Summarization Evaluation in the Absence of Human Model Summaries Using the Compositionality of Word Embeddings

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, Fang Chen
University of New South Wales, Sydney, Australia
Data61 CSIRO, Sydney, Australia
{elahehs,mohammade,wong,fang}@cse.unsw.edu.au

Abstract

We present a new summary evaluation approach that does not require human model summaries. Our approach exploits the compositional capabilities of corpus-based and lexical resource-based word embeddings to develop the features reflecting coverage, diversity, informativeness, and coherence of summaries. The features are then used to train a learning model for predicting the summary content quality in the absence of gold models. We evaluate the proposed metric in replicating the human assigned scores for summarization systems and summaries on data from query-focused and update summarization tasks in TAC 2008 and 2009. The results show that our feature combination provides reliable estimates of summary content quality when model summaries are not available.

1 Introduction

Quantifying the quality of summaries is an important and necessary task in the field of automatic text summarization. Current summary evaluation methods like manual and automated pyramid (Passonneau et al., 2005; Passonneau et al., 2013) and well-established ROUGE scores (Lin, 2004) heavily rely on multiple human-generated model summaries to assess the quality of system-generated summaries. This evaluation paradigm falls short on non-standard test sets where model summaries are not available. According to the quantitative analysis by Louis and Nenkova (2009a); Singh and Jin (2016), evaluating summaries by their comparison with the input obtains good correlations with manual evaluations. Therefore, identifying a suitable input-summary similarity metric will provide a means for model-free evaluation of summaries.

We hypothesize that comparing semantic representations of the input and summary content will lead to more accurate input-summary evaluation. Hence, we explore the effectiveness of compositionality of word embeddings in developing a model-free automatic metric to evaluate summary content quality. In particular, our approach incorporates the word embedding models trained on the Google News corpus and the WordNet lexical resource to compare centroid vectors of the input and summary. To demonstrate the effectiveness of our approach, we have conducted a set of experiments on data from query-focused and update summarization tasks in TAC[1] 2008 and 2009. The reliability of our metric is also studied conducting an error analysis. The experiment results show that quantifying the indicators of content quality by taking advantage of compositional properties of the word and sense embeddings produces summary scores which accurately replicate human assessments. It is noteworthy that our approach complements but is not intended to replace existing model-based evaluation approaches, since their reliability and strength are important for high confidence evaluations.

2 Related Work

Proposals for developing automatic summary evaluation methods (Ellouze et al., 2013; Ng and Abrecht, 2015; ShafieiBavani et al., 2017) have been put forward in the past. However, these methods are not

[1] http://www.nist.gov/tac/

Proceedings of the 27th International Conference on Computational Linguistics, pages 905–914
Santa Fe, New Mexico, USA, August 20-26, 2018.

applicable on non-standard test sets where model summaries are not available. Herein, we try to briefly review the most significant approaches that have addressed this issue. Donaway et al. (2000) proposed an alternative to model-based evaluation where a comparison of the input text with a summary can clarify how good the summary is. A summary that has higher similarity with the input text can be considered better than one with lower similarity. Radev et al. (2003) performed an automated ranking of the test documents using a search engine scenario. Their approach was motivated by the assumption that the distribution of terms in a good summary is similar to the distribution of terms in the input document.

With the same intuition, Louis and Nenkova (2009a; 2013) introduced an evaluation system (SIMetrix) that comprises multiple features to determine the quality of a summary. Their focus was on computing divergences between the probability distributions of words in the input and summary. Jensen Shannon divergence and feature regression turned out to be their best metrics. Louis and Nenkova (2009b) also presented a similar evaluation approach utilizing a collection of large number of system summaries in place of model summaries. Saggion et al. (2010); Cabrera-Diego and Torres-Moreno (2017) proposed follow-up works to SIMetrix to assess the usefulness of divergences for multilingual summarization evaluation, and the applicability of multiple divergences for evaluating summaries.

