UNED: Evaluating Text Similarity Measures without Human Assessments*

Enrique Amigó † Julio Gonzalo † Jesús Giménez ‡ Felisa Verdejo†

† UNED, Madrid
{enrique,julio,felisa}@lsi.uned.es

‡ Google, Dublin
jesgim@gmail.com

Abstract

This paper describes the participation of UNED NLP group in the SEMEVAL 2012 Semantic Textual Similarity task. Our contribution consists of an unsupervised method, Heterogeneity Based Ranking (HBR), to combine similarity measures. Our runs focus on combining standard similarity measures for Machine Translation. The Pearson correlation achieved is outperformed by other systems, due to the limitation of MT evaluation measures in the context of this task. However, the combination of system outputs that participated in the campaign produces three interesting results: (i) Combining all systems without considering any kind of human assessments achieve a similar performance than the best peers in all test corpora, (ii) combining the 40 less reliable peers in the evaluation campaign achieves similar results; and (iii) the correlation between peers and HBR predicts, with a 0.94 correlation, the performance of measures according to human assessments.

1 Introduction

Imagine that we are interested in developing computable measures that estimate the semantic similarity between two sentences. This is the focus of the STS workshop in which this paper is presented. In order to optimize the approaches, the organizers provide a training corpus with human assessments. The participants must improve their approaches and select three runs to participate. Unfortunately, we can not ensure that systems will behave similarly in both the training and test corpora. For instance, some Pearson correlations between system achievements across test corpora in this competition are: 0.61 (MSRpar-MSRvid), 0.34 (MSRvid-SMTeur), or 0.49 (MSRpar-SMTeur). Therefore, we cannot expect a high correlation between the system performance in a specific corpus and the test corpora employed in the competition.

Now, imagine that we have a *magic box* that, given a set of similarity measures, is able to predict which measures will obtain the highest correlation with human assessments without actually requiring those assessments. For instance, suppose that putting all system outputs in the magic box, we obtain a 0.94 Pearson correlation between the prediction and the system achievements according to human assessments, as in Figure 1. The horizontal axis represents the magic box ouput, and the vertical axis represents the achievement in the competition. Each dot represents one system. In this case, we could decide which system or system combination to employ for a certain test set.

Is there something like this magic box? The answer is yes. Indeed, what Figure 1 shows is the results of an unsupervised method to combine measures, the *Heterogeneity Based Ranking* (HBR). This method is grounded on a generalization of the heterogeneity property of text evaluation measures proposed in (Amigó et al., 2011), which states that the more a set of measures is heterogeneous, the

*This work has been partially funded by the Madrid government, grant MA2VICMR (S-2009/TIC- 1542), the Spanish government, grant Holopedia (TIN2010-21128-C02-01) and the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreement nr. 288024 (LiMoSINe project).

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 454–460,
Montréal, Canada, June 7-8, 2012.

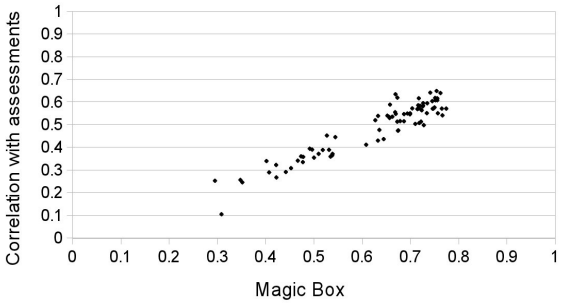

Figure 1: Correspondence between the magic box information and the (unknown) correlation with human assessments, considering all runs in the evaluation campaign.

more a score increase according to all the measures is reliable. In brief, the HBR method consists of computing the heterogeneity of the set of measures (systems) for which a similarity instance (pair of texts) improves each of the rest of similarity instances in comparison. The result is that HBR tends to achieve a similar or higher correlation with human assessments than the single measures. In order to select the most appropriate single measure, we can meta-evaluate measures in terms of correlation with HBR, which is what the previous figure showed.

We participated in the STS evaluation campaign employing HBR over automatic text evaluation measures (e.g. ROUGE (Lin, 2004)), which are not actually designed for this specific problem. For this reason our results were suboptimal. However, according to our experiments this method seem highly useful for combining and evaluating current systems. In this paper, we describe the HBR method and we present experiments employing the rest of participant methods as similarity measures.

