Mikolov, W.-T. Yih, and G. Zweig, “Linguistic regularities incontinuous space word representations,” pp. 9–14, 2013

The main finding of the paper was the improved syntactical and semantic regularity capturing that recurrent neural networks offer over previously used methods. The authors claim that the main advantage of using distributed representations (word embeddings), over n-gram models, is its ability to generalise and as such carry over improvements to similar words.

When referring to regularities between words, syntax refers to the practical (visual) connections. For example, the relationship between the singular and plural from of a noun: ‘apple’ to ‘apples’. RNNs can better recognise these granular regularities than n-grams because the latter is limited to a representation of words, whereas RNNs enable context over slices of characters. In comparison, semantics refers to the ‘behind the scenes’ (non-visual) connections via meaning. For example, the relationship between the class inclusion to singular collection: ‘clothing’ to ‘shirt’. The recurrent element of RNNs is what benefits capturing of these regularities, the residual factor of neutron outputs allows context to be represented within embeddings.

Representing words as vectors allows the use of standard mathematical operations, for example finding the offset between two embeddings can provide abstract theoretical regularities between words. Where the offset between embeddings of ‘apple’ and ‘apples’ is likely to be very similar to the offset between ‘car’ and ‘cars’. In this case, the offset is essentially a representation of plurality. When there is existing word for a predicted word embedding, use of cosine similarity allows one to find the most similar embedding for which there is a linguistic representation.