

Part I

Conclusion

Chapter 1

Conclusion

Current research in natural language understanding relies on creating computer manipulatable representations of natural language for purposes of making inferences about meaning.

While, the normal machine learning adage that given enough data and a model with sufficiently high representational capacity any problem can be solved always applies, 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.

The research presented here on linear combinations of embeddings has shown that this simple input representation technique is surprisingly powerful. This surprising power comes from the value of surface level information in many practical natural language understanding tasks. The word content being the most obvious of these. This word content is presented in a dense and informative form in a LCOWE; in a way that captures and preserves lexical similarity information. While the LCOWE loses word order information, it very well preserves the word content. That content proved more useful for the tasks considered in this research.

We considered a number of tasks to identify the utility of this representation in for different uses. ?? investigated classifying paraphrases as means to investigate 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 for the important category of sentence representations. ?? demonstrated a method that could partially recover bags of words. ?? extended this to attempt 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.

1.1 Future work

1.1.1 Adversarial Test cases

It is worth consideration, that adversarial test cases allow advancement of the state of the art to increase the capacity of models to represent all possible inputs. However, understanding how common they are is essential.

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 adversarial such a challenging 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.

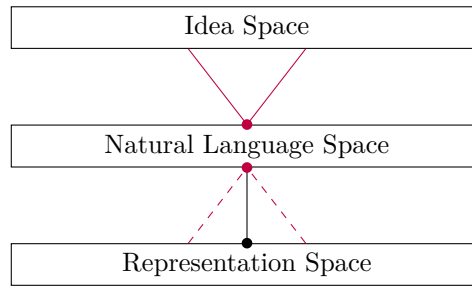


Figure 1.1: The representation space is a computationally manipulate representation of the idea space. The natural language utterances come from points in the idea space; though due to ambiguity we can only truly hope to estimate distributions when the interpret them. A single point embedding as an approximation to a distribution with a single tight peak.

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 excellent ability to capture word content; Pure language models have no ability to capture word content, but excellent ability to capture word order. Language modelling based tasks 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 order representation.

1.3 Some reflections upon semantic spaces

We can consider that there is a true semantic space of ideas. When speaking, this space is projected down to natural languages space. To again quote Webber: “A sentence is a group of words expressing a complete thought.”, it is not a complete thought, only the *expression* of one. This projection from idea to utterance is imperfect – it is lossy. Many ideas are expressed the same way, and language thus has a lot of ambiguity. When we try to understand the meaning of a natural language utterance we are trying to find the point in idea space that the speaker intend. Some times the natural language space alone, is enough to recover a good idea of the the point in idea space the speaker intends, but other times it is not.

The preimage of a point in natural language space (e.g. a sentence), is a probability distribution over idea space that could have lead to that utterance – $P(\text{meaning} \mid \text{utterance})$. This distribution could be combined with other factors (in a Bayesian way); either from that natural language context, or the enviroment more broadly. For example, to use a meaning that centres around word sense: we can identify two (of the many) senses of the word **apples**: one in reference to the fruit, the other in reference to the computers made by the eponymous company. Thus, on its own the sentence **Apples are good.** suggests a distribution with at least two peaks in idea space. Combine that utterance, with the context of being in a computer store, rather than a grocer, and the probability of one peak can be increased, though the other not entirely removed. Further around each peak remains adjacent closely related possible meanings. For example the statement could be in relation to only computers, or also to other products. idea space is a continuous space, with every utterance corresponding to a unique point. It is an uncountably large space. In contrast natural language space is countably large, being composed of finite length combinations of of symbols taken from a finite alphabet. An uncountable number of points in idea space are projected down to a single point in natural language space.

When designing a embedding method, for sentences, words or other structures, we seek to define a representation space that has good In particular it should have a continuous mapping from to and from embedding space. A neighbourhood in representation space, should correspond to a neighbourhood in idea space. ?? investigated this for sentence embeddings. By taking points in natural language space known to come from very near by points in idea space, that is to say paraphrases, and checking that they belong to near points in embedding space.

As each point in natural language space is defines a distribution over idea space of what may be meant; and representation space is attempting to be in correspondence to idea space; it is such that each point in natural language space should project to a distribution over embedding space. Instead, most methods project to natural language points to single points in embeddings space. This is viable when the region in idea space that the natural language point could have come from

I would really like to talk about how word embeddings are useful for caption evaluation here.

is small – in particular that the distribution in idea space is narrow variance and mono-modal. In that case the single point estimate in embedding space is an useful approximation.

This has particularly clear utility for word sense embeddings, which are defined by multimodal distributions, with large peaks for each homonym, and smaller nearby peaks for each polyseme. Furthermore we can't rule out the speaker using the word incorrectly or metaphorically which gives rise to nonzero values elsewhere in idea space. Word-sense embeddings produce multiple sense embeddings – ideally one corresponding to each peak in idea space. We know these peaks are only rough approximations to the true point in idea space for a given usage of a word. ?? attempts find other points in the embeddings space, that better corresponds to the true point in idea space for the particular use.

The color understanding task considered in ?? is interesting. While it is a typical natural language understanding system, which takes a point in natural language space (a color name), moves through a representation space (the output of one of the input modules: SOWE, CNN, or RNN) using supervision to output something from meaning space. In this case the meaning space is explicit, it is the HSV color space. Using point estimation it outputs a point in meaning space.

Unsupervised methods, in particular word embeddings, but also more generally, are ungrounded. They are based only on natural language space observations. The goal is not to capture meaning in this space, but rather to create a space that is a good input to a supervised system that can learn a good correspondence from natural language space to idea space. It is for this reason that principles such as Firth's distributional hypothesis, that words occurring in similar contexts have similar meaning, are so useful as a basis. While it does not allow the encoding of meaning, it allows thus encoding of similarity of meaning. This is ideally suited for creating a space that will make a good source representation for a supervised method applied for natural language understanding task. Were that task accomplished with a neural network, the later hidden layers, or the fine-tuned embeddings would form a grounded representation of idea space.

While the research presented in this dissertation has made use of the idea that we are working with a sample from a distribution over a proxy for meaning space, it is our belief that further advancements would benefit from fully considering word embeddings and other objects from representation spaces, not as discrete points but as distributions.

Bibliography

- Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (Oct. 2014). “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734.
- Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni (2018). “What you can cram into a single $\&! \#^*$ vector: Probing sentence embeddings for linguistic properties”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136.