Inference corpus (MultiNLI) [Williams et al., 2018]. The SNLI contains 570,000 human-written English sentences, while the MultiNLI contains 433,000 sentences but covers a more diverse range of texts. SNLI and MultiNLI are currently the most popular corpora for training automated RTE/NLI systems. Both SNLI and MultiNLI use the three-way classification format of the task.

As with PI, the work on RTE and NLI is mostly for English. Notable exceptions are the XNLI corpus [Conneau et al., 2018], a machine-translated portion of MultiNLI and the SPARTE corpus [Peñas et al., 2006] for RTE in Spanish, created from question-answering corpora.

State-of-the-art in RTE: The development of the automated RTE/NLI systems follows a similar trend as the development of the automated PI systems. The first RTE systems used manually engineered features and simple similarity metrics. Then, there was a paradigm shift towards various Deep Learning architectures, such as autoencoders, LSTMs, and CNNs. And finally, the current state-of-the-art are Transformer based architectures [6].

With the state-of-the-art systems approaching human level performance on the datasets, many researchers have tried to **analyze the workings** of the different RTE and NLI systems. Gururangan et al. [2018] discovered the presence of annotation artifacts that enable models that take into account only one of the texts (the hypothesis) to achieve 67% (SNLI) and 52.3-53.9% (MultiNLI) accuracy, which is substantially higher than the majority baselines of 34-35%. Glockner et al. [2018] showed that models trained with SNLI fail to resolve new pairs that require simple lexical substitution. For example the models have problems determining that *"holding a saxophone"* contradicts *"holding an electric guitar"*. The human annotators indicate a contradiction in this example, as the annotation guidelines instruct them to assume that the same event is referred to by both texts. Naik et al. [2018] created label-preserving adversarial examples and concluded that automated NLI models are not robust. Wallace et al. [2019] introduced universal triggers, that is, sequences of tokens that fool models when concatenated to any input.

All of these findings indicate that the existing RTE and NLI datasets are much simpler than what native speakers are capable of. Furthermore, the datasets contain many annotation artifacts and the systems trained on them are not robust to adversarial examples. Therefore, despite the high performance achieved on the datasets, the general problem of RTE and NLI is far from resolved.

[6] The official ACL page for the RTE Challenge (`https://aclweb.org/aclwiki/Recognizing_Textual_Entailment`), the official SNLI corpus page (`https://nlp.stanford.edu/projects/snli/`), the official MultiNLI corpus page (`https://www.nyu.edu/projects/bowman/multinli/`), and the GLUE benchmark page (`https://gluebenchmark.com/leaderboard`) contain the full leaderboard of RTE systems for a variety of corpora.

Semantic Textual Similarity

Task format and definition: Semantic Textual Similarity (STS) is framed as a regression task. In STS, a human or an automated system needs to determine the degree of similarity between two given texts on a continuous scale from 0 to 5. The practical definition for Semantic Similarity in STS is *"how similar two sentences are to each other according to the following scale:*
***[5] Completely equivalent**, as they mean the same thing.*
***[4] Mostly equivalent**, but some unimportant details differ.*
***[3] Roughly equivalent**, but some important information differs/missing.*
***[2] Not equivalent**, but share some details.*
***[1] Not equivalent**, but are on the same topic*
[0] On different topics.
Examples for each semantic similarity from 0 to 5 can be seen in 3.

(3) **Similarity 5:**
The bird is bathing in the sink.
Birdie is washing itself in the water basin.
Similarity 4:
In May 2010, the troops attempted to invade Kabul.
The US army invaded Kabul on May 7th last year, 2010.
Similarity 3:
John said he is considered a witness but not a suspect.
"He is not a suspect anymore." John said.
Similarity 2:
They flew out of the nest in groups.
They flew into the nest together.
Similarity 1:
The woman is playing the violin.
The young lady enjoys listening to the guitar.
Similarity 0:
John went horse back riding at dawn with a whole group of friends.
Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

STS Corpora: The most popular corpora for the STS task are the datasets from the yearly STS competition [Agirre et al., 2012]. While the STS corpora are not large in size, they come from a variety of domains and their coverage is extended every year. Unlike the tasks of PI and RTE, the competition in STS includes non-English texts (Arabic, Spanish, Turkish). Also unlike PI and RTE, at the time this dissertation was begun there were no large scale corpora explicitly designed for STS.

State-of-the-art in STS: The development of the automated STS systems is similar to that of the automated systems for PI and RTE. The system architecture transitions from feature based through Deep Learning based systems, and finally to the current state of the art, which are transformer based systems[7].

1.1.2 Typologies of Textual Meaning Relations

In the context of the empirical tasks of Paraphrase Identification (PI), Recognizing Textual Entailment (RTE), and Semantic Textual Similarity (STS), the corresponding meaning relations are typically considered atomic. That is, the researchers in these areas make several assumptions about the data and the task:

- Each pair of texts has a single label corresponding to it. The label is one of a pre-defined set.

- The label applies to the whole text pair and cannot be expressed (decomposed) as a combination of more simple phenomena.

- Each pair of texts is processed the same way by the human annotators and the automated systems. It has the same complexity as any other pair in the dataset and it contributes the same weight to the evaluation of the model.

These assumptions are made to facilitate the definition and evaluation of the empirical tasks. However, several researchers working on Paraphrasing, Textual Entailment, and Semantic Similarity have questioned the applicability of these simplifications and have provided counter examples, such as Examples 4 and 5:

(4) **Sentence 1:** All **kids** receive the same education .
Sentence 2: All **children** receive the same education .

(5) **Sentence 1:** All **kids** receive the same education .
Sentence 2: The same education is provided to all **children** .

In both Examples 4 and 5, the two texts have approximately the same meaning and they can be labeled as "paraphrases". In the context of PI, these two examples

[7] The official page for the STS challenge[8] and the GLUE benchmark page (`https://gluebenchmark.com/leaderboard`) contain the full leaderboard of STS systems for a variety of corpora.