Tracing Semantic Variation in Slang Summary

Summary of Tracing Semantic Variation in Slang

Abstract

- Slang semantic variation is under explored in NLP
- Two Views:
  - Slang semantic variation is driven by culture-dependent communicative needs
  - The other view suggests the desire to foster semantic distinction may have led to the historical emergence of community-specific slang senses
- The study explored these theories and tested them against historical slang dictionary entries, with a focus in the UK and US over the past 2 centuries.
- Their models are able to predict the regional identity of emerging slang word meanings from historical slang records
- They offer empirical evidence that both communicative need and semantic distinction play a role in the variation of slang meaning yet their relative importance fluctuates over the course of history.
- Their work offers an opportunity for incorporating historical cultural elements into the natural language processing of slang.

Introduction

- The focus on slang has increased in NLP over recent years, with papers and systems proposed for automatic detection, generation, and interpretation of slang.
- However, these approaches don't account for the semantic variation of slang among different groups of users, which is a defining characteristic of slang.

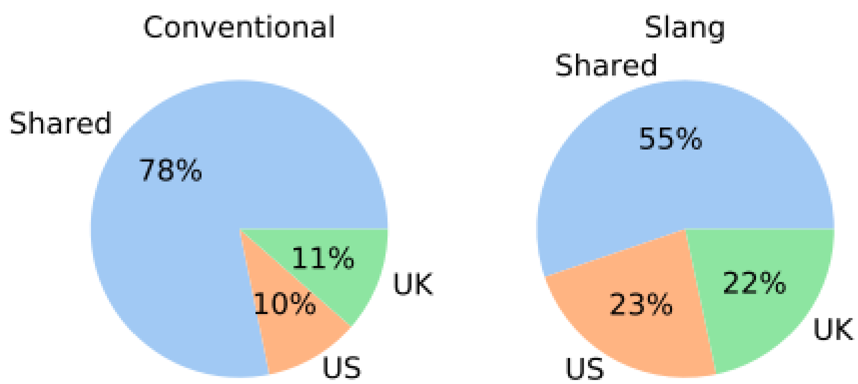

Figure 1: Distribution of regional identities among sense entries found in the English Wiktionary. See Appendix A for the detailed experimental setup.

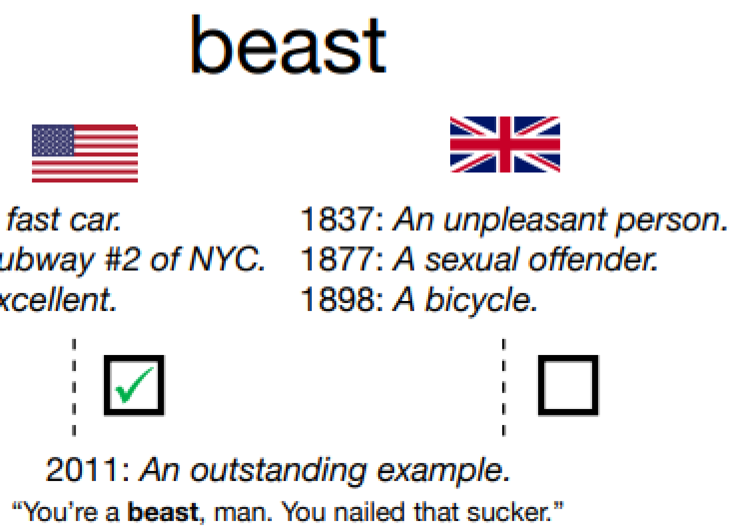

Figure 2: Illustration of semantic variation in the slang word *beast*, with senses recorded in American and British English respectively. We develop slang semantic variation models to trace the regional identity of a new emerging slang sense given its historical meanings and usages from different regions.

- They define semantic variation in slang as how a slang term might take on different meanings in different communities.
- Community characteristics (community size, network density, etc.) are relevant factors of the strength of semantic variation.
- However, it's not clear how slang senses vary among communities and what the driving force behind the variation is.

- As an initial step to model slang semantic variation, they considered a regional inference task. Given an emerging slang sense for a slang word, they infer which region it might have originated from based off of historical meanings and usages.
- A model capturing slang semantic variation should be able to trace or infer regional identities of emerging slang meanings over time.

