Big Data Small Data, In Domain Out-of Domain, Known Word Unknown Word: The Impact of Word Representations on Sequence Labelling Tasks

Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou,
Nathan Schneider & Timothy Baldwin

7 July 2015

Australian
National
University

What we want to know about word embeddings

Research questions

RQ1: Are word embeddings better than baseline approaches of one-hot unigram features and Brown clusters?

RQ2: Do word embeddings require less training data than one-hot unigram features? If so, to what degree can word embeddings reduce the amount of labelled data?

RQ3: What is the impact of updating word embeddings in sequence labelling tasks, both empirically over the target task and geometrically over the vectors?

RQ4: What is the impact of these word embeddings (with and without updating) on both OOV items (relative to the training data) and out-of-domain data?

RQ5: Are some word embeddings better than others in a sequence labelling context?

Experiments Setup

Fix the experiments conditions

- **Pre-trained word embeddings**
  - Corpora
    - UMBC [**?**], 48.1GB
    - One Billion [**?**], 4.1GB
    - English Wikipedia, 49.6GB
  - Embedding dimensions: $d \in \{25, 50, 100, 200\}$
  - Context window size: $m \in \{1, 5, 10\}$
- **Selected Word Representations**
  - Brown clustering [**?**]
  - CBOW [**?**],
  - Skipgram [**?**],
  - Global vectors [**?**]
- **Four sequence labeling tasks**
  - POS-Tagging,
  - Chunking,
  - NER,
  - MWE (Multi-word Expressions identification)

Training, development and test data

	Training	Development	*In-domain* Test	*Out-of-domain* Test
POS tagging	WSJSec. 0-18	WSJSec. 19–21	WSJSec. 22–24	EWT
Chunking	WSJ	WSJ(1K sentences)	WSJ(CoNLL-00 test)	Brown
NER	Reuters(CoNLL-03 train)	Reuters(CoNLL-03 dev)	Reuters(CoNLL-03 test)	MUC7
MWE	EWT(500 docs)	EWT(100 docs)	EWT(123 docs)	—

RQ1: Word embeddings vs. one-hot unigram features vs. Brown clusters?

RQ2: Do word embeddings require less training data than one-hot unigram features?

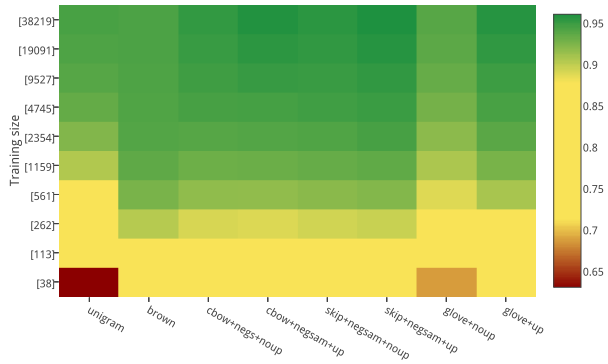

Figure : POS tagging(Acc)

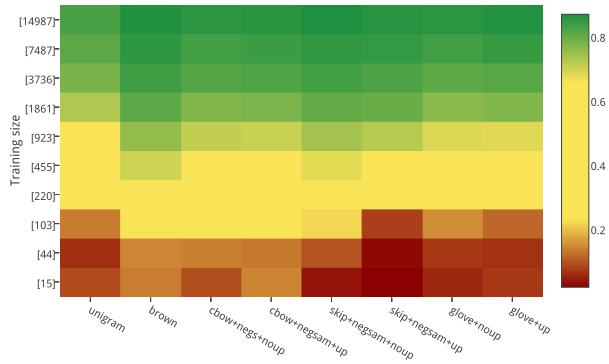

Figure : NER(F1)

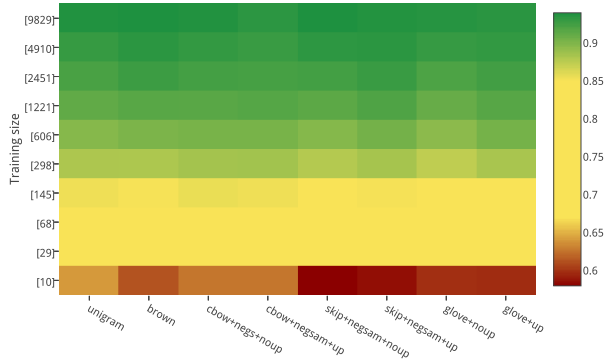

Figure : Chunking(F1)

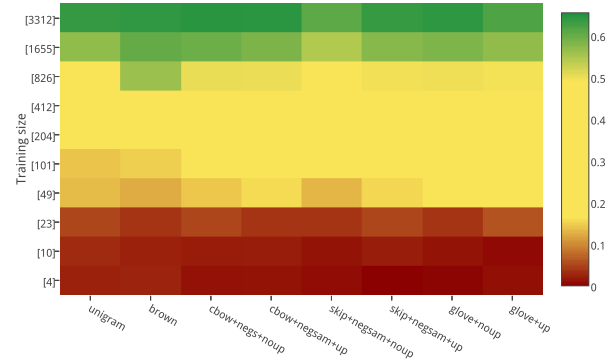

Figure : MWE(F1)

RQ3: What is the impact of updating word embeddings geometrically over the vectors?