Alternatively, we assume that the way of representing the input and summary is a key factor in high performance prediction of manual metrics. To this end, we exploit the compositional capabilities of word embeddings to design our features of content quality based on the continuous vector representation of words and senses. We then present a quantitative analysis of our features for characterizing the relation between the input and summary content in the absence of model summaries. We have finally evaluated our approach through two levels of granularity on two years data in TAC for query-focused and update summarization tasks. We have also compared our approach with model-based ROUGE, and model-free SIMetrix as an input-summary evaluation metric. The results demonstrate that the Support Vector Regression (SVR) of our features achieves the best correlation with manual judgments.

3 Data and Evaluation Metrics

We carry out our experiments on the query-focused and update summarization tasks from TAC 2009 with 44 inputs as our test set, and from TAC 2008 with 48 inputs as the development set. These datasets consist of two sets of 10 news documents for each input: (i) set *A* for initial summaries; (ii) set *B* for update summaries. Both *A* and *B* are on the same general topic but *B* contains documents published later than those in *A*. The update summary of set *B* is created assuming that the user is aware of what exists in set *A*. There are also four human-crafted model summaries for each input in each document set. A maximum of 100 words summary that addresses the information required by the given query statement (consisting of a title and narrative) has been produced by each of the 53 and 58 automatic summarizers participated in TAC 2009 and 2008, respectively. An example query statement is shown here:

Title: Barack Obama
Narrative: Track the increase in Barack Obama's popularity, visibility, support, and activities.

Content and linguistic quality are two conventional factors in evaluation of summary quality. Herein, we focus on the problem of automatic evaluation of content quality. Hence, we assess the performance of our metrics in replicating manual correlations of pyramid and responsiveness. It is noteworthy that responsiveness incorporates at least some aspects of linguistic quality.

Pyramid: This evaluation method (Passonneau et al., 2005) is a content assessment measure which compares content units in a system summary to weighted content units in a set of model summaries. It uses multiple human models from which annotators identify semantically defined Summary Content Units (SCU). Each SCU is assigned a weight equal to the number of human model summaries that express that SCU. An ideal maximally informative summary would express a subset of the most highly weighted SCUs, with multiple maximally informative summaries being possible. The pyramid score for a system summary is equal to the ratio between the sum of weights of SCUs expressed in a summary (again identified manually) and the sum of weights of an ideal summary with the same number of SCUs.

Four human summaries provided by NIST for each input and task were used for the pyramid evaluation at TAC.

Responsiveness: This is a measure of overall quality combining both content and linguistic quality. Summaries must present useful content in a structured fashion in order to better satisfy the user's need. Assessors directly assigned scores on a scale of 1 (poor) to 5 (very good) to each summary. These assessments are done without reference to any model summaries.

Linguistic Quality: This measure ranks summaries in a 5-point scale indicating how well a summary satisfied the factors of linguistic quality (i.e., grammaticality, non-redundancy, referential clarity, focus, structure and coherence). In our work, we do not evaluate linguistic quality.

4 Features for Summary Evaluation

We propose five classes of features to assess the quality of summary content in the absence of model summaries: (i) *Distributional Semantic Similarity*; (ii) *Topical Relevance*; (iii) *Query Relevance*; (iv) *Coherence*; and (v) *Novelty*. Before computing the features, all words in input documents, summaries, and queries are converted to lower case and stop-word filtered. We experiment with two variants of word embeddings as the basic building block to design our features:

Corpus-based Word Embeddings: We utilize the 300-dimensional embeddings for 3M words and phrases trained on Google News[2], a corpus of $\sim 10^{11}$ tokens, using word2vec CBOW (Mikolov et al., 2013). Word2vec learns a vector representation for each word using a neural network language model. It also allows to learn complex semantic relationships using simple vectorial operators, such as $vec(king) - vec(man) + vec(woman) \approx vec(queen)$. Stemming is not performed to make the word embeddings discover the linguistic regularities of words with the same root.

