# Assessing the Quality of Human-Generated Summaries with Weakly Supervised Learning

**Joakim Olsen** and **Arild Brandrud Næss**
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
joakiol@stud.ntnu.no, arild.naess@ntnu.no

**Pierre Lison**
Norwegian Computing Center
Oslo, Norway
plison@nr.no

## Abstract

This paper explores how to automatically measure the quality of human-generated summaries, based on a Norwegian corpus of real estate condition reports and their corresponding summaries. The proposed approach proceeds in two steps. First, the real estate reports and their associated summaries are automatically labelled using a set of heuristic rules gathered from human experts and aggregated using weak supervision. The aggregated labels are then employed to learn a neural model that takes a document and its summary as inputs and outputs a score reflecting the predicted quality of the summary. The neural model maps the document and its summary to a shared "summary content space" and computes the cosine similarity between the two document embeddings to predict the final summary quality score. The best performance is achieved by a CNN-based model with an accuracy (measured against the aggregated labels obtained via weak supervision) of $89.5\%$, compared to $72.6\%$ for the best unsupervised model. Manual inspection of examples indicate that the weak supervision labels do capture important indicators of summary quality, but the correlation of those labels with human judgements remains to be validated. Our models of summary quality predict that approximately $30\%$ of the real estate reports in the corpus have a summary of poor quality.

## 1 Introduction

Many types of reports incorporate human-generated summaries that seek to highlight the most important pieces of information described in the full document. This is notably the case for *real estate condition reports*, which are long, technical reports presenting the current condition (as it is known to the seller) of a property for sale, including the general state of each room, known damages and defects, and key technical aspects such as the heating, plumbing, electricity and roof. Despite the rich amount of information contained in these real estate reports, several surveys have shown that many buyers of real estate do not read the full documents but rather concentrate on the summaries (Sandberg, 2017). However, professionals regard the quality of these summaries as varying greatly, from good to very poor. Actors in the real estate market have suggested that this information deficit may play an important role in the reported $10\%$ of Norwegian real estate transactions ending in conflict (Huseiernes Landsforbund, 2017).

In this work we explore ways of automatically measuring the quality of such summaries, using a corpus of $96\,534$ real estate condition reports and their corresponding summaries. Although there exists a substantial body of work on summary evaluation (Lloret et al., 2018), previous work has largely focused on automatically generated summaries, often by comparing those generated summaries to reference summaries written by humans. The automated evaluation of human-generated summaries, however, has received little attention so far.

This paper presents an approach to automatically evaluate the quality of human-generated summaries when no manually labelled data is available. Instead, we rely on a set of heuristic rules provided by domain experts to automatically annotate a dataset of summaries (each coupled to their full-length document) with quality indicators. Those annotations are subsequently aggregated into a single, unified annotation layer using weak supervision (Ratner et al., 2017, 2019), based on a generative model

that takes into account the varying coverage and accuracy of the heuristic rules.

Although one could in theory directly use the labels obtained through weak supervision as quality indicators for the summaries, such an approach has a number of limitations. Most importantly, heuristic rules are only triggered under certain conditions, and may therefore "abstain" from providing a quality score on some summaries. For instance, we may have a rule stating that, if the full report describes a major defect or damage in the bathroom, then a summary that fails to mention this defect should be labelled as being of poor quality. This rule will only label summaries that meet this specific condition, and abstain from generating a prediction in all other cases. Some heuristic rules may also depend on the availability of external data sources that are not available at prediction time. For instance, one can exploit the fact that an insurance claim has been raised on the real estate as an indicator that the summary may have omitted to mention some important defects or damages. Needless to say, this heuristic can only be applied on historical data, and not on new summaries.

To address those shortcomings, we use the aggregated labels obtained via weak supervision as a stepping stone to train a neural model whose task is to assess the quality of a summary in respect to its full-length document. The neural model embeds both the document and its summary into a dedicated semantic space (referred to as the *summary content space*) and computes the final quality score using cosine similarity. As real estate condition reports are often long documents (10 pages or more), we conduct experiments with models based not only on embeddings of entire documents, but also on embeddings of sections, sentences and words.

The paper makes three contributions:

1. A framework to automatically (a) associate summaries with quality indicators based on expert-written rules, and (b) aggregate those indicators using weak supervision.

2. A neural model that predicts the summary quality by embedding both the document and its corresponding summary into a common summary content space, and then computing the similarity between the two vectors. The neural model is trained using the weakly supervised labels as described above.

3. An evaluation of this approach on a large corpus of Norwegian real estate condition reports and their associated summaries.

As detailed in Section 4, this weak supervision approach is able to outperform unsupervised methods based on Latent Semantic Analysis (Deerwester et al., 1990) or Doc2Vec embeddings (Le and Mikolov, 2014) – by a large margin. Although the approach is evaluated on a specific corpus of real estate reports, the proposed methodology can be applied to any type of summaries, provided human experts are able to specify heuristics to assess the summary quality in the target domain.

## 2 Related Work

### 2.1 Summary evaluation

Summary evaluation has so far been mostly studied in relation to the task of automatic text summarization, i.e., the automated generation of summaries conditioned on the full document (Rush et al., 2015; Cheng and Lapata, 2016; Gambhir and Gupta, 2017; Cao et al., 2018; Fernandes et al., 2019). However, few papers have investigated how to evaluate the quality of human-generated summaries such as the short summaries associated with real estate condition reports.