Theoretical Hypotheses

- They considered two theoretical hypothesis for characterizing regularity in slang semantic variation: communicative need and semantic variation.
- **Communicative Need** - how frequently a meaning needs to be communicated or expressed. They estimated communicative need based on usage frequencies from Google Ngram over the past two centuries. Certain things may be talked about more in one region over another. We might expect these differential needs to drive meaning differentiation in slang terms.
  - For example, a US specific slang sense for beast describes the subway line #2 of the New York City transit network, most likely due to the specific need for communicating that information in the US (as opposed to the UK).
- **Semantic Distinction** - language that is used to show and reinforce a group identity. Under this hypothesis, slang senses develop independently in each region and form a semantically cohesive set of meanings that reflect the cultural identity of a region. As a result, emerging slang senses are more likely to be in close semantic proximity with historical slang senses from the same region.
  - Here we operationalize semantic distinction by models of semantic chaining from work on historical word meaning extension, where each region develops a distinct chain of related regional senses over history
- We evaluate these theories using slang sense entries from Green's Dictionary of Slang over the past two centuries. Analysis on GDoS entries is appropriate because...
  1. A more diverse set of topics is covered compared to domain-specific slang found in online communities (e.g., Reddit, Twitter, etc.), and
  2. The region and time metadata associated with individual sense entries support a diachronic analysis on slang semantic variation.
- To preview our results, we show that both communicative need and semantic distinction are relevant factors in predicting slang semantic variation, with an exemplar-based chaining model offering the most robust results overall.
- Meanwhile, the relative importance of the two factors is time-dependent and fluctuates over different periods of history.
- It is worth nothing that communicative need and semantic distinction may not be completely orthogonal. In fact, differences in communicative need may drive semantic

distinction. However, we consider these hypotheses as alternative ones because they are motivated by different functions.

Data Preprocessing

- They collected slang lexical entries from Green's Dictionary of Slang (GDoS), a historical English slang dictionary covering more than two centuries of slang usage
- Each word entry in GDoS is associated with one or more sense entries. A sense entry contains a definition sentence and a series of references. Each reference contains a region tag, a date tag, and a sentence indicating the origin of the reference. In some cases, the reference contains an example of how the slang is used in context

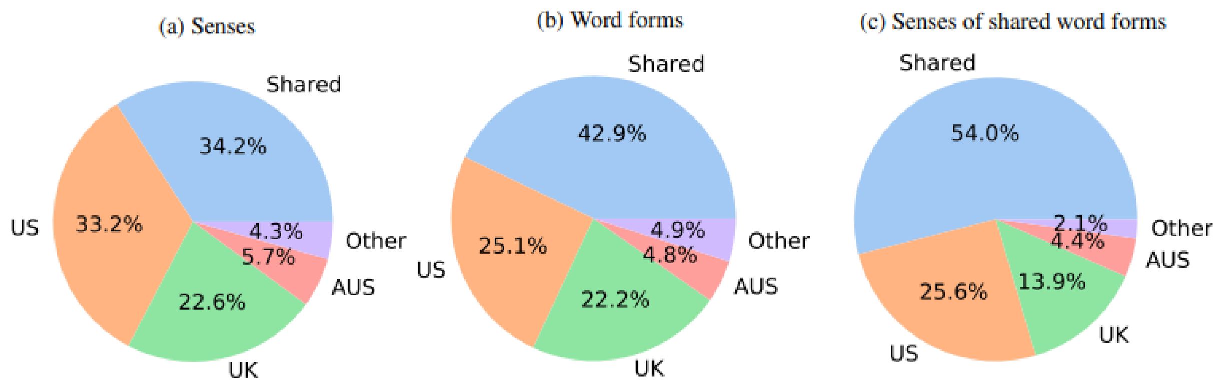

Figure 3: The distribution of GDoS slang senses and word forms across different regions. A word or sense is considered shared if two or more distinct region tags can be found in the constituent references.

- They collected all sense entries with at least one valid reference. A reference is considered valid if both its region tag and date tag are not missing nor invalid.
- For each reference, we automatically extract the associated context sentence and consider one to be valid if it contains precisely one exact occurrence of the word in the sentence.
- If a valid context sentence is found then it is attached to the corresponding reference
- The resulting sense entry may have none or more than one context sentences. In the latter case, they selected the context sentence with the earliest time tag to be associated with the sense entry, so that it best represents the usage context of when the sense first emerged.
- The earliest time tag is considered the time of emergence for a sense entry.
- They filtered all abbreviation entries as those don't create a new meaning.
- In the case of a homonym (multiple word entries for the same word form), they collapsed all entries into a single word entry.
- All of this preprocessing script is in the github repo