Lexical Resource-based Word Embeddings: We use WordNet (Fellbaum, 1998) to measure the lexico-semantic similarity between the input and its summary. Since the constraints of WordNet lexical resource can be formalized as constraints on embeddings, we can use embeddings of non-word data types (i.e., senses). Specifically, we compute the embedding of a word by averaging the embeddings of its senses in WordNet. For example, the vector of the word *suit* is modeled as the average of a vector representing *lawsuit* and a vector representing *business suit*.

We obtain the sense embeddings using the pre-trained model by Rothe and Schütze (2015), that lives in the same vector space as the pre-trained word2vec by Mikolov et al. (2013). Their model is an autoencoder neural-network that takes word embeddings and learns sense embeddings based on the following intuitions: (i) a word's embedding is the sum of the embeddings of its senses; and (ii) the senses related by WordNet relations (e.g., hypernymy, antonymy, similarity) have similar embeddings. Considering WordNet relations also helps to compute embeddings for senses in WordNet which are not in the word2vec vocabulary.

We further assume that the probability of a word sense is in proportion to its frequency in WordNet. Hence, the probability that a sense $\mathcal{S}_{ij}$ is the meaning of the word w_i, is the ratio of the frequency of that sense $freq(\mathcal{S}_{ij})$ to the total frequency of the word. If the frequency of a word sense is 0 in WordNet, we set it to 1. Finally, the embedding of word w_i is computed[3] as a weighted average of its senses $\mathcal{S}_{ij}, 1 \leq j \leq n$, where the weights represent the probability of senses:

$$\vec{w_i} = \frac{\sum_{\mathcal{S}_{ij} \in \mathcal{S}yn(w_i)} freq(\mathcal{S}_{ij}) \times \vec{\mathcal{S}_{ij}}}{n \sum_{\mathcal{S}_{ij} \in \mathcal{S}yn(w_i)} freq(\mathcal{S}_{ij})} \quad (1)$$

4.1 Distributional Semantic Similarity

A good summary must satisfy both *coverage* and *diversity* properties. For clarity, summary sentences should cover a sufficient non-redundant amount of information from the original input text. Diversity

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

[3]Words were stemmed before inferring their embeddings.

property is also fundamental especially for multi-document summarization. Moreover, one would expect good summaries to be characterized by low distance between probability distributions of words in the input and summary, and by high similarity with the input. Hence, we design this feature based on the geometric meaning of the centroid vector of a document using the compositional properties of the word embeddings (Mikolov et al., 2013). The main idea is to give a distributed representation of words/senses in the input and its summary, and compare their centroid vectors to realize how much the summary content works as a pseudo-input and condenses the meaningful information of the input.

The centroid embedding $\vec{T}$ of a text $T = \{t_1, t_2, ..., t_n\}$, is the sum of the embeddings of tokens of T divided by the number of tokens n. Based on the problem, we can also assign a weight $\mathcal{W}$ to each token in T (Figure 1). Accordingly, the centroid embedding for each summary sentence $\vec{s_j}$ is computed by averaging the embeddings of all words comprising the sentence (Radev et al., 2004). Similarly, we construct a centroid vector for each document, $\vec{d_i}$, in the input document set. To better assess the *informativeness* of the summary content, we assign higher weights to specialized words in a document by considering the Inverse Document Frequency (IDF) scores of words:

$$\vec{d_i} = \frac{\sum_{w_j \in d_i} \vec{w_j} \times TF(w_j, d_i) \times IDF(w_j)}{n \sum_{w_j \in d_i} TF(w_j, d_i) \times IDF(w_j)} \tag{2}$$

where n is the number of words in document d_i, and $\vec{w_j}$ is the embedding of word w_j. $TF(w_j, d_i)$ stands for the term frequency of w_j in d_i. The IDF scores are computed on the whole document set.

