# 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, never-the-less 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 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 it 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 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 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.

While it is clear that such linear combinations of embedding representation spaces are not perfect for representing all meanings, they are pragmatically very capable. They do not encode any information about word order. It is thus clear that there exist 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 a 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

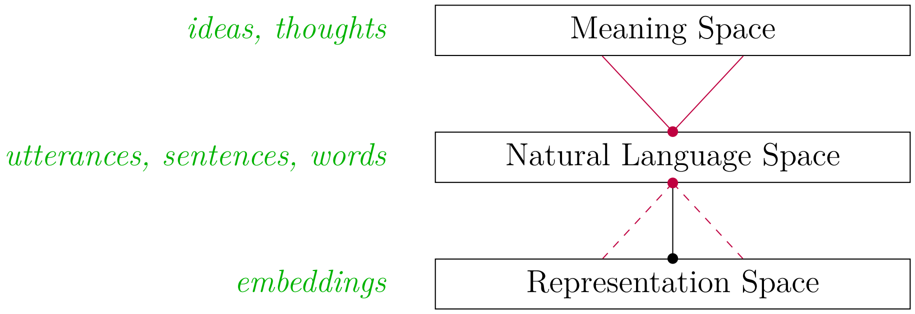

Figure 1.1: The representation space is a computationally manipulate representation of the meaning space. The natural language utterances come from points in the meaning space; though due to ambiguity we can only truly hope to estimate distributions when the interpret them. A single point embedding as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

that such sentences are nearby in a representational of semantic space.

## 1.1 Some reflections upon semantic spaces

We can consider that there is a true semantic space of ideas: a meaning space. When speaking, this space is projected down to natural languages space. Which we represent using an embedding in representation space, with the hope that this representation can be related to the meaning space. This is shown in the diagram in Figure 1.1.

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 meaning space that the speaker intends. Some times the natural language space alone is enough to recover a good idea of the the point in meaning 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 meaning space that could have lead to that utterance – $P(meaning \mid utterance)$. This distribution could be combined with other factors (in a Bayesian way); either from that natural language context, or the environment 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 meaning 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. Meaning 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 symbols taken from a finite alphabet. An uncountable number of points in meaning 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 properties for reflection relationships in the meaning space in a way that computationally manipulatable using simple operations (like sums). 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 meaning space. **??** investigated this for sentence embeddings. By taking points in natural language space known to come from very nearby points in meaning space, that is to say paraphrases, and checking that they belong to near points in embedding space.

As each point in natural language space defines a distribution over meaning space of what may be meant; and representation space is attempting to be in correspondence to meaning space; it is such that each point in natural language space should project to a distribution over embedding space. Instead, most methods project natural language points to single points in embeddings space. This is viable when the region in meaning space that the natural language point could have come from is small – in particular when the distribution in meaning space has narrow variance and is 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 meaning space. Word-sense embeddings produce multiple sense embeddings – ideally one corresponding to each peak in meaning space. We know these peaks are only rough approximations to the true point in meaning space for a given usage of a word. **??** attempts find other points in the embeddings space, that better corresponds to the true point in meaning space for the particular use.

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 meaning space. While we would not normally think of the SOWE sentence representation space as one for which there would be an easy alignment to the meaning space, **??** showed that it was. A strong point in its favour is that it directly benefits from word embeddings. While themselves ungrounded, word embeddings are excellently suited for creating a representation space, as they have an internal consistency which makes it easy to apply supervision to give grounded meaning representations. It's great strength comes from Firth's distributional hypothesis, that words occurring in similar contexts have similar meaning. While this does not allow the encoding of meaning itself, it does allow the 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 on words. Were that task accomplished with a neural network, the later hidden layers, or the fine-turned embeddings would form a grounded representation of meaning space. Our results show that that strength is carried forth into linear combinations of such embeddings.

The color understanding task considered in **??** is interesting. 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. Notably however, the meaning space is *very well grounded* to the HSV color space. We can, for many purposes, say for this natural language understanding task, the color space *is* the meaning space. Using point estimation it outputs a point in meaning space, reflecting (in some sense) the most reasonable guess of the meaning. Using distribution estimation it outputs a distribution over the meaning space, fully reflecting the knowledge we have to infer the meaning. The fact that even on the subset of the testing data where word order was ambiguous, SOWE was the best performing model highlights an important notion. Word order ambiguity is just one amongst many sources of ambiguity in any representation of natural language. In the color case, it boils down to the additional ambiguity of being unable to encode the word order difference between `bluish green` and `greenish blue` being negligible compared to the inherent ambiguity in the meaning of either. Both phrases give rise to a large and overlapping distribution across meaning space.

In cases where there are multiple reasonable word orderings, this means that multiple points in the true meaning space, correspond to a single point in the representational LCOWE space. However, this is not exceptional: many sentences have two or more interpretations, a humorous example being an accidental pun. Thus even in a space that fully captures the natural language features, a single point in that representational space corresponds to two points in meaning space; as the single point in natural language space could have come from either point in meaning space. As such, ambiguity from loss of word order is not an unique and unsalvageable problem. If we thus had a distribution over meaning space, corresponding to the interpretation of a SOWE, it would have two peaks corresponding to two different word orders. While such a discussion is purely theoretical as we do not have any way to generate such a distribution over true meaning space, it remains interesting for cases where have a space that we can treat as being the meaning space (e.g. the HSV space for colors). As we can use other contextual information define prior and thus decrease distributions associated with other ambiguities, we can use language models to provide a prior over those peaks; based on the likelihood of word orders. There exists a trivial extension of the work presented in ????, where the mixed integer programming model is constrained to give the second (and so forth) most likely solution, together with it's probability. It is however not computationally practical, nor useful without a better meaning space representation.

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 random variables with a linked distribution. This however comes with significant challenges in that manipulating the very high dimensional distributions that are required for this is computationally challenging

ere does
para-
ph belong?

?? and ?? both consider represent contagious linguistic structures – sentences and shore phrases respectively, as input representations. Further, the output of the output modules discussed in ?? is a grounded representation space, though that work did not examine it directly. ?? considers the representation of word senses, and it navigates this representation space to find new representations which better describe a particular use of a word. ?? is more atypical: the features sets considered for the point of view detection task, characterize how a particular named entity token was used throughout a chapter. It is thus not a representation of the chapter meaning, as it varies for the different named entities. It is in fact a representation of how that named entity is related to the events described in the chapter. It seems like the MOWE feature set used in ?? it is vastly insufficient to represent such information given it is only a mean of the immediately adjacent words – nothing like the whole chapter's contents. However, as was demonstrated by the classical feature set considered in the same chapter, which primarily consisted of adjacent part of speech tag counts, and which also achieved very good results, surface information encodes a surprising amount of semantic depth.

## 1.2 Future work

### 1.2.1 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 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.

## 1.3 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.

I would rea
like to talk
about how
word embe
dings are u
ful for capt
evaluation
here.