Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity

Nils Reimers†, Philip Beyer‡, Iryna Gurevych†§
†Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
‡ zeb.rolfes.schierenbeck.associates GmbH
§ Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
http://www.ukp.tu-darmstadt.de

Abstract

Semantic Textual Similarity (STS) is a foundational NLP task and can be used in a wide range of tasks. To determine the STS of two texts, hundreds of different STS systems exist, however, for an NLP system designer, it is hard to decide which system is the best one. To answer this question, an intrinsic evaluation of the STS systems is conducted by comparing the output of the system to human judgments on semantic similarity. The comparison is usually done using Pearson correlation. In this work, we show that relying on intrinsic evaluations with Pearson correlation can be misleading. In three common STS based tasks we could observe that the Pearson correlation was especially ill-suited to detect the best STS system for the task and other evaluation measures were much better suited. In this work we define how the validity of an intrinsic evaluation can be assessed and compare different intrinsic evaluation methods. Understanding of the properties of the targeted task is crucial and we propose a framework for conducting the intrinsic evaluation which takes the properties of the targeted task into account.

1 Introduction

Semantic Textual Similarity (STS) is the foundational NLP task of determining the degree of semantic similarity between two texts. Most STS systems compute the similarity score between two texts on a fixed scale, for example a scale between 0 and 5, with 0 indicating the semantics are completely independent and 5 indicating semantic equivalence. In recent years, the number and quality of systems that rate the STS between texts have increased, as has the number of tasks where such systems are used.

Textual similarity is an active research field and was part of several shared tasks. In 2012, the pilot *Semantic Textual Similarity (STS) Task* (Agirre et al., 2012) was established at the *Semantic Evaluation (SemEval)* workshop. Further shared tasks on text similarity were part of SemEval 2013 (Agirre et al., 2013), SemEval 2014 (Agirre et al., 2014), SemEval 2015 (Agirre et al., 2015), and SemEval 2016 (Agirre et al., 2016). For the latest shared task on semantic textual similarity at SemEval 2016, 43 teams were submitting 119 different systems, depicting the large interest in this field.

STS is a foundational NLP technique, however, STS systems are seldom used for the sole purpose of measuring the similarity of two texts. Often they are used in a larger context. Examples for such tasks can be found in the field of Automatic Essay Grading (Attali et al., 2006), Plagiarism Detection (Potthast et al., 2012), Automated Text Summarization (Barzilay and Elhadad, 1997), Question Answering (Lin and Pantel, 2001), or Link Discovery (He, 2009). In this paper we call a task that heavily depends on the output of an STS system an *STS based task*. These tasks are often strongly dependent on the quality of the STS system they use, but they might apply further steps as well.

Given this large number of different STS systems, it is hard for an NLP system designer to decide which STS system should be implemented and used for a specific task. As such tasks often strongly depend on the quality of the STS system, the NLP system designer likes to use the most suitable system. To support the NLP system designer in this decision, the quality of STS systems is most often compared in an *intrinsic evaluation*. In the SemEval shared tasks on STS, the participating systems were asked to

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 87–96, Osaka, Japan, December 11-17 2016.

return a continuously valued similarity score given two texts. Performance is assessed by computing the Pearson correlation between machine assigned semantic similarity scores and human judgments (Agirre et al., 2016). Systems with a high Pearson correlation coefficient are considered as "good" STS systems and would often be the first choice for the system designer of an STS based task.

Usage of the Pearson correlation is common practice despite the fact that Agirre et al. (2013) state in the discussion of the results of the SemEval 2013 task about STS: "Evaluation of STS is still an open issue" and that beside the Pearson correlation "... other alternatives need to be considered, depending on the requirements of the target application." Up to our knowledge, no one published so far results whether Pearson correlation is a good method to evaluate the performance of different STS systems.