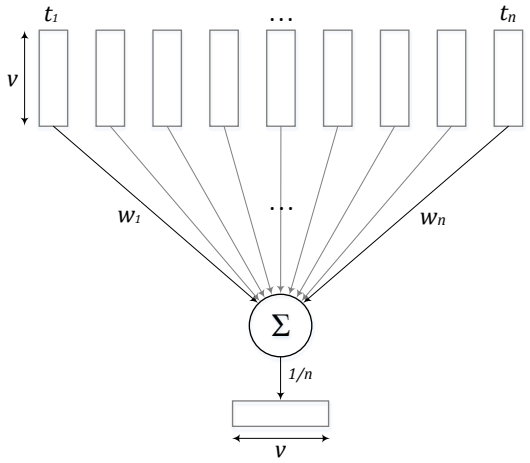

Figure 1: The weighted centroid embedding of text $T = \{t_1, t_2, ..., t_n\}$

Finally, we compare summary sentences and the input documents using the Word Mover's Distance (WMD) algorithm (Kusner et al., 2015). WMD measures the total distance the centroid embeddings of summary sentences and the input documents have to travel to become identical. Accordingly, we measure the dissimilarity degree between two sets of embedding vectors, $D = \{\vec{d_1}, ..., \vec{d_n}\}$ and $S = \{\vec{s}_1, ..., \vec{s}_m\}$, by calculating the minimum amount of summing up individual distances (travel costs) that centroid embeddings of the documents in D need to travel to reach the embeddings of sentences in S:

$$WMD(D, S) = \min_{F \geq 0} \sum_{\vec{d_i} \in D} \sum_{\vec{s_j} \in S} F_{\vec{d_i}\vec{s_j}} \times dist(\vec{d_i}, \vec{s_j}) \tag{3}$$

subject to,

$$\sum_{\vec{d_i} \in D} F_{\vec{d_i}\vec{s_j}} = \frac{1}{|S|}, \forall \vec{s_j} \in S, \sum_{\vec{s_j} \in S} F_{\vec{d_i}\vec{s_j}} = \frac{1}{|D|}, \forall \vec{d_i} \in D$$

where $F \in \mathbb{R}^{V \times V}$ with V as the vocabulary size, is a flow matrix which indicates how much probability mass should flow (or travel) from document centroid embedding $\vec{d_i}$ in set D to sentence embedding $\vec{s_j}$

in set S, and vice versa. $dist(\vec{d_i}, \vec{s_j})$ denotes the individual distance (or travel cost) between $\vec{d_i}$ and $\vec{s_j}$: $dist(\vec{d_i}, \vec{s_j}) = \|\vec{d_i} - \vec{s_j}\|_2$.

4.2 Topical Relevance

Topic features serve as a basis for evaluating topical relevance of a summary to the input documents. Herein, we aim to find the distribution of the most probable topics embodied in the input document set, and their relevance to the summary sentences. To this end, we use Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003; Arora and Ravindran, 2008) to determine the topics that characterize every document set. LDA is a generative model for documents to determine topic compositions of words and document mixtures of topics (represented by a probability distribution over topics), by assigning words to topics within documents. Hence, in the context of text modeling, the topic distribution provides an underlying semantic representation of the documents and can be useful in evaluating the summaries. Using weighted topic compositions, we measure the similarity of summary sentences with the most important topics identified in the document set.

We use Gibbs sampling (Griffiths, 2002) for inference in the topic model with concentration parameters $\alpha = 0.1$ and $\beta = 0.01$. We also set the number of topics $K = 10$ for each document set. Formally, each topic is defined as $\mathcal{T}_i = \{p_1, p_2, ..., p_n\}$, where p_j is the probability distribution of word w_j. We consider top $m = 30$ words and their probabilities to build a centroid as the representative of each topic. The embedding vector for word w_j is then multiplied with its normalized probability $\mathcal{P}_j$, and the weighted vectors are averaged to build a topic centroid representation:

$$\vec{\mathcal{T}_i} = \frac{1}{m} \sum_{j=1}^{m} \mathcal{P}_j \vec{w_j}, \quad where\ \mathcal{P}_j = \frac{p_j}{\sum_{i=1}^{m} p_i} \tag{4}$$

Finally, we use WMD to measure the dissimilarity degree between the centroid embeddings of summary sentences and those of the topics for evaluating topical relevance of the summary content.

