EXPLORING DIFFERENT WORD EMBEDDINGS FOR UNSUPERVISED PART-OF-SPEECH TAGGING

MSc dissertation

Felicia Liu

November 14, 2016

Overview

- Introduction to thesis topic
- Research objective
- Methods & Approach
- Results

Introduction to research topic

Exploring word embeddings for unsupervised part-of-speech tagging.

POS tagging Assigning labels to individual words from a given text that denote its syntactical function

Unsupervised Only use unlabeled data

Word embeddings Vector representations of words, trained to encode a word's features

Use Gaussian Hidden Markov Model for POS tagging.

Research objective

- Use word embeddings to improve POS tagging.
- Encode relevant information in these embeddings, such as syntax or morphology.

How do we create such word embeddings?

- Small context size window
- Character-level word embeddings

Methods & Approach

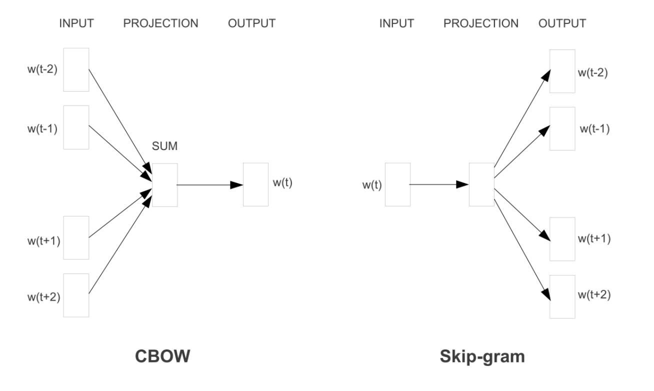
- Use one data set to train all embeddings.
- Train embeddings with vector sizes 20, 50, 100, and 200, vary other parameters
- Use V-measure to assess performance with gold-standard labeled text.

Methods & Approach

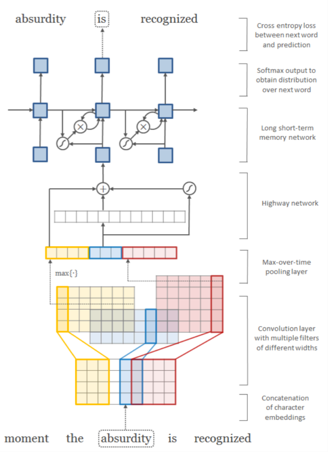
Word embeddings used in this work include

- SENNA (Collobert et al., 2011)
- GloVe (Pennington et al., 2014)
- word2vec (Mikolov et al., 2013)
- Structured word2vec (Ling et al., 2015)
- Character-level embeddings (Kim et al., 2016)

WORD2VEC MODEL



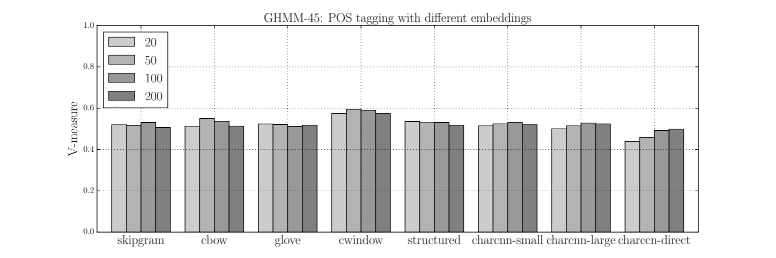
Character-level embeddings model



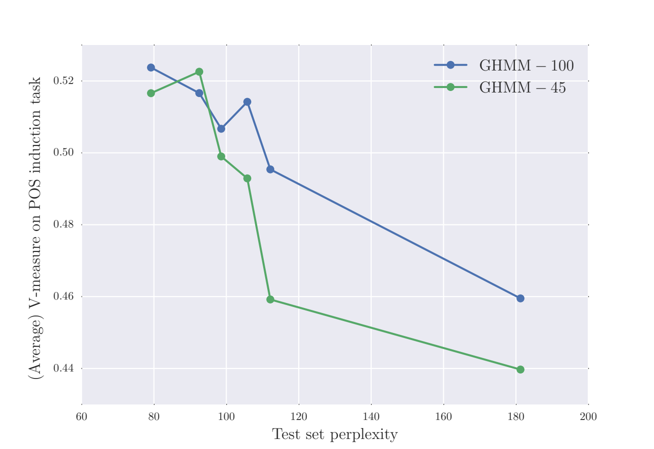
Other experiments

- Varying the level of tokenisation of input data
- Reducing dimensionality of embeddings obtained through the character-level model
- Combining word- and character-level embeddings