Data Analysis

- Begin by analyzing entries collected from GDoS to quantify semantic variation.
- For each sense entry, we determine its regional identity using the region tags associated with each reference.
    - Note: there may be more than 1 valid region tags associated with each sense entry. In these cases, they considered the sense entry to be a shared sense across all constituent region tags. Otherwise, the sense entry is considered unique to its region.
    - Likewise, a word entry is considered shared if two or more distinct region tags can be found among any of its sense's references
- They observed substantial lexical variation within the data where more than half of the word forms are regional. While most of the sense entries are also regional, many of them may be associated with regional word forms. In this case, the variation is caused by difference in lexical choice and does not entail semantic variation. We control for lexical variation by only considering sense entries associated with shared word forms.
    - Even after controlling for lexical variation, they observed roughly half of the senses remain regional. And most of these are captured by the US and UK, with Australian slang also making an appearance.
    - They focused their study by modeling the two most prevalent groups (US and UK)

Predictive Task

- They used regional inference task to model semantic variation.
- Given an emerging slang sense s for a word w, infer the region $r \in R$ from which the emerging sense originates. Here, R is the set of regions being considered (so {US, UK})
- The semantic variation model V is defined as follows:
    - $P(r) \propto V(s, w, r)$
    - Here, the semantic variation model V captures the likelihood of observing the emerging slang sense s expressed using word w within region r and can be either generative or discriminative in nature
- The target region can be predicted by maximizing the likelihood:
    - $r^* = \arg\max V(s, w, r)$
- An effective semantic variation model should prefer regions that are more likely for the new sense to emerge.

Modeling Communicative Need

- We first describe a set of semantic variation models V inspired by the communicative need principle. We operationalize communicative need using frequency statistics from historical corpora originated from each region.
- First, we propose a form need model that considers the frequency of the slang word form w:

- $$\mathcal{V}_{\text{form\_need}}(\cdot) \propto f_{r,s_t-\alpha:s_t}(w)$$
    - The function f is the frequency of observing word w from region r within a time window α, which precedes the sense's time of emergence st.
    - The form need model does not take into account any semantic information from the emerging sense s and simply estimates whether the word form w is more prevalent in one region.
- The semantic need model incorporates semantic info by checking the frequency of all content words within the definition sentence sd of sense s
    - $$\mathcal{V}_{\text{semantic\_need}}(\cdot) \propto \sum_{c \in \text{content}(s_d)} f_{r,s_t-\alpha:s_t}(c)$$
- The context need model is informed by the usage context sentence sc of sense s:
    - $$\mathcal{V}_{\text{context\_need}}(\cdot) \propto \sum_{c \in \text{content}(s_c) \setminus w} f_{r,s_t-\alpha:s_t}(c)$$
    - We remove the word w since it is not part of the context.
- The context need model checks the communicative context to estimate contextual relevance with respect to each region. Both of the above models can also be framed as a majority vote model instead of taking a sum of frequencies:
    - $$\mathcal{V}(\cdot) \propto \sum_{c} \mathbb{1}_{\max_r f_{r,s_t-\alpha:s_t}(c)} f_{r,s_t-\alpha:s_t}(c)$$
    - We find the majority vote scheme to be robust in our experiments as frequency counts of common words could otherwise dominate the estimates.

Semantic Distinction Models

- They modeled semantic variation using historical slang senses associated with the word w in a region r that emerged before st , denoted as Sw,st,r. Under this paradigm, the semantic variation model V can be specified as follows:
    - $$\mathcal{V}_{\text{distinction}}(\cdot) \propto g(s, \mathcal{S}_{w,s_t,r})$$
- Function g can be viewed as a classifier that measures the categorical similarity between the emerging sense s and historical senses from region r. We model g generatively using semantic chaining models from historical word sense extension.
- They adapted three prominent variants of semantic chaining
    1. The one nearest neighbor (onenn) model that only considers the most similar historical sense
    2. The mean exemplar model that accounts for all historical senses

3. The prototype model which collapses all historical sense into a single prototypical sense

- When performing chaining, each sense is represented by embedding its corresponding definition sentence sd using a sentence embedder E:

$$g_{\text{onenn}}(\cdot) = \max_{s' \in \mathcal{S}_{w,s_t,r}} sim(E(s_d), E(s'_d))$$

$$g_{\text{exemplar}}(\cdot) = \overline{\sum_{s' \in \mathcal{S}_{w,s_t,r}} sim(E(s_d), E(s'_d))}$$

$$g_{\text{prototype}}(\cdot) = sim\left(E(s_d), \overline{\sum_{s' \in \mathcal{S}_{w,s_t,r}} E(s'_d)} \right)$$

- The similarity between two sense embeddings is computed using negative exponentiated distance with a learnable kernel width parameter h:

$$sim(e, e') = \exp\left(- \frac{||e - e'||_2^2}{h} \right)$$

Experiments

- Will include this later