

A Corpora Based Toy Model for DisCoCat

Stefano Gorgioso

University of Oxford

The Plan

What we assume:

- (i) an abstract corpus, as a set/sequence of sentences
- (ii) each sentence annotated with a constituent structure tree
 - we consider context-free grammars à la Chomsky.

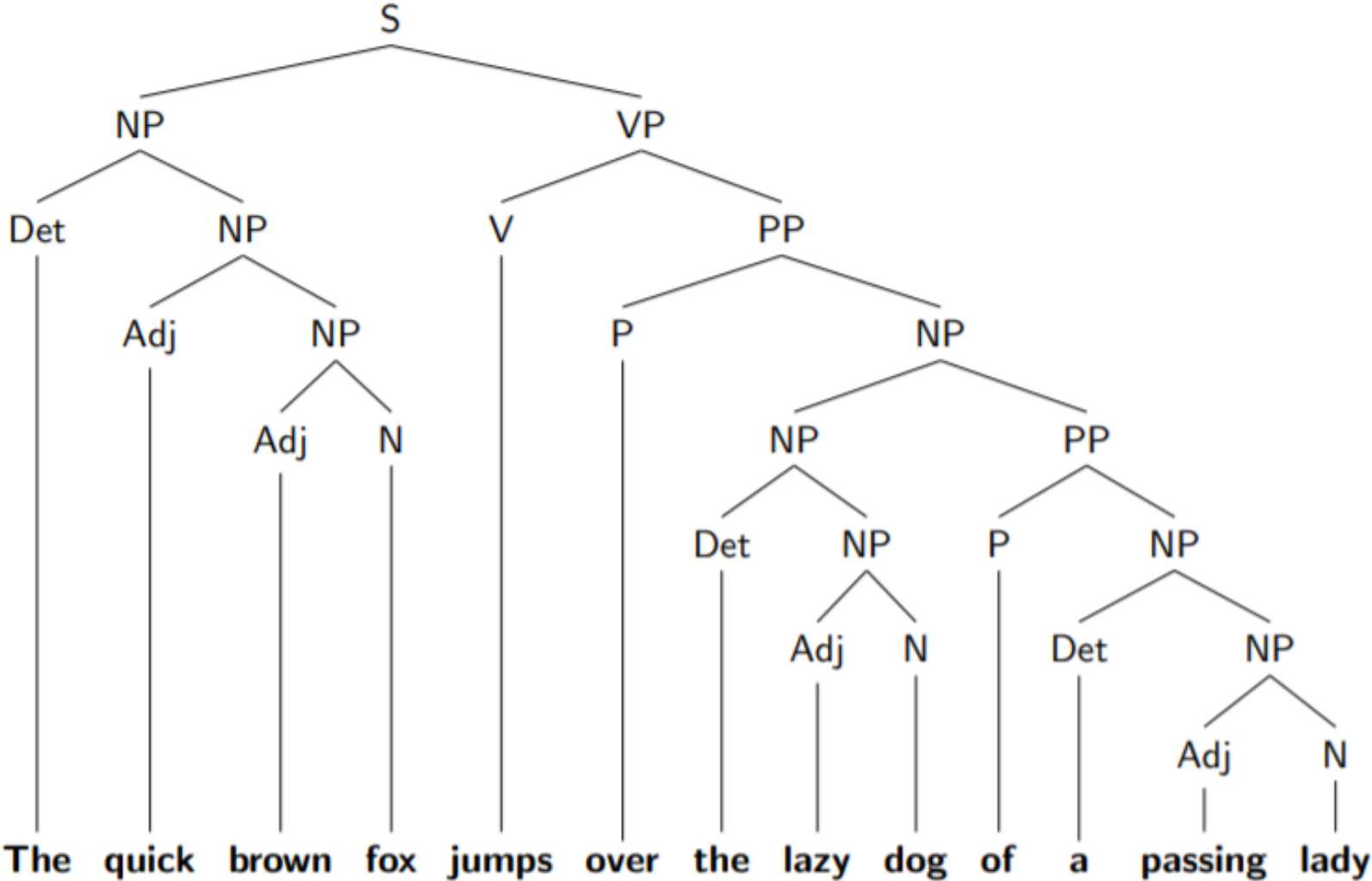
What we do with it:

- (i) obtain a toy pregroup grammar from the annotated corpus
 - entirely object-oriented, no sentence type
- (ii) obtain semantics in a category of R -semimodules¹
 - any involutive commutative semiring R , but here we focus on \mathbb{N}

Our semantics are free/minimal, in a certain sense explained later.

¹Free and finite-dimensional.

