How Much is Said in a Tweet?

A Multilingual, Information-theoretic Perspective

Graham Neubig & Kevin Duh {neubig,kevinduh}@is.naist.jp

Research Questions

1. How much information is contained in a tweet?
2. Does information content vary from language to language?
3. If so, is it due to language characteristics or behavioral factors?

Methodology

Data

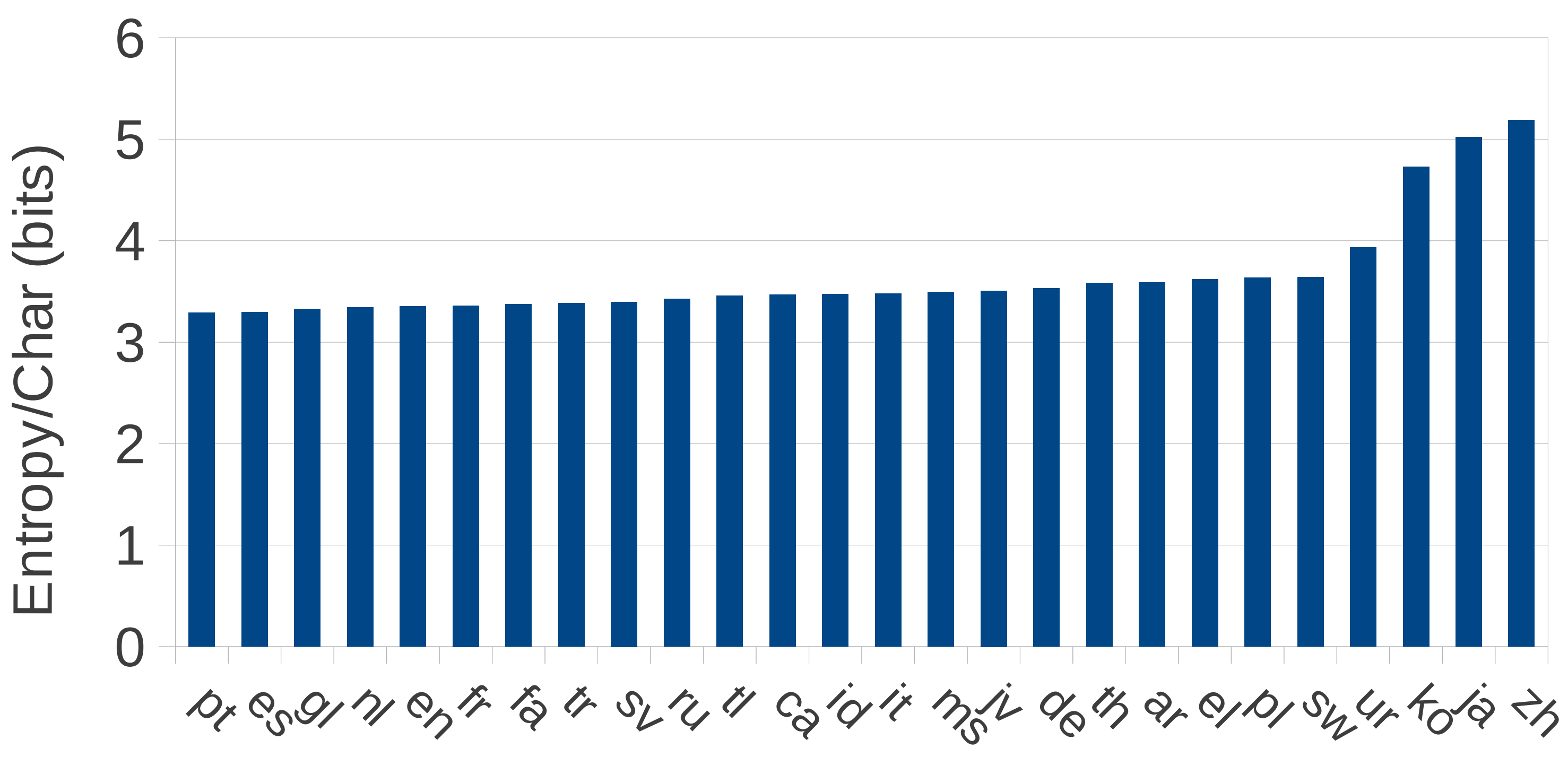
- Collected 6 weeks of tweets in 2012
- Language ID by langid.py → 92M tweets with 95%+ confidence
- Final dataset: 26 languages with more than 50k tweets