Results on POS tagging



Influence of perplexity on results



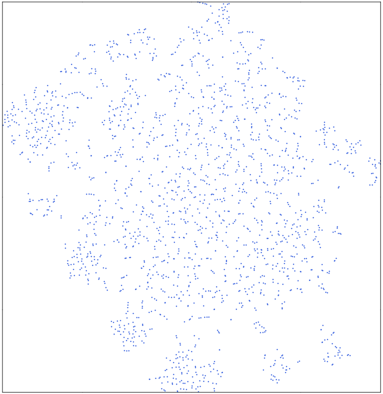
Nearest neighbours 1

word	skip-100	CBOW-50	GloVe-20	cwin-50	struct-20	SENNA	Large-100	Small-100
	woman	woman	true	woman	woman	letter	peterson	peterson
	persons	child	man	child	firm	case	persons	pension
person	anyone	settlor	result	man	child	honor	pearson	persons
	patient	spouse	real	nation	hobby	ages	emerson	pearson
	victim	persons	woman	girl	demon	problem	patterson	poisson
	kings	hussa	prince	prince	queen	lord	ping	ping
	queen	theodric	alexander	queen	lord	queen	ming	qing
king	harthacanute	kings	charles	captain	lady	emperor	qing	kind
	mordha	sillok	edward	bishop	prince	titles	kind	bing
	monarch	culen	henry	lord	grace	fighting	ring	kong
	spain	belgium	rome	spain	cuba	santa	franc	trance
	italy	italy	germany	belgium	luxembourg	composition	franco	franc
france	belgium	spain	spain	italy	guatemala	germany	frances	franco
	germany	bordeaux	portugal	austria	portugal	italy	franca	ordnance
	luxembourg	luxembourg	japan	poland	peru	sweden	francs	franca
	grayish	grayish	lime	yellowish	whitish	violet	resisted	yiddish
	greyish	greenish	honey	grayish	wavy	territorial	yiddish	swedish
reddish	yellowish	blackish	trout	bluish	feathered	academia	revised	rush
	greenish	purplish	yellow	mottled	coppery	aggregate	registered	dish
	bluish	irides	fat	red	bulbous	tore	swedish	irish
	robert	robert	robert	william	william	robert	richards	orchard
	walter	philip	george	harold	harold	peter	richardson	richards
richard	francis	william	francis	robert	stephen	reportedly	orchard	richardson
	hugh	ralph	james	charles	albert	david	richland	richland
	arthur	john	thomas	hugh	john	william	archaic	richmond

