Evaluating and Improving Automatic Speech Recognition using Severity

Ryan Whetten
Boise State University
1910 W University Dr
Boise, ID 83725
ryanwhetten@boisestate.edu

Casey Kennington
Boise State University
1910 W University Dr
Boise, ID 83725
caseykennington@boisestate.edu

Abstract

A common metric for evaluating Automatic Speech Recognition (ASR) is Word Error Rate (WER) which solely takes into account discrepancies at the word-level. Although useful, WER is not guaranteed to correlate well with human judgment or performance on downstream tasks that use ASR. Meaningful assessment of ASR mistakes becomes even more important in high-stake scenarios such as healthcare. We propose 2 general measures to evaluate the severity of mistakes made by ASR systems, one based on sentiment analysis and another based on text embeddings. We evaluate these measures on simulated patient-doctor conversations using 5 ASR systems. Results show that these measures capture characteristics of ASR errors that WER does not. Furthermore, we train an ASR system incorporating severity and demonstrate the potential for using severity not only in the evaluation, but in the development of ASR. Advantages and limitations of this methodology are analyzed and discussed.

1 Introduction

Automatic Speech Recognition (ASR) is the task of processing human speech into text, but no ASR is perfect and certain types of errors can cause potential problems. ASR has drastically improved over the past decade and has changed the way many people interact with computers in applications such as voice search, dictation, and virtual assistants (Yu and Deng, 2016; Alharbi et al., 2021). It is common practice to evaluate ASR by calculating the word error rate (WER) which can be calculated by counting the number of words that need to be substituted (S), deleted (D), and inserted (I) to go from a ground-truth transcription to the output of an ASR. This count is then divided the total number of words in the ground-truth transcription (N) similar to Levenshtein et al. (1966), often written as $(S + I + D)/N$. Essentially, WER treats each discrepancy between the ground-truth transcription and the ASR output **equally**.

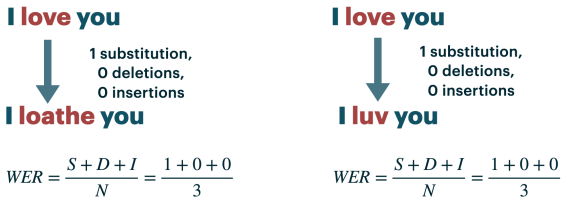


Figure 1: WER calculations for the “I love you" example. S, D, I represent the number of substitutions, deletions, and insertions to go from the ground-truth transcription to the output of an ASR. N represents the total number of words in the ground-truth transcription.

However, not all ASR errors are equal. Previous work studying ASR errors has shown that WER is not always well correlated with human judgement or performance in a given downstream task, such as in information retrieval, natural language understanding or named entity recognition (Garofolo, 1999; Galibert et al., 2016; Wang et al., 2003; Riccardi and Gorin, 1998; Kim et al., 2021a,b). For example, take the sentence “I love you,” and suppose an ASR system produces “I loathe you.” This results in a WER of 0.33. Now suppose another ASR system predicts “I luv you.” This too results in a WER of 0.33 (see example in Figure 1). Although the WERs are equal, compared to the ground-truth “I love you,” the mistake “luv” is less severe than the mistake “loathe,” which gives off the opposite meaning from the ground-truth sentence.

Being able to understand the severity of ASR errors becomes even more critical in high-stake scenarios such as healthcare where ASR has been used since the 1970s (Johnson et al., 2014). In healthcare research, transcriptions are used in a wide variety of tasks such as in the automated detection of dementia (Farzana et al., 2022), in estimating scores of standardized cognitive health screening tests (Farzana and Parde, 2020), and in in the prediction and explanation of diagnosis (Ngai

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 79–91
July 13, 2023 ©2023 Association for Computational Linguistics

and Rudzicz, 2022). However, these works operate on the basis of having an intelligible and accurate transcript. The purpose of this research is to develop a method for systematically measuring and understanding the quality of ASR systems, especially in high-stake settings like healthcare, going beyond WER by looking at the difference in meaning between the ground-truth and ASR output.

In this work, we propose two methods for the automatic rating of the severity of errors in ASR transcriptions by using 1) the difference in sentiment ratings and 2) cosine distances between text embeddings of the output of the ASR and the ground-truth human transcription. We compare them with human-labeled severity scores. Text embeddings prove to be correlate better with human labels of severity than WER. This work also shows sentiment ratings, text embeddings, and WER capture different aspects of mistakes in transcriptions and shows advantages and limitations of each method. Lastly, we demonstrate the potential for severity to be used in the development of ASR systems. We conclude by discussing limitations and future areas of research.

2 Related Work

To overcome the limitations of WER, other measures for ASR evaluation have been proposed such the Match error rate (MER), and Word information lost (WIL) (Morris et al., 2004). In the context of information retrieval, metrics based on named-entity word error rate correlates higher with retrieval performance than WER (Garofolo, 1999). However, these methods are still based on the literal word correctness and do not take into account semantics.

