where the *word* and *sense* components are represented by BOW A and B, respectively): **a) no expansion** (A={*word*}, B={*sense*}), **b) expand right (R)** (A={*word*}, B={*sense* gloss & example}), **c) expand left (L) & right (R)** (A={*word* glosses & examples}, B={*sense* gloss & example}). After applying the Align method, we obtain measures JNC, LIN, RES and $WTV1$; VectorSum results in $WTV2$.

phrase2word. As this lexical level also suffers from low context, we adapt the above variations, where the *phrase* and *word* components are represented by BOW A and BOW B, respectively. Thus, we have: **a) no expansion** (A={*phrase*}, B={*word*}), **b) expand R** (A={*phrase*}, B={*word* glosses and examples}), **c) expand L & R** (A={*phrase* glosses & examples}, B={*word* glosses and examples}). We extract the same measures as for *word2sense*.

sentence2phrase. For this variation, we use only corpus based measures; BOW A represents the *sentence* component, B, the *phrase*. Since there is sufficient context available, we follow the **no expansion** variation, and obtain metrics $WTV1$ (by applying Align) and $WTV2$ (using VectorSum).

paragraph2sentence. At this level, due to the long context that entails one-to-many mappings between the words in the *sentence* and those in the *paragraph*, we use a text clustering technique prior to calculating the features' weights.

[1] https://code.google.com/p/word2vec/

a) no clustering. We use only corpus based measures, where the *paragraph* represents BOW A, and the *sentence* represents BOW B. Then we apply Align and VectorSum, resulting in $WTV1$ and $WTV2$, respectively.

b) paragraph centroids extraction. Since the longer text contains more information compared to the shorter one, we extract k topic vectors after K-means clustering the left context.[2] These centroids are able to model topics permeating across sentences, and by comparing them with the word vectors pertaining to the short text, we seek to capture how much of the information is covered in the shorter text. Each word is paired with the centroid that it is closest to, and the average is computed over these scores, resulting in $WTV3$.

c) sentence centroids extraction. Under a different scenario, assuming that one sentence covers only a few strongly expressed topics, unlike a paragraph that may digress and introduce unrelated noise, we apply clustering on the short text. The centroids thus obtained are able to capture the essence of the sentence, so when compared to every word in the paragraph, we can gauge how much of the short text is reflected in the longer one. Each centroid is paired with the word that it is most similar to, and we average these scores, thus obtaining $WTV4$. In a way, methods b) and c) provide a macro, respectively micro view of how the topics are reflected across the two spans of text.

3.3 Meta-learning

The measures of similarity described above provide a single score per each *long text - short text* pair in the training and test data. These scores then become features for a meta-learner, which is able to optimize their impact on the prediction process. We experimented with multiple regression algorithms by conducting 10 fold cross-validation on the training data. The strongest performer across all lexical levels was Gaussian processes with a radial basis function (RBF) kernel. Gaussian processes regression is an efficient probabilistic prediction framework that assumes a Gaussian process prior on the unobservable (latent) functions and a likelihood function that accounts for noise. An individual classifier[3] was trained for each lexical level and applied to the test data sets.

[2]Implementation provided in the Scikit library (Pedregosa et al., 2011), where k is set to 3.

[3]Implementation available in the WEKA machine learning software (Hall et al., 2009) using the default parameters.

4 Evaluations & Discussion

Our system participated in all cross-level subtasks under the name *SimCompass*, competing with 37 other systems developed by 20 teams.

Figure 1 highlights the Pearson correlations at the four lexical levels between the gold standard and each similarity measure introduced in Section 3, as well as the predictions ensuing as a result of meta-learning. The left and right histograms in each subfigure present the scores obtained on the train and test data, respectively.

In the case of *word2sense* train data, we notice that expanding the context provides additional information and improves the correlation results. For corpus-based measures, the correlations are stronger when the expansion involves only the right side of the tuple, namely the *sense*. We notice an increase of 0.04 correlation points for WTV1 and 0.09 for WTV2. As soon as the *word* is expanded as well, the context incorporates too much noise, and the correlation levels drop. In the case of knowledge-based measures, expanding the context does not seem to impact the results. However, these trends do not carry out to the test data, where the corpus-based features without expansion reach a correlation higher than 0.3, while the knowledge-based features score significantly lower (by 0.16). Once all these measures are used as features in a meta learner (*All*) using Gaussian processes regression (GP), the correlation increases over the level attained by the best performing individual feature, reaching 0.45 on the train data and 0.36 on the test data. SimCompass ranks second in this subtask's evaluations, falling short of the leading system by 0.025 correlation points.

Turning now to the *phrase2word* subfigure, we notice that the context already carries sufficient information, and expanding it causes the performance to drop (the more extensive the expansion, the steeper the drop). Unlike the scenario encountered for *word2sense*, the trend observed here on the training data also gets mirrored in the test data. Same as before, knowledge-based measures have a significantly lower performance, but deep learning-based features based on word2vec (WTV) only show a correlation variation by at most 0.05, proving their robustness. Leveraging all the features in a meta-learning framework enables the system to predict stronger scores for both the train and the test data (0.48 and 0.42, respectively). Actually, for this variation, SimCompass

