1 Task Definition

A Synthesis procedure usually consists of a set of material operations that need to be performed to arrive at the desired product. Each of these operations has a corresponding set of arguments that are the materials and apparatus used and resulting product.

1.1 Synthesis procedure S

The first step is to segment the text into a list of operation-argument tuples called *actions* E_S.

Symbol	Description	Value
S	Synthesis procedure	
E_S	List of actions (operation-argument tuples)	$E_S = \{e_1 = (op_1, \mathbf{a}_\mathrm{i}), ...e_n = (op_n, \mathbf{a}_\mathrm{n})\}$
e_i	Action	
op_i	Operation	
$\mathbf{a}_\mathrm{i}$	List of arguments (each argument is a_{ij})	
a_{ij}	j^{th} argument of Operation op_i	$a_{ij} = (t_{ij}^{syn}, t_{ij}^{sem}, S_{ij})$
t^{syn}	Syntactic type	$t^{syn}(a) \in T^{syn} = \{DOBJ, PP\}$
DOBJ	Direct Object	
PP	Prepositional Phrase	
t^{sem}	Semantic type	$t^{sem}(a) \in T^{sem} = \{material, apparatus, other\}$
S_{ij}	String spans	$S_{ij} = \{s_{ij}^1, ..., s_{ij}^{\lvert S_{ij}\rvert}\}$
s_{ij}^k	k^{th} span in the j^{th} argument of operation op_i	

1.2 Connections C

Given a segmented synthesis procedure, we can build graph connections. A *connection* identifies the origin of a string span as either the output of a previous operation or a new material being introduced into the procedure.

Symbol	Description	Value/Type
C	Set of Connections.	Each connection is a six tuple $(o, i, j, k, t^{syn}, t^{sem})$
o	Origin index	If a span introduces a new material or apparatus, $o = 0$; else is is a valid Operation index in R, $\forall(o, i, j, k, t^{syn}, t^{sem}) \in C, o \in \{\mathbb{Z} \lvert 0 \leq o \leq \lvert E_S \rvert\}$
i	Destination index	
k	Span number in the argument	

Given a recipe R, a set of connections C is valid for R if there is a one to one correspondence between spans in R and connections in C.

Consider the following synthesis recipe as an example:
$Bi(NO_3)_3 5H_2O$ (1mmol)was dissolved in 2ml of diluted nitric acid ($2molL^{-1}$) in a beaker, and 20ml of CTAB (1.5mmol) solution was added into the above solution with continuous stirring. After that,20ml of NaOH (8mmol) solution was added into the mixed solution. The final solution was stirred for 30min and then transferred to a stainless steel Teflon-lined autoclave with the volume of 60ml. The autoclave was sealed and maintained at $120\deg C$ for 12h. After the reaction, the autoclave was cooled to room temperature naturally. The product was collected by centrifugation, washed several times with distilled water and absolute ethanol,and dried under vacuum at $60\deg C$ for 6h.

TO DO- Add final action graph example.

2 Connection Model definition

2.1 Connection Prior Model

Symbol	Description	Value
$\mathbf{d_i}$	Destination subset $\mathbf{d_i} \subset C$ Set of all connections that terminate at i i.e. have i as the destination index	
$\mathbf{g_i}$	Operation signature for an operation op_i given a destination set $\mathbf{d_i}$ $os(\mathbf{d_i}) = g_i$	Consists of two parts 1. **type:** $\{t^{syn} \vert \exists(o, i, j, k, t^{syn}, material)\} \in \mathbf{d_i}$ 2. **leaf:** true iff $(0, i, j, k, t^{syn}, t^{sem}) \in \mathbf{d_i}$
$os(\mathbf{d_i})$	Deterministic function that returns the Operation signature of a destination subset	$os(\mathbf{d_i}) = g_i$
$P(os(\mathbf{d_i}))$	Operation signature model	Multinomial distribution over the possible Operation signatures
$\mathbb{1}(o \rightarrow s_{ij}^k)$	Indicator function.	Value is 1 if there is a connection between index o and the span s_{ij}^k
c_p	List of connections that form a destination subest	$c_p \in \mathbf{d_i}$
c_1^{p-1}	Set of Connections that are prior to c_p in the list	$c_1^{p-1} = (c_1, ..., c_{p-1})$
d_1^{i-1}	Set of Destination subsets that are prior to d_i	$d_1^{i-1} = (d_1, ..., d_{i-1})$
$P(\mathbb{1}(o \rightarrow s_{ij}^k) \vert os(\mathbf{d_i}), d_1^{i-1}, c_1^{i-1})$	**Connection origin model**	The probability of an origin for a span conditioned on the Operation signature and all previous connections

