LING 334 - Introduction to Computational Linguistics

Week 5

Linguistic Structure, NLP “Tasks”, and Annotation
The Basic Approaches of Linguistics

It’s all over the place! Low consensus field.

This makes some sense -
language has many parts and purposes.

Descriptivism

Maybe the one thing we can all agree on:
the object of study is how and what language is,
rather than what it “should be” (prescriptivism)
Descriptivism

Origins with Pāṇini, Sanskrit linguist ~400BC

Contrast with “experts,” Strunk and White etc.
(these are cultural norms and conventions)

Key (modern) ideas:
● Language change is normal and expected
● Everyone has a “dialect”
● There are very few cross-linguistic universals
Traditional Levels of Structure

- **Phonetics**: sounds
- **Phonology**: ordering of sounds
- **Morphology**: words and word parts
- **Syntax**: ordering of words
- **Semantics**: propositional meaning
- **Pragmatics**: non-propositional meaning

Small to big units:
But there are many more...

Reference pointing out things with words
Prosody suprasegmental sounds like pitch
Discourse sequences between large units
Social Meaning social implicature of variation
The Concept of a “Task” in NLP

Research in NLP is often framed as solving a particular “task”, e.g. improving performance at some problem

Very frequent sort of task in traditional NLP:
  Given free text or speech audio,
  automatically generate a representation of some part of its linguistic structure
Phonetics

The physical production and perception of speech sounds

Unit of analysis: speech sound

NLP Tasks:
Speech synthesis
Automated transcription

https://dood.al/pinktrombone/
Phonology

The systematic organization of speech sounds
Unit of analysis: phoneme

Questions include:
● Which set of sounds does a language use?
● What rules constrain their orderings?

Example: /P/ aspiration
● ‘pin’ - the ‘p’ sound has a puff of air $[p^h]$ 
● ‘spin’ - it doesn’t $[p]$
Morphology

The structure and constituent parts of words

Unit of analysis: morpheme

(smallest meaning-bearing unit)

Morphemes can be:

Free can stand alone, words like ‘cat’ and ‘banana’

Bound can’t stand alone, word-parts like ‘un-’ and ‘-est’

NLP Tasks:

● Morphological Segmentation (very important in synthetic langs!)
● Lemmatization and Inflection
Syntax

The systematicity of word orderings

“The sloth ate the cupcake.” \(\neq\) “The cupcake ate the sloth.”

* “Cupcake sloth ate the the.”

NLP Tasks:

- Syntactic Parsing
- Downstream applications, e.g.:
  - Machine Translation
  - Semantic Similarity
Semantics

The propositional (e.g., literal) meanings of words and larger units (frequently sentences)

Table =

We’re digging into semantics next week!
Pragmatics

The beyond-propositional meanings of words and larger units

Among the many possibilities:

Implicature  “I’m sad.” “Here’s a popsicle.”
Performatives “I now pronounce you X and Y.”
Deference  “Please follow me, your majesty.”

Information Structure

NLP Tasks:
Many social/applied!
Reference

What entity in the world does a linguistic expression point out? Includes pronouns, honorifics, naming and nicknaming

Winograd Schema Challenge:
“The goose wouldn’t fit in the boat because it was too big.”
“The goose wouldn’t fit in the boat because it was too small.”

NLP Tasks:
- Coreference Resolution
- Named Entity Recognition
Discourse

The relations between clauses and propositions

NLP Tasks:

- Discourse Parsing
- Argumentation Mining
Social Meaning

Many sorts of complex socially enmeshed meaning-making:

- Sentiment and stance
- Regional variation
- Identity performance
- Memes and spread of ideas

Each can be an NLP Task!
Data in Linguistics

Introspection, and/or “native speaker intuitions”

Collected observations of language in use (e.g. corpora)

Laboratory data (experimentally collected or manipulated)

All of the above potentially augmented with *annotations*
Linguistic Annotations

To train a relevant model, we need training data
   So, we hand-label some!

