

CLINGDINGS

- How do you determine the worth of a language?
- November 6th 2019: Hai, Peng
- Alan Ridel
- Hai Hu 02-19-2020
- Zeeshan 02-19-2020

How do you determine the worth of a language?

How do you determine the worth of a language?

Arle Lommen October 30 2019

2019 is the UN year of indigenous languages.

- Highly idealized language ideals
 - Everyone be able to use their own language as they see fit.
 - Obviously this isn't exactly how this works.

Every language has an intrinsic value

However, in a world of limited resources, not all 7,000 of the

world's languages can be invested in.

less than 1% of content is translated into another language

To cover 100% of content this would take about 20 million translators. This is only for one additional language though.

to cover all 135 economically important languages we would need 2 billion translators.

Let's look at other views of value

the number of speakers does not determine the value of a language for a business

- Though this may be exactly the information that is relevant to NGO's or religious organizations.

Maybe we can look at GDP?

- Could look at GDP per capita to get a sense of the wealth of the individual **I'm not sure why wealth per individual is

Internet adoption rate

- More relevant to today's globally connected tech powered companies.

Pre 2019 CSA was selecting 50? language with online relevance (e.g. usage by communities online).

Calculated number of speakers for each country/territory.

Used a zero-sum approach (no accounting for multi-lingualism).

Assigned languages to four tiers basd on cumulative market research.

This measure is called eGDP or electronic GDP, this is not a measure of ecommerce.

after 2019,

- Added multilingualism
- they expanded from 300 locales to 500 locales.
- Added model of income inequality to help scale GDP (e.g. if 12% of a country's pop is online, those people probably have higher GDP than the average of the whole population)

November 6th 2019: Hai, Peng

Product harm report evaluation

Product harm crises are when products cause incidents and lead to issues and then the public response produces negative publicity for the company and the government body in charge of regulation

Two issues for issuing a recall

- Delayed announcement of recall
 - Food recalls take an average of 57 days after discovery
 - Automotive recalls average of 306 days later (US))
- Low recall completion (small proportion of products that should be recalled are recalled)

Legal wiggle room

- Have to explain how recall was discovered
- What steps were taken to determine whether recall should be done
- free to determine how they release the information and how much information they release

Research questions

do recall communication

examples differ across industries

Hypothesis

- Longer the recall takes the worse the company is viewed
- The more steps taken the more optimistic the more favorably the company is viewed

the idea is that these shape
the way that the company
frames their response.

- The model needs to account for year effects, firm effects, etc.
- dependent variable is linguistic variables
- independent variable is number of steps taken by the company and time taken to report

argument structure is

crucial for previous
research, in addition to
subjectivity measures
difference emerges in
number of content words
(nouns, verbs adjectives
and adverbs)
word (lexical complexity)

- MATTR (moving average ttr)
- STTR mean ttr for every 100 words
- CTr (corrected ttr) $\text{types} / \sqrt{2 * \text{\#tokens}}$

Structural complexity
(length of t-unit +

dependency length (what is a tunit?)

Reading ease score takes into consideration number of syllables per word and number of words per sentence.

However, the number of syllables per word is hard to reliably calculate.

Alan Ridel

We know that there were a large number (25,000 books published during the victorian era) of books, we have a lot of information about gender and year level stats.

no corpus that exists reflects the population of published novels during this period effectively.

The Chadwyck-Healey corpus is particularly bad, 50% of the data comes from male authors published before 1876 even though this was only 15% of the population.

Random sampling of the population is not really possible because we don't actually have a complete database of all novels published during the victorian era.

instead we do quota sampling.

We divide up the population into categories based on year and gender and manually encode a randomly selected chapter.

- Not a representative sample
 - overrepresents authors who wrote more than one novel
 - over represents novels published in multiple volumes

Maybe there's a bias in which things were published or which types of genres tend to do multi volume things

The solution is to use post-stratification as a way to do analysis of granular distinctions after the fact:

- e.g. novels published by women in 1940
- novels involving trains

Hai Hu 02-19-2020

Building a natural language inference dataset in Chinese

What is NLI?

when you have to determine whether a hypothesis contradicts, entails from or is neutral towards a premise.

Issues with SNLI

Turkers do not want contradiction to go both ways.

Bias in hypotheses

If you train on SNLI on just the hypotheses, you get better than majority baseline.

There's bias in the hypotheses One thing is that sleeps contradicts almost any other action. Additional heuristics in the dataset probably introduced by the Turkers probably exist. By creating synthetic data that goes against the heuristics, the result is very very poor performance (19% accuracy for BERT was the best).

XNLI:

- 15 languages
- translated from SNLI/MNLI
 - bad quality translation, lots of things that just don't translate well

Our chinese NLI

- undergrads instead of turkers
- told to write 3 neutral, 3 contradiction, 3 entail as a way of getting them to introduce more variety.
- Students still apply heuristics.
- Issues that emerged:
 - phone call transcriptions are bad
 - use of questions in premises was confusing

Todo

- how to get more variation in hypotheses?
- one annotator only writes Entailments not C/N

Zeeshan 02-19-2020

Internship at Amazon and forthcoming thesis

What is transfer learning?

- Transfer learning is a a variety of different things. For a taxonomy read Ruder 2019.
- pretraining of word embeddings is probably the most famous form of transfer learning.

Multi-task learning

Hard vs soft parameter sharing Hard parameter sharing literally shares some of the initial layers and then has task specific layers towards the end.

Soft parameter sharing uses soem method of regularization to force common layers for the two tasks to be close to eachother.