There are two factors defining the quality of intrinsic evaluations: The used dataset with the human judgments on similarity and the used evaluation measure to compare system outputs with the judgments. In this work, we will concentrate on the used evaluation measure. We studied three STS based tasks with different properties and evaluated 14 different STS systems. As the first task, we selected a classification task on text reuse using the Wikipedia Rewrite Corpus (Clough and Stevenson, 2011), as the second task a binary classification task of news article pairs on their relatedness, and as the third task one on detecting a related news article in a large corpus.

In our three examined tasks, we noticed that the Pearson correlation was misleading and especially ill-suited to predict the best STS system for the task. The performance of the STS systems in the intrinsic evaluation using Pearson correlation had no resemblance to their performance in the three different STS based tasks, i.e. for an NLP system designer the results of the intrinsic evaluation could be discarded. Other evaluation measures were much better in predicting which STS systems will perform well. In our experiments we could not observe that a single evaluation measure consistently produced the best predictions. The requirements on the STS systems for different tasks are too distinct, that a single evaluation measure could cope with all those. We thus claim that understanding the properties of the task and mapping them to the desired properties of the evaluation measure is crucial when selecting a measure for an intrinsic evaluation. Therefore, we propose in section 4 a new framework on the intrinsic evaluation of STS systems by taking the requirements of the target task into account.

This publication is based on the thesis of Beyer (2015). Some details in this paper are ommited for brevity and can be found online.[1]

2 Limitations of the Pearson Correlation and Alternative Evaluation Measures

Figure 1 depicts the output of four hypothetical STS systems in comparison to the gold standard derived from human judgment. These four distributions, also known as Anscombe's quartet, all have the same Pearson correlation coefficient of 0.816. By comparing only the Pearson correlation, all systems would be judged as equally good.

Pearson correlation is especially sensitive to non-linear relations, for example as depicted in the upper-right scatter plot, and to outliers, as depicted in the bottom scatter plots. In the scatter plot in the down-left corner, a single outlier is sufficient to disturb an otherwise perfect correlation. In the down-right corner the opposite is the case, a single outlier is sufficient to produce a high correlation of 0.816 even though there is no relationship between all other outputs and the human judgments. It is obvious that a human would judge the quality of these four STS systems quite differently, even though all four systems achieve the same Pearson correlation coefficient of 0.816.

2.1 Different STS Based Tasks Require Different Evaluation Measures

The usage of the Pearson correlation for the evaluation of STS systems has been questioned before. Zesch (2010) lists the limitations that the Pearson correlation is sensitive to outliers, that it can only measure a linear relationship, and that the two variables need to be approximately normally distributed. To overcome these limitations, Zesch recommends to use Spearman's rank correlation coefficient. The Spearman's rank correlation does not use the actual values to compute a correlation, but the ranking of

[1] https://www.ukp.tu-darmstadt.de/publications/details/?no_cache=1&tx_bibtex_pi1[pub_id]=TUD-CS-2015-12076

