Automatic lexicon induction via multi-sequence alignment

Regina Barzilay and Lillian Lee

regina@cs.columbia.edu llee@cs.cornell.edu

October 9, 2001

Lexical Choice

Lexicalization — choosing words to communicate semantic predicates.

- Find all possible verbalizations for a semantic concept;
- Select the best verbalization among them.

Input	Output
supports(Canada, resolution11)	Canada supports UN resolution 11
	The resolution was supported by Canada
	Canada expressed support for it

Building a Lexical Chooser

- Lexical choice is always domain-dependent.
- In traditional generation:
  - lexical choice is performed manually;
  - the lexical component is not readily adaptable.
- We aim to learn the lexical chooser from data(multiple verbalizations).

Our Domain

Input: Output of the theorem prover Nuprl.

$$\boxed{\begin{array}{c} x : \mathbb{N} \;\vdash\; \exists y : \mathbb{N}. y > x \\ \text{BY } \texttt{Witness} \ulcorner x+1 \urcorner \\ x : \mathbb{N} \;\vdash\; x+1 > x \end{array}}$$

- Nuprl Tactics – proof strategies (variable instantiation, proof by contradiction, case split).
- Nuprl Terms – mathematical expressions($x + 1$, $y > x$).

Our Goal

Given a parallel corpus of semantic input and its verbalization, learn translations of semantic units into English phrases.

$x : \mathbb{N} \vdash \exists y : \mathbb{N}. y > x$	Instantiate 'y' with 'x+1'
BY `Witness` $\ulcorner x + 1 \urcorner$	Choose 'x+1' as the value of 'y'.
$x : \mathbb{N} \vdash x + 1 > x$	Substitute 'y' with 'x+1'.

Our Corpus

- Corpus: parallel corpus of proofs and their verbalizations (Holland-Minkley et al 1999).
- Each proof has multiple, human-generated verbalizations (from 2 to 20).
- The language exhibits high variability:

$Split$	There are ? possible choices
	We will proceed by cases on the value of ?
	We got ? cases
	split on the ? cases of
	This produces ? cases

Alignment

- Our task: align semantic representation and its English realization.
- Alignment is used for different tasks.
  - Key feature: use of multi-sequence alignment.

Task Challenges

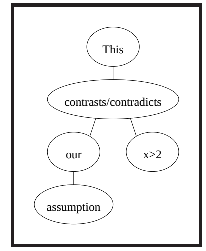
- English text contains anaphoric expressions, paraphrases, aggregations.
- Human verbalizations often contain insertions and omissions.
- We can not expect a big corpus.
- The verbalization of one semantic unit may vary in granularity from word level to sentence level.
  - Terms are either not verbalized or translated 1:1.
  - Tactics are verbalized as clause level templates.

Multiple Sequence Alignment

- A direct binary alignment between semantic input and a text verbalization may not work.
- Comparison between semantic input and several verbalizations makes commonalities more evident. Semantic input: $a \neq 0,\ b \neq 0 \vdash\ a \times b \neq 0$.
  - Given a and b as in the theorem statement , prove that $a \times b \neq 0$.
  - Suppose $a \neq 0$ and $b \neq 0$. Prove that their product is also nonzero.