Constituent Structure Trees

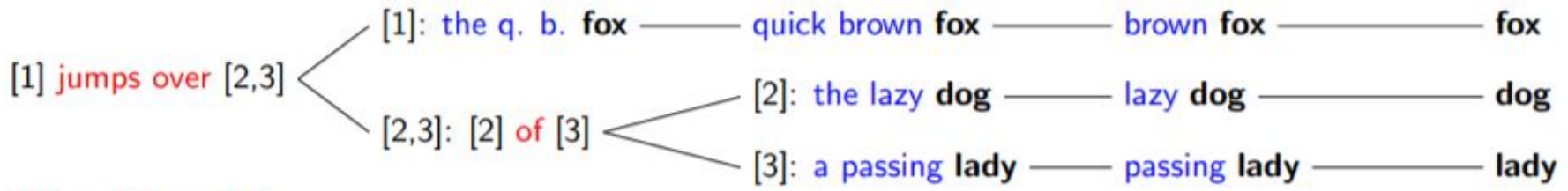
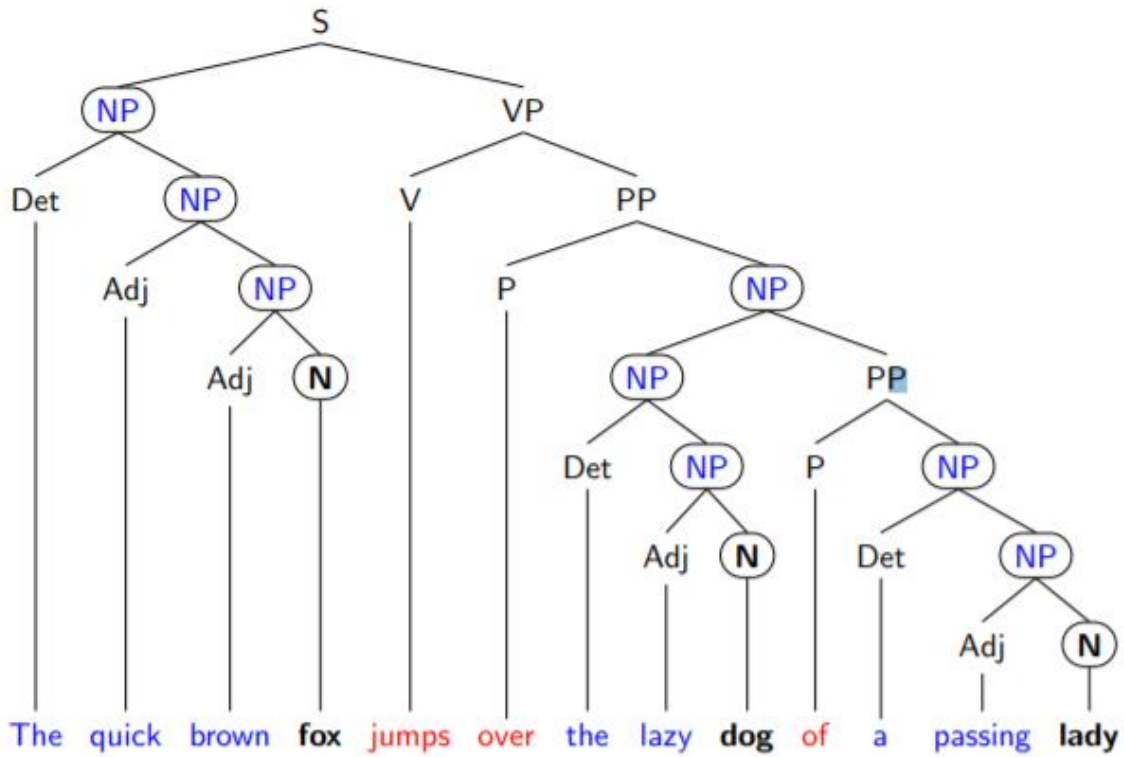


The Pregroup Grammar

A single atomic type n for *objects*, together with:

- (i) object words, of type n
 - the nouns in the corpus sentences
- (ii) modifying words, of type $n^r \cdot n$ or $n \cdot n^l$
 - the words modifying nouns/NPs into other NPs
- (iii) interaction fragments, of type $n^r \cdot n \cdot n^l$
 - sentence fragments connecting noun phrases

Objects and their Interactions



Interaction words
 Modifying words
 Interaction fragments

The Category of R -Semimodules

To obtain our semantics, we consider:

- (i) an involutive, commutative semiring R
- (ii) the category $R\text{-Mod}$ of free finite-dim R -semimodules
 - objects in the form R^X , for finite sets X
 - morphisms $R^X \rightarrow R^Y$ are $Y \times X$ R -valued matrices

Many examples of interest are in this form:

- finite sets and relations, for $R = \mathbb{Bool}$
- finite-dim real/complex vector spaces, for $R = \mathbb{R}, \mathbb{C}$
- finite-dim convex cones², for $R = \mathbb{R}^+$
- finite multi-sets and “multi-relations”, for $R = \mathbb{N}$

²Including probability distributions and stochastic maps.