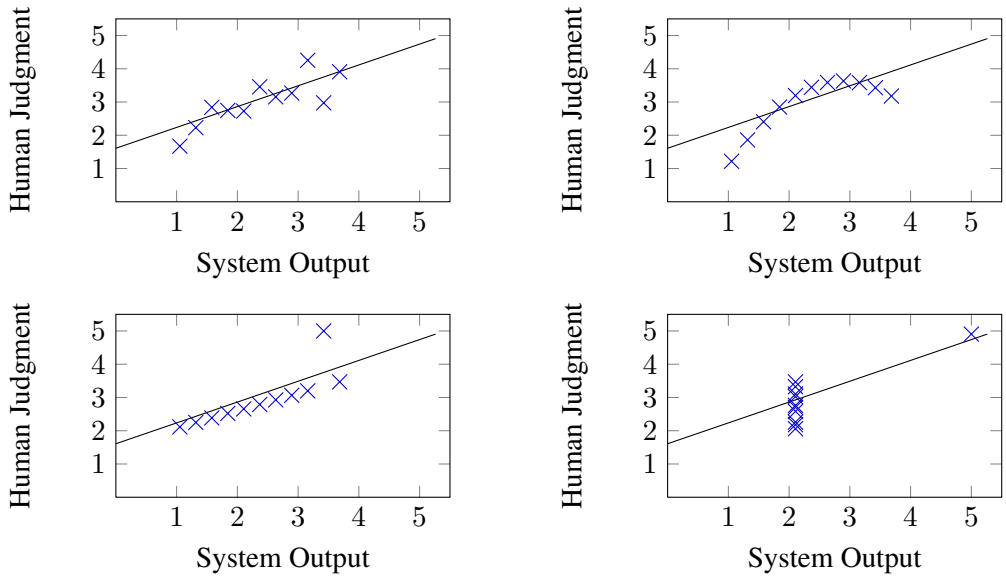

Figure 1: Anscombe's quartet with four different distributions. All distributions have a Pearson correlation coefficient of 0.816. All four STS systems would therefore be considered equally good.

the values. It is therefore not sensitive to outliers, non-linear relationships, or non-normally distributed data. However, most intrinsic evaluations of STS systems only report the Pearson correlation.

Depending on the STS based task, only some characteristics of the STS systems are important. For plagiarism detection, documents are often pre-filtered using an STS system and only documents with a score larger a certain threshold are passed for further inspection. For this task, only the decision whether the score is above the threshold is of importance, less the precise value. For the task of finding the top 10 most similar documents in a corpus for a search query, it is important that the STS system works well in distinguishing similar from dissimilar documents and is able to spot the most similar documents. Working perfectly on dissimilar documents and achieving the gold standard ranking for those is far less important than being able to find the few documents with high similarity. On the other hand, selecting semantically different sentences is important in automatic text summarization, therefore the STS system should work well to detect dissimilar text pairs.

It is unlikely that one evaluation measure can cope well with these different requirements. For Pearson and Spearman's rank correlation for example, all system outputs contribute equally, even though that for several tasks only some scores are relevant, often pairs that are especially similar or especially dissimilar. Using Pearson or Spearman's rank correlation therefore bears the risk that systems working especially well for the desired properties are missed. Hence, we study the following alternative evaluation measures for the intrinsic evaluation:

- The **normalized Cumulative Gain (nCG)** can be used to evaluate the ranking quality of STS scores (Järvelin and Kekäläinen, 2000). Let $\vec{m}$ be the vector with the gold STS values ranked by the STS system highest to lowest, i.e. at position 1 is the most similar pair according to the STS system and m_1 is the human judgment of this pair. The Cumulative Gain at k is defined as $CG_k = \sum_{i=1}^{k} m_i$. To normalize this value, it is divided by the so called ideal Cumulative Gain iCG_k which is the maximal value of CG_k for a perfect system. The normalized Cumulative Gain (nCG) is then defined as $nCG_k = \frac{CG_k}{iCG_k}$. Note that nCG is always equal 1 when k is equal to the number of text pairs.

- The **normalized Discounted Cumulative Gain (nDCG)** applies a discount factor to the normalized Cumulative Gain (Kekäläinen, 2005). It makes the assumption that similar text pairs are more important than less similar text pairs. An STS system which would score well with respect to the nDCG measure is well suited to find the most similar pairs in a corpus while it would be less suited to find the less similar, most distinct pairs. It is defined as $nDCG_k = \frac{DCG_k}{iDCG_k}$ with $DCG_k =$

$m_1 + \sum_{i=2}^{k} m_i / \log_2(i)$ and $iDCG_k$ the ideal Discounted Cumulative Gain. We can compute the $nDCG$ either for all text pairs or up to some position making an especially strong emphasis on the most similar pairs. Note: In case dissimilar pairs are more important for the targeted task, nCG and $nDCG$ can simply be modified by reversing the order of vector $\vec{m}$.