Lloret et al. (2018) provide an overview of evaluation metrics for text summarization, focusing on three quality criteria: *readability*, *non-redundancy* and *content coverage*. Although readability and non-redundancy are important criteria to evaluate automatic text summarization systems, they are less relevant for assessing human-generated summaries written by professionals. The criteria of content coverage is, however, relevant in both contexts, and will be the main focus of this paper.

Metrics for summary evaluation can be divided in three overarching groups (Cabrera-Diego and Torres-Moreno, 2018; Ermakova et al., 2019):

1. Manual evaluation based on human judgments, where participants fill questionnaires to rate the summary quality according to a number of criteria (Nenkova and Passonneau, 2004; Saggion et al., 2010).

2. Automatic evaluation from overlap-measures with reference summaries written by human experts (Lin, 2004; Conroy and Dang, 2008; Giannakopoulos, 2013; Zhang et al., 2020). One popular metric based on this idea is ROUGE (Lin, 2004), which is computed from

the proportion of $n$-grams that are observed in both the generated output and the reference summaries.

3. Automatic evaluation without reference summaries, typically using measures of divergence between the generated summary and the source document (Torres-Moreno et al., 2010; Louis and Nenkova, 2013; Cabrera-Diego and Torres-Moreno, 2018).

The evaluation method proposed in this paper fits into the last category, as we do not require the availability of reference summaries. However, contrary to divergence-based metrics, the summary quality is estimated here on the basis of heuristic rules provided by human experts.

### 2.2 Document similarity

The proposed approach is also related to models of semantic similarity, as the purpose of our summary evaluation is to assess the extent to which the criteria of content coverage is satisfied.

There is a vast body of existing work on how to measure the semantic similarity between documents. This topic is also the focus of various benchmarks, such as the Microsoft Research Paraphrase (MSRP) corpus (Dolan et al., 2004) and the Semantic Textual Similarity (STS) benchmark (Cer et al., 2017), both expressed as pairs of short documents. The ACL Anthology Network (Radev et al., 2009) is also used for measuring semantic similarity between articles in Liu et al. (2017). Gong et al. (2019) investigates how to measure similarity between documents of varying sizes.

Document similarity can be computed from topic models based on, e.g., Latent Dirichlet Allocation (Blei et al., 2003; Rus et al., 2013; Liu et al., 2017), or through document embeddings (Le and Mikolov, 2014; Lau and Baldwin, 2016; Liu et al., 2017; Cer et al., 2017; Gong et al., 2019; Vrbanec and Meštrović, 2020). Contextual word representations such as BERT, XLNet or GPT-3 (Devlin et al., 2018; Yang et al., 2019; Brown et al., 2020), can also be used to derive document embeddings and have been shown to improve performance on document similarity benchmarks (Reimers and Gurevych, 2019; Li et al., 2020), notably on the MSRP corpus and the STS benchmark.

Of particular relevance to this paper is the text matching approach of Zhong et al. (2020) in which the source document and potential summaries are matched in a semantic space. Their approach is, however, optimised for the problem of extracting summaries, while our focus is on evaluating existing, human-generated summaries, using expert-written rules as quality indicators.

### 2.3 Weak supervision

The key idea behind weak supervision is to label data points using a combination of weak (noisy) supervision signals instead of relying on a single gold standard. Those supervision signals are typically expressed as *labeling functions*, which may take the form of heuristic rules, lookups in external knowledge bases, machine learning models, or even annotations from crowd-workers. The result of those labeling functions are then aggregated using a generative model that estimates the accuracy (and possible correlations) of each function. Once aggregated, the (probabilistic) labels can be employed to train any type of machine learning model using supervised learning. One key benefit of weak supervision frameworks lies in their ability to inject *expert knowledge* to learn data-driven models in situations when data is scarce or non-existent (Hu et al., 2016; Wang and Poon, 2018).

Weak supervision makes it possible to leverage external knowledge sources to automatically label data points instead of relying exclusively on hand-annotated data. An early application of this idea is distant supervision (Mintz et al., 2009; Ritter et al., 2013), where knowledge bases are used to automatically label documents with specific categories. One popular approach for weak supervision is the Snorkel framework, which was first introduced by Ratner et al. (2016), and later expanded by Ratner et al. (2017) and Ratner et al. (2019).

Weak supervision frameworks have been applied to a number of NLP tasks, from named entity recognition to relation extraction and dialogue state tracking (Bach et al., 2019; Bringer et al., 2019; Hancock et al., 2019; Lison et al., 2020; Safranchik et al., 2020). There is, however, little work with weak supervision related to document similarity or summary quality evaluation.

## 3 Approach

The approach adopted in this paper is divided in two steps. We first define and apply a set of *labeling functions* to the dataset, allowing us to derive binary (good/bad) quality indicators on the summaries in relation to their full-length reports. Those

quality indicators are then aggregated into a single, probabilistic measure of summary quality using weak supervision. The dataset and labeling functions are described in Sections 3.1 and 3.2.