Kafle and Huenerfauth (2017) incorporate semantics into a measure by including a weighted distance between Word2Vec (Mikolov et al., 2013a) vectors for misspelled words. They show that their measure correlates better with the perception of people who are Deaf or Hard of Hearing and mention using sentiment analysis in future work. Kim et al. (2021a,b), similar to this work, take into account semantics by running the ground-truth reference and ASR output through RoBERTa and obtain an embedding for each by computing the mean of all the output vectors of the RoBERTa model. The ground-truth and ASR output are compared by using the cosine distance between their embeddings. Roux et al. (2022) uses similar methods and look at the POSER (the Part of Speech Error Rate) and EmbER (Embedding Error Rate), a WER that is weighted by the semantic similarity of incorrectly transcribed words. Our work supports the results of these works in a healthcare context as well as explores the idea of using sentiment analysis. Furthermore, our work is the first that we know of to include some approximation of the semantic meaning into the training regime of ASR.

3 Methods

When it comes to understanding and automatically rating the seriousness of errors in ASR, one needs to have a method for systematically analyzing the difference in *meaning* between two phrases or sentences. While philosophically what a body of text truly means is a difficult question to answer, we can capture some essence of the *meaning* of an utterance using sentiment analysis and text embeddings.

Sentiment analysis is the task of detecting the attitude, emotions, or polarity of a given text. These algorithms usually take in a string as input and output a prediction from -1 to 1 based on how negative or positive the text is. Because these algorithms vary, we use 3 different sentiment analyzers from 3 different widely-used NLP libraries NLTK, FLAIR, and TextBlob (TB) (Hutto and Gilbert, 2014; Akbik et al., 2019; Loria et al., 2018). This is a naive method of capturing the *meaning* of text because two texts can have different meanings and both have similar sentiment. Although overly simplistic the purpose of using sentiment is to create and test a baseline measure that captures aspects of text besides discrepancies in spelling.

Another method for numerically capturing the *meaning* of natural language is to use sentence-level embeddings (phrases or sentences are converted into n-dimensional vectors). There are a variety of methods for embedding words that range from simple rule-based methods to methods that involve machine learning (Mikolov et al., 2013a; Pennington et al., 2014; Peters et al., 2018). Similarly, methods have been developed for embedding more than just single words (Le and Mikolov, 2014; Reimers and Gurevych, 2019). Whether for individual words or phrases, with good embeddings, the more semantically similar words or phrases are, the closer they should be in the n-dimensional vector-space (Mikolov et al., 2013a,b).

We use 4 readily available pre-trained models provided by SentenceTransformers[1] to com-

[1] https://www.sbert.net/docs/pretrained_models.html

pute text embeddings[2]: bert-base-nli-mean-tokens (BertNLI), all-MiniLM-L6-v2 (MiniLM), all-mpnet-base-v2 (MPNET), and all-distilroberta-v1 (DisRob) (Reimers and Gurevych, 2019; Wang et al., 2020; Song et al., 2020; Sanh et al., 2019). These models are selected due to their performance in semantic similarity tasks. These models are able to compute embeddings fast and are not too big, ranging from 80 MB to 420 MB.

3.1 ASR Systems

We use 5 ASR systems for experimentation in order to collect and obtain results from a variety of architectures. We choose these architectures because of their availability, performance and because they can be run locally (which means one would not have to deal with potential issues with sending sensitive data over the internet to a cloud ASR system). The five we use are Mozilla's DeepSpeech based on (Amodei et al., 2016) with the version 0.9.3 model and scorer, Meta's Wav2Vec2 (W2V2) (Baevski et al., 2020), CMU's Pocket Sphinx with the 'en-us' model and dictionary (Huggins-Daines et al., 2006), Alpha Cephei's Vosk using the vosk-model-en-us-0.22 model, and OpenAI's Whisper (Radford et al., 2022). See Section A.1 for more details.

4 Data

For the purposes of experimenting in a healthcare scenario, a dataset of simulated patient-physician medical interviews is used (Fareez et al., 2022). This dataset contains 272 audio files with transcripts (about 7 to 20 minutes or from 800 to 2200 words). The conversations are categorized into the following cases/subjects: respiratory, musculoskeletal, cardiac, dermatological, and gastrointestinal diseases. The majority of simulations were respiratory cases (78.7%).

The files are split into non-silent intervals using librosa[3] setting the threshold of silence to 60 decibels. This results in over 39,600 non-silent intervals, which we will call utterances. Of these, we take a sample of 110 utterences and run them through the ASR systems and get the ground-truth transcription from the corresponding transcription files. These were run through the ASR systems to create a list of 550 pairs of strings, one string being the ground-truth and the other coming from an ASR. One of the audio files contained no speech and was removed for a final total of 545 pairs. The transcripts were normalized by removing speaker identification notes "P:" and "D:" for patients and doctors (found in the ground-truth), making all letters lowercase, and by removing any special characters and punctuation except for apostrophes (as these could be important in distinguishing words like "its" and "it's".

150 of these pairs of transcripts were given to 3 medical school students who were asked to rate each pair with either 0, 1, or 2 (2 being a severe error, 1 being a not so severe error, and 0 being a very minor error or perfect transcription). The exact instruction given and a few examples of the data are provided in Figure 4.

4.1 Data Validation: Do Raters Agree?

Previous work suggests that the severity of errors in transcription is a difficult task where there is not very good consensus among raters (Luzzati et al., 2014). Prior to developing a measure that rates errors in the same way a human would, it first needs to be shown that humans do have some methodology or consistency amongst each other when it comes to rating the severity of errors.