- **Accuracy** is a common evaluation measure for many tasks. However, as the STS scores are continuously valued, it is unclear how to compute it. One option is to define arbitrary bins and check whether the human judgments and the computed STS scores fall in the same bin. This requires that the minimal and maximal value of the STS systems is known and that there is a linear relationship between the STS scores and the human judgments. For the SemEval shared tasks, the systems were supposed to produce an output between 0 and 5 identical to the scale for the human judgments. In this work we set arbitrary borders at 1.5 and 3.5 with the intention to detect pairs with high similarity, e.g. for plagiarism detection, or pairs with low similarity, e.g. for detecting distinct sentences to be used in summarization. Accuracy^{low} describes the accuracy for scores below 1.5, and Accuracy^{high} describes the accuracy for scores higher than 3.5.

- Besides accuracy, the **$\mathbf{F_1}$-score** is a commonly used measure in several NLP tasks. Similar to accuracy, we have the challenge to define meaningful bins. As before, we set arbitrary borders at 1.5 and 3.5. F_1^{low} describes the F_1-score for low similarity pairs with a score below 1.5, and F_1^{high} describes the F_1-score for high similarity pairs with a score higher than 3.5.

Besides these evaluation measures, we define further measures by combining those using the harmonic mean $hmean(a, b) = 2ab/(a + b)$ and the unweighted macro average $macro\_avg(a, b) = (a + b)/2$. We also define the measure nDCG$_{AvgRank}$ which is the average of nDCG$_3$, nDCG$_5$, and nDCG$_{10}$ and nCG$_{AvgRank}$ which is the average of nCG$_3$, nCG$_5$, and nCG$_{10}$. These two measures only evaluate the top 3, 5, and 10 most similar text pairs according to the STS systems.

2.2 Impact of the Evaluation Measure on the Ranking

The evaluation measure of the intrinsic evaluation of STS systems can have a huge impact on the ranking of different STS systems. For the SemEval 2012 shared task on Semantic Textual Similarity (Agirre et al., 2012), 88 different STS systems were submitted by the participants. The official evaluation measure of this shared task used the Pearson correlation. We took the predicted STS scores and ranked the systems according to the alternative evaluation measures described in section 2. The results are depicted in Table 1. Using Spearman's rank correlation instead of Pearson correlation changed the positions of the STS systems on average by 6.6. The largest observed difference was 21 positions, i.e. a system that performed quite well according to the Pearson correlation coefficient achieved a mediocre result when the Spearman rank correlation coefficient is used. An even larger difference was observed when comparing Pearson correlation to the other presented evaluation measure like nDCG. The ranking of the STS systems according to the Pearson correlation was on average 19.0 positions different than the ranking according to nDCG.

The results show that the used evaluation measure plays an important role in defining the ranking of different STS systems in the intrinsic evaluation. A system which is assessed to be good using one evaluation measure, for example Pearson correlation, might perform badly according to another evaluation measure, for example Spearman's rank correlation.

3 Evaluation of the Predictiveness of Different Evaluation Measures

When the STS system is a crucial component of an STS based task, we expect that STS systems achieving good results in an intrinsic evaluation should in general lead to better results in the task and STS systems with weak results should in general lead to worse results for the STS based task. This allows us to define how *predictive* an intrinsic evaluation is:

> *Given the ranking of STS systems in an intrinsic evaluation as well as the ranking of the systems in an extrinsic evaluation, we say the intrinsic evaluation has **high predictiveness** when the*

Mean Absolute Difference	Spearman	nDCG$_{all}$	nDCG$_{AvgRank}$	Accuracy	F_1
Pearson	6.6	19.0	29.4	12.5	12.5
Spearman	-	19.6	29.1	12.0	15.2
nDCG$_{all}$		-	20.6	21.1	20.8
nDCG$_{AvgRank}$			-	31.8	26.3
Accuracy				-	14.3

Table 1: Mean Absolut Difference between ranks of submissions for the shared task on Semantic Textual Similarity at SemEval 2012 using different evaluation measures.