Then, using those aggregated labels as targets, we learn a neural model that maps the reports and summaries to a common *summary content space*. The resulting embeddings should reflect only key semantic information that is relevant for measuring summary quality, so that it can be measured by the cosine similarity in this space. The neural architecture and associated document embedding methods are defined in Sections 3.3 and 3.4.

Assessing the summary quality using a neural model instead of relying directly on the quality indicators derived from the labeling functions has two major advantages. First, the neural model can generalise to all possible report/summary pairs, while aggregated labels may be absent for some summaries, as the rules are only triggered when specific conditions are met. Second, some labeling functions depend on external resources that may be unavailable at prediction time. For instance, one labeling function relies on whether the buyer has filed an insurance claim, which is a piece of information that is only available for historical data, and requires us to "peek into the future".

### 3.1 Dataset

The corpus contains 96 534 real estate condition reports, each containing the following parts:

i) Textual descriptions of various parts of the real estate (e.g., rooms) along with a textual assessment of their physical condition.

ii) Condition degrees ("tilstandsgrad" or TG) for parts of the real estate, in the range 0–3, where 0 indicates perfect condition (for new buildings) and 3 a seriously deteriorated condition, due to a major damage or defect.

iii) Metadata for the real estate and the condition report – e.g., size, building year, the author of the report, date of assessment, etc.

iv) The summary.

We consider (i) as constituting the full-length report, denoted $\mathbf{r}$, while the summary text (iv) will be denoted $\mathbf{s}$. The metadata (ii)–(iii) is used only by the weak supervision model. The average report length is 1287 words (standard deviation: $\pm$627 words), while the average summary length is 183 words (standard deviation: $\pm$138 words).

### 3.2 Labeling Functions

A collection of 22 labeling functions was specified in cooperation with domain experts. Each function has two possible output values, depending on whether it implies a bad summary, denoted by ($-$) or a good summary, denoted by ($+$). If the rule condition is not met, the rule abstains from suggesting an output (Ratner et al., 2017). The full list of labeling functions is the following:

1. Summary shorter than 50 words. ($-$)
2. Summary longer than 400 words. ($-$)
3. TG3 for the bathroom, but no mention of the bathroom in summary. ($-$)
4. TG3 for the kitchen, but no mention of the kitchen in summary. ($-$)
5. TG3 for the roof, but no mention of the roof in summary. ($-$)
6. TG2 or TG3 for the bathroom, with mention of the bathroom in summary. ($+$)
7. TG2 or TG3 for the kitchen, with mention of the kitchen in summary. ($+$)
8. TG2 or TG3 for the roof, with mention of the roof in summary. ($+$)
9. Correction of TG in the bathroom, but no mention of the bathroom in summary. ($-$)
10. Correction of TG in the kitchen, but no mention of the kitchen in summary. ($-$)
11. Correction of TG on the roof, but no mention of the roof in summary. ($-$)
12. Summary with long words readability score (LIKS) above 55. ($-$)
13. Summary with unique words readability score (OVR) above 96. ($-$)
14. An insurance claim has been raised on the real estate after the transaction. ($-$)
15. Written by an agent with insurance claims on more than 7.5% of her reports. ($-$)
16. Written by an agent with LIKS-score higher than 55 on more than 40% of her reports. ($-$)
17. Written by an agent with OVR-score higher than 96 on more than 40% of her reports. ($-$)
18. Written by an agent with fewer than 10 reports that year. ($-$)
19. Fewer than 20% of the words in the summary are found in the report. ($-$)
20. Fewer than 3% of the words in the report are found in the summary. ($-$)
21. More than 70% of the words in the summary are also found in the report. ($+$)
22. More than 20% of the words in the report are also found in the summary. ($+$)

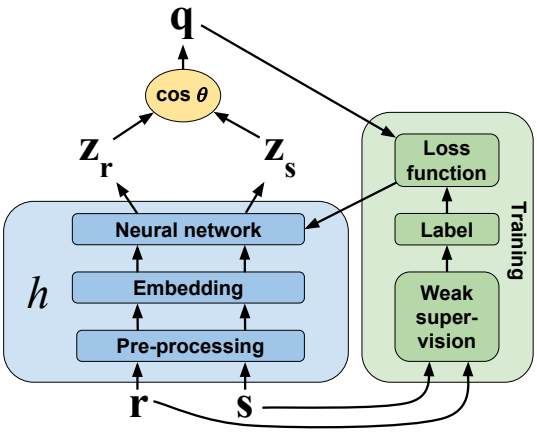


Figure 1: General model architecture $q(\mathbf{r}, \mathbf{s})$.

For a given summary, let $y$ be the unknown true label, with possible values $-1$ (bad) and 1 (good), and let $\boldsymbol{\lambda}$ be the outputs of the labeling functions. By applying these to the real estate condition reports, a generative label model $P_{\boldsymbol{\mu}}(y \mid \boldsymbol{\lambda})$ can be estimated in a fully unsupervised fashion, as described by Ratner et al. (2019). We then obtain labels $y^+ = P_{\boldsymbol{\mu}}(y = 1 \mid \boldsymbol{\lambda}) \in [0, 1]$, indicating the probability that a given summary is good.

