Subcategorising DisCoCat for Creative Light Verb Construction Generation

Lin de Huybrecht
Artificial Intelligence Lab
Vrije Universiteit Brussel, Belgium
lin.de.huybrecht@vub.be

Abstract

The DisCoCat (Distributional Compositional Categorical) framework models language meaning in terms of both grammar and semantics, thereby offering the transparency required to model human cognition, which neural methods are currently lacking. Furthermore, DisCoCat incorporates the information flow between words regardless of their position in a sentence. Additionally, it offers the ability to compare the meaning of sentences with a different grammatical structure, which is often impossible in other compositional distributional models. DisCoCat has been mainly used for modelling the semantics of sentences, but not for creative natural language generation (CNLG). In this project, we aim at improving the state of the art in NLG techniques which are grounded in cognitive and linguistic theory. Moreover, we focus on generating and interpreting light verb constructions using an instantiation of the DisCoCat framework.

Introduction

Whilst tools that use statistical methods, like ChatGPT[1] and GitHub Copilot[2] are capable of seemingly spectacular tasks, they come with significant disadvantages (Maruyama 2022). Large Language Models and other statistical methods for natural language processing require a huge quantity of training data and an extensive amount of computing power. Furthermore, they tend to lack transparency in multiple ways. First, their internal representations and mechanisms are often unclear. Second, the systems are often closed in terms of availability of the source code, provenance of their training data and extensive documentation. This lack of transparency often makes it impossible to explain or verify the output of the systems. However, as will be discussed later on in this proposal, transparency is key when one wants to asses the creativity of such a computational system. Maruyama (2022) argues that the aforementioned problems could be resolved with *categorical AI*, which integrates statistical AI with symbolic AI. The latter is precisely what Coecke, Sadrzadeh, and Clark (2010) proposed with the DisCoCat framework (Distributional Compositional Categorical framework). The overall goal of DisCoCat is to represent sentences in such a way that they can be compared together, even if they have a different grammatical structure, whilst taking into account both their surface form and grammar. DisCoCat has been shown to exceed the performance of other NLP approaches (Grefenstette and Sadrzadeh 2011; Kartsaklis and Sadrzadeh 2013). In this project, we will improve the state of the art in creative natural language generation, expanding the Distributional Compositional Categorical Framework for creative natural language generation. We will apply this to the use case of *light verb constructions* (Figure 1). A light verb construction is a construction wherein the verb does not carry most of the meaning (e.g., "Alice did a revision of her paper."). Its meaning is "lightend", i.e., made less heavy than in a full verb construction (e.g., " Alice revised her paper.").

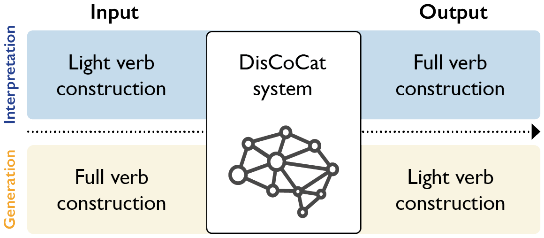

Figure 1: Diagram of this project's system.

Background

In distributional semantics, it is assumed that the meaning of an expression can be derived from the meaning of its constituents using syntax rules. Hence, word meanings are obtained by extracting co-occurrence counts from text corpora using an n-word co-occurrence window. However, the meanings that result from it do not account for connections between words that fall outside of the n-word window. For example, when using a five-word window in the text "Papers that I read which were interesting", there will be no link between *papers* and *interesting*, even though the text could be rephrased as "I read interesting papers". Hence the approximation to a semantic space given by the co-occurrence statistics does not suffice to accurately model the order and combination of words in real language. One should also use syntactical properties of text. This results in *compositional* distributional semantics (Clark and Pulman 2007). Here, the goal is to combine distributional word vectors into sentence vectors so that sentence vectors can be used to com-

[1] https://openai.com/blog/chatgpt/

[2] https://github.com/features/copilot

pare sentences in the same way as word vectors. Several methods for computing these sentence vectors have been proposed, e.g., addition and multiplication (Mitchell and Lapata 2008), the tensor product (Clark and Pulman 2007) and circular convolution (Plate 1991). Coecke et al. (2020) proposed the DisCoCat (Distributional Compositional Categorical) framework, which uses category theory (Eilenberg and MacLane 1945) to represent both grammar and semantics and the relation between them. In this framework, sentences are modelled by using one category for the semantics of language (e.g., *FVect*, the category of finite-dimensional vector spaces) and one for the grammar (e.g., a pregroup, *Preg*, or any other categorial grammar). The meanings contained in the semantics space are obtained using distributional methods. The types of the categorial grammar for the grammar space can be obtained using a part-of-speech-tagger. The key here is that the two categories share a common structure (FVect and Preg are both compact closed). This will ensure that there exists a mapping from the grammar category to the semantics category so that we end up with a new category, i.e., the combination of the grammar and semantics categories. This results in a truly compositional model.