2 Definitions

2.1 Similarity measures

In (Amigó et al., 2011) a novel definition of similarity is proposed in the context of automatic text evaluation measures. Here we extend the definition for text similarity problems in general.

Being Ω the universe of texts d, we assume that a similarity measure, is a function $x : \Omega^2 \longrightarrow \Re$ such that there exists a decomposition function $f : \Omega \longrightarrow \{e_1..e_n\}$ (e.g., words or other linguistic units or relationships) satisfying the following constraints; (i) maximum similarity is achieved only when the text decomposition resembles exactly the other text; (ii) adding one element from the second text increases the similarity; and (iii) removing one element that does not appear in the second text also increases the similarity.

$$f(d_1) = f(d_2) \leftrightarrow x(d_1, d_2) = 1$$

$$(f(d_1) = f(d'_1) \cup \{e \in f(d_2) \setminus f(d_1)\}) \rightarrow x(d'_1, d_2) > x(d_1, d_2)$$

$$(f(d_1) = f(d'_1) - \{e \in f(d_1) \setminus f(d_2)\}) \rightarrow x(d'_1, d_2) > x(d_1, d_2)$$

According to this definition, a random function, or the inverse of a similarity function (e.g. $\frac{1}{x(d_1 d_2)}$), do not satisfy the similarity constraints, and therefore cannot be considered as similarity measures. However, this definition covers any kind of overlapping or precision/recall measure over words, syntactic structures or semantic units, which is the case of most systems here.

Our definition assumes that measures are granulated: they decompose text in a certain amount of elements (e.g. words, grammatical tags, etc.) which are the basic representation and comparison units to estimate textual similarity.

2.2 Heterogeneity

Heterogeneity (Amigó et al., 2011) represents to what extent a set of measures differ from each other. Let us refer to a pair of texts $i = (i_1, i_2)$ with a certain degree of similarity to be computed as a *similarity instance*. Then we estimate the Heterogeneity $H(\mathcal{X})$ of a set of similarity measures $\mathcal{X}$ as the probability over similarity instances $i = (i_1, i_2)$ and $j = (j_1, j_2)$ between distinct texts, that there exist two measures in $\mathcal{X}$ that contradict each other. Formally:

$$H(\mathcal{X}) \equiv P_{\substack{i_1 \neq i_2 \\ j_1 \neq j_2}}(\exists x, x' \in \mathcal{X} | x(i) > x(j) \wedge x'(j) < x'(i))$$

where $x(i)$ stands for the similarity, according to measure x, between the texts i_1, i_2.

3 Proposal: Heterogeneity-Based Similarity Ranking

The heterogeneity property of text evaluation measures (in fact, text similarity measures to human references) introduced in (Amigó et al., 2011) states that the quality difference between two texts is lower bounded by the heterogeneity of the set of evaluation measures that corroborate the quality increase. Based on this, we define the *Heterogeneity Principle* which is applied to text similarity in general as: *the probability of a real similarity increase between random text pairs is correlated with the Heterogeneity of the set of measures that corroborate this increase:*

$$P(h(i) \geq h(j)) \sim H(\{x|x(i) \geq x(j)\})$$

where $h(i)$ is the similarity between i_1, i_2 according to human assessments (gold standard). In addition, the probability is maximal if the heterogeneity is maximal:

$$H(\{x|x(i) \geq x(j)\}) = 1 \Rightarrow P(h(i) \geq h(j)) = 1$$

The first part is derived from the fact that increasing Heterogeneity requires additional diverse measures corroborating the similarity increase. The direct relationship is the result of assuming that a similarity increase according to any aspect is always a positive evidence of true similarity. In other words, a positive match between two texts according to any feature can never be a negative evidence of similarity.

As for the second part, if the heterogeneity of a measure set is maximal, then the condition of the heterogeneity definition holds for any pair of distinct documents ($i_1 \neq i_2$ and $j_1 \neq j_2$). Given that all measures corroborate the similarity increase, the heterogeneity condition does not hold. Then, the compared texts in (i_1, i_2) are not different. Therefore, we can ensure that $P(h(i) \geq h(j)) = 1$.