4.3 Query Relevance

To measure the relevance degree of the summary content to the given query, we calculate the query embedding vector $\vec{Q}$ by averaging the embeddings of all words in the query narrative. Similarly, the centroid embedding vector for each summary $\vec{S}$ is also constructed. We further measure the cosine similarity between these vectors to formulate query relevance:

$$sim(\vec{S}, \vec{Q}) = \frac{\vec{S} \cdot \vec{Q}}{\| \vec{S} \| \| \vec{Q} \|} \tag{5}$$

4.4 Coherence

Coherence measures the degree to which a sequence of summary sentences represents a logical flow of thought. We compute the similarity between embeddings of adjacent summary sentences using cosine similarity. It results in $n-1$ comparisons for a summary of n sentences. While similarity between sentences is beneficial for coherence, very high similarity reflects redundancy in the summary. Given that, we combine the *mean* and *standard deviation* of the cosine similarity scores by training a simple linear regression model on our development set. In this way, we measure the trade-off between continuity and redundancy as the coherence feature.

4.5 Novelty

We would like our evaluation model to move beyond assessing initial summaries by giving a simple feature of Novelty to better evaluate update summaries. This feature rewards the update summary consisting novel words that do not exist in initial document set D_A, but are semantically related to update document set D_B. The relevancy of these words in update summary S_j, to the documents in set *B*, is measured using the cosine similarity between the embeddings of novel words and the centroid embedding of the

whole document set *B*. We use the bag-of-words representation of the summary and the document sets while defining novel words. We finally measure the degree of novelty ($\mathcal{N}$) as:

$$\mathcal{N}(S_j) = \frac{1}{|S_j|} \sum_{w_i \in S_j | w_i \notin D_A} sim(\vec{w_i}, \vec{D_B}) \tag{6}$$

where $|S_j|$ is the total number of unique words in the update summary S_j. For S_j without any novel words, $\mathcal{N}(S_j) = 0$.

5 Feature Combination with SVR

We combine all the above features using a Support Vector Regression (SVR) model to predict the summary quality. We first transform the proposed features into a standard vector notation. Each summary S_i is represented by a feature vector $X = \{x_1, x_2, ..., x_n\}$ where n is the number of features. SVR model aims to learn a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, which will be used to predict the content evaluation score for each summary $y \in \mathbb{R}$ given a feature vector $X \in \mathbb{R}^n$. In particular, given l training instances $(X_1, y_1), ..., (X_l, y_l)$, the SVR model is learnt by solving the following optimization problem (Vapnik, 1999); W is a vector of feature weights; ϕ is a function that maps feature vectors to a new vector space of higher dimensionality to allow non-linear functions to be learnt in the original space; $C > 0$ and $\epsilon > 0$ are given.

$$\min_{W,b,\xi,\xi^*} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{l} \xi_i + C\sum_{i=1}^{l} \xi^* \tag{7}$$

subject to (for $i = 1, ..., l$):

$$\begin{aligned} W^T \cdot \phi(X_i) + w_0 - y_i &\leq \epsilon + \xi_i \\ y_i - W^T \cdot \phi(X_i) - w_0 &\leq \epsilon + \xi^* \\ \xi_i &\geq 0 \\ \xi^* &\geq 0 \end{aligned}$$

The goal is to learn a linear (in the new space) function, whose prediction (value) $W^T \cdot \phi(X_i) + w_0$ for each training instance X_i will not to be farther than ϵ from the target (correct) value y_i. Since this is not always feasible, two slack variables ξ_i and ξ_i^* are used to measure the prediction's error above or below the target y_i. The objective (7) jointly minimizes the total prediction error and $\|W\|$, to avoid overfitting. The utilized SVR is implemented in Scikit-learn (Pedregosa et al., 2011). We use the default parameter settings, (*kernel='rbf', degree=3, gamma='auto', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=-1*) without further optimization.