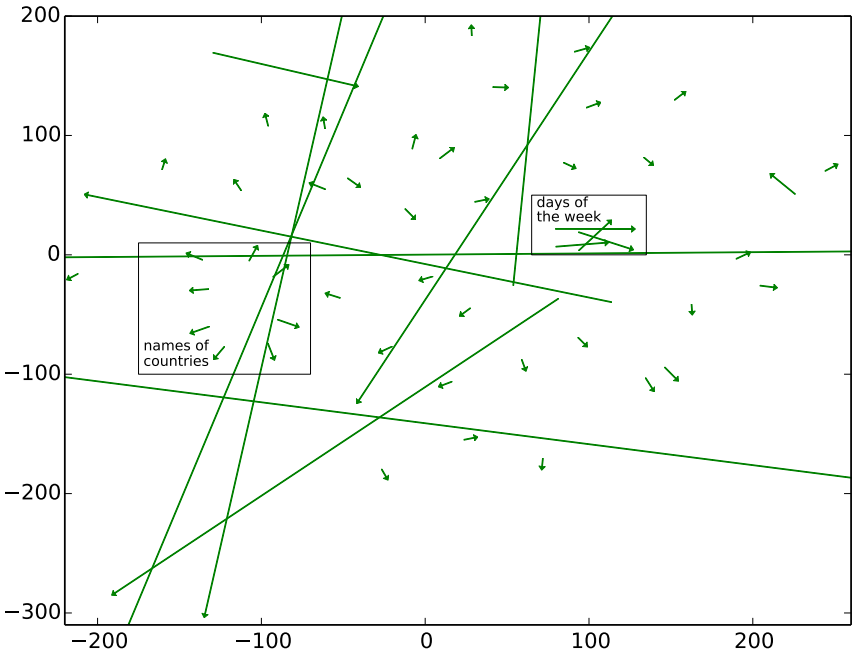

Figure : Chunkingwith SKIP-GRAM

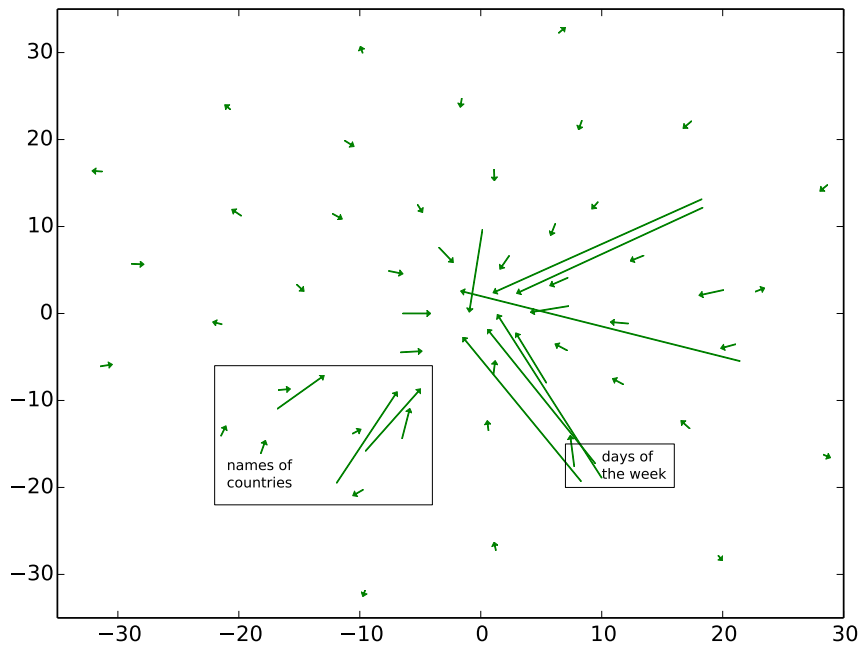

Figure : NERwith SKIP-GRAM

RQ4: What is the impact of these word embeddings (with and without updating) on both OOV items (relative to the training data) and out-of-domain data?

RQ5: Are some word embeddings better than others in a sequence labelling context?