The proposal in this paper consists of ranking similarity instances by estimating, for each instance i, the average probability of its texts (i_1, i_2) being closer to each other than texts in a different instance j:

$$\mathrm{R}(i) = \mathrm{Avg}_j(P(h(i) \geq h(j)))$$

Applying the heterogeneity principle we can estimate this as:

$$\mathrm{HBR}_{\mathcal{X}}(i) = \mathrm{Avg}_j(H(\{x|x(i) \geq x(j)\}))$$

We refer to this ranking function as the *Heterogeneity Based Ranking* (HBR). It satisfies three crucial properties for a measure combining function:

1. HBR is independent from measure scales and it does not require relative weighting schemes between measures. Formally, being f any strict growing function:

$$\mathrm{HBR}_{x_1..x_n}(i) = \mathrm{HBR}_{x_1..f(x_n)}(i)$$

2. HBR is not sensitive to redundant measures:

$$\mathrm{HBR}_{x_1..x_n}(i) = \mathrm{HBR}_{x_1..x_n,x_n}(i)$$

3. Given a large enough set of similarity instances, HBR is not sensitive to non-informative measures. Being x_r a random function such that $P(x_r(i) > x_r(j)) = \frac{1}{2}$, then:

$$\mathrm{HBR}_{x_1..x_n}(i) \sim \mathrm{HBR}_{x_1..x_n,x_r}(i)$$

The first two properties are trivially satisfied: the $\exists$ operator in H and the score comparisons are not affected by redundant measures nor their scales properties. Regarding the third property, the heterogeneity of a set of measures plus a random function x_r is:

$$H(\mathcal{X} \cup \{x_r\}) \equiv$$

$$P_{\substack{i_1 \neq i_2 \\ j_1 \neq j_2}}(\exists x, x' \in \mathcal{X} \cup \{x_r\} | x(i) > x(j) \wedge x'(j) < x'(i)) =$$

$$H(\mathcal{X}) + (1 - H(\mathcal{X})) * \frac{1}{2} = \frac{H(\mathcal{X}) + 1}{2}$$

That is, the heterogeneity grows proportionally when including a random function. Assuming that the random function corroborates the similarity increase in a half of cases, the result is a proportional relationship between HBR and HBR with the additional measure. Note that we need to assume a large enough amount of data to avoid random effects.

4 Official Runs

We have applied the HBR method with excellent results in different tasks such as Machine Translation and Summarization evaluation measures, Information Retrieval and Document Clustering. However, we had not previously applied our method to semantic similarity. Therefore, we decided to apply directly automatic evaluation measures for Machine Translation as single similarity measures to be combined by means of HBR. We have used 64 automatic evaluation measures provided by the ASIYA Toolkit (Giménez and Màrquez, 2010)[1]. This set includes measures operating at different linguistic levels (lexical, syntactic, and semantic) and includes all popular measures (BLEU, NIST, GTM, METEOR, ROUGE, etc.) The similarity formal constraints in this set of measures is preserved by considering lexical overlap when the target linguistic elements (i.e. named entities) do not appear in the texts.

We participated with three runs. The first one consisted of selecting the best measure according to human assessments in the training corpus. It was the INIST measure (Doddington, 2002). The second run consisted of selecting the best 34 measures in the training corpus and combining them with HBR, and the last run consisted of combining all evaluation measures with HBR. The heterogeneity of measures was computed over 1000 samples of similarity instance pairs (pairs of sentences pairs) extracted from the five test sets. Similarity instances were ranked over each test set independently.

In essence, the main contribution of these runs is to corroborate that Machine Translation evaluation measures are not enough to solve this task. Our runs appear at the Mean Rank positions 42, 28 and 77. Apart of this, our results corroborate our main hypothesis: without considering human assessment or any kind of supervised tunning, combining the measures with HBR resembles the best measure (INIST) in the combined measure set. However, when including all measures the evaluation result decreases (rank 77). The reason is that some Machine Translation evaluation measures do not represent a positive evidence of semantic similarity in this corpus. Therefore, the HBR assumptions are not satisfied and the final correlation achieved is lower. In summary, our approach is suitable if we can ensure that all measures (systems) combined are at least a positive (high or low) evidence of semantic similarity.

But let us focus on the HBR behavior when combining participant measures, which are specifically designed to address this problem.