Following the evaluation metrics used in Luzzati et al. (2014), we use Cohen's Kappa (Cohen, 1960) and Fleiss' Kappa (Fleiss, 1971), to measure at inner-annotator agreement. However, these metrics do not take in to account that the data is ordinal (i.e. a discrepancy in ratings of values 0 and 1 is treated the exact same as a discrepancy in values of 0 and 2 even though the latter discrepancy is greater than the former (Falotico and Quatto, 2015)). Therefore, since the nature of these ratings is ordinal, we also look at the Kendall's rank correlation coefficient to measure the quality of the ordinal association between two given raters (Kendall, 1938).

We calculate a Fleiss' Kappa values of 0.452 and Cohen's Kappa scores that range from 0.420 to 0.567, which shows moderate agreement between raters (Table 1). The Kendall's correlation coefficient between raters indicated a strong correlation between rater ranging from 0.662 to 0.727 (Table 1). Considering the subjectivity of the task, the moderate Kappa values and high correlation values suggest that there is reliable consistency among raters. Having shown reliable consistency among raters, the following sections describe the experiments.

[2] All these models compute sentence embeddings, not word embeddings. We refer to these as text embeddings because many of the utterances in the data are not full sentences.

[3] https://librosa.org/

	Rater 1	Rater 2	Rater 3
Rater 1	-	0.416	0.567
Rater 2	0.727	-	0.440
Rater 3	0.718	0.662	-

Table 1: Inner-annotator agreement and correlation between raters measured by Cohen's Kappa (upper right quadrant) and Kendall's correlation coefficient (lower left quadrant).

5 Exp. 1: Testing Severity Scores

In this experiment, we test if sentiment analyzers and/or text embeddings can rate errors similarly to the way humans would in a healthcare setting. We calculate the WER and various *severity scores* (defined in Section 5.1) using the sentiment analyzers and the language models for text embeddings. We compare the severity scores by measuring the correlation between the severity scores and the mode human rating.

5.1 Severity Scores

Given a sentiment analyzer $s(x)$ that outputs a value between -1 and 1, we can take the absolute value of the difference in sentiment and use this as model to represent the *difference in meaning* or severity. This can be expressed by the following:

$$\text{Severity}(x, y) = |s(x) - s(y)|$$

where x and y are a pair of ground-truth and ASR output strings. This results in a rating on the range [0, 2], where 0 would be two phrases that have the exact same sentiment rating and a rating of 2 would represent the most severe error possible, having sentiments and polar extremes.

For text embeddings, knowing that similar embeddings should be semantically closer, we can represent the *difference in meaning* as one minus the cosine of their embeddings.

$$\text{Severity}(x, y) = 1 - cosine(x, y)$$

This results in a rating on the range [-1, 1]. However, it is common practice to bound the vectors in the positive space which would result in range of [0, 1]. Because we are looking at the dissimilarity, a value of 0 would represent two strings that are the same and a value close to 1 would represent a two strings that are very different semantically.

5.2 Experiment 1 Results

The correlations in Table 2 show WER has a correlation with human ratings of severity of 0.43. All the embedding scores correlated better with human ratings, ranging from 0.53 to 0.59. In contrast, all of the sentiment severity scores were less correlated than WER, ranging from 0.29 to 0.34.

WER	FLAIR	MiniLM	DisRob
0.43	0.34	0.55	**0.59**

Table 2: Correlation between human rating of severity to WER, and severity based sentiment (FLAIR), and severity based on text embeddings (MiniLM and DisRob). Severity scores based on text embedding correlate the best with human ratings. For full table see Table 5 in the appendix.

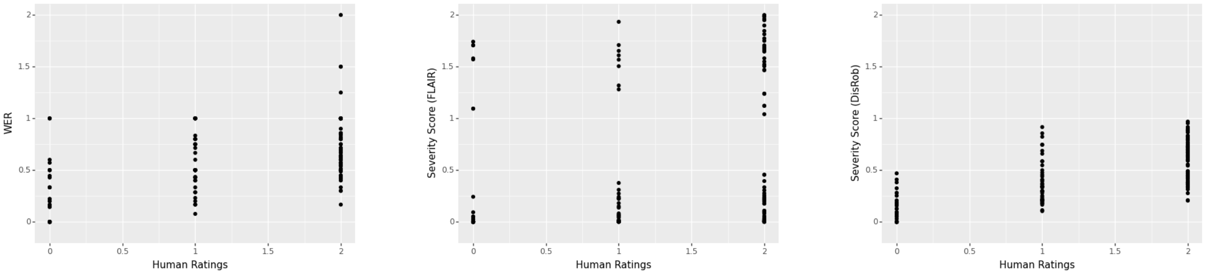
This is shown graphically in Figure 2. Subplot 2c (from DisRob) shows that embeddings do the best job of clustering ASR errors with the same human rating together (i.e. with a similar cosine distance). In contrast, the WERs and sentiment scores shown in subplots 2a and 2b are more spread out, having some severe errors with relatively low WER/sentiment score and some non-severe errors with a relatively high WER/sentiment score. These correlations show that text embeddings can be better suited for the automatic evaluation of severity in ASR errors than WER and supports findings from Kim et al. (2021b).

