Directionality of Inference Rules

- The inference rules learned by DIRT (and similar systems) are presumed to be bidirectional, which is ok for true paraphrases but not necessarily other rules.
- For example, DIRT learned: X eats Y <=> X likes

  X eats Y usually implies that X likes Y

  X likes Y does not imply that X eats Y

  So this rule has directionality: Y eats Y => X likes Y

  Y likes Y ≠> X eats Y
- **Textual entailment** systems also aim to recognize both equivalence relations (paraphrasing) and directional entailment (when one relation implies another).

LEDIR

- [Bhagat et al., 2007] created a system called LEDIR that determines whether an inference rule is bidirectional or unidirectional.
- LEDIR uses the **distributional hypothesis** combined with **selectional preferences** to determine which rule is more general.

  ***CLAIM**: Relational selectional preferences can be used to automatically determine the plausibility and directionality of an inference rule.*
- DIRT's inference rules have relatively low precision, so LEDIR also uses this approach to filter implausible rules.

Applications

Question Answering: if a question asks about Y, and X => Y, then the answer may be found in a context about X.

Q: Do monkeys like bananas?

Text: Monkeys frequently eat bananas. => YES!

Summarization: *extractive summarization* systems select sentences that are useful for a summary.

Information Retrieval: paraphrases are useful for query expansion.

Knowing directionality can help choose more specific or general sentences.

Inference Rule Directionality

Problem: given the inference rule p_i <=> p_j, determine which type of inference is most appropriate:

1. p_i <=> p_j
2. p_i => p_j
3. p_i <= p_j
4. No plausible inference

Directionality Hypothesis: *If two binary semantic relations tend to occur in similar contexts, and the first one occurs in significantly more contexts the second, then the second most likely implies the first and not vice versa.*

The key idea is to identify which of the two relations is more general.

Selectional Preferences

- "The **selectional preferences** of a predicate is the set of semantic classes that its arguments can belong to [Wilks 1975]."
- [Pantel et al., 2007] defined the **relational selectional preferences (RSPs)** of a binary relation *p* as the set of semantic classes C(*x*) and C(*y*) that can occur in the *x* and *y* positions.
- LEDIR applies the distributional hypothesis to a semantic representation of contexts, to abstract away from specific words. Instead of sets of words in the X and Y positions, LEDIR uses sets of semantic classes.

Joint Relational Model (JRM)

- Given a relation p and a large corpus, we extract all instances <x, p, y> of the relation and identify the semantic classes of x and y, C(x) and C(y) respectively.
- Create a triple database of: <C(x), p, C(y)>
- Rank the candidates using point-wise mutual information (PMI).

$$PMI(c_x|p, c_y|p) = \log_2 \left[\frac{P(c_x, c_y|p)}{P(c_x|p) * P(c_y|p)} \right]$$

Independent Relational Model (IRM)

- The joint model accumulates frequencies for pairs that occur together. However, requiring specific pairs will likely lead to sparse data issues (i.e., there are many pairs were may not see, even though they are possible).
- The IRM creates a database for each path and slot separately. The independent RSPs for a relation p are:

<C(*x*), p, *> and <*, p, C(*y*)>

All combinations of these X and Y sets are then generated to have the same form as the joint RSPs:

<C(*x*), p, C(*y*)>

Overlap Coefficient

- Given a candidate inference rule, LEDIR uses an **overlap cofficient** measure to assess the similarity between the selectional preferences of the two paths.
- Given 2 vectors A and B, the overlap coefficient is:

$$sim(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

- Given two paths and their selectional preferences:

$$sim(p_i, p_j) = \frac{|\langle C_x, p_i, C_y\rangle \cap \langle C_x, p_j, C_y\rangle|}{\min(|C_x, p_i, C_y|, |C_x, p_j, C_y|)}$$

Plausibility and Directionality

They decomposed their task into two subproblems:

(1) determining whether an inference is plausible

If $sim(p_i, p_j) \geq \alpha$ → inference is plausible
Else → inference is not plausible

(2) determining the directionality of plausible rules

If $\frac{|\ C_x, p_i, C_y\ |}{|\ C_x, p_j, C_y\ |} \geq \beta$ then we conclude $p_i <= p_j$

else if $\frac{|\ C_x, p_i, C_y\ |}{|\ C_x, p_j, C_y\ |} \leq 1/\beta$ then we conclude $p_i => p_j$

else we conclude $p_i <=> p_j$

Gold Standard Data

- They randomly sampled 160 inference rules produced by DIRT on 1Gb of text (but discarded 3 rules with nominalizations).
- Two human annotators labeled these 157 inference rules with one of four labels:

  $p_i <=> p_j$ $p_i => p_j$ $p_i <= p_j$ no plausible inference

  57 inference rules were used for development, and 100 were reserved as a blind test set.
- The annotators were given 10 randomly selected instances of each rule to provide example contexts.
- Inter-annotator agreement yielded $\kappa = 0.63$
  The annotators adjudicated the disagreements.

Semantic Classes and Text Collection

They experimented with two sources of semantic classes.

- WordNet : a cut at depth four produced a set of 1287 semantic classes.
- CBC clustering algorithm: 1628 semantic classes (clusters)

An AP newswire text collection containing 31 million words was used to obtain the probability statistics.

The Minipar parser was applied to these texts.

Baselines

Several baselines were evaluated to compare with LEDIR.

For each candidate inference rule:

- **B-random** : randomly assigns one of the four possible tags.
- **B-frequent** : assigns the most frequently occurring tag in the gold standard.
- **B-DIRT** : assigns the bidirectional <=> tag (because this is what DIRT assumed).

Results on the Test Set

They experimented with many parameter settings using the 57 rules in the development set. The best parameters were then applied to the test set.

Model		α	β	Accuracy (%)
B-random		-	-	25
B-frequent		-	-	34
B-DIRT		-	-	25
JRM	CBC	0.15	2	38
	WN	0.55	2	38
IRM	**CBC**	**0.15**	**3**	**48**
	WN	0.45	2	43

Table 1: Summary of results on the test set

Error Analysis

A confusion matrix showed that LEDIR does a good job at identifying the direction of rules, but often does not recognize that a rule is not plausible.

		GOLD STANDARD			
		⇔	⇒	⇐	NO
SYSTEM	⇔	16	1	3	7
	⇒	0	3	1	3
	⇐	7	4	22	15
	NO	2	3	4	9

Table 2: Confusion Matrix for the best performing system, IRM using CBC with α=0.15 and β=3.

Summary

- LEDIR applies the distributional hypothesis to a semantic space by explicitly representing the semantic classes associated with the two slots in a binary relation.
- The directionality of an inference rule can then be determined by comparing the relatively generality of their slot fillers.
- This approach is not necessarily sufficient to filter implausible rules, though. Generating accurate inference rules is still a challenging problem.