[1] http://www.lsi.upc.edu/ nlp/Asiya

5 Experiment with Participant Systems

5.1 Combining System Outputs

We can confirm empirically in the official results that all participants runs are positive evidence of semantic similarity. That is, they achieve a correlation with human assessments higher than 0. Therefore, the conditions to apply HBR are satisfied. Our goal now is to resemble the best performance without accessing human assessments neither from the training nor the test corpora. Figure 2 illustrates the Pearson correlation (averaged across test sets) achieved by single measures (participants) and all peers combined in an unsupervised manner by HBR (black column). As the figure shows, HBR results are comparable with the best systems appearing in the ninth position. In addition, Figure 4 shows the differences over particular test sets between HBR and the best system. The figure shows that there are not consistent differences between these approaches across test beds.

The next question is why HBR is not able to improve the best system. Our intuition is that, in this test set, average quality systems do not contribute with additional information. That is, the similarity aspects that the average quality systems are able to capture are also captured by the best system.

However, the best system within the combined set is not a theoretical upper bound for HBR. We can prove it with the following experiment. We apply HBR considering only the 40 less predictive systems in the set (the rest of measures are not considered when computing HBR). Then we compare the results of HBR regarding the considered single systems. As Figure 3 shows, HBR improves substantially all single systems achieving the same result than when combining all systems (0.61). The reason is that all these systems are positive evidences but they consider partial similarity aspects. But the most important issue here is that combining the 40 less predictive systems in the evaluation campaign

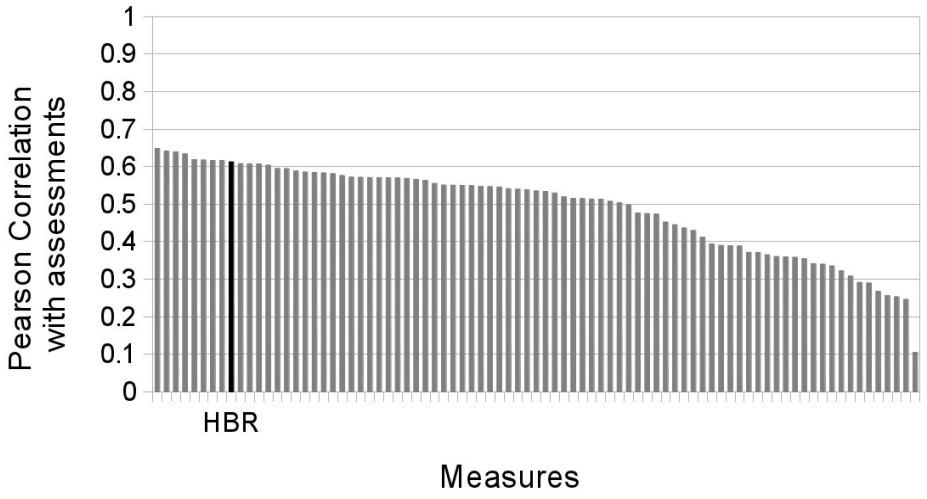

Figure 2: Measures (runs) and HBR sorted by average correlation with human assessments.

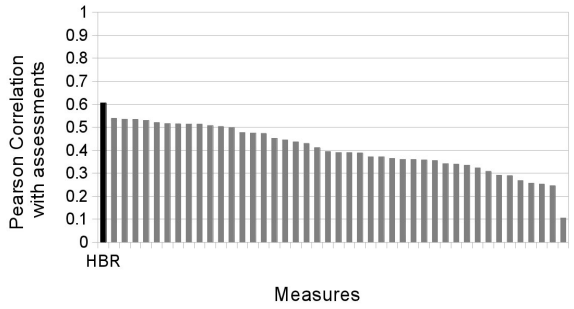

Figure 3: 40 less predictive measures (runs) and HBR sorted by average correlation with human assessments.

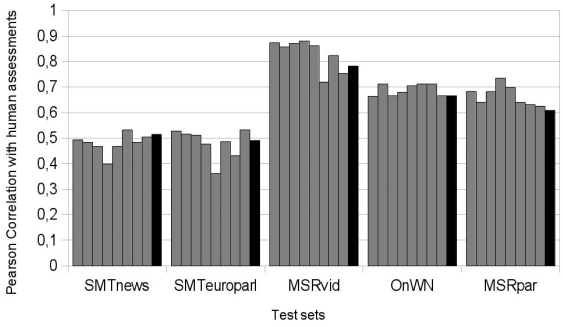

Figure 4: Average correlation with human assessments for the best runs and HBR.

is enough to achieve high final scores. This means that the drawback of these measures as a whole is not what information is employed but how this information is scaled and combined. This drawback is solved by the HBR approach.

