A Reflective View on Text Similarity

Daniel Bär, Torsten Zesch, and Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Computer Science Department, Technische Universität Darmstadt
Hochschulstrasse 10, D-64289 Darmstadt, Germany
www.ukp.tu-darmstadt.de

Abstract

While the concept of similarity is well grounded in psychology, text similarity is less well-defined. Thus, we analyze text similarity with respect to its definition and the datasets used for evaluation. We formalize text similarity based on the geometric model of conceptual spaces along three dimensions inherent to texts: structure, style, and content. We empirically ground these dimensions in a set of annotation studies, and categorize applications according to these dimensions. Furthermore, we analyze the characteristics of the existing evaluation datasets, and use those datasets to assess the performance of common text similarity measures.

2 Formalization

In psychology, similarity is well formalized and captured in formal models such as the set-theoretic model (Tversky, 1977) or the geometric model (Widdows, 2004). In an attempt to overcome the traditionally loose definition of text similarity, we rely on a conceptual framework based on conceptual spaces (Gärdenfors, 2000). In this model, objects are represented in a number of geometric spaces. For example, potential spaces related to countries are political affinity and geographical proximity. In order to adapt this model to texts, we need to define explicit spaces (i.e. dimensions) suitable for texts. Therefore, we analyzed common NLP tasks with respect to the relevant dimensions of similarity, and then conducted annotation studies to ground them empirically.

Table 1 gives an overview of common NLP tasks and their relevant dimensions: structure, style, and content. Structure thereby refers to the internal developments of a given text, e.g. the order of sections. Style refers to grammar, usage, mechanics, and lexical complexity (Attali and Burstein, 2006). Content addresses all facts and

1A famous 19th century Russian writer of realist fiction and philosophical essays
Table 1: Classification of common NLP tasks with respect to the relevant dimensions of text similarity: structure (str), style (sty), and content (c)

| Task                        | str | sty | c  |
|-----------------------------|-----|-----|----|
| Authorship Classification   | ✓   | ✓   | ✓  |
| Automatic Essay Scoring     | ✓   | ✓   | ✓  |
| Information Retrieval       | ✓   | ✓   | ✓  |
| Paraphrase Recognition      | ✓   | ✓   |   |
| Plagiarism Detection        | ✓   | ✓   |   |
| Question Answering          | ✓   | ✓   |   |
| Short Answer Grading        | ✓   | ✓   |   |
| Summarization               | ✓   | ✓   |   |
| Text Categorization         | ✓   | ✓   |   |
| Text Segmentation           | ✓   | ✓   |   |
| Text Simplification         | ✓   | ✓   |   |
| Word Sense Alignment        | ✓   | ✓   |   |

Taking this dimension-centric view on text similarity also opens up new perspectives. For example, standard information retrieval usually considers only the content dimension (keyword overlap between query and document). However, a scholar in digital humanities might be interested in texts that are similar to a reference document with respect to style and structure, while texts with similar content are of minor interest. In this paper, we only address dimensions inherent to texts, and do not consider dimensions such as user intentions.

2.1 Empirical Grounding

In order to empirically ground the proposed dimensions of text similarity, we conducted a number of exemplary annotation studies. The results show that annotators indeed distinguish between different dimensions of text similarity.

Content vs. Structure  In this study, we used the dataset by Lee et al. (2005) that contains pairwise human similarity judgments for 1,225 text pairs. We selected a subset of 50 pairs with a uniform distribution of judgments across the whole similarity range. We then asked three annotators: “How similar are the given texts?” We then computed the Spearman correlation of each annotator’s ratings with the gold standard: $\rho_{A_1} = 0.83$, $\rho_{A_2} = 0.65$, and $\rho_{A_3} = 0.85$. The much lower correlation of the annotator $A_2$ indicates that a different dimension might have been used to judge similarity.