In this experiment, we only study the correlation between the proposed severity scores and human ratings. However, it is common for WER to be averaged across all the utterances in a test dataset. The average WER becomes a single value that is used as a metric to gauge the overall performance of ASR systems. Since the severity scores correlate with human labels, we test these measures to see if they can be used in a similar manner to average WER to gauge the performance of ASR.

6 Exp. 2: Severity in ASR Evaluation

This experiment demonstrates the potential of using the severity measures in metrics for the overall evaluation and comparison of ASR systems. To test this, we propose three metrics (Section 6.1). We use the average WER and each metric to evaluate the performance of the 5 ASR systems.

These metrics are compared with the average WER and with each other across the ASR systems.



(a) WER

(b) Sentiment (Flair)

(c) Embeddings (DisRob)

Figure 2: Graphs comparing human ratings of severity (x-axis) to WER, FLAIR and DisRob. DisRob (embeddings) do the best job of clustering ASR errors with the same human rating together.

6.1 Metric 1: MAE of Difference in Sentiment

We propose the mean absolute error of the differences in sentiment (MAE-DS). Given a sentiment analyzer $s(x)$ that outputs a value between -1 and 1, we can express the MAE-DS in the formula below:

$$\frac{1}{n}\sum_{x,y\in C}|s(x)-s(y)|$$

where *C* is a corpus of pairs of ground-truth transcriptions and ASR predictions and *n* is the number pairs. *x* and *y* are a set of ground-truth and predicted utterances from C.

The output of this metric will range from 0 to 2, and will be simple to interpret. For example, a MAE-DS of 0.5 indicates that, in terms of sentiment, the ASR's output is, on average, 0.5 off of the ground-truth.

6.2 Metric 2: MSE of Difference in Sentiment

The second metric we propose is the mean squared error of the differences in sentiment (MSE-DS). Similar to the first metric, given a sentiment analyzer $s(x)$ we can write this in the formula below:

$$\frac{1}{n}\sum_{x,y\in C}(s(x)-s(y))^2$$

where, *C* is a corpus of pairs of ground-truth transcriptions and ASR predictions and *n* is the number pairs. *x* and *y* are a set of ground-truth and predicted utterances from C.

The range of this metric is from 0 to 4. Because sentiment scores are squared, this will penalize more heavily the ASR errors that have a greater distance in sentiment from the ground-truth.

6.3 Metric 3: Sentence Similarity using Language Models

For the third metric, we propose using the mean of the cosine distance. This can be written with the following formula:

$$\frac{1}{n}\sum_{x,y\in C}1-cosine(x,y)$$

where *C* is a corpus of pairs of ground-truth and ASR output and *n* is the number pairs. *x* and *y* are a set of embeddings of a given ground-truth and ASR output from C.

6.4 Experiment 2 Results

The results are summarized in Table 3. For all the metrics, the lower the value the better. Generally, results are consistent, no matter which metric used, the majority show that Whisper has the best performance followed by Vosk. Following these in performance are DeepSpeech2 (DS2), Wav2Vec2 (W2V2), and PocketShpinx (PS). We go into more depth and show to what extent these metrics agree and where the metrics disagree, in order to gain insights about what information these metrics are capturing.

From Vosk to Whisper there is decrease in WER of about 0.034, and the average decrease in the cosine distance over the 4 language models is quite small, around 0.009. In another example, the decrease in WER from DeepSpeech2 to Vosk is 0.175 while the average decrease cosine distance over the 4 language models is greater at 0.176. These differences show that the rate of improvement in the severity (cosine distance) is not necessarily related to rate of improvement in WER (i.e. one metric can improve greatly while the other not so much).

To further demonstrate the differences of these metrics, we look at specific examples where WER

Base Measure	Metric	DeepSpeech2	PockSphinx	Vosk	Whisper	Wav2Vec2
WER	Average WER	0.482	0.910	0.307	**0.273**	0.525
Sentiment	NLTK_mae	0.127	0.241	0.062	**0.056**	0.127
	FLAIR_mae	0.620	0.700	0.324	**0.322**	0.516
	TB_mae	0.111	0.181	0.050	**0.029**	0.120
Sentiment	NLTK_mse	0.057	0.141	0.022	**0.020**	0.048
	FLAIR_mse	0.981	1.132	**0.459**	0.473	0.788
	TB_mse	0.051	0.086	0.026	**0.010**	0.044
Cosine Distance	MiniLM	0.361	0.649	0.171	**0.153**	0.403
	BertNLI	0.188	0.398	**0.079**	0.093	0.181
	MPNET	0.400	0.688	0.193	**0.180**	0.400
	DisRob	0.388	0.676	0.189	**0.172**	0.406

Table 3: Results of Experiment 2. The top row shows each of the ASR systems. The following row shows the WER. The labels in the first column that end in mae and mse are the mean absolute error and the mean squared error of the difference in sentiment scores respectively. The last for rows are the average cosine distance.

and measures of severity disagree. We do this by analyzing the most severe errors according to one measure while another measure is kept relatively low. We first look at the most severe according to FLAIR sentiment scores while keeping WER below 0.5 (examples of this are shown in the first 6 rows in Table 6, in the appendix). We then look at the most *severe* according to cosine distance while still keeping WER below 0.5 (shown in the middle group of 6 in Table 6). Finally, we look at the most *severe* according to WER while keeping the cosine distance below 0.5 (the last 6 rows in Table 6). These edge case examples show advantages and limitations of WER, sentiment scores, and scores based on text embeddings.

