S2 Distributional Representations
Summary Report

Group 4

April 2025

Statement

This report has been compiled based on collective notes from our group discussion sessions.

1 Vector space models

1.1 Notion of Meaning

Meaning Representation

Distributional models represent meaning through vector spaces, where words are mapped to points based on their co-occurrence with other words in large corpora. This approach reflects a Wittgensteinian view of meaning, where meaning is rooted in usage—how words are employed in context. The vector space representation helps organize words into clusters, grouping together words that share similar contextual relationships. For example, words like ”car,” ”sedan,” and ”van” might be grouped because they are commonly used in similar contexts.

Semantic Relations

These models primarily capture **semantic similarity** (e.g., ”car” is more similar to ”sedan” than to ”dog”) and **semantic relatedness** (e.g., ”bee” is related to ”honey”). These relationships arise because words that co-occur in similar contexts are assumed to share meaning. However, distributional models struggle with more fine-grained relations such as **synonymy**, **antonymy**, and **hyponymy**. For example, distributional models may place antonyms like ”hot” and ”cold” close together because they appear in similar syntactic structures or topic contexts, despite their opposite meanings. Humans intuitively recognize word relations not just through context but also through **general knowledge** and **logic**—hot is the opposite of cold, whereas distributional representations only see them as words occurring in similar contexts.

1.2 Differences with model-theoretic

Based on Frege's theory of **reference and sense**, the reference is its truth value, while its sense is the thought that is expressed. *However*, as Lenci & Sahlgren(2023, pp.23-24) point out, the idea of sense and reference is as different as possible from the notion of distributional representation that is defining lexical meaning through **contextual use, co-occurrence patterns, and statistical regularities** in language data. Also, the logical representation is afforded by whatever system.

1.3 Pros and Cons of Distributional Representations

We discuss a lot on the benefits, challenges and limitations for representing meaning of natural language using distributional representation.

1.3.1 Benefits

Flexibility in meaning: Distributional models allow for more flexibility in terms of meaning, capturing semantic similarity and relatedness, beyond just truth or falsehood. This contrasts with a one-hot encoding or localist representation, where the relationship between words is not captured. *For example*, searching for "Gothenburg motels" should bring up results related to "Gothenburg hotels," despite not explicitly using the word "hotel."

Less work for linguists: The empirical nature of distributional representations reduces the amount of manual work needed from linguists.

Scalability: These models scale well as they rely on vast amounts of data and can grow with the available resources.

1.3.2 Challenges

Definition of context: One major challenge is defining what context actually means. Is context the whole sentence, or just the words immediately to the left and right of a word? Enriched context (such as using longer windows or syntactic dependencies) has been shown to improve results, but it also increases the risk of data sparsity.

Domain adaptation: Another challenge is that the same models may not work across all domains. There is a need for adaptation, as a model trained on one domain may not perform well in another due to the differences in the contextual use of language.

Data and energy limitations: While distributional models can scale, they are limited by the availability of training data and computational energy. This can create bottlenecks when processing large datasets.

Data quality: It's not just the quantity of data that is limited but also its quality, especially when trying to capture nuances like emotions, tone, or laughter. This is complicated further because the data often comes from internet interactions, which may not fully represent real-world language use.

Real-world language use: The challenge lies in capturing true real-world language, especially when emotions, tone, or non-verbal cues like gaze or attention are involved. These subtleties may be missed or inadequately captured in a corpus.

1.3.3 Limitations and Dangers

Data bias: Distributional semantics is based on large corpora, and if the data is biased or does not contain a diverse range of examples (e.g., across genders or social classes), models might learn biased associations. For example, the association of ”doctor” with ”man” and ”nurse” with ”woman” may emerge, reinforcing stereotypes.

Surface-level understanding: A fundamental limitation is that meaning does not come from form alone. Distributional models can learn patterns of word co-occurrence, but these patterns are not grounded in real-world experience. As highlighted by E. M. Bender and A. Koller, the idea of ”meaning is use” refers to language usage in real contexts. While machines may learn statistical patterns, they do not truly understand meaning—they only capture surface-level patterns.

Corpora limitations: The corpora used for training distributional models struggle with grounding, meaning they cannot fully represent the complexities of real-world interactions (such as tone, emotional cues, etc.).

Inadequate capture of nuanced expressions: Emotional judgments, laughter, or other subtle human expressions are often missing or poorly represented in the corpora, limiting the ability of models to understand these aspects.

1.4 Computational Resources and Tools

The creation of distributional representations requires large datasets, such as **Wikipedia** or **news corpora**, and tools like **Word2Vec**, **GloVe**, and **BERT**. These models rely on co-occurrence matrices, neural embeddings, and dimension reduction techniques to build meaningful word representations.

1.5 Scope of application

Suitable tasks: Textual semantic similarity, paraphrase detection, machine translation, question answering. These tasks benefit from distributional representations since they rely on understanding the semantic relationship between words or sentences. **Information retrieval, sentiment anal-**

ysis, Named Entity Recognition (NER): These tasks can be efficiently addressed using distributional models because they focus on word relationships and patterns of co-occurrence.