Measuring Information

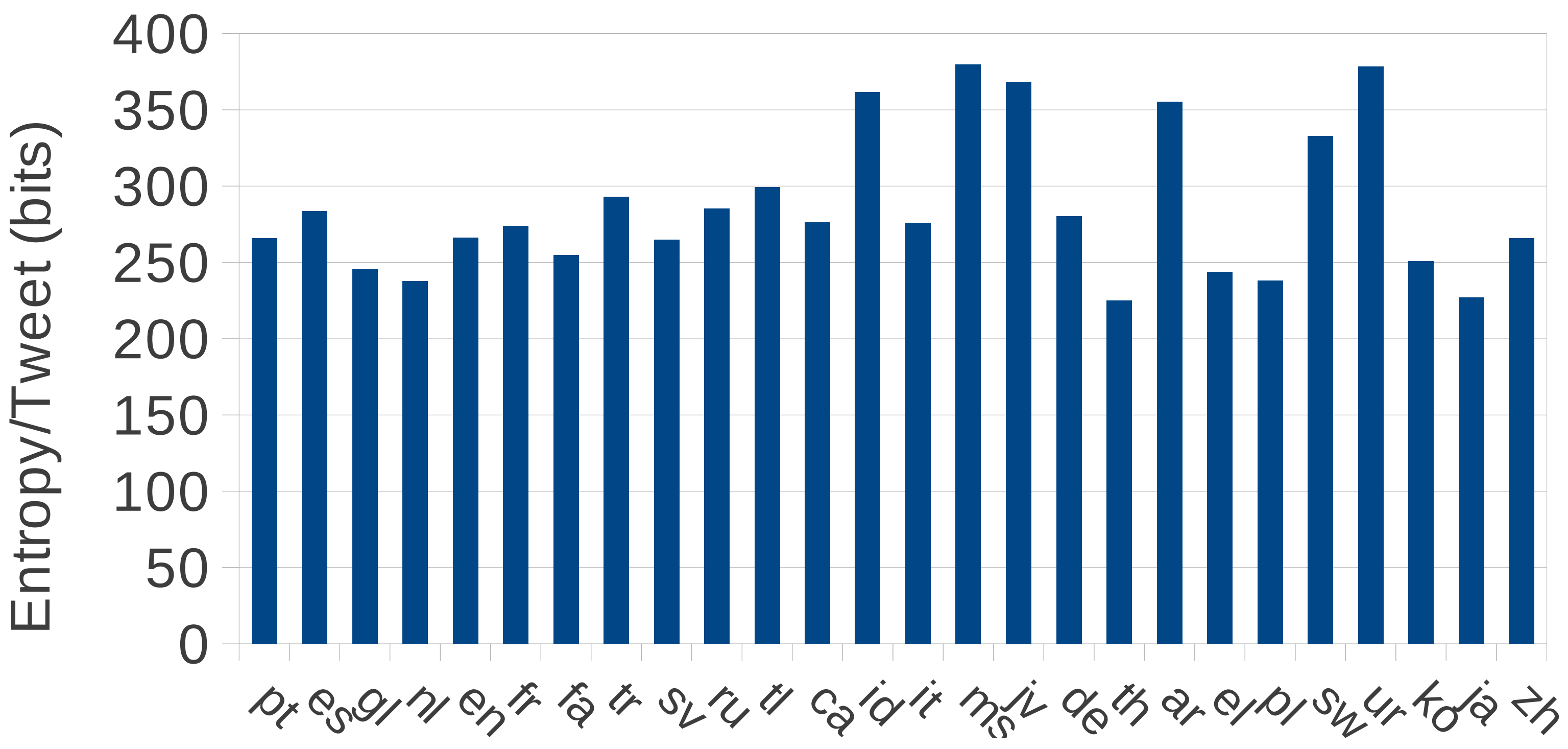
- Entropy = -log Probability(string), by character 7-gram
- Report average entropy of 10-fold cross-validation