6.5 Advantages and Limitations

All of the examples mentioned in this section are shown in Table 6 in the appendix.

WER has the main advantage of being simple and consistent. Unlike sentiment or text embeddings, there are not multiple models. The main limitation of WER is that, because it is not based on any *understanding* or model of the language, there are severe errors that have a relatively low WER, and vice versa, there are non-severe errors that have a high WER such as *a multivitamin* vs. *a multi vitamin*.

Sentiment has strong limitations due to the fact that these algorithms are designed to only measure how positive or negative a text is. Sentiment proved to be sensitive to misses in disfluencies like *um* or *uh*. This is highlighted in the example, *uhm it started last night* vs *and it started last night*, where there was a strong difference in sentiment of 1.707. This can be an advantage or a limitation depending on the scenario. Many ASR systems overlook disfluencies, but, for example in human robot interactions, spoken dialogue systems, or in the prediction of dementia status, disfluencies can be vital to understanding and performance (Baumann et al., 2017; Clark and Tree, 2002; Farzana et al., 2022; Lopez-de Ipiña et al., 2017; Mueller et al., 2018).

There is the also a limitation on the accuracy of the model. In the examples *any previous surgeries* vs. *any previous surgery* or *uh i smoke about a pack a day* vs. *uh smoke about a pack of day*, there is a high difference in sentiment yet the only difference is in missing the pronoun *i* or the plural of *surgery*, which should not affect sentiment greatly.

Despite these limitations, sentiment is able to catch some severe errors where the WER is relatively low. In the example where *crystal meth* becomes *crystal mud* or where *chest pain* becoming *chatting* the WER is 0.125 and 0.333 respectively, but the difference in sentiment is very high at 1.858 and 1.889 respectively.

Text embeddings are limited by the performance of the model, like sentiment, yet capture more than just polarity of a given text. Knowing that many of these models are trained in a self-supervised manor using the context in the training text, we can see how the embeddings in the example of *my parents* and *our friends* would be similar. Both of these phrases could occur in with similar surrounding text; they have the same grammatical structure (a possessive adjective followed by a noun) and parents and friends are both human relationships.

Another limitation on these models is the amount of text they can handle. Anything above the model's limit gets truncated, and consequently, loses the meaning of truncated text. Although utterances are commonly short in ASR training data, the character limitation on these models could affect performance on longer utterances.

Despite these limitations, text embeddings were able to capture well the differences in meaning. Text embeddings were able to give a high score to the examples where *crystal meth* becomes *for sunlight* and where *chest pain* becomes *testing* when WER and sentiment scores were relatively low. Text embeddings were also able to give low ratings for different writings of the word *okay* and numbers (*ok* vs. *okay*, or *uh thirty eight degrees* vs. *38 degrees*) when WER were high.

7 Exp. 3: Using Severity to Improve ASR

Up to this point results show that 1) there is reliable consistency among human raters, 2) the cosine distance of text embeddings correlates better with human labels of severity than WER, and 3) using sentiment or text embeddings in a metric for the overall evaluation of ASR captures different information than WER. With these results established, the purpose of this experiment is to test if an automatic measure of severity can be used in more than just the evaluation of ASR, but in the training regime as well.

Previous work done in the study of ASR errors involves approaches to automatically detect errors using word and text embeddings (and even other features such as acoustic/prosodic features), (Ghannay et al., 2015, 2018, 2020) and to automatically repair errors in specific cases (such as in certain homophones in French) (Dufour and Estève, 2008). However, instead of ASR error detection or repair which happens post-prediction, our approach is to include severity into the training of an ASR system in an attempt to reduce the number of errors (measured by WER) and to reduce the overall severity of the errors produced (measured by the average cosine distance from Section 6.3). To do this, we incorporate *severity* into the loss function during training of an ASR.

7.1 Using Severity in the Loss Function

It is common for ASR systems that involve neural networks to be trained using the Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006). This algorithm allows one to work with data where both inputs and outputs can vary in length such as in handwriting recognition and speech recognition. Taking advantage of dynamic programming methods, given and input of audio X and a ground-truth transcript Y, CTC can calculate efficiently $p(Y|X)$.

To incorporate *severity* into the loss function, the cosine distance is used as a weight in the loss function. To calculate this weight, w, the cosine distance is limited in the range from a near-zero number, 1.0×10^{-7}, to 1. This is shown in Equation 1, where w is the weight that represents the *severity* between the ground-truth, Y_{truth}, and the output of the ASR, Y_{pred}. This weight is multiplied by the CTC loss value to get the final loss (Equation 2).

This results in a function where the original CTC loss is scaled down, along with the gradients during training proportional to semantic similarity between the ground-truth and ASR output.

$$w = 1 - \max(1.0 \times 10^{-7}, \cos(Y_{truth}, Y_{pred})) \quad (1)$$

$$L = w * \text{CTC} \quad (2)$$

We will refer to this proposed loss function as a CTC-by-Cosine loss function. For this experiment, we use the *all-MiniLM-L6-v* (MiniLM) to generate the embeddings for the ground-truth and ASR predictions.