6 Experiments and Results

Reporting correlations with manual evaluation metrics is the norm for validating automatic metrics. We use Spearman correlation metric to study the predictive power of our automatic features in replicating manual correlations of pyramid and responsiveness. Hence, we compare the rankings of systems against the human scores assigned to systems. The correlations[4] of our metrics are reported at two levels of granularity:

System Level (MACRO): The average score for a system is computed over the entire set of test inputs using both manual and automatic evaluations. The correlations between ranks assigned to systems by these average scores are indicative of the strength of our features to predict overall system rankings on the test set.

Analyzing the macro level results on TAC 2008 (Table 1), we find that the variants of distributional similarity and the topical relevance features produce system rankings very similar to those produced by

[4]Significance values for the correlations are produced using the AS 89 algorithm (Best and Roberts, 1975).

	QUERY - MACRO		QUERY - MICRO		UPDATE - MACRO		UPDATE - MICRO	
Features	**Pyr.**	**Resp.**	**Pyr.**	**Resp.**	**Pyr.**	**Resp.**	**Pyr.**	**Resp.**
Corpus-based Dist. Similarity	-0.887	-0.748	75.0	72.9	-0.833	-0.761	77.1	70.8
LexRes-based Dist. Similarity	-0.871	-0.723	72.9	68.8	-0.828	-0.755	77.1	68.8
Corpus-based Topical Relevance	-0.803	-0.696	72.9	70.8	-0.777	-0.720	75.0	72.9
LexRes-based Topical Relevance	-0.799	-0.705	70.8	70.8	-0.759	-0.735	72.9	70.8
LexRes-based Query Relevance	0.624	0.590	58.3	58.3	0.599	0.576	62.5	56.3
Corpus-based Query Relevance	0.615	0.547	56.3	52.1	0.613	0.576	60.4	56.3
LexRes-based Novelty	-	-	-	-	0.537	0.502	54.2	50.0
Corpus-based Novelty	-	-	-	-	0.530	0.500	58.3	45.8
Corpus-based Coherence	0.361	0.375	37.5	39.6	0.352	0.358	41.7	35.4
LexRes-based Coherence	0.353	0.362	35.4	37.5	0.349	0.358	37.5	37.5
Support Vector Regression	0.895	0.786	79.2	75.0	0.872	0.808	87.5	77.1
SIMetrix JS divergence	-0.880	-0.736	72.9	72.9	-0.827	-0.764	85.4	75.0
SIMetrix regression	0.867	0.705	77.1	66.7	0.789	0.605	81.3	58.3
ROUGE-1 recall (4 models)	0.859	0.806	97.9	95.8	0.912	0.865	97.9	95.8
ROUGE-2 recall (4 models)	0.905	0.873	100	91.7	0.941	0.884	100	91.7

Table 1: Input-summary evaluation on the query focused and update summarization tasks from TAC 2008 data: MACRO level Spearman correlations, all results are significant ($p < 0.05$); MICRO level percentage of inputs with significant correlations ($p < 0.05$).

human. Other features, on the other hand, are less predictive of content quality. Distributional similarities also outperform SIMetrix, which proves the importance of semantic representation of the input and summary for comparison purposes in summary content evaluation. Overall, our feature regression obtains the best correlations with both types of manual scores, and even outperforms ROUGE-1 regarding pyramid for query-focused task. The usefulness of novelty feature is also reflected in high SVR correlation results for the update summarization task.

Overall ROUGE correlation is evidence that the model summaries provide information that is unlikely to ever be approximated by exploring the input alone. However, our features can provide reliable estimates of system quality when averaged over a set of test inputs. We also observe that corpus-based models mostly outperform their corresponding lexical resource-based models. A possible reason is the higher coverage of Google News word2vec model comparing to the WordNet-based sense embedding model. For example, some words like proper nouns (e.g., *'Barak Obama'*) are not covered in WordNet. However, replacing a word's embedding by the sum of the embeddings of its senses could generally improve the quality of embeddings (Rothe and Schütze, 2015). That is why our SVR performs well by leveraging WordNet senses for more precise word embeddings, and involving Google News to complement the WordNet coverage.