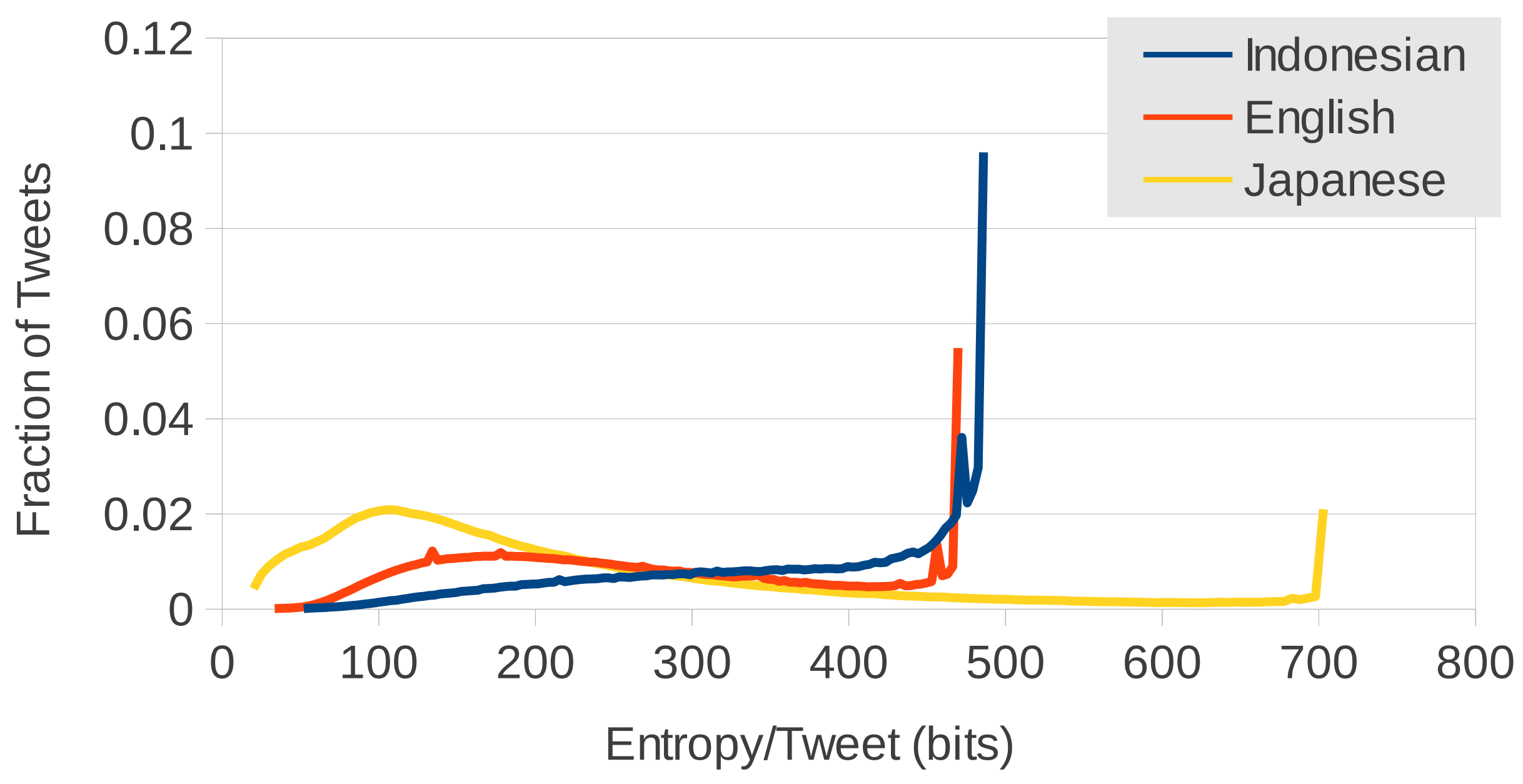
Finding 1: Multilingual Comparison

Languages with larger character sets (e.g. Japanese, Chinese) contain more information per character, as expected.



But language characteristics have little correlation with information per tweet, as authors do not use full 140chars.