Traditionally, most commonly done by experts

Today, frequently done with crowdsourcing as well

Which is more appropriate depends on the task!
   See relevant readings re: wisdom of the crowd -
   Naive annotators can do a great job!
Annotation Schemes

An annotation scheme or ontology instantiates a theory of language.

Example - Part of Speech Tagging:

36 Penn Treebank Tags

Implicit Proposal: these are what’s important

I/PRP love/VBP eating/VBG noodles/NNS

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
Annotation Schemes (cont.)

Frequently developed over multiple rounds of piloting

Common tradeoff between specificity and speed/expense/scale

Do I want 40 categories and 400 annotations, or 5 categories and 4,000 annotations?

Zipf’s Law - vanishing returns as we get many categories
Annotation Evaluation

Linguistic categories are purely abstract human creations!

There is no ground truth. (rut roh)

So we usually evaluate with **Inter-annotator Agreement**

Have some proportion of the data annotated by multiple people

Obtain a measurement of consistency - how often do people make the same judgment?
Inter-Annotator Agreement

Common - Cohen’s Kappa

Compare the expected agreement to the actual:

\[ \kappa \equiv \frac{p_o - p_e}{1 - p_e} \]

- \( p_o \) = probability of observed agreement
- \( p_e \) = probability of expected agreement
Inter-Annotator Agreement

Say we have a task with two labels, POS and NEG, and two annotators, A and B - count up each category:

<table>
<thead>
<tr>
<th></th>
<th>Annotator A</th>
<th>Annotator B</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>45</td>
<td>15</td>
</tr>
<tr>
<td>NEG</td>
<td>25</td>
<td>15</td>
</tr>
</tbody>
</table>

22
Inter-Annotator Agreement

Get totals:

<table>
<thead>
<tr>
<th>Annotator</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>45</td>
<td>15</td>
</tr>
<tr>
<td>NEG</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>total</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>N = 100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Inter-Annotator Agreement

They agreed 60% of the time

\[ p_o = \frac{(45 \text{ POS} + 15 \text{ NEG})}{100 \text{ total}} = 60\% \]

<table>
<thead>
<tr>
<th></th>
<th>POS</th>
<th>NEG</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>45</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>NEG</td>
<td>25</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>total</td>
<td>70</td>
<td>30</td>
<td>N = 100</td>
</tr>
</tbody>
</table>
Inter-Annotator Agreement

Probability of expected is trickier - calculate expected freq for each category:

\[
E_{freq} = \frac{\text{row}_{total} \times \text{col}_{total}}{N}
\]

<table>
<thead>
<tr>
<th></th>
<th>POS</th>
<th>NEG</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POS</strong></td>
<td>45</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td><strong>NEG</strong></td>
<td>25</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>70</td>
<td>30</td>
<td>N = 100</td>
</tr>
</tbody>
</table>
Inter-Annotator Agreement

Now we can get $p_e$:

$$p_e = \frac{42 \text{ POS exp} + 12 \text{ NEG exp}}{100} = 0.54$$

<table>
<thead>
<tr>
<th></th>
<th>POS</th>
<th>NEG</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>45</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>NEG</td>
<td>25</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>total</td>
<td>70</td>
<td>30</td>
<td>N = 100</td>
</tr>
</tbody>
</table>
Inter-Annotator Agreement

And calculate Kappa: \[
\frac{p_o - p_e}{1 - p_e} = \frac{0.6 - 0.54}{1 - 0.54} = 0.13
\]

<table>
<thead>
<tr>
<th></th>
<th>POS</th>
<th>NEG</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>45</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>NEG</td>
<td>25</td>
<td>15 (12)</td>
<td>40</td>
</tr>
<tr>
<td>total</td>
<td>70</td>
<td>30</td>
<td>N = 100</td>
</tr>
</tbody>
</table>
Interpretation of Agreement Metrics

Usually scaled 0.0 - 1.0:
What counts as good?

Differing opinions!

Ultimately, it’s made up, so it depends on the task