> *ranking in the intrinsic evaluation is similar to the ranking in the extrinsic evaluation. We say the **predictiveness** is **low**, when the ranking in the intrinsic evaluation is disconnected from the ranking of the systems in the extrinsic evaluation.*

In an ideal situation, the ranking of the STS systems in the intrinsic evaluation would be identical to the ranking in the extrinsic evaluation. As we cannot expect this, we need to define an objective measure on how similar these two rankings are. We can measure the resemblance of two rankings using the Mean Absolute Difference (MAD), the Mean Squared Difference (MSD), or the Spearman's rank correlation coefficient ρ. Given the rankings IR of the n STS systems in the intrinsic evaluation and the rankings ER in the extrinsic evaluation with IR_i corresponding to the rank of the i-th STS system in the intrinsic evaluation and ER_i corresponding to the rank in the extrinsic evaluation, the values are defined as:

$$\rho(IR, ER) = \frac{\text{cov}(IR, ER)}{\sigma_{IR}\sigma_{ER}}$$

$$\text{MSD}(IR, ER) = \frac{1}{n}\sum_{i=1}^{n}(IR_i - ER_i)^2$$

$$\text{MAD}(IR, ER) = \frac{1}{n}\sum_{i=1}^{n}|IR_i - ER_i|$$

where $\text{cov}(IR, ER)$ is the covariance of the rankings and σ_{IR}, σ_{ER} are the standard deviations of the rank variables. An intrinsic evaluation with high predictiveness would have MAD and MSD values close to 0 and a Spearman's correlation ρ close to 1. We would consider an intrinsic evaluation of STS system useful, when it scores well on MAD, MSD, and ρ values for a large range of STS based tasks.

3.1 Experiments

To assess the predictiveness of different evaluation measures presented in section 2, we chose three STS based tasks and evaluated 14 different STS systems. We used the implementation for these systems from the publically available framework DKPro Similarity[2]. For each STS system, we computed the score in an intrinsic evaluation. For the intrinsic evaluation, we used the datasets provided for the SemEval 2012 task on Semantic Textual Similarity (Agirre et al., 2012). The intrinsic evaluation was performed for 16 different evaluation measures described in section 2. We then compared the ranking of the different STS systems in the intrinsic evaluation with the ranking of those in the STS based task, which allows to compute the *predictiveness* of the (intrinsic) evaluation measure.

As our first STS based task, we chose the task of text reuse detection. Clough and Stevenson (2011) presented the Wikipedia Rewrite Corpus, a dataset with 95 documents, each containing an answer to one

[2] `https://dkpro.github.io/dkpro-similarity/`

of five questions about computer science. The answers employ different levels of reuse of a Wikipedia article. The degree of reuse was split in one of four categories: `near copy`, `light revision`, `heavy revision`, and `non plagiarized`. The performance for this task is evaluated by calculating the accuracy. To map the continuous output of the STS systems to the four categories, we used the One Rule (OneR) classifier (Holte, 1993) with optimized bucket sizes as well as a logistic regression classifier. The OneR classifier chooses a simple decision boundary for the different classes. Both have been evaluated using 10-fold cross-validation and the classifier with the better result was chosen.

The second and third STS based task uses a newly created corpus compiled from the German newspaper DIE ZEIT and ZEIT Online[3]. For most of the articles, the authors added two links to related articles that provide further information on the same news topic. The second task is a binary classification task with the goal to identify whether two articles are related or not. The ground truth is the original choice from the journalist. The OneR classifier was used to map the continuous STS score to the binary decision. Results were evaluated using 10-fold cross-validation.

The third STS based task tries to detect the two articles that are related to the target article in a set of articles from ZEIT Online. For the target article and each article in the set, we compute the STS score. The article in the set with the highest STS score was selected. We compared if this article is one of the related articles chosen by the author. Accuracy was computed for 100 randomly selected documents.

3.2 Results

We evaluated 14 different STS systems for the three presented tasks. For the first STS based task on text reuse detection, the best STS system achieved an accuracy of 70%, while the worst achieved an accuracy of 43%. For the second STS based task on deciding whether two news articles are related, the best STS system achieved an accuracy of 77%, while the worst achieved an accuracy of 48%. And for the third STS based task on finding the related article out of a set of articles, the best STS system achieved an accuracy of 67%, while the worst achieved an accuracy of 6%.