The probability of a set of connections C is given by the product of each of the destination subsets:

$$P(C) = \prod_i P(\mathbf{d_i} | \mathbf{d}_1, ...\mathbf{d}_{\text{i-1}})$$

This decomposes into:

$$P(\mathbf{d_i} | \mathbf{d}_1, ...\mathbf{d}_{\text{i-1}}) = P(os(\mathbf{d_i})) \prod_{c_p \in \mathbf{d_i}} P(\mathbb{1}(o \rightarrow s_{ij}^k) | os(\mathbf{d_i}), d_1^{i-1}, c_1^{i-1})$$

2.1.1 Operation Signature Model

$P(os(\mathbf{d_i}))$ is a multinomial distribution over the possible Operation signatures. Operation signature g_i for an operation op_i given a destination set $\mathbf{d_i}$ consists of two parts - "type" and "leaf". If the origin index is 0 for all connections in $\mathbf{d_i}$, then op_i is a leaf. The "type" of Operation op_i is given by the syntactic type and semantic types of the connections in $\mathbf{d_i}$.

For example, in the synthesis recipe mentioned earlier, Operation signature for "dissolved" action is ({DOBJ, PP}, true) and "added" is ({DOBJ}, false)

2.1.2 Connection origin model

$P(\mathbb{1}(o \rightarrow s_{ij}^k) | os(\mathbf{d_i}), d_1^{i-1}, c_1^{i-1})$ is the Connection origin model. It is a multinomial distribution that models the probability of an origin of a span conditioned on the Operation signature and all previous connections.

$$\mathbb{1}(o \rightarrow s_{ij}^k) = \begin{cases} 1 & \text{if there is a connection from action with indexoto the span } s_{ij}^k \\ 0 & \text{otherwise} \end{cases}$$

For example, if g_i is a leaf, the origin of $s_i j^k$ must be 0.

2.2 Synthesis Recipe Model

Symbol	Description	
$\mathbf{h_i}$	A set of all previous actions called history	$\mathbf{h_i} = (e_1, ...e_i)$
$P(t_{ij}^{syn}, t_{ij}^{sem} \mid C, \mathbf{h_i})$	Argument types model Ensures that syntactic and semantic types of the argument match the syntactic and semantic types of the incoming connections to spans of that argument	Probability is 1 if all types match; 0 otherwise
$P(S_{ij} \mid t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h_i})$	**String span model** Models how likely it is to generate a particular string span given the types of its encompassing argument, the connections and history.	
$origin(s, C)$	Origin index of connection in C to the span s	$origin(s, C) = o \leftrightarrow \exists(o, i, j, k, t^{syn}, t^{sem}) \in C)$

Given a set of connections C for a recipe R, the model models how actions of the recipe interact and gives the probability of generating a set of recipe actions $E_S = \{e_1 = (v_1, a_1), ..., e_n = (v_n, a_n)\}$. Intuitively, R is more likely given C if the destinations are a good text representations of the origins. For example, a string span "autoclave" is more likely to refer to the action "autoclave was cooled to" than "added to mixed solution"