7.2 Architecture

The system we implement is based on DeepSpeech2 (Amodei et al., 2016), where the input is spectrogram from audio files and the output is the probability distribution of over a set of characters at each time step. The set of characters consists of all the letters of the English alphabet along with the following characters: apostrophe, questions mark, exclamation mark, and blank symbol.

The system starts with two 2D convolutional layers with kernels [11, 41] and [11, 21], both with 32 filters, batch normalization, and is passed through a ReLU activation function after each layer. After the convolutions, there are five bidirectional gated recurrent layers (GRU) each with 512 units with a dropout layer with a rate of 0.5 after each recurrent layer except for the last one.

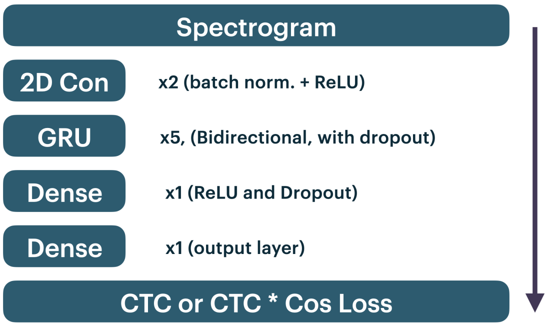


Figure 3: Components of ASR System from Exp. 3.

After the last recurrent layer there are two dense layers. The first one maintains the same size as the recurrent layers and is passed through a ReLU layer and a dropout layer (with a rate of 0.5). The second dense layer is the output layer with softmax as the activation function. Adam is used for optimization with a learning rate of 1.0×10^{-4}. Figure 3 depicts the core components of this model.

This results in a system of about 26M parameters. This is relatively small compared to other ASR systems. For example, DeepSpeech2 has 38M parameters, the base version of Wav2Vec2 has 95M parameters, and the base version of Whisper contains 74M parameters. However, the purpose of this experiment is not to achieve state of the art performance with a novel architecture, it is to test on a smaller scale the plausibility of using severity in the development of ASR.

7.3 Data and Training Regime

Because the 109 utterances of the simulated patient doctor conversation files is insufficient to train an ASR system, for this experiment we use the LJ Speech Dataset which consists of "13,100 short audio clips of a single speaker reading passages from 7 non-fiction book" (Ito and Johnson, 2017).

We train 2 systems on the first 90% percent of the data, withholding the last 10% for validation. The baseline system has the architecture described above and uses only the CTC loss function. The second system uses the exact same architecture and training regime, but uses the CTC-by-Cosine loss function for the last 5 epochs.

7.4 Experiment 3 Results

Results show improvements in both severity (average cosine distance) and WER when incorporating severity into the loss function. From the baseline to the CTC-by-Cos system, there is a relative decrease of about 85% in severity and WER on the training dataset, from 0.058 WER and 0.051 average cosine distance to a 0.008 WER and 0.006 average cosine distance. For the validation data there was a relative decrease of about 18% in severity and WER on the validation dataset, from 0.268 WER and 0.249 average cosine distance to a WER and average cosine distance of 0.219 and 0.201 respectively (see Table 4). This improvement in performance suggests that there is potential to use severity in the development of ASR to decrease both the overall severity and WER. To the best of our knowledge, this is the first work demonstrates the value of using semantics in the training of ASR, working towards the areas of future work mentioned in Kim et al. (2021a,b).

Model	Base	CTC-by-Cos
Train WER	0.058	**0.008**
Train COS	0.051	**0.006**
Val WER	0.268	**0.219**
Val COS	0.249	**0.201**

Table 4: Performance of base and CTC-by-Cos systems on both training (Train) and validation (Val) datasets. COS is the average cosine distance from Section 6.3.

8 Conclusion

Automatic Speech Recognition (ASR) is becoming an increasingly important tool from personal use to the medical field. However, Word Error Rate (WER), a common metric for evaluation, only takes into account word discrepancies (i.e. Figure 1). In this study we 1) compare WER and measures of severity based on sentiment and text embeddings to human labels of severity in a healthcare setting, 2) use these measures in metrics to evaluate the overall quality of mistakes in transcriptions, and 3) incorporate severity into the training of WER.

Results show that 1) cosine distance of text embedding correlates better with human ratings than WER, 2) these measures based on sentiment and text embeddings capture different qualities in ASR errors and can overcome limitations of WER, and 3) incorporating severity into the training of an ASR system increased performance, lowering the overall severity and WER significantly. In future work, we will experiment with different architectures, data, and methods for using semantics in the training of ASR systems.

Limitations

Aside from the limitations of these proposed measures mentioned in Section 6.5, we acknowledge other limitations here. The conclusion drawn here were based on a limited amount of raters and data. While we believe the data to be fairly representative, the results are consistent with other work, more raters from the medical field, more audio data, and in a variety of contexts should be used to make these empirical results more concrete.

We also acknowledged limitations of Experiment 3 (Section 7). We only experiment with one architecture and one dataset. While mathematically using severity as signal (in a way acting as an adjustable learning rate) proportional to the semantic distance seems reasonable, to make more conclusive results on this methodology, a variety of architectures, seeds, and audio data would make results more conclusive.

Ethics Statement

We believe this work to have been conducted and to contribute in an honest, non-discriminate, and professional manor, and to our best knowledge and reflection, we believe we are in full compliance with the ACL Ethics Policy. We note that in future work, if using non-simulated/non-public medical data, care must be taken to protect the privacy of the people involved in full compliance with HIPPA and any IRBs that review these studies.