The Category of R -Semimodules

Some desirable features of the category of R -semimodules:

- (i) $R\text{-Mod}$ is a \dagger -symmetric monoidal category
- (ii) $R\text{-Mod}$ is compact closed, with self-dual objects
- (iii) $R\text{-Mod}$ has classical structures associated to canonical bases

The image shows two diagrammatic equations. The first equation defines the multiplication (represented by a circle with three legs) as a sum over $x \in X$ of a diagram with two upward-pointing triangles (labeled x) and one downward-pointing triangle (labeled x). The second equation defines the comultiplication (represented by a circle with one leg) as a sum over $x \in X$ of a diagram with one downward-pointing triangle (labeled x).

$$\text{Multiplication} := \sum_{x \in X} \text{Diagram with two upward triangles and one downward triangle}$$
$$\text{Comultiplication} := \sum_{x \in X} \text{Diagram with one downward triangle}$$

The Distributional Part

We construct our semantic space from the corpus:

- (i) consider the set X of all word instances in all sentences

$$X = \{(\underline{s}, j) \mid \underline{s} \text{ sentence, } j \text{ index of word instance in } \underline{s}\}$$

- (ii) take R^X as the semantic space
- (iii) embed a word w as the indicator function of all its instances

$$\begin{array}{c} \downarrow \\ w \end{array} := \sum_{s_j=w} \begin{array}{c} \downarrow \\ \underline{s}, j \end{array}$$

The Compositional Part

Modifier words are mapped to projectors:

$$M_u := \text{diagram} \quad \text{where} \quad \text{diagram} := \sum_{(\underline{s}, j) \in m_u} \text{diagram}$$

The diagram on the left shows a circle with a vertical line passing through its center. A curved line connects the right side of the circle to the top vertex of an inverted triangle labeled m_u . The diagram on the right shows an inverted triangle labeled m_u with a vertical line passing through its top vertex. The summation symbol \sum is positioned between the two diagrams, with the subscript $(\underline{s}, j) \in m_u$ below it. To the right of the summation is another inverted triangle labeled \underline{s}, j with a vertical line passing through its top vertex.

We define m_u to be the set of instances of object words which appear in objects modified by u . For example, from our sentence we'd have:

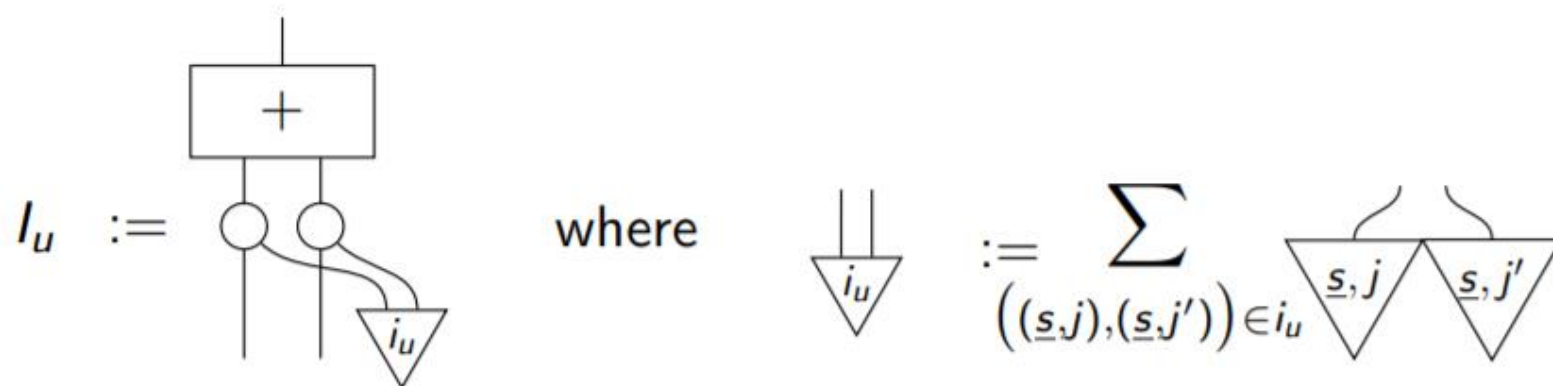
$$\begin{aligned} \{\text{fox}\} &\subseteq m_{\text{quick}} \\ \{\text{lady}\} &\subseteq m_{\text{passing}} \\ \{\text{fox}, \text{dog}\} &\subseteq m_{\text{the}} \end{aligned}$$

We automatically get some logic out of operator algebra³.