What does DisCoCat offer that most large language models do not? First of all, it is transparent and meaning-aware (Coecke et al. 2022). Once the grammar and semantics categories are instantiated, every step in the process of constructing meaning representations is interpretable, which is not the case in most deep learning methods. The relations between word meanings are explicitly modelled in DisCoCat (Coecke 2017) and the grammaticality of sentences is verified via type-reductions. Furthermore, DisCoCat has been demonstrated to outperform non-compositional models and n-gram models (Grefenstette 2013) on word-sense disambiguation tasks and verb disambiguation tasks (Wijnholds 2020).

Research Objectives

The overall objective of this research project is to study how DisCoCat can be leveraged for CNLG. More specifically, we are interested in studying light verb constructions from a semantics perspective.

Critical analysis of DisCoCat meaning models We aim at establishing the theoretical foundations for extending DisCoCat for creative NLG. We conduct a thorough study of existing instantiations of the DisCoCat framework in terms of their properties and affordances.

Subtyping methods for light verb construction in DisCoCat Categorising language at the level of nouns, verbs and adjectives is insufficient for NLG due to two distinct aspects of language: syntax and semantics. For example, "the table thinks" is syntactically correct, but not semantically. Therefore, we need a more detailed subcategorisation. In this research project, we will study a more specific sub-problem, i.e., that of light verb constructions. In this objective, we aim at formalising the computational mechanisms needed to build an AI system that uses spectral knowledge representation to represent the aforementioned subcategorisations. This will allow for transparent reasoning about language semantics.

Generating Novel and Creative Light Verb constructions using DisCoCat We aim at building novel generation methods using the new representations obtained in the previous objective. We build a system (Figure 1) that is able to map full verb constructions to light verb constructions (*light verb construction generation*) and vice versa (*light verb construction interpretation*). Additionally, the system will be extended to create new light verb constructions given some input which uses full verb constructions.

Evaluation of artefacts created by the system We aim at demonstrating that the outputs of our work constitute an improvement on the state of the art in Computational Creativity Theory, and/or evaluates why they fail to do so.

Methodology

Synthetic literature survey of relevant domains This project differs from many doctoral projects in that it has a strong emphasis on combining different ideas from multiple research fields: CC, AI, NLG, quantum mathematics (QM), (cognitive) linguistics, etc. Therefore, we explicitly incorporate a work package on conducting a literature survey of these areas into the work plan.

Extending the DisCoCat model for creative NLG DisCoCat has been primarily used for modelling the semantics of linguistic expressions. We will analyse existing instantiations of the DisCoCat framework in order to extend it for CNLG. Additionally, we design the representations for the subcategorisation that enables the generation of language that is meaningful in terms of syntax and semantics.

Creative text generation with the semantic meaning model We replicate the software of (Wijnholds 2020) to gain a deeper understanding of his methodology. We will use this implementation as a basis for expanding with subcategorisations.

Evaluating the system and its artefacts in the context of Computational Creativity Theory We evaluate the internal meaning representations and the generative process. For the latter, specify new value measures for creativity in the context of creating in the light verb construction domain.

Dissemination and writing the PhD thesis We gather external comments and feedback by means of conferences, journal reviews and publications. Finally, we write the PhD thesis based on our published and submitted work.

Expected results

- An instantiation of the DisCoCat framework that features subcategorisations of grammatical types at a level that enables natural language generation such that the output makes sense on both a grammatical and semantic level.
- A new method for CNLG applied to the generation of light verb constructions.
- A thorough evaluation of (the output of) the system using techniques from the CC literature.

References

Clark, S., and Pulman, S. 2007. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, 52–55.

Coecke, B.; de Felice, G.; Meichanetzidis, K.; and Toumi, A. 2020. Foundations for Near-Term Quantum Natural Language Processing.

Coecke, B.; Felice, G. d.; Meichanetzidis, K.; and Toumi, A. 2022. How to Make Qubits Speak. In *Quantum Computing in the Arts and Humanities: An Introduction to Core Concepts, Theory and Applications*. Springer International Publishing. 277–297.

Coecke, B.; Sadrzadeh, M.; and Clark, S. J. 2010. Mathematical foundations for a compositional distributional model of meaning. 36(1-4):345–384.

Coecke, B. 2017. From Quantum Foundations via Natural Language Meaning to a Theory of Everything. In *The Incomputable: Journeys Beyond the Turing Barrier*, Theory and Applications of Computability. Springer International Publishing. 63–80.

Eilenberg, S., and MacLane, S. 1945. General theory of natural equivalences. 58:231–294.

Grefenstette, E., and Sadrzadeh, M. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 1394–1404. ACL.

Grefenstette, E. 2013. Category-Theoretic Quantitative Compositional Distributional Models of Natural Language Semantics.

Kartsaklis, D., and Sadrzadeh, M. 2013. Prior Disambiguation of Word Tensors for Constructing Sentence Vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1590–1601. ACL.

Maruyama, Y. 2022. Categorical Artificial Intelligence: The Integration of Symbolic and Statistical AI for Verifiable, Ethical, and Trustworthy AI. In *Artificial General Intelligence*, Lecture Notes in Computer Science, 127–138. Springer International Publishing.

Mitchell, J., and Lapata, M. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, 236–244.

Plate, T. 1991. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'91, 30–35. Morgan Kaufmann Publishers Inc.

Wijnholds, G. 2020. A Compositional Vector Space Model of Ellipsis and Anaphora.