In summary, the main conclusion that we can extract from these results is that, in the absence of human assessments, HBR ensures a high performance without the risk derived from employing potentially biased training corpora or measures based on partial similarity aspects.

6 An Unsupervised Meta-evaluation Method

But HBR has an important drawback: its computational cost, which is $\mathcal{O}(n^4 * m)$, being n the number of texts involved in the computation and m the number of measures. The reason is that computing H is quadratic with the number of texts, and the method requires to compute H for every pair of texts. In addition, HBR does not improve the best systems.

However, HBR can be employed as an unsupervised evaluation method. For this, it is enough to compute the Pearson correlation between runs and HBR. This is what Figure 1 showed at the beginning of this article. For each dot (participant run), the horizontal axis represent the correlation with HBR (magic box) and the vertical axis represent the correlation with human assessments. This graph has a Pearson correlation of 0.94 between both variables. In other words, without accessing human assessments, this method is able to predict the quality of

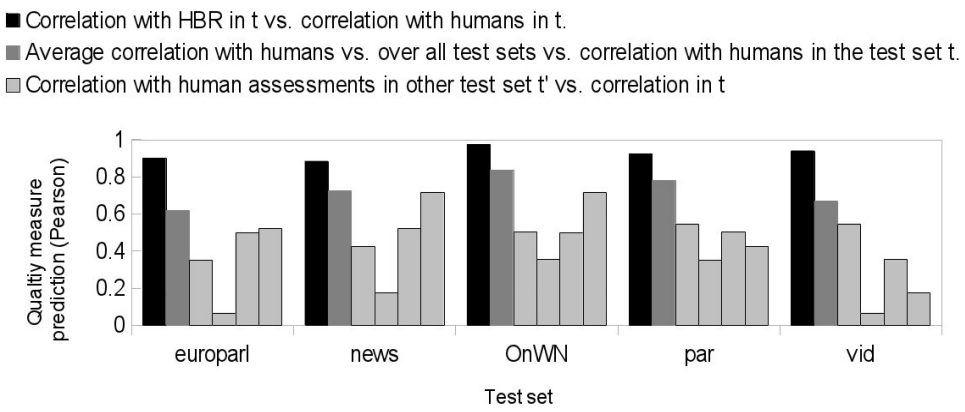

Figure 5: Predicting the quality of measures over a single test set.

textual similarity system with a 0.94 of accuracy in this test bed.

In this point, we have two options for optimizing systems. First, we can optimize measures according to the results achieved in an annotated training corpus. The other option consists of considering the correlation with HBR in the test corpus. In order to compare both approaches we have developed the following experiment. Given a test corpus t, we compute the correlation between system scores in t versus a training corpus t'. This approach emulates the scenario of training systems over a (training) set and evaluating over a different (test) set. We also compute the correlation between system scores in all corpora vs. the scores in t. Finally, we compute the correlation between system scores in t and our predictor in t (which is the correlation system/HBR across similarity instances in t). This approach emulates the use of HBR as unsupervised optimization method.

Figure 5 shows the results. The horizontal axis represents the test set t. The black columns represent the prediction over HBR in the corresponding test set. The grey columns represent the prediction by using the average correlation across test sets. The light grey columns represents the prediction using the correlation with humans in other single test set. Given that there are five test sets, the figure includes four grey columns for each test set. The figure clearly shows the superiority of HBR as measure quality predictor, even when it does not employ human assessments.

7 Conclusions

The Heterogeneity Based Ranking provides a mechanism to combine similarity measures (systems) without considering human assessments. Interestingly, the combined measure always improves or achieves similar results than the best single measure in the set. The main drawback is its computational cost. However, the correlation between single measures and HBR predicts with a high confidence the accuracy of measures regarding human assessments. Therefore, HBR is a very useful tool when optimizing systems, specially when a representative training corpus is not available. In addition, our results shed some light on the contribution of measures to the task. According to our experiments, the less reliable measures as a whole can produce reliable results if they are combined according to HBR.

The HBR software is available at http://nlp.uned.es/∼enrique/

References

Enrique Amigó, Julio Gonzalo, Jesus Gimenez, and Felisa Verdejo. 2011. Corroborating text evaluation results with heterogeneous measures. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 455–466, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd Inter-*

national Conference on Human Language Technology, pages 138–145.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Chin-Yew Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.