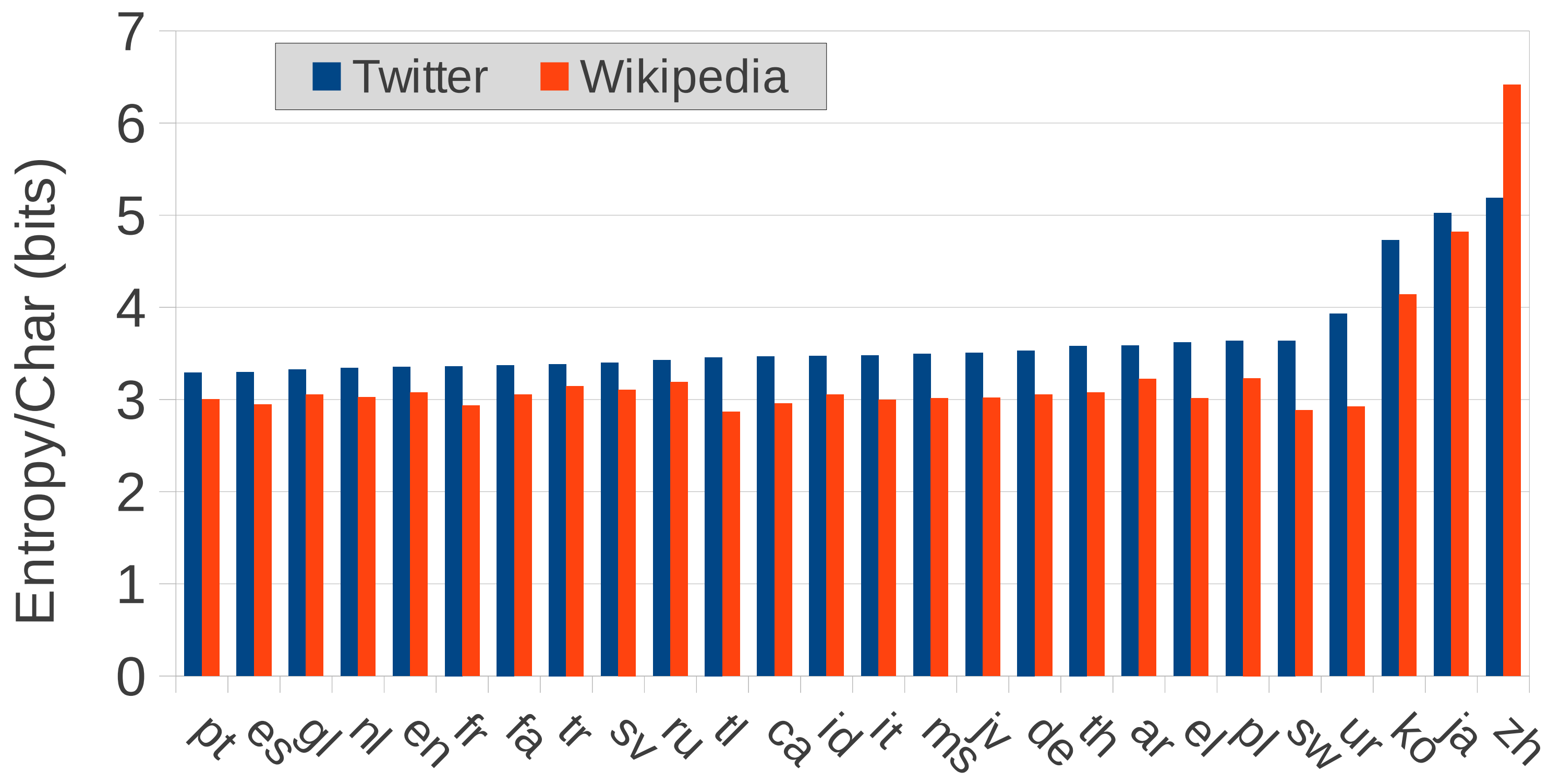
Behavioral factors, e.g. propensity to quote, are better predictors of information per tweet in a language.