Acknowledgements

We would like to thank the reviewers for their time in thoroughly reviewing this work. We would also like to thank Andrew Whetten for his insights and comments about dissimilarity measures and Josue Torres-Foncesca for his help in setting up the lab computer. We would also like to thank the annotators from ICOM for their time and expertise.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9:131858–131876.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2017. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, pages 421–432.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Richard Dufour and Yannick Estève. 2008. Correcting asr outputs: specific solutions to specific errors in french. In *2008 IEEE Spoken Language Technology Workshop*, pages 213–216. IEEE.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):1–7.

Shahla Farzana, Ashwin Deshpande, and Natalie Parde. 2022. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 37–48, Dublin, Ireland. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2020. Exploring mmse score prediction using verbal and non-verbal cues. In *INTERSPEECH*, pages 2207–2211.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Olivier Galibert, Mohamed Ameur Ben Jannet, Juliette Kahn, and Sophie Rosset. 2016. Generating task-pertinent sorted error lists for speech recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1883–1889, Portorož, Slovenia. European Language Resources Association (ELRA).

John S Garofolo. 1999. 1998 trec-7 spoken document retrieval track overview and results john s. garofolo, ellen m. voorhees, cedric gp auzanne, vincent m. stanford, bruce a. lund national institute of standards and technology (nist) information technology laboratory. In *Broadcast News Workshop'99 Proceedings*, page 215. Morgan Kaufmann.

Sahar Ghannay, Yannick Esteve, and Nathalie Camelin. 2015. Word embeddings combination and neural networks for robustness in asr error detection. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1671–1675. IEEE.

Sahar Ghannay, Yannick Estève, and Nathalie Camelin. 2018. Task Specific Sentence Embeddings for ASR Error Detection. In *Interspeech 2018*, Hyderabad, India. ISCA.

Sahar Ghannay, Yannick Estève, and Nathalie Camelin. 2020. A study of continuous space word and sentence representations applied to asr error detection. *Speech Communication*, 120:31 – 41.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W Black, Mosur Ravishankar, and Alexander I Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14(1):1–14.

Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 165–174.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L Seltzer. 2021a. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. *arXiv preprint arXiv:2104.02138*.

Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael L Seltzer. 2021b. Evaluating user perception of speech recognition system quality with semantic distance metric. *arXiv preprint arXiv:2110.05376*.

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

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Karmele Lopez-de Ipiña, Unai Martinez-de Lizarduy, Pilar M Calvo, Blanca Beitia, Joseba Garcia-Melero, Miriam Ecay-Torres, Ainara Estanga, and Marcos Faundez-Zanuy. 2017. Analysis of disfluencies for automatic detection of mild cognitive impartment: a deep learning approach. In *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pages 1–4. IEEE.

Steven Loria et al. 2018. textblob documentation. *Release 0.15*, 2(8).

Daniel Luzzati, Cyril Grouin, Ioana Vasilescu, Martine Adda-Decker, Eric Bilinski, Nathalie Camelin, Juliette Kahn, Carole Lailler, Lori Lamel, and Sophie Rosset. 2014. Human annotation of asr error regions: Is "gravity" a sharable concept for human annotators? In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3050–3056.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Kimberly D Mueller, Rebecca L Koscik, Bruce P Hermann, Sterling C Johnson, and Lyn S Turkstra. 2018. Declines in connected language are associated with very early mild cognitive impairment: Results from the wisconsin registry for alzheimer's prevention. *Frontiers in Aging Neuroscience*, 9:437.

Hillary Ngai and Frank Rudzicz. 2022. Doctor XAvIer: Explainable diagnosis on physician-patient dialogues and XAI evaluation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 337–344, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Giuseppe Riccardi and Allen L Gorin. 1998. Stochastic language models for speech recognition and understanding. In *ICSLP*.

Thibault Bañeras Roux, Mickael Rouvier, Jane Wottawa, and Richard Dufour. 2022. Qualitative evaluation of language model rescoring in automatic speech recognition. In *Interspeech*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pretraining for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, pages 577–582. IEEE.

Dong Yu and Li Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

A Appendix

A.1 Summary of ASR Systems

Mozilla's DeepSpeech2 (DS2) is an implementation of Amodei et al. (2016). In this architecture, Recurrent Neural Networks take in spetrograms from an audio file and are trained to output text[4].

Meta's Wav2Vec2 (W2V2) is a model proposed by Baevski et al. (2020). Unlike DeepSpeech, this architecture operates directly on the raw audio data instead of spretrograms. The model is trained first in a semi-supervised method on many hours of unlabeled speech data and then is fine tuned on labeled data. This model is made easily accessible by HuggingFace [5].