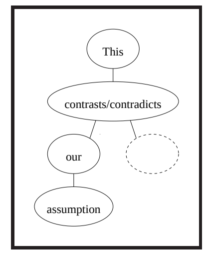
Our approach

Align several verbalizations of the same input



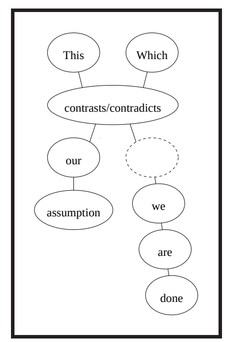
↓

Align semantic input and substitute terms



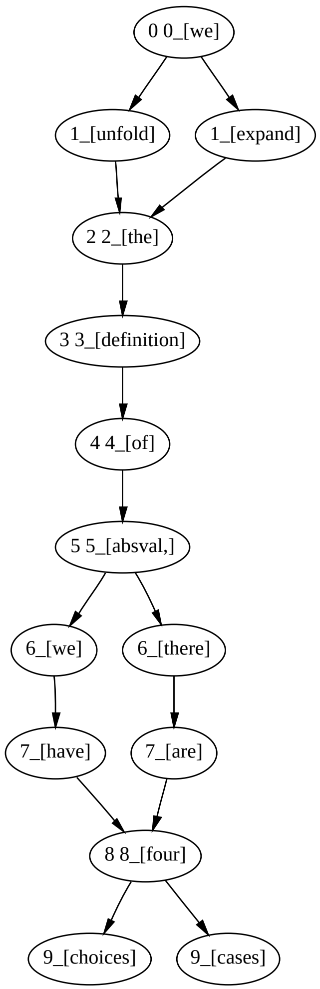
↓

Align template lattices from different instances to refine template lattice



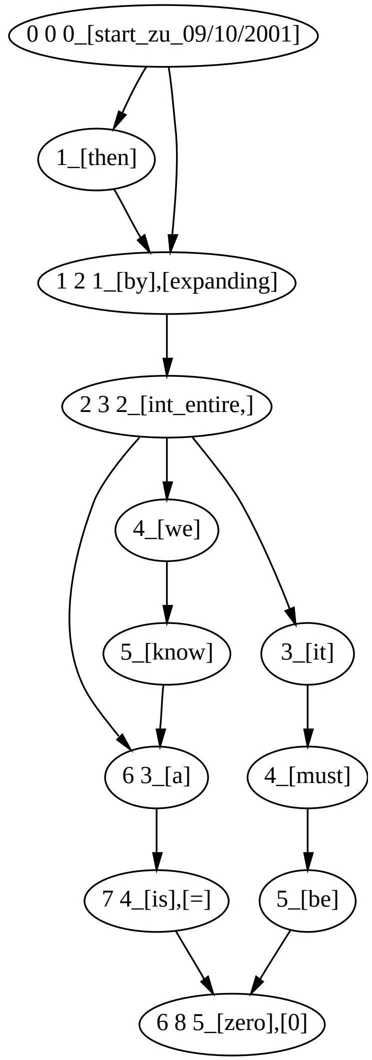
Alignment of Parallel Verbalizations

Inferring a paraphrasing dictionary:



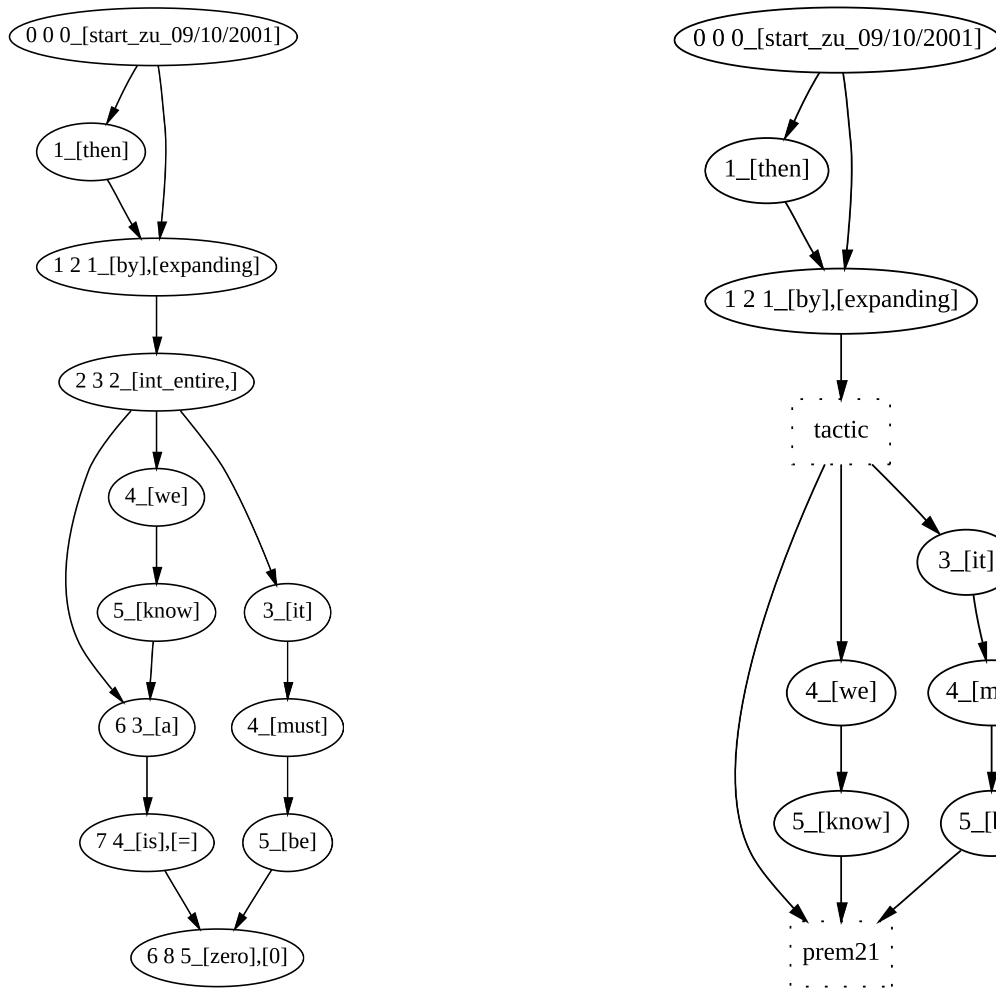
(conclusion, result)	(applying, by)	(Z, integers)
(arbitrary, integers)	(–, monus)	(assumptions, supposition)
(need to, want to)	(is, =)	(analysis, split)
(arbitrary, in)	(by, expanding)	(monus, ndiff)
(and, such)	(,, and)	(and, or)
(done, equal)	(arbitrary, z)	(., and)
(integers, n)	(assumption, fact)	(:, be)
(given, let)	(conclude, know)	(assume, suppose)
(:, that)	(0, zero)	(by, unfolding)
(let, suppose)	(— —, ndiff)	(are, be)
(need to, should)	(case, otherwise)	(expanding, unfolding)
(a, any)	(assume, suppose)	(assume, let)
(prove, show)	(contracts, contradicts)	(else, otherwise)