To further investigate this issue, we asked the annotators about the reasons for their judgments. $A_1$ and $A_3$ consistently focused only on the content of the texts and completely disregarded other dimensions. $A_2$, however, was also taking structural similarities into account, e.g. two texts were rated highly similar because of the way they are organized: First, an introduction to the topic is given, then a quotation is stated, then the text concludes with a certain reaction of the acting subject.

Content vs. Style  The annotators in the previous study only identified the dimensions content and structure. Style was not addressed, as the text pairs were all of similar style, and hence that dimension was not perceived as salient. Thus, we selected 10 pairs of short texts from Wikipedia (WP) and Simple Wikipedia\(^2\) (SWP). We used the first paragraphs of WP articles and the full texts of SWP articles to obtain pairs of similar length. Pairs were formed in all combinations (WP-WP, SWP-WP, and SWP-SWP) to ensure that both similarity dimensions were salient for some pairs. For example, an article from SWP and one from WP about the same topic share the same content, but are different in style, while two articles from SWP have a similar style, but different content.

We then asked three annotators to rate each pair according to the content and style dimensions. The results show that WP-WP and SWP-SWP pairs are perceived as stylistically similar, while WP-SWP pairs are seen similar with respect to their content.

2.2 Discussion

The results demonstrate that humans indeed distinguish the major dimensions of text similarity. Also, they seem intuitively able to find an appropriate dimension of comparison for a given text collection. Smith and Heise (1992) refer to that as perceived similarity which “changes with changes in selective attention to specific perceptual properties.” Selective attention can be modeled using dimension-specific similarity measures. The scores for all dimensions are computed in parallel, and then summed up for each text pair.\(^3\) Thereby, we automatically obtain the discriminating dimension (see Figure 1). $A$, $B$, and $C$ are documents of

\(^2\)Articles written in Simple English use a limited vocabulary and easier grammar than the standard Wikipedia.

\(^3\)The last step requires all measures to be normalized.
### 3 Evaluation Datasets

Four datasets are commonly used for evaluation (see Table 2). They contain text pairs together with human judgments about their perceived similarity. However, none of those datasets has yet undergone a thorough analysis with respect to the dimensions of text similarity encoded therein.

#### 3.1 30 Sentence Pairs

Li et al. (2006) introduced 65 sentence pairs which are based on the noun pairs by Rubenstein and Goodenough (1965). Each noun was replaced by its definition from Collins Cobuild English Dictionary (Sinclair, 2001). The dataset contains judgments from 32 subjects on how similar in meaning one sentence is to another. Li et al. (2006) selected 30 pairs to reduce the bias in the frequency distribution (30 Sentence Pairs, henceforth).

We conducted a re-rating study to evaluate whether text similarity judgments are stable across time and subjects. We collected 10 judgments per pair asking: “How close do these sentences come to meaning the same thing?”

The Spearman correlation of the aggregated results with the original scores is $\rho = 0.91$. We conclude that text similarity judgments are stable across time and subjects. It also indicates that humans indeed share a common understanding on what makes texts similar.

In order to better understand the characteristics of this dataset, we performed another study. For each text pair we asked the annotators: “Why did people agree that these two sentences are (not) close in meaning?” We collected 10 judgments per pair in the same crowdsourcing setting as before. To our surprise, the annotators only used lexical semantic relations between terms to justify the similarity relation between texts. For example, the text pairs about tool/implement and cemetery/graveyard were consistently said to be synonymous. We conclude that – in this setting – humans reduce text similarity to term similarity.

