Acronym Translation: A Corner Case for Standard Machine Translation Systems

A research project by Richard Yue and Kenneth Church

ABSTRACT

- Translating acronyms is hard
- Task: predict Google correctness
- Examples:
  - *« De nombreux facteurs de risque participent au développement de cette pathologie, parmi lesquels les **acides gras trans** (**AGT**). »*
  - *« […]une diminution d'expression de 12 gènes mutés dans **l'anémie de Fanconi** (**AF**) »*

Input	
French short form	**French long form**
AGT	Acides gras trans
AF	Anémie de Fanconi

Output		
Google Translation	**Correct short form**	**English long form**
TGA (incorrect)	TFA	Trans fatty acids
AF (incorrect)	FA	Fanconi Anemia

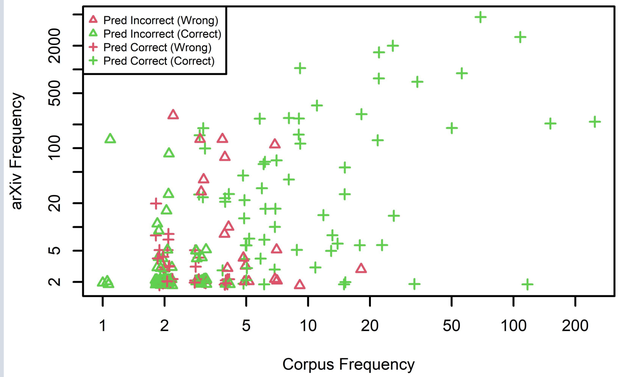
400 English and French acronyms

DATA SET

Item	Metric
Our Corpus	13.4k abstracts
Tokens	6.2M
Types	66k
Acronyms	200 analyzed
Google errors	Estimated 10k

RESULTS

$$Google\ is\ correct \sim \log(arXivfreq) + \log(corpusfreq)$$



CONCLUSION

- Google Translate works well for frequent acronyms
- We use logistic regression to predict Google correctness as a function of two frequencies