Chapter 1

Conclusion

Current research in natural language understanding relies on the creation of representations of natural language that can be readily manipulated by computer algorithms for purposes of making inferences about meaning. This thesis has focused on one particular type of representation: linear combinations of embeddings. This is a very simple representation, closely related to a bag of words. There is a machine learning adage: that given enough data and a model with sufficiently high representational capacity any problem can be solved. However, we seem to have found a sweet spot, where a model seemingly without sufficiently high representational capacity, nevertheless performs excellently on tasks with the amount of data that we have. It seems clear that there will always exist low-medium resource settings where linear combinations of embeddings will remain an ideal method for many practical problems.

The research presented here on linear combinations of embeddings has shown that this simple input representation technique is surprisingly powerful. This power is related to the fact that surface level information plays a significant role in practically giving human understandable meaning to a natural language utterance. Word content is the most obvious surface level information, and is effectively captured by a LCOWE. The LCOWE represented this in a dense, but informative vector. While the LCOWE loses word order information, it preserves the aggregated content very well, making it very useful for the tasks considered in this research.

We considered a number of tasks to identify the utility of this representation. **??** investigated classifying paraphrases as a means to investigate the quality of SOWE as a sentence embedding method. **??** defined models for color estimation from short phrases. **??** considered if we could use weighted combinations of sense embeddings to better capture the sense used in a particular example. **??** considered taking the mean of the embeddings adjacent to named entity tokens across a fictional text as a feature to characterize how the named entity token was being used. We followed up these practical demonstrations of capacity, with further investigations into what can be recovered from the SOWE in the important area of sentence representations. **??** demonstrated a method that could partially recover bags of words from a given SOWE. **??** extended this work by attempting to order those bags of words into sentences. This demonstrated that a surprising amount of information is still available in the summed embeddings; which helps to explain why they work so well.

Linear combinations of embedding are not perfect for representing all meaning, as they do not encode any information about word order. It is thus clear that there exists sentences and phrases that are ambiguous when represented this way. However, we note that such sentences are rare: often there is only one likely ordering, particularly in any given text with a restricted domain. Most sentences are relatively short; multiple similarly likely word ordering occur more often in longer sentences. Many reorderings are paraphrases, or near paraphrases, particularly when done at the clause level. Though some orderings, such as noun swaps of nouns with similar ontological classification (e.g. Agents, Objects) do exist at almost all lengths: many are paraphrases `The banana is next to the orange` *vs.* `The orange is next to the banana`; and others are similar in meaning: `The banana is to the left of`

`the orange` *vs.* `The orange is to the left of the banana`. It is desirable that such sentences are nearby in a representational of the semantic space.

1.1 Future work

1.1.1 Adversarial Test cases

A limitation of the LCOWE representations is that they have no ability to represent word order. This is in-contrast to RNNs and other commonly used neural network based representations of multi-word natural language input. It is possibly to construct adversarial test cases, that no LCOWE can succeed on. This can be done by selecting sentences with multiple reasonable word orders with very different meanings. It is worth consideration, that such adversarial test cases allow advancement of the state of the art to increase the capacity of models to represent all possible inputs. However, they do not necessarily advance the practical state of the art in representing real inputs that occur in a particular domain. Thus it is is essential to understand how common such adversarial test cases are in practice.

Future work in this area requires not just the construction of adversarial examples; but of the determination of how common they are in practice. Adversarial examples are not ubiquitous in real world tasks. It is important not to succeed on only these cases, while failing on the more common simple cases.

It is also important to consider how challenging an adversarial test case is. In **??**, the ordered task which was to make predictions for colors for which the different words in the name could appear in different orders to describe different colors. For example `bluish green` and `greenish blue` are different colors. However, they are very *similar* colors. As such the error from discarding word order, is less than the error from using a more complicated model such as an RNN. Such a more complex model is harder to train, and those practical difficulties can dominate over a small amount of theoretical lack of capacity.

1.1.2 Language Models and Orderless Representations

There is a complementary aspect to LCOWE and language models. While LCOWE have no capacity to handle word order, but they have an excellent ability to capture word content; whereas pure language models have no ability to capture word content, but have an excellent ability to capture word order. Language modelling based models incorporating a representation stage, such as encoder-decoders (Cho et al. 2014), do not capture word content as well as LCOWE (Conneau et al. 2018). They do, however, have state of the art order representation.