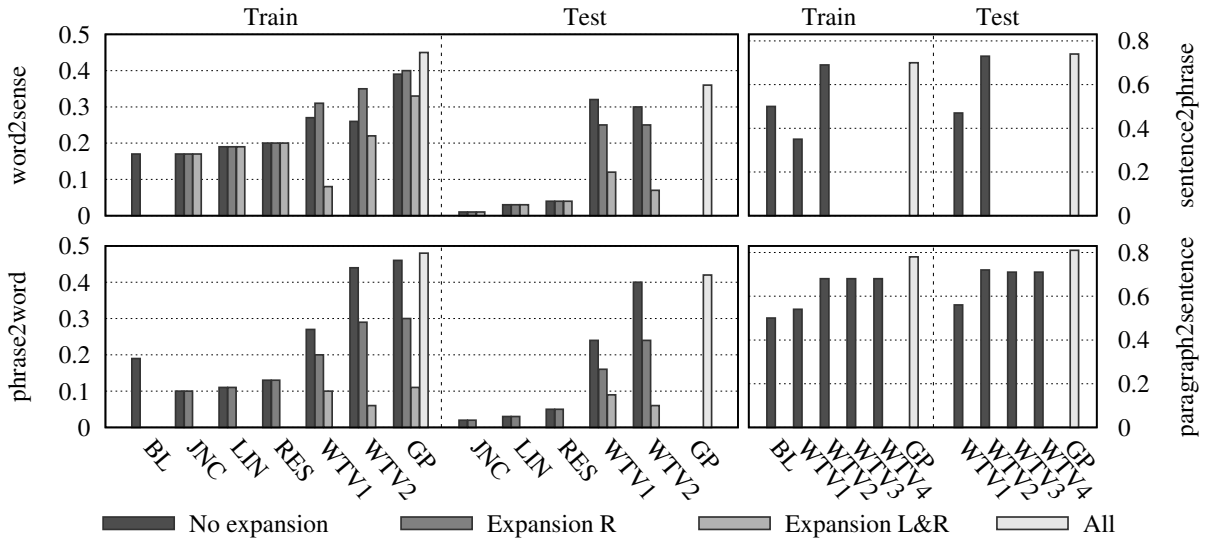

Figure 1: Pearson correlation of individual measures on the train and test data sets. As these measures become features in a regression algorithm (GP), prediction correlations are included as well. BL represents the baseline computed by the organizers.

obtains the highest score among all competing systems, surpassing the second best by 0.10.

Noticing that expansion is not helpful when sufficient context is available, for the next variations we use the original tuples. Also, due to the reduced impact of knowledge-based features on the training outcome, we only focus on deep learning features (*WTV1*, *WTV2*, *WTV3*, *WTV4*).

Shifting to *sentence2phrase*, *WTV2* (constructed using VectorSum) is the top performing feature, surpassing the baseline by 0.19, and attaining 0.69 and 0.73 on the train and test sets, respectively. Despite also considering a lower performing feature (*WTV1*), the meta-learner maintains high scores, surpassing the correlation achieved on the train data by 0.04 (from 0.70 to 0.74). In this variation, our system ranks fifth, at 0.035 from the top system.

For the *paragraph2sentence* variation, due to the availability of longer contexts, we introduce *WTV3* and *WTV4* that are based on clustering the left and the right sides of the tuple, respectively. *WTV2* fares slightly better than *WTV3* and *WTV4*. *WTV1* surpasses the baseline this time, leaving its mark on the decision process. When training the GP learner on all features, we obtain 0.78 correlation on the train data, and 0.81 on test data, 0.10 higher than those attained by the individual features alone. SimCompass ranks seventh in performance on this subtask, at 0.026 from the first.

Considering the overall system performance, SimCompass is remarkably versatile, ranking among the top at each lexical level, and taking the first place in the SemEval Task 3 overall evaluation with respect to both Pearson (0.58 average correlation) and Spearman correlations.

5 Conclusion

We described SimCompass, the system we participated with at SemEval-2014 Task 3. Our experiments suggest that traditional knowledge-based features are cornered by novel corpus-based word meaning representations, such as word2vec, which emerge as efficient and strong performers under a variety of scenarios. We also explored whether context expansion is beneficial to the cross-level similarity task, and remarked that only when the context is particularly short, this enrichment is viable. However, in a meta-learning framework, the information permeating from a set of similarity measures exposed to varying context expansions can attain a higher performance than possible with individual signals. Overall, our system ranked first among 21 teams and 38 systems.

Acknowledgments

This material is based in part upon work supported by National Science Foundation CAREER award #1361274 and IIS award #1018613 and 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

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 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)*, Montreal, Canada.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 Shared Task: Semantic Textual Similarity, including a Pilot on Typed-Similarity. In *The Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2013. DKPro Similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 121–126, Sofia, Bulgaria.

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

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

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

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

Thad Hughes and Daniel Ramage. 2007. Lexical semantic knowledge with random graph walks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Prague, Czech Republic.

Aminul Islam and Diana Zaiu Inkpen. 2006. Second order co-occurrence PMI for determining the semantic similarity of words. In *Proceedings of the Fifth Conference on Language Resources and Evaluation*, volume 2, pages 1033–1038, Genoa, Italy, July.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of the conference on Recent Advances in Natural Language Processing RANLP-2003*, Borovetz, Bulgaria, September.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceeding of the International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 Task 3: Cross-Level Semantic Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet sense similarity for word sense identification. In *WordNet, An Electronic Lexical Database*. The MIT Press.

Michael E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*, Toronto, June.

Chin-Yew Lin and Eduard 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.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, Madison, Wisconsin.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *American Association for Artificial Intelligence (AAAI-2006)*, Boston, MA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality . In *NIPS*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *NAACL HLT*, pages 746–751, Atlanta, GA, USA.

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

Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul 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)*.

Kishore. Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Quebec, Canada. Morgan Kaufmann Publishers Inc.

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

1995. *Roget's II: The New Thesaurus*. Houghton Mifflin.

Gerard Salton and Michael E. 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.

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 2(32).

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

Peter 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.