We compared the performance of the different STS systems in the three tasks with their performance in the intrinsic evaluation. We expect that the intrinsic evaluation allows us to distinguish between well performing STS systems and bad performing STS systems. Table 2 shows the Spearman ranking coefficient $\rho(IR, ER)$ between the performance of the STS measures in the intrinsic evaluation versus their performance on the STS based tasks. A coefficient close to 1 indicates that the ranking of the system in the intrinsic evaluation was similar to its performance in the STS based task. The values for the two other predictiveness indicators, *Mean Absolute Difference* and *Mean Squared Difference* can be found are nearly identical to the Spearman rank coefficient. Thus, we omit them for brevity.

3.3 Discussion

In the three studied STS based tasks, there was no correlation between the performance of STS systems on the intrinsic evaluation using Pearson or Spearman rank correlation and their performance in the STS based tasks. In two cases, the correlation $\rho(IR, ER)$ between the intrinsic ranking IR and extrinsic ranking ER was even negative, indicating that STS systems that performed well in the intrinsic evaluation performed especially poorly in STS based tasks. From an engineering perspective this raises serious doubts about the value of an intrinsic evaluation that uses Pearson correlation.

Using other measures than Pearson correlation for the intrinsic evaluation however enabled a much better prediction of the performance of the STS systems for the STS based task. A strong STS system in such an intrinsic evaluation was also able to perform well in the STS based task. It is interesting to note that other STS based tasks were especially good predictors, i.e. an STS system performing well in task 1 was also performing well in task 2 and task 3, even though the characteristics of these tasks were very distinct. In all three tasks, the same STS system achieved the best result. However, in none of the performed intrinsic evaluations achieved this system the best place and was placed on 2nd to 6th place depending on the used evaluation measure. The system that achieved the best place in various intrinsic

[3] `http://www.zeit.de`

Intrinsic Evaluation Measure	Task 1		Task 2		Task 3	
	ρ	Rank	ρ	Rank	ρ	Rank
nDCG$_{AvgRank}$	0.504	1	0.380	6	0.338	5
nCG$_{AvgRank}$	0.504	1	0.380	6	0.338	5
F_1^{low}	0.497	3	0.717	2	0.611	2
nDCG	0.431	4	0.238	10	0.238	8
hmean(F_1^{low}, F_1^{high})	0.427	5	0.722	1	0.614	1
hmean(Pearson,F_1)	0.264	6	0.686	3	0.536	3
hmean(Spearman,F_1)	0.163	7	0.594	4	0.439	4
macro_avg(F_1^{low}, F_1^{high})	-0.053	8	0.422	5	0.289	7
hmean(Pearson, nCG$_{AvgRank}$)	-0.136	9	0.339	8	0.130	9
hmean(Spearman, nCG$_{AvgRank}$)	-0.216	10	0.277	9	0.089	10
Accuracy^{low}	-0.277	11	0.057	14	-0.062	13
Pearson correlation	-0.326	12	0.198	11	-0.031	11
Spearman's rank correlation	-0.343	13	0.172	12	-0.040	12
hmean(Accuracy^{low},Accuracy^{high})	-0.370	14	0.031	15	-0.113	15
Accuracy^{high}	-0.378	15	0.062	13	-0.102	14
F_1^{high}	-0.524	16	-0.123	16	-0.283	16
Task 1: Text reuse classification			0.70		0.82	
Task 2: Binary classification of article pairs	0.70				0.91	
Task 3: Related article detection	0.82		0.91			

Table 2: The Spearman rank correlation $\rho(IR, ER)$ between the intrinsic ranking IR and the extrinsic ranking ER for the three evaluated STS based tasks: (1) Text reuse classification, (2) binary classification of article pairs, and (3) related article detection. A ρ-coefficient close to 1 means a large correlation between the performance in the intrinsic evaluation and the performance in the STS based task. The *Rank* depicts the ranking, highest to lowest, of the ρ-coefficients for each task.

evaluations performed quite poorly for the STS based tasks only achieving the 6th, 9th, and 12th place, respectively, out of 14 tested systems.

