USING N-GRAM SIMILARITY METRICS FOR RELATION CLUSTERING

Stephen Mayhew, Nicholas Kamper

mayhewsw@rose-hulman.edu
kampernj@rose-hulman.edu
Rose-Hulman Institute of Technology

1. INTRODUCTION

2. EXPERIMENTS

Use HMM for the actual clustering. Dead wall.

HMM for scoring (Multiple sequence alignment). Dead wall.

Decided to ditch HMM altogether.

How about using N-grams? (Considered using Google data) Started writing our own code in Java (because of legacy code)

Going slowly, tried several libraries: NLTK, OpenNLP, KYLM, and some other small open source projects.

Finally, we chanced on Python ngram similarity library that uses ngrams to predict similarity between two strings.

"The NGram class is a set that supports searching for its members by N-Gram string similarity." [1]

3. REFERENCES

[1] Michel Albert, Graham Poulter, and open source contributors. Python NGram Similarity Library. *http://packages.python.org/ngram/*