As the text pairs are originally based on term pairs, we computed the Spearman correlation between the text pair scores and the original term pair scores. The very high correlation of $\rho = 0.94$ shows that annotators indeed judged the similarity between terms rather than texts. We conclude

---

**Table 2: Statistics for text similarity evaluation datasets**

| Dataset                              | Text Type / Domain          | Length in Terms (#) | # Pairs | Rating Scale | # Judges |
|--------------------------------------|-----------------------------|---------------------|---------|--------------|----------|
| 30 Sentence Pairs (Li et al., 2006) | Concept Definitions        | 5–33 (11)           | 30      | 0–4          | 32       |
| 50 Short Texts (Lee et al., 2005)   | News (Politics)             | 45–126 (80)         | 1,225   | 1–5          | 8–12     |
| Computer Science Assignments         | Computer Science            | 1–173 (18)          | 630     | 0–5          | 2        |
| Microsoft Paraphrase Corpus          | News                        | 5–31 (19)           | 5,801   | binary       | 2–3      |

---

$^*$Same question as in the original study by Li et al. (2006).
We used Amazon Mechanical Turk via CrowdFlower.
Table 3: Results on the 30 Sentence Pairs dataset

| Measure                     | $r$  | $\rho$ |
|-----------------------------|------|--------|
| Cosine Baseline             | .81  | .83    |
| Term Pair Heuristic         | .83  | .84    |
| ESA (Wikipedia)             | .61  | .77    |
| ESA (Wiktionary)            | .77  | .82    |
| ESA (WordNet)               | .75  | .80    |
| Kennedy and Szpakowicz (2008) | .87  | .79    |
| LSA (Tsatsaronis et al., 2010) | .84  | .87    |
| OMIOITIS (Tsatsaronis et al., 2010) | .86  | .89    |
| STASIS (Li et al., 2006)    | .82  | .81    |
| STS (Islam and Inkpen, 2008) | .85  | .84    |

Table 4: Results on the 50 Short Texts dataset. Statistically significant improvements in bold.

| Measure                             | $r$  |
|-------------------------------------|------|
| Cosine Baseline                     | .56  |
| ESA (Wikipedia)                     | .46  |
| ESA (Wiktionary)                    | .53  |
| ESA (WordNet)                       | .59  |
| ESA (Gabrilovich and Markovitch, 2007) | .72  |
| LSA (Lee et al., 2005)              | .60  |
| WikiWalk (Yeh et al., 2009)         | .77  |

Evaluation Results

Evaluation Results Table 3 shows the results of state of the art similarity measures obtained on this dataset. We used a cosine baseline and implemented an additional baseline which disregards the actual texts and only takes the target noun of each sentence into account. We computed their pairwise term similarity using the metric by Lin (1998) on WordNet (Fellbaum, 1998). Our heuristic achieves Pearson $r = 0.83$ and Spearman $\rho = 0.84$. The block of results in the middle shows our implementation of Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) using different knowledge sources (Zesch et al., 2008). The bottom rows show scores previously obtained and reported in the literature. None of the measures significantly outperforms the baselines. Given the limitation of encoding rather term than text similarity and the fact that the dataset is also very small (30 pairs), it is questionable whether it is a suitable evaluation dataset for text similarity.

3.2 50 Short Texts

The dataset by Lee et al. (2005) comprises 50 relatively short texts (45 to 126 words) which contain newswire from the political domain. In analogy to the study in Section 3.1, we performed an annotation study to show whether the encoded judgments are stable across time and subjects. We asked three annotators to rate “How similar are the given texts?” We used the same uniformly distributed subset as in Section 2.1. The resulting Spearman correlation between the aggregated results of the annotators and the original scores is $\alpha = 0.05$, Fisher Z-value transformation.

Evaluation Results Table 4 summarizes the results obtained on this dataset. We used a cosine baseline, and our implementation of ESA applied to different knowledge sources. The results at the bottom are scores previously obtained and reported in the literature. All of them significantly outperform the baseline. In contrast to the 30 Sentence Pairs, this dataset encodes a broader view on the content dimension of similarity. It obviously contains text pairs that are similar (or dissimilar) for reasons beyond partial string overlap. Thus, the dataset might be used to intrinsically evaluate text similarity measures.