### 3.3 Summary Quality Model

Let $\mathcal{R}$ denote the set of all possible reports and summaries, and let $\mathcal{Z}$ be the summary content space. We define the summary quality model as a function $q(\mathbf{r}, \mathbf{s})$ comparing two document embeddings:

$$q(\mathbf{r}, \mathbf{s}) = \cos\,\mathrm{sim}\big(h(\mathbf{r}), h(\mathbf{s})\big) = \cos\,\mathrm{sim}(\mathbf{z_r}, \mathbf{z_s}),$$

where $h : \mathcal{R} \to \mathcal{Z}$ is a learned mapping from texts (full reports or summaries) to vectors. The general architecture is illustrated in Figure 1.

The training objective for $h$ should be such that a good (bad) summary should yield a high (low) cosine similarity. We also want our models to return quality scores distributed over the entire cosine domain $[-1, 1]$, and we find that the standard cross-entropy loss tends to push the values towards the edges. Instead, we use a variation of the cosine embedding loss function, given by

$$l\big(q(\mathbf{r}, \mathbf{s}), y\big) = \begin{cases} \max\big(0, \tau_{\text{good}} - \cos\,\mathrm{sim}(\mathbf{z_r}, \mathbf{z_s})\big), & y = 1 \\ \max\big(0, \cos\,\mathrm{sim}(\mathbf{z_r}, \mathbf{z_s}) - \tau_{\text{bad}}\big), & y = -1, \end{cases}$$

where $\tau_{\text{good}}$ and $\tau_{\text{bad}}$ are thresholds on the quality scores of good/bad summaries. A loss of zero is obtained if good summaries have a quality score higher than $\tau_{\text{good}}$ or if bad summaries have a quality score lower than $\tau_{\text{bad}}$. The model will thereby not perform better by pushing the quality of summaries above $\tau_{\text{good}}$ or below $\tau_{\text{bad}}$, which encourages the model to return scores on a larger part of the cosine domain $[-1, 1]$. We find experimentally that $\tau_{\text{good}} = 0.2$ and $\tau_{\text{bad}} = -0.2$ result in models with an appropriate distribution of values.

The weak supervision labels $y^+$ are expected to be noisy. We follow Ratner et al. (2019) in using a noise-aware version of our loss function $l\big(q(\mathbf{r}, \mathbf{s}), y\big)$ for training, which we define by

$$\begin{aligned} l^*\big(q(\mathbf{r}, \mathbf{s}), y^+\big) &= E_{y \sim P_{\boldsymbol{\mu}}(y|\boldsymbol{\lambda})}\Big[l\big(q(\mathbf{r}, \mathbf{s}), y\big)\Big] \\ &= y^+ \cdot l\big(q(\mathbf{r}, \mathbf{s}), 1\big) + (1 - y^+) \cdot l\big(q(\mathbf{r}, \mathbf{s}), -1\big). \end{aligned} \tag{1}$$

Having defined the general model architecture and its training procedure, we now detail various solutions to express the mapping $h$.

### 3.4 Document embeddings

#### 3.4.1 LSA and Doc2vec

We start with unsupervised baseline models, and experiment with both Latent Semantic Analysis (Deerwester et al., 1990) and Doc2vec (Le and Mikolov, 2014), for their ability to easily embed arbitrarily long documents. We train LSA and Doc2vec on the training set (ignoring the quality labels, as those techniques are self-supervised). These models can be described in Figure 1 by removing the training and neural network components, and by using LSA or Doc2vec for the embeddings. We use a dimensionality of 500 for LSA and 100 for Doc2vec.

#### 3.4.2 FFN-based models

Our first supervised model for $h$ is a feed-forward network. We first embed the reports and summaries with LSA or Doc2vec (both of dimension 500) as described above, and add a feed-forward transformation of those vectors which is optimised on the basis of the embedding loss function. The network weights are shared for both the full report $\mathbf{r}$ and the summary $\mathbf{s}$. The architecture becomes as illustrated in Figure 1 by inserting LSA or Doc2vec into the embedding component and a feed-forward network into the neural-network component. We refer to the resulting models as LSA+FFN and Doc2vec+FFN.

We employ the ReLu activation function in all layers except the last, which is linear (i.e., has no activation function). By using only a single feed-forward layer, this model architecture becomes

equivalent to a linear transformation of the LSA or Doc2vec embeddings. We refer to the resulting models as LSA+LinTrans and Doc2vec+LinTrans.

The hidden layers have 1000 units, and the final layer 100 units. LSA+FFN and Doc2vec+FFN respectively use two and three hidden layers.

#### 3.4.3 LSTM-based models

The second model for the function $h$ mapping reports and summaries to the summary content space is an LSTM network. LSTMs are commonly used over word embeddings, but this approach is hard to scale due to the length of real estate condition reports. Instead, we split the reports into sections, and summaries into sentences and use LSA or Doc2vec to embed each, giving a sequence of vectors for each report and summary, and train the LSTM on these. A final, fully connected linear layer is placed on the LSTM output. In Figure 1 the pre-processing component now includes the splitting of sections/sentences, the embedding component is LSA or Doc2vec, and the neural-network component is the LSTM. We refer to the resulting models as LSA+LSTM and Doc2vec+LSTM. We use a single, unidirectional LSTM layer with a cell dimensionality of 100, along with 100 units in the final dense layer.