4 Proposal of an Evaluation Framework for Semantic Textual Similarity

On a well-designed and representative dataset, an STS system should show similar behavior in the intrinsic evaluation as it will show for real world data of STS based tasks. For example, in case the STS system is well suited to find the most similar text pairs in the intrinsic evaluation set, then it will likely also be suitable to find the most similar text pairs for other datasets. This STS system would then be useful for tasks where finding the most similar text pairs is essential.

However, different STS based tasks have different requirements on STS systems and, therefore, different properties of STS systems are important. After studying the most common STS based tasks, we propose the following three dimensions to classify the requirements of an STS based task:

- **Cardinality** describes how many texts are compared to how many others. It consists of two sub categories *1:1* and *1:n*. *1:1* in this context means that exactly one text is compared with exactly one other text and only the result of this single comparison is of interest. *1:n* means that one text will be compared with a whole set of other texts and the results of these comparisons will be used in some way. The third option, *m:n*, would theoretically be possible, but no example of this was found.

- **Set of Interest** describes which of the elements of the result set will be used. It has three sub categories: *All*, *k-best*, and *Threshold*. *All* in this context means that all results of all comparisons will be used in some form. *k-best* describes the case where only the "k" best results will be used in some way. And *Threshold* is used when only results over a certain threshold will be used.

- **Information** describes the type of information from the result set that is of interest. It has three sub categories: *Value*, *Rank*, and *Classification*. The case where the actual value of the result of a comparison is of interest falls in the category *Value*. *Rank* on the other hand is used if only the rank of each comparison is used in some way. *Classification* means that a simple classification, for example texts are similar or not, is used.

The first STS based task in section 3.1 on text reuse is an example for a task with cardinality *1:1*, as one text, the answer, is compared to only one other text, the Wikipedia article. The STS score is classified into one of four categories, hence, the *Information* of interest is *Classification*. Tasks of such type can only tolerate minimal variety in the STS scores, i.e. texts of similar similarity should be mapped to similar scores independent of other factors like text length etc. Otherwise, the classification into categories doesn't work well.

The third STS based task in section 3.1 on detecting the related articles in a set of articles is an example for a task with cardinality *1:n*, set of interest *k-best* and information *Rank*. For this task, one document is compared to a set of other documents and the user is interested in the most similar pair. STS systems for this task should be good at ranking text pairs according to their similarity.

With the three dimensions 18 different combinations are possible. However, some of these combinations can be disregarded because they are not plausible. For any combination that involves a *Cardinality* of *1:1* only a *Set of Interest* of *All* is useful, because the result set contains only one result. In addition, the *Information* can't be *Rank*. Overall, only nine combinations are plausible. All nine possible combinations with examples of STS based tasks are described in detail in (Beyer, 2015).

For these nine plausible combinations, we propose in Table 3 an evaluation measure for the intrinsic evaluation that should capture the requirements of the target task. The proposed evaluation measures take similar characteristics into account that are required for the task. An NLP system designer could use this framework to determine the requirements of his task. Instead of selecting the STS system with the best Pearson correlation, the system designer would use the selected evaluation measure to run his own ranking of the STS systems to spot potentially strong STS systems for his task. The reasoning for the individual choices is given in (Beyer, 2015).

Requirements	**Proposed Evaluation Measure for Intrinsic Evaluation**
(1:1, All, Classification)	harmonic mean of F_1-score for low and high similarity pairs
(1:1, All, Value)	Pearson correlation
(1:n, All, Rank)	nDCG or Spearman rank correlation
(1:n, All, Classification)	harmonic mean of F_1-score for low and high similarity pairs
(1:n, All, Value)	Pearson correlation
(1:n, k-best, Value)	harmonic mean of nCG_k and Pearson correlation
(1:n, k-best, Rank)	nDCG_k
(1:n, Threshold, Value)	harmonic mean of F_1-score for low and high similarity pairs and Pearson correlation
(1:n, Threshold, Rank)	harmonic mean of F_1-score for low and high similarity pairs and Spearman rank correlation

Table 3: This Semantic Textual Similarity Framework proposes an evaluation measure for intrinsic evaluation based on the requirements of the target task.