Multiple Sentence Alignment



Matching with Semantic Input

A=0, `int_entire`



Need for Merging Tactic Verbalizations

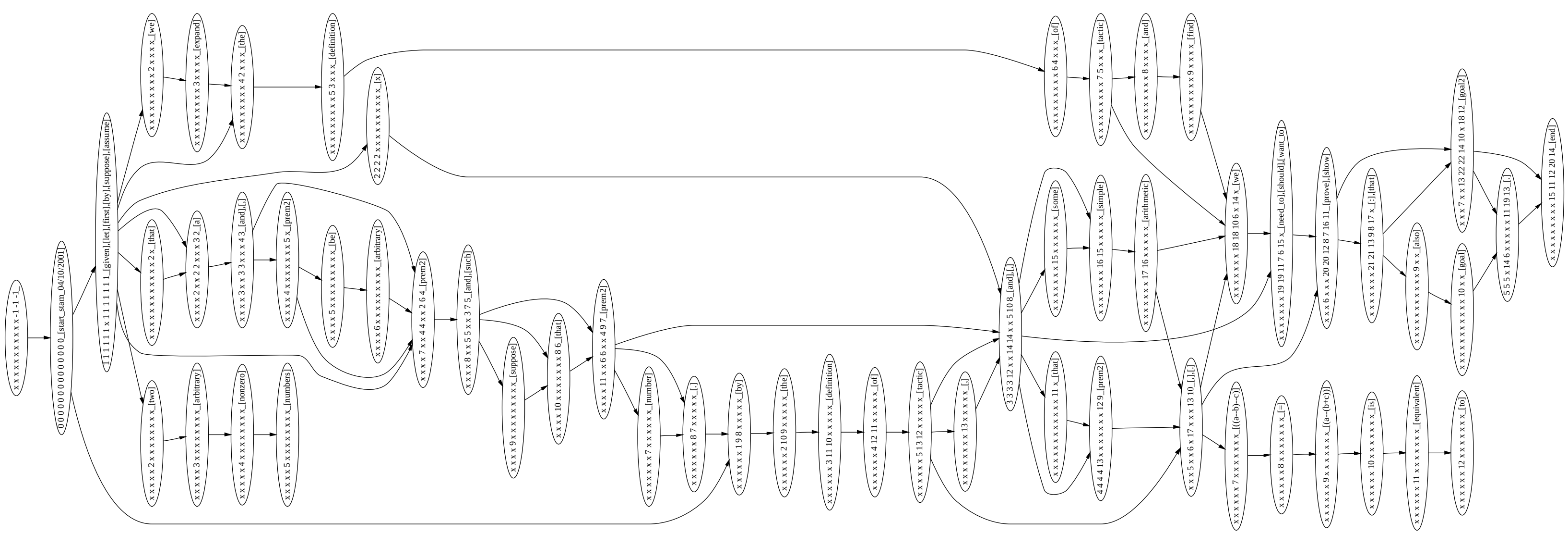
- Clean from context.
  - We can apply the lemma to get [goal], because x is a natural number.
  - From [assumption] and the lemma, we get [goal], and we are done.
  - We first apply the lemma, and then use the assoc-max, to get [goal].
- Covers for cases of imperfect match with semantics.

  Nuprl Input: $x > 1$, $y > 1$,

  Verbalization: x and $y > 1$.

Merging Tactic Verbalizations

Apply the multisequence alignment algorithm to template lattices.



From a Lattice to a Sentence

- Find a consensus path.
- Template generation.
  - Let $\boxed{\text{assumption1}}$ and $\boxed{\text{assumption2}}$, $\boxed{\text{assumption3}}$. Show that $\boxed{\text{goal}}$.
  - Assume $\boxed{\text{assumption1}}$, $\boxed{\text{assumption2}}$ and $\boxed{\text{asuumption3}}$. We need to show that $\boxed{\text{goal}}$.
  - Given $\boxed{\text{assumption1}}$ and $\boxed{\text{assumption2}}$, show that $\boxed{\text{goal}}$.
  - Assume $\boxed{\text{assumption1}}$ such that $\boxed{\text{assumption2}}$. We want to prove that $\boxed{\text{goal2}}$.

From a Lattice to a Sentence

- Prune a lattice from “uncommon” paths.
- Insert in the lattice’s slots lattices for arguments.
- Linearize based on contextual constraints and language model.

Plans for Evaluation

- Compare template generated proofs with proofs produced by a traditional generator with a manually built lexical component. (Holland-Minkley et al 1999)
- Apply the method on other domains (e.g., a basketball domain).

Conclusions

- The proposed algorithm is a promising method for automating the lexicalization component.
- Multisequence alignment lattice is a powerful representation for verbalizations of semantic predicates. This representation can be easily integrated within a statistical generation system.