³As long as R satisfies the additive cancellation law.

The Compositional Part

Interaction fragments are mapped to binary operations. First we construct the operation for single-word fragments:

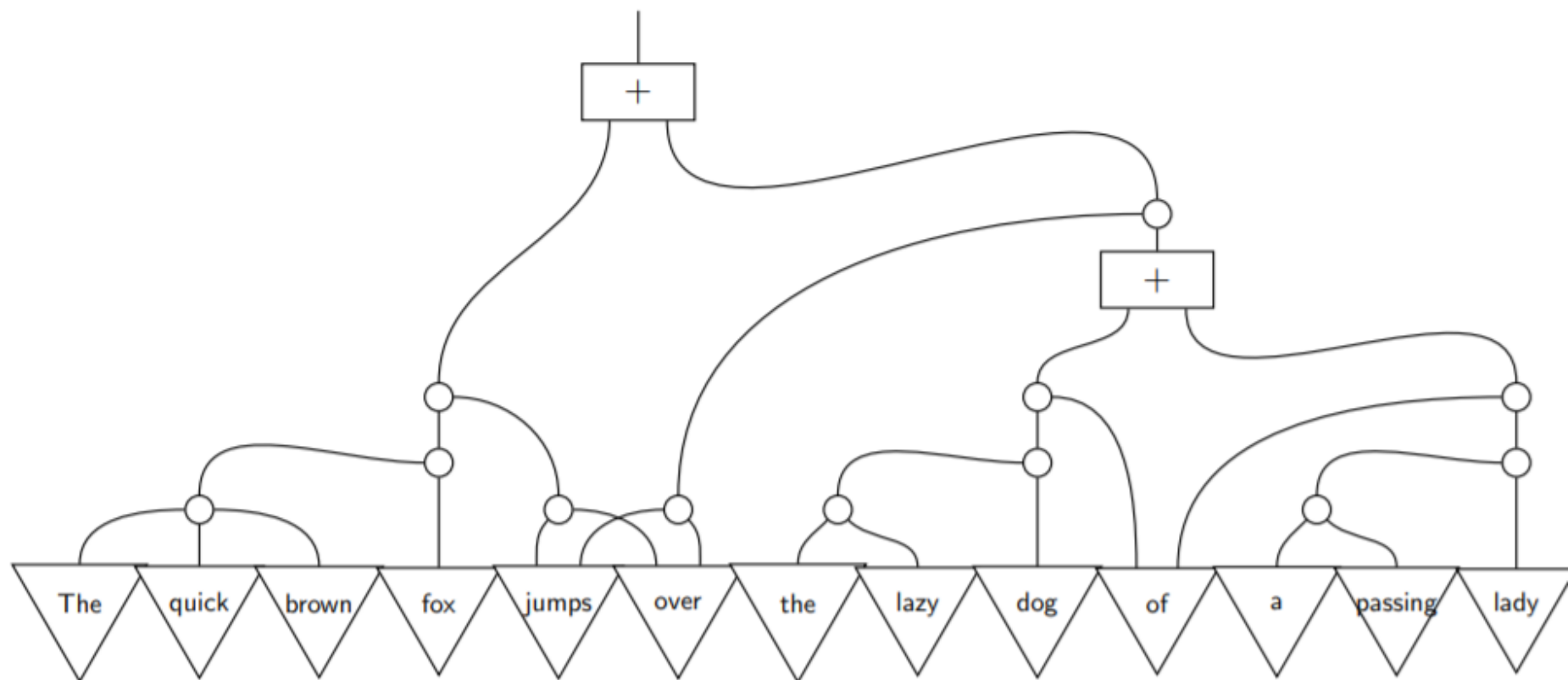


We define i_u to be the set of pairs of instances which appear in objects put into relation by u . For example, from our sentence we'd have:

$$\{(\text{fox,dog}), (\text{fox,lady})\} \subseteq i_{\text{jumps}} \tag{0.1}$$

The End Result

Here is the resulting⁴ semantics for our sentence:



⁴After some applications of the spider theorem, to group modifier words together.

Future work

A lot of things to do!

- (i) Toy model needs a number of improvements
 - treatment of personal/possessive pronouns
 - treatment of conjunctions
- (ii) More sophisticated choice of semiring
 - encoding of polarity, modality and inflection
- (iii) Compressing the free model to obtain concrete models
 - change of semiring + linear compression \Rightarrow more semantics?
- (iv) CPM construction (possibly iterated)
 - treatment ambiguity and entailment
- (v) Enriched/higher order categories
 - encode simplicial structure extracted from the corpus

Thank you for your attention!