#### 3.4.4 Convolutional models

The final model for $h$ is a convolutional neural network with word embeddings as inputs. Those word embeddings are estimated either by Word2vec (dimension: 100) or a neural embedding layer (dimension: 500), both trained on the training set of the corpus. We use 1D convolutions with window size $\in \{2, 3, 5, 7, 10\}$ and a number of filters equivalent to the word embedding dimension. We then apply a maximum pooling to obtain a single output vector, fed to a final, fully-connected linear layer.

One benefit of convolutional neural networks is their scalability when processing long documents. The convolutional model detects local text patterns that are especially predictive for the summary quality, thereby providing a good mapping to the summary content space. In Figure 1 the pre-processing component now includes tokenisation, the embedding component is the embedding layer or Word2vec, and the neural-network is the CNN. We refer to the resulting models as EmbLayer+CNN and Word2vec+CNN.

| No. | Cov. | Overlap | Conflict | Acc. |
|---|---|---|---|---|
| 1 (−) | 10.4 % | 96.2 % | 22.1 % | 100 % |
| 2 (−) | 7.9 % | 91.1 % | 82.3 % | 10.9 % |
| 3 (−) | 5.1 % | 90.2 % | 27.5 % | 71.5 % |
| 4 (−) | 2.4 % | 95.8 % | 50.0 % | 58.5 % |
| 5 (−) | 2.6 % | 92.3 % | 30.8 % | 78.0 % |
| 6 (+) | 36.9 % | 76.4 % | 46.1 % | 74.9 % |
| 7 (+) | 11.6 % | 93.1 % | 47.4 % | 97.3 % |
| 8 (+) | 25.1 % | 83.7 % | 46.2 % | 82.0 % |
| 9 (−) | 7.6 % | 84.2 % | 22.4 % | 73.5 % |
| 10 (−) | 5.1 % | 90.2 % | 45.1 % | 60.8 % |
| 11 (−) | 8.1 % | 82.7 % | 34.6 % | 72.9 % |
| 12 (−) | 11.8 % | 92.4 % | 42.4 % | 73.4 % |
| 13 (−) | 10.7 % | 93.5 % | 26.2 % | 100 % |
| 14 (−) | 1.8 % | 83.3 % | 55.6 % | 47.9 % |
| 15 (−) | 1.6 % | 93.8 % | 43.8 % | 71.5 % |
| 16 (−) | 10.8 % | 88.9 % | 48.1 % | 57.9 % |
| 17 (−) | 10.0 % | 91.0 % | 37.0 % | 100 % |
| 18 (−) | 5.4 % | 85.2 % | 48.1 % | 58.9 % |
| 19 (−) | 3.4 % | 76.5 % | 14.7 % | 83.4 % |
| 20 (+) | 6.3 % | 85.7 % | 49.2 % | 63.3 % |
| 21 (−) | 7.1 % | 94.4 % | 11.3 % | 100 % |
| 22 (+) | 6.2 % | 91.9 % | 48.4 % | 100 % |

Table 1: Analysis of the 22 labeling functions when applied to the real estate condition report corpus.

## 4 Evaluation

### 4.1 Weak Supervision Labels

Table 1 shows for each labeling function its coverage (as a percentage of the full corpus), the proportion of overlaps with at least one other labeling function, the proportion of conflicts with at least one other labeling function, and its accuracy estimated through the aggregated label model.

The weak supervision model abstains from labeling $15.9\%$ of the summaries, giving us a labeled dataset of $M_{\text{lab}} = 81\,195$ samples. Figure 2 shows a histogram of the resulting probabilistic labels, $y_m^+ = P_{\boldsymbol{\mu}}(y_m = 1 \mid \boldsymbol{\lambda}_m)$ for $m = 1, \ldots, M_{\text{lab}}$, where each $y_m^+$ is the probability of summary $m$ being of high quality. We observe many summaries for which $y_m^+ \approx 0$ or $y_m^+ > 0.7$. The labels seem otherwise quite evenly distributed on the probability range $[0, 1]$, and their average is $0.493$, which indicates that the dataset is well balanced and does not require oversampling. We split the labeled dataset of 81 195 samples in the ratio 8:1:1, yielding a training set of $64\,955$ samples and validation and test sets of $8\,120$ samples each.

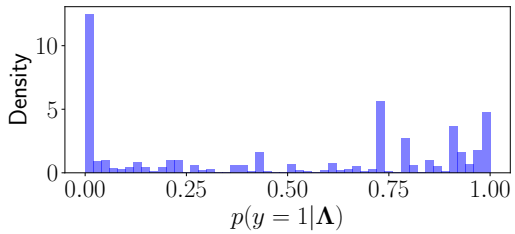


Figure 2: Histogram showing the distribution of labels from the weak supervision model.

| Model | Loss | Acc. | $F_1$ |
|---|---|---|---|
| LSA | - | 0.726 | 0.755 |
| Doc2vec | - | 0.684 | 0.686 |
| LSA+LinTrans | 0.095 | 0.863 | 0.876 |
| Doc2vec+LinTrans | 0.101 | 0.850 | 0.863 |
| LSA+FFN | 0.080 | 0.882 | 0.893 |
| Doc2vec+FFN | **0.079** | 0.885 | 0.897 |
| LSA+LSTM | **0.079** | 0.882 | 0.895 |
| Doc2vec+LSTM | 0.080 | 0.880 | 0.891 |
| EmbLayer+CNN | 0.088 | 0.888 | 0.898 |
| Word2vec+CNN | 0.085 | **0.895** | **0.905** |

Table 2: Model performances on the test set.