Input Level (MICRO): For each individual input, we compare the rankings for the system summaries using manual and automatic evaluations. Micro-level analysis highlights the ability of an evaluation metric to assess the quality of system summaries produced for a specific input. This task is bound to be harder than system level predictions. For clarity, even with wrong prediction of rankings on a few inputs, the average scores (macro-level) for a system might not be affected.

To be in line with SIMetrix, we report the percentage of inputs for which significant correlations were obtained (Table 1). We observe that feature combination with SVR gives the best results overall, similar to our findings for the macro level. The implication is that no single feature can reliably predict good content for a particular input. Moreover, our feature regression outperforms SIMetrix. This is because our approach depends not merely on the distribution of terms in the input, and therefore provides better representation for a set of documents each describing different opinion on a given issue. For example, our topical relevance feature gives a representative vector for every important aspect of the document set. However, superiority of ROUGE performance to the rest of measures shows that model summaries generated for specific input would still give better indication of important information in the input.

6.1 Error Analysis:

In this study, we aim to assess the reliability of our metric for evaluation in the absence of human model summaries, where ROUGE cannot be used. It is noteworthy that we do not intend to directly compare the performance of ROUGE with our metric. Thereupon, we provide an error analysis to understand if our SVR and ROUGE are making errors in ordering the same systems or whether their errors are different. Since at the macro level, the correlations between our regression and pyramid scores is close to those of ROUGE-2, we further analyze their errors. We considered pairs of systems and identified the better system in each pair according to the pyramid scores. Afterwards, we recorded how often ROUGE-2 and the SVR provided the correct judgment for the pairs as indicated by the pyramid evaluation. Table 2 provides the results for all 1,653 pairs of systems at the macro level.

	SVR correct	**SVR incorrect**
ROUGE-2 correct	1,355(82.0%)	97(5.9%)
ROUGE-2 incorrect	100(6.0%)	101(6.1%)

Table 2: Error analysis: Overlap between ROUGE-2 and SVR predictions for the best system in a pair (TAC 2008, 1,653 pairs). The gold-standard judgment for a better system is computed using pyramid.

A large majority (82%) of the same pairs are correctly predicted by both ROUGE and the SVR. Another 6% of the pairs are such that both metrics do not provide the correct judgment. Therefore, ROUGE and our SVR appear to agree on a large majority of the system pairs. There is a small percentage (12%) that is correctly predicted by only one of the metrics.

6.2 Evaluation on the Test Set:

Our SVR was trained on the TAC 2008 data with pyramid scores as the target. Herein, we evaluate this metric using the TAC 2009 data (Table 3). We report the correlations obtained by ROUGE-SU4 as the official baseline measure at TAC 2009 for comparison of automatic evaluation metrics. The results indicate that the correlations are lower than on our development set. This might be caused by the different characteristics of inputs in two year's data (Louis and Nenkova, 2013). However, the SVR is consistently predictive across two years, and outperforms SIMetrix.

	QUERY - MACRO		**QUERY - MICRO**		**UPDATE - MACRO**		**UPDATE - MICRO**	
Metric	**Pyr.**	**Resp.**	**Pyr.**	**Resp.**	**Pyr.**	**Resp.**	**Pyr.**	**Resp.**
Support Vector Regression	0.80	0.75	87.5	77.1	0.77	0.65	79.2	75.0
SIMetrix JS divergence	-0.74	-0.71	84.1	75.0	-0.72	-0.61	77.3	72.7
SIMetrix Regression	0.77	0.67	81.8	65.9	0.71	0.54	75.0	52.3
ROUGE-SU4 (4 models)	0.92	0.79	95.5	81.8	0.85	0.69	100	86.4

Table 3: Input-summary evaluation on the query focused and update summarization tasks from TAC 2009 data: MACRO level Spearman correlations, all results are significant ($p < 0.05$); MICRO level percentage of inputs with significant correlations ($p < 0.05$).