However, the distribution of similarity scores in this dataset is heavily skewed towards low scores, with 82% of all term pairs having a text similarity score between 1 and 2 on a 1–5 scale. This limits the kind of conclusions that can be drawn as the number of the pairs in the most interesting class of highly similar pairs is actually very small.

Another observation is that we were not able to reproduce the ESA score on Wikipedia reported by Gabrilovich and Markovitch (2007). We found that the difference probably relates to the cut-off value used to prune the vectors as reported by Yeh et al. (2009). By tuning the cut-off value, we could improve the score to 0.70, which comes very close to the reported score of 0.72. However, as this tunc-
Measure | r
--- | ---
Cosine Baseline | .44
ESA (Mohler and Mihalcea, 2009) | .47
LSA (Mohler and Mihalcea, 2009) | .43
Mohler and Mihalcea (2009) | .45

Table 5: Results on the Computer Science Assignments dataset

ing is done directly on the evaluation dataset, it probably overfits the cut-off value to the dataset.

3.3 Computer Science Assignments

The dataset by Mohler and Mihalcea (2009) was introduced for assessing the quality of short answer grading systems in the context of computer science assignments. The dataset comprises 21 questions, 21 reference answers and 630 student answers. The answers were graded by two teachers – not according to stylistic properties, but to the extent the content of the student answers matched with the content of the reference answers.

Evaluation Results

We summarize the results obtained on this dataset in Table 5. The scores are reported without relevance feedback (Mohler and Mihalcea, 2009) which distorts results by changing the reference answers. None of the measures significantly\(^8\) outperforms the baseline. This is not overly surprising, as the textual similarity between the reference and the student answer only constitutes part of what makes an answer the correct one. More sophisticated measures that also take lexical semantic relationships between terms into account might even worsen the results, as typically a specific answer is required, not a similar one. We conclude that similarity measures can be used to grade assignments, but it seems questionable whether this dataset is suited to draw any conclusions on the performance of similarity measures outside of this particular task.

3.4 Microsoft Paraphrase Corpus

Dolan et al. (2004) introduced a dataset of 5,801 sentence pairs taken from news sources on the Web. They collected binary judgments from 2–3 subjects whether each pair captures a paraphrase relationship or not (83\% interrater agreement). The dataset has been used for evaluating text similarity measures as, by definition, paraphrases need to be similar with respect to their content.

---

\(^8\alpha = .05,\) Fisher Z-value transformation

Table 6: Results on Microsoft Paraphrase Corpus

Evaluation Results

We summarize the results obtained on this dataset in Table 6. As detecting paraphrases is a classification task, we use an additional majority baseline which classifies all results according to the predominant class of true paraphrases. The block of results in the middle contains measures that are not specifically tailored towards paraphrase recognition. None of them beats the cosine baseline. The results at the bottom show measures which are specifically tailored towards the detection of a bidirectional entailment relationship. None of them, however, significantly outperforms the cosine baseline. Obviously, recognizing paraphrases is a very hard task that cannot simply be tackled by computing text similarity, as sharing similar content is a necessary, but not a sufficient condition for detecting paraphrases.

3.5 Discussion

We showed that all four datasets encode the content dimension of text similarity. The Computer Science Assignments dataset and the Microsoft Paraphrase Corpus are tailored quite specifically to a certain task. Thereby, factors exceeding the similarity of texts are important. Consequently, none of the similarity measures significantly outperformed the cosine baseline. The evaluation of similarity measures on these datasets is hence questionable outside of the specific application scenario. The 30 Sentence Pairs dataset was found to rather represent the similarity between terms than texts. Obviously, it is not suited for evaluating text similarity measures. However, the 50 Short Texts dataset currently seems to be the best choice. As it is heavily skewed towards low similarity scores, though, the conclusions that can be drawn from the results are limited. Further datasets are
necessary to guide the development of measures along other dimensions such as structure or style.