### 4.2 Model Performance

We evaluate our models against the weak supervision labels. The model performances on the test set are given in Table 2, measured by the standard classification scores accuracy and $F_1$, and for the supervised models also by the loss function given in (1). We train the models using the Adam optimizer with a learning rate of $1 \times 10^{-4}$, reduced by a factor of $0.1$ after one third of the epochs, and again after two thirds. We also employ a dropout of 0.2 in the hidden layers.

For the computation of accuracy and $F_1$, the probabilistic labels $y^+$ and the quality measures $q(\mathbf{r}, \mathbf{s})$ are converted to binary labels; the threshold for $y^+$ is 0.5, while for $q(\mathbf{r}, \mathbf{s})$ the threshold is tuned on the validation set. We see that the supervised models outperform the unsupervised ones and that the model Word2vec+CNN achieves the best performance both in terms of accuracy and $F_1$.

It should be noted that the aggregated labels obtained with weak supervision only constitute a proxy for the ground truth. Although we expect them to provide good indications of the overall quality of the summaries in this domain, we cannot be certain of how well they correlate with human judgment, so our conclusions regarding the ability of various models to measure summary quality must remain somewhat tentative.

Figure 3 illustrates the performance of four models by showing the distributions of quality measures for samples where the weak supervision label model is confident about the label. Summaries with $y^+ \geq 0.9$ are shown in green and those with $y^+ \leq 0.1$ are shown in red. We observe that all of these models are, to some degree, able to distinguish good summaries from bad ones. The unsupervised LSA baseline does, however, have much more overlap than the other models, which reflects the poorer performance in Table 2. The distributions for the model LSA+LSTM is unexpected, in that it pushes the quality measures just below $\tau_{\text{bad}} = -0.2$ or just above $\tau_{\text{good}} = 0.2$, instead of distributing them on the complete quality range $[-1, 1]$. This behavior effectively makes it a classifier rather than a model of quality measure. We observe the same behavior for the Doc2vec+LSTM model and FFN-based models. The LinTrans and CNN-based models, on the other hand, yield a good separation of good and bad summaries, while distributing them on a large portion of the quality range, which is the behavior we seek.

Figure 4 illustrates the distribution of quality measures assigned to all of the $M = 96\,534$ samples in the corpus by the Word2vec+CNN model. By comparing this histogram to the one in Figure 2, we see that this model provides a more continuous quality measure than the labels aggregated from the labeling functions using weak supervision.

### 4.3 Summary Quality

When applied to the entire corpus of real estate condition reports and summaries, including the ones that the weak supervision model abstained from labeling, the Word2vec+CNN model finds that $35\%$ of the summaries have a quality score $q(\mathbf{r}, \mathbf{s})$ below $\tau_{\text{bad}} = -0.2$, our chosen threshold for being of poor quality, while $33\%$ are judged to be of high quality (i.e., $q(\mathbf{r}, \mathbf{s}) > \tau_{\text{good}} = 0.2$), while the remaining $31\%$ are considered mediocre. The LSA+LinTrans model find $28\%$ of the summaries to be of poor quality, and an average of the CNN and LinTrans models gives a proportion of poor summaries around $30\%$. If almost a third of the summaries of real estate condition reports are in fact of poor quality, this would bode ill for the real estate buyers that do not read the full reports.