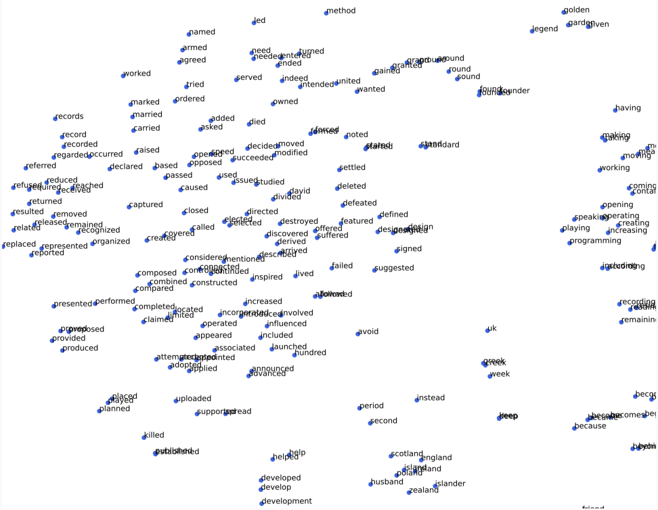
Nearest neighbours 2

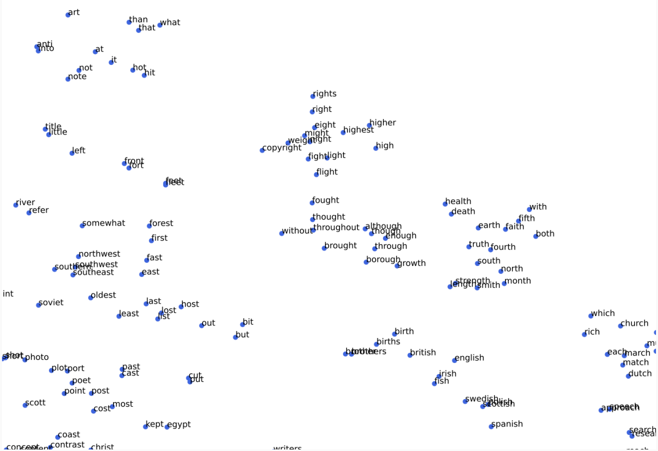
word	skip-100	CBOW-50	GloVe-20	cwin-50	struct-20	SENNA	Large-100	Small-100
inconsiderable	denominate	sphenic	endear	obligatory	destabilizing		considerable	considerable
	policyholders	discounting	materialise	signifigant	satisfactory		inconquerable	indecomposable
	signifigant	surprising	disappoint	unimpeachable	foolproof		inconsolable	inconquerable
	extortionate	denormal	congenial	unaffordable	scientifically		insufferable	incommensuable
	substantial	staggering	distasteful	explicit	realigning		imponderable	unconscionable
unsteadiness	neuralgic	equilibrioception	zelotes	decompensated	thymoma		unreadiness	unreadiness
	schizotypy	neuralgic	christopher	dyserythropoietic	mechnical		steadiness	uneasiness
	insensibility	persecutory	wck	microcornea	hemiplegic		untidiness	steadiness
	vestibulo	schizotypy	ranko	expressivity	hyperglycemic		uneasiness	untidiness
	unnaturalness	diffractive	hypernatremia	morphea	indirectness		steadfastness	sturdiness
commenting	commented	commented	insisting	insisting	insisted	hackers	committing	committing
	gloating	criticizing	sells	insists	insisting	possessed	competing	commemorating
	raved	insisting	binds	focusing	commented	corvette	commanding	commanding
	remarked	remarking	insists	concentrating	speculating	cyborg	coming	coming
	joked	discussing	enabled	insisted	insists	jtdirl	connecting	competing
comment	comments	comments	request	remark	consensus	lyon	commencement	commitment
	remarks	reply	notice	report	slump	marco	commitment	commencement
	reply	remark	finding	notice	shame	nan	competent	component
	remark	remarks	account	statement	ban	thebes	clement	competent
	quip	quip	permission	commentary	commentary	orchestras	comments	clement
unaffected	affected	obscured	disruped	affected	affected	micronesia	affected	infected
	obscured	disturbed	outdated	disrupted	disrupted	baptist	unwanted	affected
	disturbed	hindered	motivated	regulated	obscured	brett	unidentified	inflicted
	untouched	affected	unreliable	damaged	enforced	apparatus	uninhabited	effected
	evident	compromised	unstable	overwhelmed	regulated	trenton	unfinished	unidentified
affect	impair	contribute	reject	reflect	utilize	ngo	effect	effect
	contribute	impair	observe	disrupt	involve	paula	affects	affects
	depend	reflect	recognize	eclude	activate	cabin	affected	affected
	affecting	alter	deny	involve	utilise	collier	affecting	affecting
	affects	relate	arise	satisfy	overwhelm	mentally	affection	defect

t-SNE applied to character-level embeddings

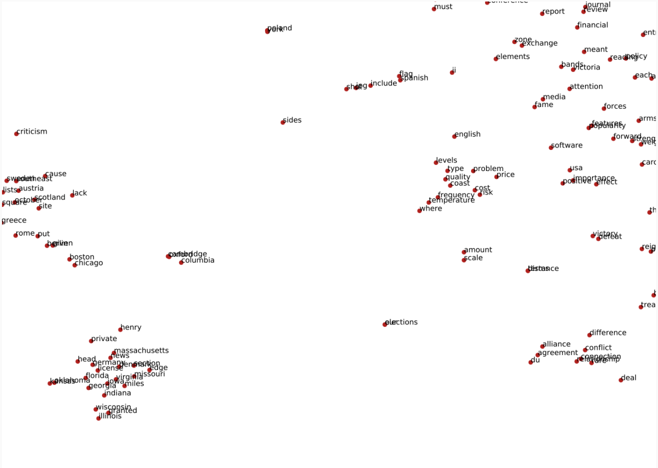


Character-level embeddings clusters 1




art
than
what
that
anti
into
at
it
not
hot
hit
note
rights
right
eight
higher
might
highest
weight
night
copyright
high
fight
light
flight
title
little
left
front
fort
feet
fleet
river
refer
fought
health
with
death
thought
fifth
somewhat
forest
without
throughout
although
though
earth
faith
enough
both
first
brought
through
truth
fourth
northwest
fast
southern
southwest
borough
growth
south
southeast
east
north
strength
length
smith
month
int
oldest
soviet
last
host
lost
list
least
which
out
bit
church
but
birth
rich
births
short
shot
photo
brothers
british
english
each
march
match
plot
port
past
irish
cast
dutch
cut
put
fish
poet
point
post
swedish
polish
scottish
scott
most
cost
speech
approach
kept
egypt
spanish
search
coast
contrast
christ
concept
writers

t-SNE applied to visualise word-level embeddings


must
report
journal
review
financial
poland
zone
exchange
meant
elements
policy
bands
victoria
each
flag
spanish
include
attention
media
fame
forces
sides
criticism
features
popularity
english
forward
software
levels
type
problem
price
quality
coast
usa
importance
positive
effect
cost
risk
frequency
temperature
where
sweden
southeast
cause
lists
austria
lack
scotland
square
october
site
greece
rome
put
boston
chicago
columbia
victory
defeat
amount
scale
distance
elections
difference
henry
private
massachusetts
news
head
germany
section
license
edge
missouri
florida
virginia
iowa
miles
georgia
indiana
wisconsin
granted
illinois
alliance
conflict
agreement
connection
du
deal

Conclusion

- Structured word embeddings do better than unstructured.
- Continuous bag-of-words (CBOW) performs better than skipgram.
- Character-level embeddings do not show competitive performance when used alone, can improve performance when combined.
- Perplexity of language model is correlated with embedding performance on POS tagging task.

Questions?