Unsuitable tasks: **Domain adaptation:** Distributional representations struggle in tasks where the domain varies significantly, as models may fail to adapt to new domains. **Real-world applications:** The models are challenged in real-world scenarios, such as robot control, image captioning, and dialogue. These tasks require a deeper understanding of the world beyond just word co-occurrence patterns. **Discourse-level phenomena:** Complex aspects like irony or deeply cultural references are not captured effectively, as these require context beyond co-occurrence patterns.

2 Compositionality

2.1 Reasons and Benefits

2.1.1 Reasons

Avoid Being Left Behind by Machines: Helps ensure that we are not left behind as machine learning technologies progress. This hybrid approach allows for leveraging the strengths of both types of representations.

Productivity in NLP: Boost productivity in natural language processing. This refers to the ability to generate new words or meanings over time, or adapt the same or similar words to different meanings depending on the context or time period.

2.1.2 Benefits

Pre- and Postprocessing: The integration allows for better pre- and postprocessing. One can either work with categories to improve distributional results or use distributional data to enhance a rule-based system, leading to better outcomes in various tasks.

Better Results for Rule-based Systems: By using distributional results, formal representations can improve the performance of rule-based systems, allowing them to capture more nuanced patterns in language use.

Security in Robotics and Applications: This combination is particularly useful for robotics and applications where security is a concern. Code generated by large language models (LLMs) might sometimes introduce security risks, especially through erroneous library imports. The hybrid approach helps mitigate such risks by combining formal checks with distributional insights.

2.2 Challenges of Hybrid Models

2.2.1 General Challenges

Division of Labor: Replacing constants with vectors (e.g., S = VP [1,2,3,4] NP [3,6,1,2]) might not be straightforward, especially in terms of how it works for both pre-processing and post-processing.
No Universal Solution: There isn't a one-size-fits-all solution for hybrid models, as it assumes we are always striving for perfection in representing the real world. The more specific challenge depends on the intended use of the system.

2.2.2 Specific Technical Challenges

Vector Addition: Vector addition does not account for syntactic structures. This leads to increased ambiguity in linguistic expressions, especially when the length of the sequence increases, as the vectors may carry more information that is not effectively captured.
Vector Multiplication: When dealing with long sequences, vector multiplication may reduce information significantly, particularly if many of the vectors contain zeros. This can lead to a loss of meaningful data.
Tensor Representation: Linguistic expressions of different lengths cannot be directly compared since they correspond to vectors with different dimensions, creating a challenge in ensuring consistency and comparison.
Size and Complexity: While circular convolution has been proposed as a solution for modeling linguistic data, it has its own risks. When compressing the tensor, it relies on components that are randomly distributed. However, in linguistic data, vectors typically have non-random structures, so this randomness may not suit all linguistic applications.

Integrating these two models presents several difficulties, particularly in defining how to combine their structures. Formal models often fail to handle the flexibility of natural language, while distributional models struggle with precision. The challenge is to create systems that leverage both strengths without overwhelming the model with unnecessary complexity.

2.3 Interpretability

2.3.1 Degree of Interpretability

Sparse Vector Representation: Sparse vector representations, such as those based on the frequency of words, are somewhat interpretable. These representations can be analyzed by inspecting the raw count data, giving insights into the relationship between terms based on frequency.

Neural Networks: Once we move beyond the first layer of neural networks, the interpretability of the representations significantly decreases. The deeper layers of the network become less transparent, making it difficult to understand how the network is processing and transforming the data.

2.3.2 Mapping Between Two Types of Representations

Combining formal and distributional representations (i.e., hybrid approaches) can increase interpretability. By integrating rule-based or symbolic representations with distributional models, we can gain a better understanding of the transformations and connections between different types of data, improving the transparency and interpretability of the system.

3 Future Directions in Language Representation

We also discuss the potential for **dynamic representations** that combine distributional representations with additional context, such as **visual**, **non-verbal**, or **social cues**. These models could offer a more comprehensive view of meaning by incorporating more diverse aspects of human communication. For example, in **robotics** or **image captioning**, the meaning of a word might depend not only on its linguistic co-occurrence but also on the visual or situational context in which it is used.

4 Conclusion

The seminar highlighted the powerful capabilities of distributional models in capturing meaning from context, but also acknowledged their significant limitations, particularly in handling nuanced semantic relations and grounding in real-world meaning. The combination of formal linguistic models and distributional approaches offers a promising direction for enhancing language understanding, allowing for greater interpretability and flexibility. However, challenges such as data bias, the need for grounding, and the complexity of combining these models remain important considerations for future research.

The future of NLP may lie in dynamic, hybrid systems that combine formal precision with the flexibility of distributional representations, offering a richer and more nuanced understanding of language.

5 Contribution Description

Denis: contributed to most of the discussion, particularly on the different factors about representing the meaning of natural language using distributional models. Including the reasons and benefits for combining formal representations with distributional ones.
Cristina: contributed to most of the discussion, focusing mainly on how meaning is represented by distributional models and how these affect the representation of natural language meaning. Provided extension on the biggest challenges

of hybrid models.
Sharon: contributed to discussions, mainly about the representation of meaning through distributional models, benefits of representing natural language meaning through distributional models.
Huoyuan: contributed to some discussions about the limitations and benefits of representations. Provided list of computational resources, tools, and methods used to create these representations. Responsible for writing the document.