Three example summaries are included in Ap-

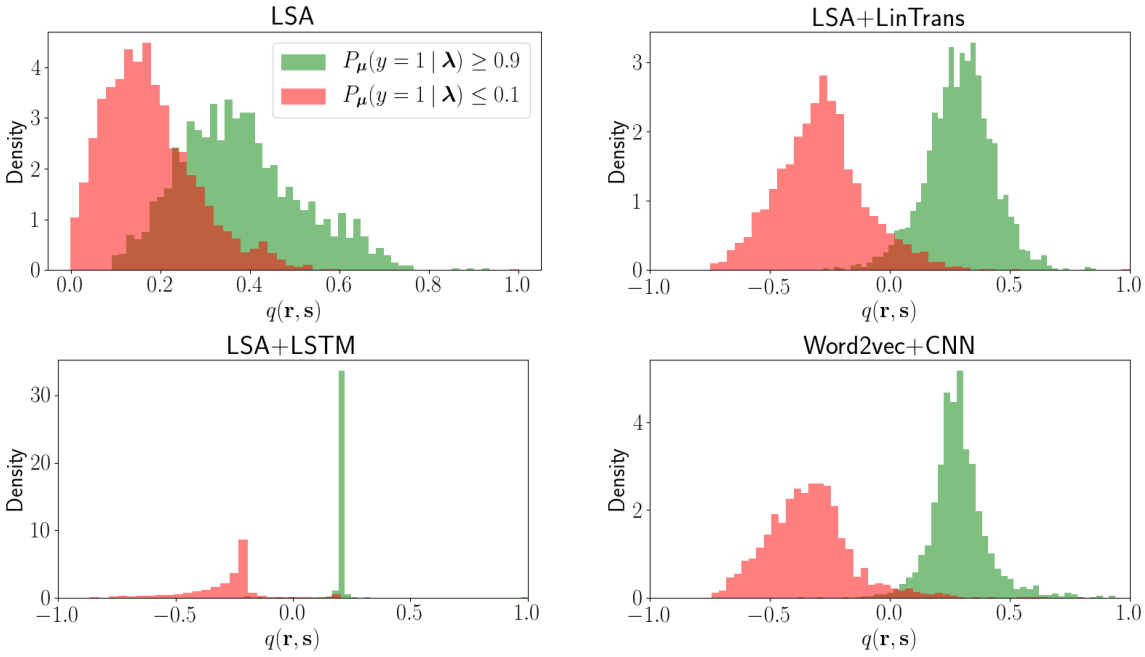


Figure 3: Normalized histograms showing the distribution of quality measures $q(\mathbf{r}, \mathbf{s})$ for summaries from the test set that the label model considers as good (shown as green) and bad (shown as red).

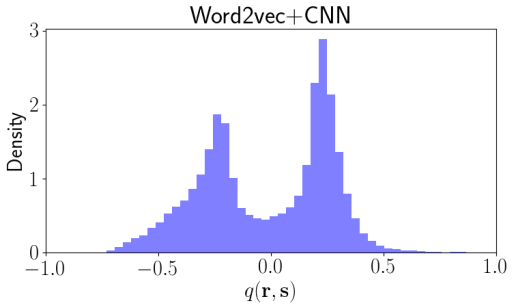


Figure 4: Normalized histogram of $q(\mathbf{r}, \mathbf{s})$ for the entire corpus of summaries.

| **Model** | **Ex. 1** | **Ex. 2** | **Ex. 3** |
|---|---|---|---|
| $P_{\mu}(y = 1 \mid \boldsymbol{\lambda})$ | 0.92 | 0 | 0 |
| LSA+LinTrans | 0.24 | −0.68 | 0.32 |
| Doc2vec+LinTrans | 0.46 | −0.54 | 0.28 |
| EmbLayer+CNN | 0.67 | −0.62 | 0.41 |
| Word2vec+CNN | 0.23 | −0.68 | 0.61 |

Table 3: Quality scores for the three example summaries given in the appendix.

pendix A. Their predicted quality measures using the weak supervision model, the LinTrans models and the CNN models are given in Table 3. We see that all models agree that the first summary is of good quality, and that the second is relatively bad. Since the first summary is quite thorough while the second is excessively short and quite uninformative, this is in line with our expectations. The third summary, however, is considered poor by the label model but quite good by the neural models. As this is also a quite thorough summary which captures the essence of its corresponding report, it would seem that the supervised models outperform in this case the labels they were trained on. We observed several such examples in the corpus, but without data from human judgments, we cannot ascertain to what extent the neural models are truly more reliable than the weak supervision labels.

## 5 Conclusion

This paper describes a novel approach to automatically assess the quality (focusing primarily on the criteria of content coverage) of human-generated summaries, using a corpus of real estate condition reports as a concrete example. The approach relies on the creation of document embeddings that are appropriate for measuring summary quality. This gives us a particular kind of semantic space (the summary content space) where summary quality can be measured by the cosine similarity between the report and its summary.

Since we have no access to "ground truth" values for the summary quality, we obtain indirect quality indicators based on a set of 22 heuristic rules gathered from human experts. Those quality indi-

cators are then aggregated into a single probability (of a summary being of high quality) using weak supervision. The aggregated probabilities are subsequently employed as targets for training neural models optimised for the task of predicting summary quality. Evaluation results show that the best neural model, based on a convolutional architecture, achieves an overall accuracy of $89.5\%$ when measuring the model output against the aggregated labels, while the best unsupervised model (LSA) only achieves an accuracy of $72.6\%$.

An important limitation of the proposed method is the reliance on indirect indicators of summary quality (as expressed by the heuristic rules) instead of human judgments. A key research question for future work is thus to examine the correlations between the quality measures derived from the labeling functions and human judgments. While the heuristic rules do not capture all aspects that may influence the overall quality of a summary, our hypothesis (yet to be validated) is that they nevertheless correlate well with human judgments. An additional benefit of these heuristic rules is their explanatory power, making it possible to provide concrete, human-readable suggestions on *how* to improve a given summary.

Although not considered in this paper, the use of document embeddings relying on contextual word representations is another interesting research question that we wish to investigate in future work.

