Qing Dou and Ashish Vaswani

1/26/2015

Experiments:

Task:

Spanish – English decipherment

(top 5k frequent word types)

Data:

Spanish Gigaword

English Gigaword

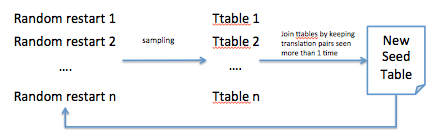
English dependency language model built using 100 million tokens

Embeddings:

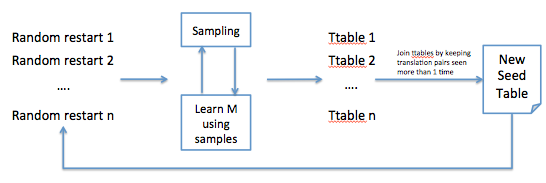
Use word2vec to build word embeddings for both English and Spanish using approx. 1 billion tokens each.

Sampling Procedure:

Baseline:



With Embeddings:



Results:

Systems:

Baseline

**Embed-25**: use vectors with 25 dimensions

**Embed-50**: use vectors with 50 dimensions

Evaluation:

Gold: build a translation table using 1m token parallel data

Decipherment Accuracy: number of correct / total

Total is the number of word types in the gold ttable and also in the top 5000 frequent word types (4229)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 100k tokens | 1m tokens | 10m tokens |
| Baseline | 1.8%(587) | 7.2% (2964) | 27%(4530) |
| Embed-25 | 5.0%(1374) | 20.0%(3347) | 53.7%(4861) |
| Embed-50 | 8.9%(1573) | 22.7%(3414) | N/A |

(numbers in brackets show the number of word types in test ttable)

Malagasy-English:

Data:

Malagasy: 1.9m parallel data + 16m web crawled news data

English: 1.8m parallel data + 400m Gigaword + 300m African news

Accuracy on top 5000 frequent word types in parallel data

|  |  |  |  |
| --- | --- | --- | --- |
|  | 100k tokens | 1m tokens | 10m tokens |
| Baseline | 1.2%(318) | 2.0% (1920) | 4.0%(4526) |
| Embed-25 | 1.8%(827) | 3.1%(2220) | 5.4%(4665) |
| Embed-50 | 1.7%(751) | 3.8%(2252) | N/A |

(numbers in brackets show the number of word types in test ttable)