An interesting combination of the two, would be an encoder model, where the coding layer, is augmented by concatenating the final RNN output, with a sum of word embeddings, for all the input words to the encoder. An example of this for a encoder-decoder is shown in Figure 1.1. This would effectively allow a bypass of the encoder RNN. A similar bypass of intermediate layers has been used in feed-forward networks including the notable neural probabilistic language model (Bengio et al. 2003). The significant advantage of bypassing the RNN encoder is that it allows the model to weight the value of the orderful representation of the RNN output, against the unordered representation of the SOWE and learn use which ever is better for the task. Further, having explicit access to the surface level features in the SOWE, should help encourage the orderful encoder to learn more important deeper features.

A coding layer featuring components from a encoder capturing order-features, and a SOWE capturing surface features can be expected to perform better at both representations than either alone. This expectation is due to the the weighting above the shared layer will train to weight each feature for what it is better at, and thus during gradient decent the weights for the encoder would be decreased for surface information that is better obtained from the SOWE. It would thus allow each part of

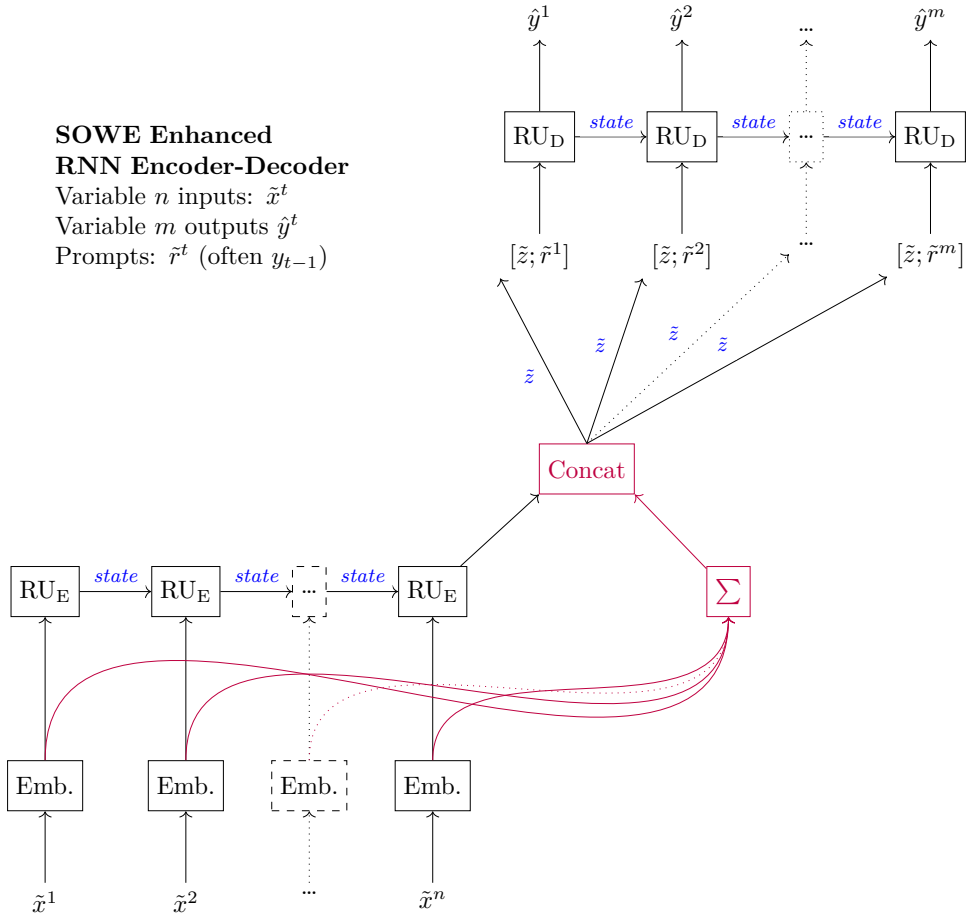

Figure 1.1: An encoder-decoder model with a SOWE encoder bypass layer added (shown in red).

the network to *focus* on what it is best at, thus creating better representations./ This thesis has shown that SOWE can excel at surface level tasks (and that more tasks that expected are surface level). On deeper tasks where structure becomes more important ordered representations out perform it (Conneau et al. 2018). By combining the two we expect to get the best of both worlds, and produce truly excellent models for natural language understanding.