The probability of a recipe given the connections is given by:

$$P(R|C) = \prod_i P(e_i|C, \mathbf{h_i})$$

Assuming Operation and arguments of an action are independent, the probability of

$$P(e_i|C, \mathbf{h_i}) = P(v_i|C, \mathbf{h_i}) \prod_j P(a_{ij}|C, \mathbf{h_i})$$

Since Operation signature g_i is defined by the set of connections, we have

$P(v_i|C, \mathbf{h_i}) = P(v_i|g_i)$ which in turn is a multinomial distribution

The probability of an argument $a_{ij} = (t_{ij}^{syn}, t_{ij}^{sem}, S_{ij})$ given connections and history is given by

$$P(a_{ij}|C, \mathbf{h_i}) = P(t_{ij}^{syn}, t_{ij}^{sem}|C, \mathbf{h_i}) P(S_{ij}|t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h_i})$$

2.2.1 Argument Types model

$$P(t_{ij}^{syn}, t_{ij}^{sem}|C, \mathbf{h_i}) = \begin{cases} 1 & \text{if all types match} \\ 0 & \text{otherwise} \end{cases}$$

TO DO- Example. After action graph is drawn.

2.2.2 String span model

This models how likely it is to generate a particular string span given the types of its encompassing argument, the connection and history.

Assuming each span is independent:

$$P(S_{ij}|t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h}_\mathrm{i}) = \prod_{s_{ij}^k \in S_{ij}} P(s_{ij}^k|t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h}_\mathrm{i})$$

This distribution is broken down into three cases. (Note that the function $origin(s, C)$ determines the origin index of the connection in C to the span s.

Part-composite model
When the encompassing argument is a material and the origin is a previous Operation i.e.

$$P(s_{ij}^k|t_{ij}^{syn}, t_{ij}^{sem}) = (material, origin(s_{ij}^k \neq 0), C, \mathbf{h}_\mathrm{i})$$

then, the probability of the spans depends on the ingredients that the span represents given the connections in C. For example, "substrate" is more likely given SiO_2 rather than H_2O
IBM Model 1 is used to model the probability of a composite destination phrase given a set of origin material tokens. Say $material(s_{ij}^k, C)$ defines the set of spans in material arguments such that there is a directed path from the arguments to s_{ij}^k, the model defines the probability of a span given the propagated material spans - $P(s_{ij}^k|material(s_{ij}^k, C))$

Raw material model
When the encompassing argument is a material but the origin index is 0, i.e.

$$P(s_{ij}^k|t_{ij}^{syn}, t_{ij}^{sem}) = (material, origin(s_{ij}^k = 0), C, \mathbf{h}_\mathrm{i})$$

there is no flow of materials into the span. A span that represents a newly introduced material (ex: $NaOH$ in the synthesis recipe example) gets higher probability than the product of a previous operation (ex: "mixed solution")
We use a naive Bayes model over the tokens in the span

$$P(s|\text{is raw}) = \prod_l P(w_l|\text{is raw}) \text{ where } w_l \text{ is the } l^{th} \text{ token in s}$$

Apparatus model
When encompassing argument is an apparatus, $P(S_{ij}^k|t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h})$ models the appropriateness of origin operation's apparatus for the destination.
When s_{ij}^k is non empty i.e. apparatus is non-implicit,

$$P(S_{ij}^k|t_{ij}^{syn}, t_{ij}^{sem}, C, \mathbf{h}) = \begin{cases} 1 & \text{if string span matches the apparatus argument of the Operation} \\ 0 & \text{otherwise} \end{cases}$$

For example, the probability of "in a beaker" conditioned on an origin with apparatus "beaker" is 1, but 0 with location "autoclave"

When s_{ij}^k is empty (apparatus is an implicit argument), a multinomial model $P(loc(origin(s_{ij}^k, C))|v_i)$ determines how likely it is that an Operation op_i occurs in the same apparatus as that of the origin Operation