Fraction of Variance analysis:
$R^2 = 1$ - residual/variance

Factors	All 26	Latin
Char Set Size	0.5%	19.3%
Char/Word	15.0%	14.6%
Twitter terms	18.5%	**71.8%**
Retweet ratio	0.5%	19.2%
Quote ratio	**43.8%**	**80.2%**

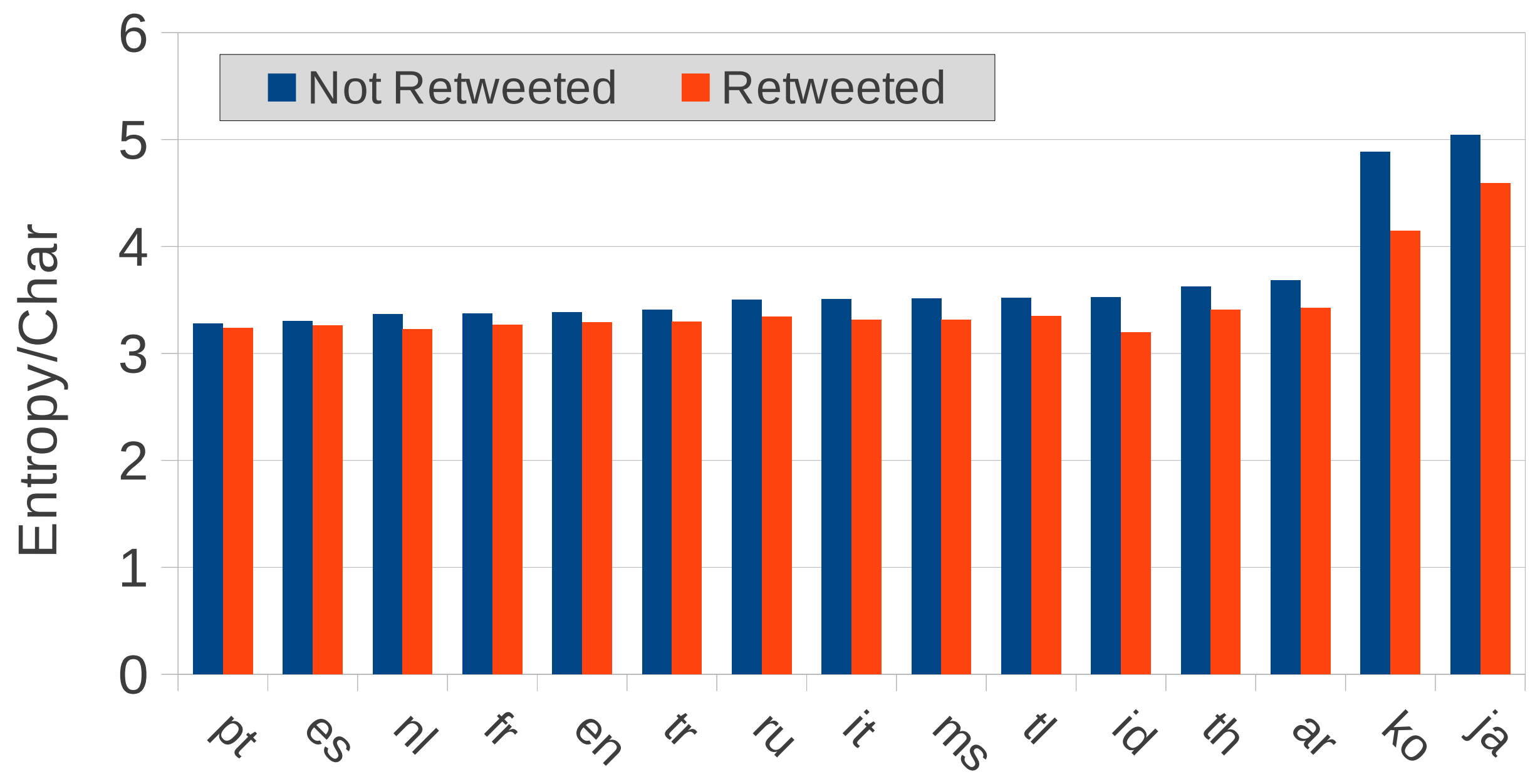
Finding 2: Comparison with Canonical Text