CMU'S PocketSphinx (PS) is one of the lighter ASRs we use (Huggins-Daines et al., 2006). PS is a light-weight ASR that is a part of the open source speech recognition tool kit called the CMUSphinx Project. This model was trained on 1,600 utterances from the RM-1 speaker-independent training corpus. Unlike the previously mentioned models, PS does not use neural networks and is instead based on traditional methods of speech recognition by using Hidden Markov Models, language models, and phonetic dictionaries.[6]

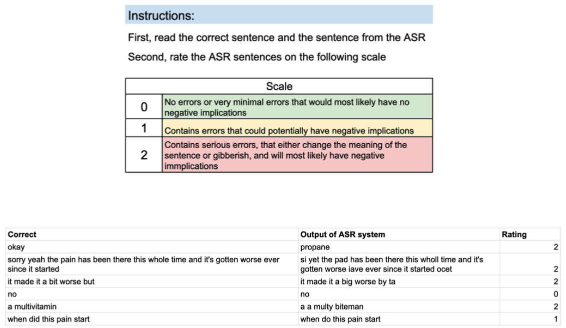
Alpha Cephei's Vosk (with the vosk-model-en-us-0.22 model) is built using Kaldi (Povey et al., 2011), and like PS, uses an acoustic model, language model, and phonetic dictionary. However unlike PS, Vosk uses a neural network for the acoustic model part of the system.[7]

OpenAI's Whisper unlike Wave2Vec2, uses a purely supervised method of training gathering 680K hours of transcribed content from the internet in 99 different languages (Radford et al., 2022). Following other architectures such as DeepSpeech2, this model takes spectrograms of audio as input, but instead of Recurrent Neural Networks, this models uses an encoder-decoder Transformer architecture based on Vaswani et al. (2017) with a variety of special tokens used to indicate which task is being performed (ex. transcription or translation). For our experiments, we use the base model[8] (consisting of 74 million parameters).

A.2 Rater Credentials

All three raters are currently enrolled in a doctoral program at the Idaho College of Osteopathic Medicine (ICOM). Experience of the members includes medical research at locations such as the Mayo Clinic and the University of Utah, work as a Spanish-English interrupter in medical clinics, work as an anesthesia technician, and holding positions such as student representative on ICOM's research committee.

A.3 Instructions given to Raters



Instructions:

First, read the correct sentence and the sentence from the ASR

Second, rate the ASR sentences on the following scale

Scale	
0	No errors or very minimal errors that would most likely have no negative implications
1	Contains errors that could potentially have negative implications
2	Contains serious errors, that either change the meaning of the sentence or gibberish, and will most likely have negative immplications

Correct	Output of ASR system	Rating
okay	propane	2
sorry yeah the pain has been there this whole time and it's gotten worse ever since it started	si yet the pad has been there this wholl time and it's gotten worse iave ever since it started ocet	2
it made it a bit worse but	it made it a big worse by ta	2
no	no	0
a multivitamin	a a multy biteman	2
when did this pain start	when do this pain start	1

Figure 4: Image showing the instructions given to raters and a few example pairs of sentences with the correct transcription on the left, the output of ASR in the middle and the human rating of severity on the right.

[4] https://deepspeech.readthedocs.io/en/latest/index.html

[5] https://huggingface.co/docs/transformers/model_doc/wav2vec2

[6] https://github.com/cmusphinx/pocketsphinx-python

[7] https://alphacephei.com/vosk/

[8] https://huggingface.co/openai/whisper-base

WER	NLTK	FLAIR	TextBlob	MiniLM	BertNLI	MPNET	DisRob
0.43	0.29	0.34	0.29	0.55	0.53	0.56	**0.59**

Table 5: Correlation between human rating of severity to WER, and severity based sentiment (NLTK, FLAIR, and TextBlob), and severity based on text embeddings (MiniLM, BertNLI, MPNET, and DisRob). Severity scores based on text embedding correlate the best with human ratings.

Ground-Truth ASR output	FLAIR	MiniLM	WER	ASR
uh i smoke about a pack a day uh smoke about a pack of day	1.929	0.104	0.250	Whis.
and how often do you use crystal meth and how often do you use crystal mud	1.858	0.371	0.125	Whis.
ok sounds like a a pretty stressful job and like a pretty stressful job	1.850	0.298	0.375	DS2
uhm it started last night and it started last night	1.707	0.138	0.200	W2V2
what they did for your heart attack what they did for your herd attack	1.617	0.546	0.143	W2V2
any previous surgeries any previous surgery	1.580	0.111	0.333	DS2
nothing has seemed to make it any... dorthins seemed to make him any...	0.003	0.692	0.364	W2V2
what they did for your heart attack what they did for your herd attack	1.617	0.546	0.143	W2V2
and how often do you use crystal meth and how often do you use for sunlight	0.010	0.512	0.250	Vosk
that you're experiencing some chest pain that you're experiencing some testing	0.049	0.469	0.333	Whis.
about the same ok and has it gotten... the same moqe and has it gotten more...	0.028	0.461	0.200	W2V2
that you're experiencing some chest pain that you're experiencing some chatting	1.889	0.456	0.333	Vosk
ok okay	1.094	0.061	1.000	DS2
a multivitamin a multi vitamin	0.000	0.150	1.000	DS2
my parents our friends	0.005	0.370	1.00	PS
i've tried uh i have tried add	1.733	0.451	1.000	Vosk
uh thirty eight degrees 38 degrees	0.161	0.177	0.750	Whis.
uh thirty eight degrees the degrees	0.007	0.324	0.750	DS2

Table 6: Examples of *severe* errors. The first 6 and second groups of 6 are based on sentiment and text embeddings respectively while WER is kept below 0.5. The last 6 are based on WER whie cosine distance of text embeddings is kept below 0.5.