An analogical model

(1) Recall from last week: two different ways of handling novel (unknown) words
   - Use rules that you have previously learned, and use it to derive your answer
   - Compare the novel item to other, known items, and base your decision on the behavior of similar items

(2) There are many different models that operate on this principle
   - Generalized Context Model (Nosofsky 1988; Nakisa, Plunkett, and Hahn 1997)
   - Analogical Model of Language (Skousen 1989; Derwing and Skousen 1994; Eddington 2000)
   And many more models use a combination of analogy and generalized knowledge
   - Connectionist/neural net models (Rumelhart and McClelland 1987, and much subsequent work)
   - Bybee (1985); Bochner (1993)
   - Dual Mechanism Model of Morphology (see Pinker 1999 and Clahsen 1999 for overviews)
   We’ll look at just one particular version, the Analogical Model of Language

(3) An example (Eddington 2000): what is the past tense of sin?
   - One possible approach: look through the lexicon of existing verbs, find all the verbs that are similar to sin (e.g., spin, sing, thin, etc.), and let them vote on how to treat the word sin
     - We could, for example, use string edit distance, as presented earlier in this course, to determine how similar sin is to other verbs, and then let those verbs vote in proportion to their similarity, or maybe some more complicated weighting of similarity, frequency, etc.
   - A different approach: examine uniformity of contexts

(4) Contexts of varying degrees of generality.
   - Verbs that are sin
   - Verbs that start with si
   - Verbs that end with in
   - Verbs that start with s and end with n
   - Verbs that start with s
   - Verbs that contain i
   - Verbs that end in n
   - etc…

(5) To keep things small and concrete, suppose the lexicon contains only the following items:

   sing  [sɪŋ]  irr.
   thin  [θɪn]  reg.
   gun  [gʌn]  reg.
   stink  [stɪŋk]  irr.
   drown  [draʊn]  reg.
   sip  [sɪp]  reg.
   load  [lod]  reg

---

1This is not the best example, for several different reasons. First, as we’ve already discussed, there is the problem that sin is an existing verb, and thus could be derived in one of two possible ways; thus, we have to be careful that our intuitions about what is a reasonable output for the word sin are not contaminated by our prior knowledge that its past tense “really is” sinned. In addition, it seems likely that sin is a denominal verb, and denominal verbs are virtually all regular in English.
A brief sermon: Notice that we’ve changed from looking at the particular changes involved in forming the past tense (e.g., [i] → [æ], ∅ → [d], etc.), and are now just broadly categorizing them as regular vs. irregular. This is a simplification of the task that one often finds in models that categorize based on existing words. Simplifications of the task are often necessary and important in letting us test things that might otherwise be beyond our current modeling capabilities, but at the same time, we should take such results with a grain of salt. The true test of a model is whether it can generate the correct output in all of its details, and it would be rash to assume that a model that can successfully mark things as “regular” vs. “irregular” can necessarily solve the harder task of generating complete outputs.

(6) “Subcontexts” of sin:

Notation: $s = \text{anything other than s}$

<table>
<thead>
<tr>
<th>Subcontext</th>
<th>Words</th>
<th>Outcomes</th>
<th>Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>sin</td>
<td>(none)</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>šin</td>
<td>[θın]</td>
<td>reg.</td>
<td>0</td>
</tr>
<tr>
<td>sin</td>
<td>(none)</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>siñ</td>
<td>[sŋ], [sip]</td>
<td>irreg, reg</td>
<td>2</td>
</tr>
<tr>
<td>šin</td>
<td>[gən], [draʊn]</td>
<td>reg, reg</td>
<td>0</td>
</tr>
<tr>
<td>šiñ</td>
<td>[stŋk]</td>
<td>irreg</td>
<td>0</td>
</tr>
<tr>
<td>siñ</td>
<td>(none)</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>šiñ</td>
<td>[lod]</td>
<td>reg</td>
<td>0</td>
</tr>
</tbody>
</table>

(7) “Supracontexts” of sin ($C =$ context)

<table>
<thead>
<tr>
<th>SupraC</th>
<th>SubC’s disagree</th>
<th>SubC’s Words</th>
<th>Outcomes</th>
<th>SupraC’s Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>sin</td>
<td>sin</td>
<td>0 (none)</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>s–i–</td>
<td>sin, šin</td>
<td>2 sip, sing</td>
<td>reg, irreg</td>
<td>2</td>
</tr>
<tr>
<td>s–n</td>
<td>šin, ši</td>
<td>0 (none)</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>–in</td>
<td>šin, šin</td>
<td>0 thin</td>
<td>reg</td>
<td>0</td>
</tr>
<tr>
<td>s––</td>
<td>šin, ši, siñ, šiñ</td>
<td>2 sip, sing</td>
<td>reg, irreg</td>
<td>2</td>
</tr>
<tr>
<td>–i–</td>
<td>šin, ši, siñ, šiñ</td>
<td>2 think, stink, sip, sing</td>
<td>irreg, irreg, reg, irreg, reg</td>
<td>6</td>
</tr>
<tr>
<td>––n</td>
<td>šin, šin, šiñ, šiñ</td>
<td>0 thin, gun, drown</td>
<td>reg, reg, reg, reg</td>
<td>0</td>
</tr>
<tr>
<td>––– ––</td>
<td>šin, šiñ, šiñ, šiñ, šiñ, šiñ</td>
<td>2 thin, gun, drown, stink, sip, load</td>
<td>reg, reg, irreg, reg, reg</td>
<td>12</td>
</tr>
</tbody>
</table>

(8) Which contexts should play a role in determining the outcome for sin?

- Proposed answer: only the homogenous ones
- A context is homogenous is there are no more disagreements in the supracontext than in its subcontexts
- Which of the contexts in (7) are homogenous

(9) Using homogenous contexts to decide on the outcome for the past tense of sin

- This is accomplished by examining pointers in each context:

<table>
<thead>
<tr>
<th>s i n</th>
<th>s i –</th>
<th>s – n</th>
<th>– i n</th>
<th>s – –</th>
<th>– – –</th>
</tr>
</thead>
<tbody>
<tr>
<td>sip</td>
<td>sing</td>
<td>thin</td>
<td>sip</td>
<td>sing</td>
<td>thin</td>
</tr>
<tr>
<td>REG</td>
<td>IRREG</td>
<td>REG</td>
<td>REG</td>
<td>IRREG</td>
<td>REG</td>
</tr>
<tr>
<td>2 reg, 2 irreg</td>
<td>1 reg</td>
<td>2 reg, 2 irreg</td>
<td>9 reg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 18 pointers: 14 point to REG, 4 point to IRREG. Probability that sin is regular: 14/18 = 78%
(10) Why calculate analogy in such a complicated way?

- Recall that our basic philosophy was simply to let words influence one another
- However, we want to make sure that the strongest effect is by groups of words that are very similar, and behave consistently
- Determining contexts and homogeneity helps to locate sets of words that share definable features, and pattern together consistently

References