4 Conclusions

In this paper, we reflected on text similarity as a foundational technique for a wide range of tasks. We argued that while similarity is well grounded in psychology, text similarity is less well-defined. We introduced a formalization based on conceptual spaces for modeling text similarity along explicit dimensions inherent to texts. We empirically grounded these dimensions by annotation studies and demonstrated that humans indeed judge similarity along different dimensions. Furthermore, we discussed common evaluation datasets and showed that it is of crucial importance for text similarity measures to address the correct dimensions. Otherwise, these measures fail to outperform even simple baselines.

We propose that future studies aiming at collecting human judgments on text similarity should explicitly state which dimension is targeted in order to create reliable annotation data. Further evaluation datasets annotated according to the structure and style dimensions of text similarity are necessary to guide further research in this field.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Klaus Tschira Foundation under project No. 00.133.2008. We thank György Szarvas for sharing his insights into the ESA similarity measure with us.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2.0. Journal of Technology, Learning, and Assessment, 4(3).

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In Proc. of the 20th International Conference on Computational Linguistics.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In Proc. of the 3rd Inl. Workshop on Paraphrasing, pages 17–24.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proc. of the 20th Inl. Joint Conference on Artificial Intelligence, pages 1806–1811.

Peter Gärdenfors. 2000. Conceptual Spaces: The Geometry of Thought. MIT Press.

Nelson Goodman. 1972. Seven strictures on similarity. In Problems and projects, pages 437–446. Bobbs-Merrill.

David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13(3):111–117.

Amiral Islam and Diana Inkpen. 2008. Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. ACM Transactions on Knowledge Discovery from Data, 2(2):1–25.

Alistair Kennedy and Stan Szpakowicz. 2008. Evaluating Roget’s Thesaurus. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 416–424.

Michael D. Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In Proceedings of the 27th Annual Conference of the Cognitive Science Society, pages 1254–1259.

Yuhua Li, David McLean, Zuhair Bandar, James O’Shea, and Keeley Crockett. 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8):1138–1150.

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

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In Proceedings of the 21st National Conference on Artificial Intelligence.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text Semantic Similarity for Automatic Short Answer Grading. In Proc. of the Europ. Chapter of the ACL, pages 567–575.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. 2006. Paragraph Recognition via Dissimilarity Significance Classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 18–26.

Daniel Ramage, Anna N. Rafferty, and Christopher D. Manning. 2009. Random Walks for Text Semantic Similarity. In Proceedings of the Workshop on Graph-based Methods for Natural Language Processing, pages 23–31.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. Communications of the ACM, 8(10):627–633.

John Sinclair, editor. 2001. Collins COBUILD Advanced Learner’s English Dictionary. HarperCollins, 3rd edition.

Linda B. Smith and Diana Heise. 1992. Perceptual similarity and conceptual structure. In B. Burns, editor, Percepts, Concepts, and Categories. Elsevier.

George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. 2010. Text relatedness based on a word thesaurus. Journal of Artificial Intell. Research, 37:1–39.

Amos Tversky. 1977. Features of similarity. In Psychologica Review, volume 84, pages 327–352.

Stephen Wan, Dras Mark, Robert Dale, and Cécile Paris. 2006. Using dependency-based features to take the “paraphrase” out of paraphrase. In Proc. of the Australasian Language Technology Workshop, pages 131–138.

Dominic Widdows. 2004. Geometry and Meaning. Center for the Study of Language and Information.

Eric Yeh, Daniel Ramage, Christopher D. Manning, Enkeo Agirre, and Aitor Soroa. 2009. WikiWalk: Random walks on Wikipedia for Semantic Relatedness. In Proceedings of the Workshop on Graph-based Methods for Natural Language Processing, pages 41–49.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Using Wiktorary for Computing Semantic Relatedness. In Proc. of the 23rd AAAI Conf. on AI, pages 861–867.

Yitao Zhang and Jon Patrick. 2005. Paraphrase Identification by Text Canonicalization. In Proc. of the Australasian Language Technology Workshop, pages 160–166.