## References


Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Chris Ré, and Rob Malkin. 2019. Snorkel DryBell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, pages 362–375. Association for Computing Machinery.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Eran Bringer, Abraham Israeli, Yoav Shoham, Alex Ratner, and Christopher Ré. 2019. Osprey: Weak supervision of imbalanced extraction problems without code. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*, pages 1–11. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Luis Adrián Cabrera-Diego and Juan-Manuel Torres-Moreno. 2018. Summtriver: A new trivergent model to evaluate summaries automatically without human references. *Data & Knowledge Engineering*, 113:184–197.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

John M. Conroy and Hoa Trang Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 145–152, Manchester, UK. COLING 2008 Organizing Committee.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on*

*Computational Linguistics*, pages 350–356. Association for Computational Linguistics.

Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information processing & management*, 56(5):1794–1814.

Patrick Fernandes, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Structured neural summarization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in ACL 2013 MultiLing workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria. Association for Computational Linguistics.

Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. 2019. Document similarity for texts of varying lengths via hidden topics. *arXiv preprint arXiv:1903.10675*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2410–2420, Berlin, Germany. Association for Computational Linguistics.

Huseiernes Landsforbund. 2017. Konfliktnivået ved bolighandel må ned. [Online; accessed 3-February-2021].

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. Proceedings of Machine Learning Research.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online. Association for Computational Linguistics.

Ming Liu, Bo Lang, Zepeng Gu, and Ahmed Zeeshan. 2017. Measuring similarity of academic articles with semantic profile and joint word embedding. *Tsinghua Science and Technology*, 22(6):619–632.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: An overview. *Language Resources and Evaluation*, 52(1):101–148.

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

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology network. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore. Association for Computational Linguistics.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. 2019. Training complex models with multi-task weak supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4763–4771.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3574–3582. Curran Associates Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378.

Vasile Rus, Nobal Niraula, and Rajendra Banjade. 2013. Similarity measures based on latent dirichlet allocation. In *Computational Linguistics and Intelligent Text Processing*, pages 459–470. Springer.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Esteban Safranchik, Shiying Luo, and Stephen Bach. 2020. Weakly supervised sequence tagging from noisy rules. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5570–5578.

Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-Morales. 2010. Multilingual summarization evaluation without human models. In *COLING 2010: Posters*, pages 1059–1067, Beijing, China. COLING 2010 Organizing Committee.

Tor Sandberg. 2017. Kjøper dyre boliger i blinde. *Dagsavisen*. [Online; accessed 3-February-2021].

Juan-Manuel Torres-Moreno, Horacio Saggion, Iria da Cunha, Eric SanJuan, and Patricia Velázquez-Morales. 2010. Summary evaluation with and without references. *Polibits*, (42):13–20.

Tedo Vrbanec and Ana Meštrović. 2020. Corpus-based paraphrase detection experiments and review. *Information*, 11(5):241.

Hai Wang and Hoifung Poon. 2018. Deep probabilistic logic: A unifying framework for indirect supervision. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1891–1902, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

## Appendix A. Example Summaries

1. Enebolig fra 1978 som er holdt vedlike og har god standard, tatt alder i betraktning. Den er noe påkostet over tid ellers er det originalt. Det er valmtak med bordtak. Renner og nedløp. Bindingsverkvegger som er isolert med stående panel og murforblending. Vinduer med karm og ramme i tre med isolerglass. Massiv utgangsdør i teak. Det er leca grunnmur og støpt dekke. Dreneringen er fra byggetiden. Innvendig er det panel og plater i himling, gulv har fliser, belegg, laminat, tepper og parkett. Baderom med fliser på gulv og vegger med sanitær utstyr som er fra byggetiden. Det er eget wc rom og dusjkabinett i fyr-rom og wc med servant i vaskerom. Eik kjøkkeninnredning med profiler på overskap og underskap fra byggetiden. Sentralfyr for olje og strøm som er ca 10 år. Oljetank under terrasse. Elektrisk anlegg med skrusikringer. Garasje fra 1986 den er oppført med støpt dekke, leca ringmur, stående kledning. Valmtak med betongstein, renner og nedløp i plastbelagt stål. Det er 2 stk leddporter. Det er registrert vanlig elde og bruksslitasje på eiendommen.

2. Boligen ligger i et etablert boligområde, med kort vei til skole, barnehage og forretning. Det er gjort bemerkninger som bør utbedres, som våtrom og oppgraderinger pga. normal bruksslitasje. Forøvrig les rapport.

3. Bolig bygget i år 2005 med gjeldende forskrifter fra byggeår. (Plan og bygningsloven fra 1985, revidert i 1997. Teknisk forskrift -97.) Boligen og garasje fremstår som normalt vedlikeholdt. Malte flater på alle vegger og himlinger i oppholdsrom. Keramiske fliser på gulv og vegger i bad. Keramiske fliser på gulv i vaskerom. Vedovn i stue med inndekning fra år 2010. Gruset

område rundt boligen. Stor terrasse på oppside med støpte fundamenter. Garasje med plass til to biler. Keramiske fliser på vegger og gulv i bad. TG2 grunnet alder. Keramiske fliser på gulv i vaskerom. Vegger platet med malt tapet. TG2 grunnet alder. Adkomstdør trenger justering. TG2 Platon grunnmursplate. Manglende topp-list. Dette kan samle fukt mot grunnmur. Løv og barnåler bak platonplate ble registrert ved befaring. Rensing og festing av plate anbefales. TG3 Ett nedløp i front av bolig ikke tilkoblet drensrør. TG2.