Project Report

Leroy Wang¹, Gibbs Geng¹Lavinia Dunagan¹
¹UW

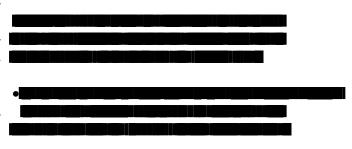
Approach Our team developed a **Speed-Oriented N-gram System** for this project. Our system (**SONS**) only looks at the last N characters (we call this an N-gram) of each input string. **SONS** has stored a large dictionary that has the following {key: value} pairs: { N-gram: top 3 most frequent characters following the N-gram}. For each N-gram, **SONS** will check if it is in **SONS**'s dictionary. If it is, **SONS** will output the corresponding 3 most frequent characters given the N-gram.



Data Our training data comes from 3 sources: *Europarl Corpus*,

Our **SONS** was trained on the following 21 European languages using *Europarl Corpus*:

• English, Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Dutch Polish, Portuguese, Romanian, Slovak, Slovene, Swedish



Implementation

(It's worth noting that our Checkpoint 3 system was extraordinarily fast - it takes 30 ms to process 180k multilingual test instances while achieving 34% accuracy, when testing on our machine)

Training When our SONS was learning its 1-gram dictionary, it will only include characters that appear more than 5 times in the training data.