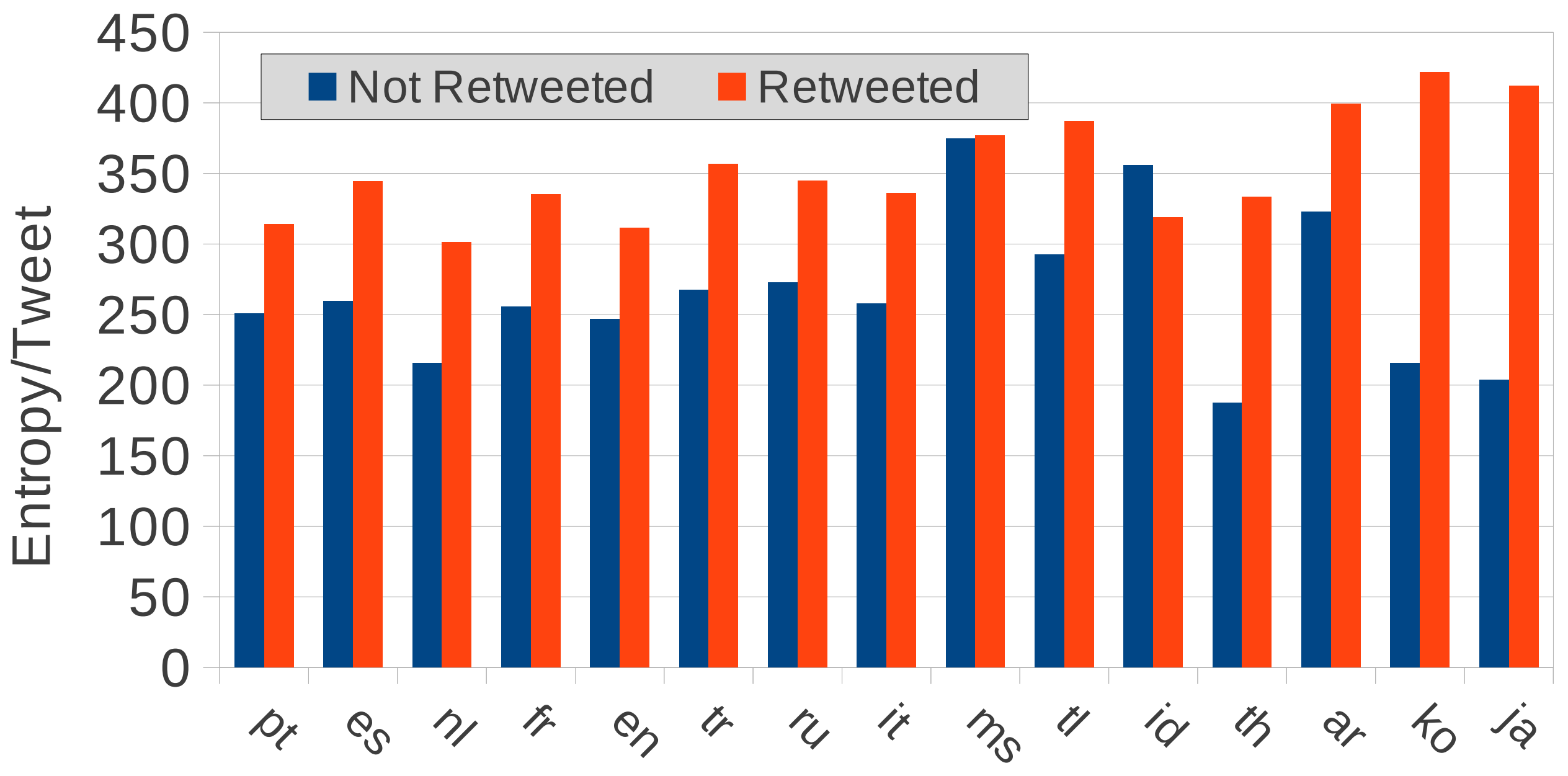
Tweets generally contain more information per character compared to canonical Wikipedia text.
- reasons: abbreviations and less consistent writing style
- exception in Chinese: more characters (63% new) in Wiki



Finding 3: Retweet vs Non-retweet

Retweets have more information per tweet, but less information per character.





Ideas for Future Work

- Learning how authors compress information under constraints
- Measuring information in multiple tweet discourse
- Analyzing tweets that cross linguistic/geographic boundaries