An extensive evaluation of this framework is topic of our future research. The focus of this paper was establishing the need of alternative evaluation measures besides Pearson correlation and how to assess the quality of intrinsic evaluations. We encourage future work by others on this topic to find an intrinsic evaluation that meets the diverse needs of STS based tasks.

5 Conclusion and Future Work

In this paper we demonstrated the challenges of the intrinsic evaluation of STS systems. We introduced the concept of *predictiveness*: An STS system performing well in an intrinsic evaluation should also perform well for STS based tasks. This notion of predictiveness allows us to compare different evaluation measures besides the commonly used Pearson correlation. For three studied tasks we could observe that the predictiveness of an intrinsic evaluation with Pearson correlation is fairly low or even negative. We presented other evaluation measures which had a much higher predictiveness, i.e. those methods could predict much better which STS systems perform well in the STS based tasks. Based on this, we proposed a framework how to evaluate STS scores that take the requirements of the target task into account.

For future intrinsic evaluations of STS systems we find it crucial that not only the Pearson correlation is published, but additionally the STS scores generated by the systems can be downloaded. This allows to compute other evaluation measures, for example Spearman's ranking correlation, nDCG, or F_1-score. It also allows NLP system designers to select an evaluation measure for the intrinsic evaluation that captures the important characteristics needed by their target task.

In our experiments we could observe that the predictiveness of other extrinsic evaluations is high. Systems performing well on the task of text reuse of English Wikipedia articles also did well for detecting related articles on German news articles despite the fact of a different language, a different text genre and a completely different task. An alternative evaluation method of STS systems could be to test those on a broad range of different STS based tasks. The design of these STS based tasks must be standardized and the impact of components or features besides the STS system should be reduced to a minimum.

Acknowledgements

This work has been supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1, by the German Institute for Educational Research (DIPF) and by the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus program under the promotional reference 01-S12054.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 385–393, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of SemEval-2016*, pages 509–523, San Diego, California, June. Association for Computational Linguistics.

Yigal Attali, Jill Burstein, Yigal Attali, Jill Burstein, Michael Russell, Design Thomas Hoffmann, Yigal Attali, and Jill Burstein. 2006. Automated Essay Scoring with E-Rater V.2. *Journal of Technology, Learning, and Assessment*.

Regina Barzilay and Michael Elhadad. 1997. Using Lexical Chains for Text Summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Philip Beyer. 2015. Proposal for a STS Evaluation Framework for STS based Applications. Master's thesis, Technische Universität Darmstadt, March. Available online: `https://www.ukp.tu-darmstadt.de/publications/details/?no_cache=1&tx_bibtex_pi1[pub_id]=TUD-CS-2015-12076`.

Paul Clough and Mark Stevenson. 2011. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24.

Jiyin He, 2009. *Link Detection with Wikipedia*, pages 366–373. Springer Berlin Heidelberg, Berlin, Heidelberg.

Robert C. Holte. 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, 11(1):63–90.

Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 41–48, New York, NY, USA. ACM.

Jaana Kekäläinen. 2005. Binary and Graded Relevance in IR evaluations-Comparison of the Effects on Ranking of IR Systems. *Inf. Process. Manage.*, 41(5):1019–1033, September.

Dekang Lin and Patrick Pantel. 2001. Discovery of Inference Rules for Question-answering. *Nat. Lang. Eng.*, 7(4):343–360, December.

Martin Potthast, Tim Gollub, Matthias Hagen, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. 2012. Overview of the 4th International Competition on Plagiarism Detection. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*.

Torsten Zesch. 2010. *Study of Semantic Relatedness of Words Using Collaboratively Constructed Semantic Resources*. Ph.D. thesis, Technische Universität, Darmstadt.