Overall results also show that correlations with pyramid scores are higher than those with responsiveness. The reason is that our features mainly evaluate summary content. Responsiveness judgments, on the other hand, are based on both content and linguistic quality. Nevertheless, our SVR performs better than SIMetrix in replicating responsiveness scores. This might be advantaged by considering coherence as a linguistic quality feature. Hence, a natural extension of our work would be considering more linguistic quality features along with content evaluations.

7 Conclusion and Future Work

We have presented an effective model-free summary content evaluation approach that exploits the compositional properties of word and sense embeddings to develop a variety of features for input-summary comparisons. The results show that the strength of different features varies considerably, and their combination provides reliable estimates of summary content quality when model summaries are not available. This lends further support to our proposal to use semantic representation of the input and summary contents for the model-free summary content evaluation.

Our ongoing work includes considering distributional and relational semantics together (Fried and Duh, 2014; Verga and McCallum, 2016; Rossiello, 2016) for different sentence representations, and using more complex neural language models (Le and Mikolov, 2014; Zhang and LeCun, 2015; Jozefowicz et al., 2016) for the comparison.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. The first author was supported by the "Australian Government Research Training Program Scholarship".

References

Rachit Arora and Balaraman Ravindran. 2008. Latent Dirichlet Allocation based multi-document summarization. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, pages 91–97. ACM.

DJ Best and DE Roberts. 1975. Algorithm AS 89: the upper tail probabilities of Spearman's rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(3):377–379.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Luis Adrián Cabrera-Diego and Juan-Manuel Torres-Moreno. 2017. Summtriver: A new trivergent model to evaluate summaries automatically without human references. *Data & Knowledge Engineering*.

Robert L Donaway, Kevin W Drummey, and Laura A Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the 2000 North American Chapter of the Association for Computational Linguistics - Applied Neural Language Processing Conference: Workshop on Automatic Summarization*, pages 69–78. Association for Computational Linguistics.

Samira Ellouze, Maher Jaoua, and Lamia Hadrich Belguith. 2013. An evaluation summary method based on a combination of content and linguistic metrics. In *Recent Advances in Natural Language Processing*, pages 245–251. Citeseer.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Daniel Fried and Kevin Duh. 2014. Incorporating both distributional and relational semantics in word representations. *arXiv preprint arXiv:1412.4369*.

Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. *Proceedings of the International Conference on Computational Linguistics*, pages 911–926.

Tom Griffiths. 2002. Gibbs sampling in the generative model of Latent Dirichlet Allocation.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches out: Proceedings of the Association for Computational Linguistics Workshop*, volume 8.

Annie Louis and Ani Nenkova. 2009a. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 306–314. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2009b. Predicting summary quality using limited human input. In *Text Analysis Conference*.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. *arXiv preprint arXiv:1508.06034*.

Rebecca J Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the pyramid method in DUC 2005. In *Proceedings of the Document Understanding Conference*.

Rebecca J Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Short Papers*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Dragomir R Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 375–382. Association for Computational Linguistics.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Gaetano Rossiello. 2016. Neural abstractive text summarization. In *Proceedings of the Doctoral Consortium of AI*IA 2016 co-located with the 15th International Conference of the Italian Association for Artificial Intelligence*, pages 70–75.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *arXiv preprint arXiv:1507.01127*.

Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, and Eric SanJuan. 2010. Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1059–1067. Association for Computational Linguistics.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2017. A semantically motivated approach to compute ROUGE scores. *arXiv preprint arXiv:1710.07441*.

Abhishek Singh and Wei Jin. 2016. Ranking summaries for informativeness and coherence without reference summaries. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*, pages 104–109.

Vladimir Naumovich Vapnik. 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999.

Patrick Verga and Andrew McCallum. 2016. Row-less universal schema. *arXiv preprint arXiv:1604.06361*.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.