Identify Reviewers from their Comments

Mengyi Sun
Department of Ecology and Evolutionary Biology
University of Michigan, Ann Arbor
mengysun@umich.edu

Shang Zhang
Department of Physics
University of Michigan, Ann Arbor
zhshang@umich.edu

Abstract

The peer-review process is quintessential for ensuring the validity of scientific research and provides assessment about the importance of scientific publications. By definition, the peer-review process should be critical and rigorous. To guarantee this, it is important to ensure reviewers can express their opinions without the concerns of targeted harassment from the authors. Therefore, in most peer review process, reviewers are anonymous. Yet, reviewer comments unavoidably provide some information about the identities of the reviewers. In our work, we showed that once we can narrow down the reviewers to a reasonable amount of candidates, simple author attribution algorithms trained on the public-available corpus from the candidates allow identifying the reviewers with relatively high accuracy. Our work has important implications for privacy protection during and after the peer-review process.

1 Introduction

The importance of the peer-review process in scientific research can hardly be overemphasized. Peer review mainly serves two functions(Kelly and Adeli, 2014): 1) it assesses the validity and importance of scientific publication; 2) it helps the authors of the publication improve their work. For the well-functioning of peer review, most of the reviewing processes are so-called single-blind peer review processes. Namely, reviewers' identities are anonymous to the authors. This policy serves to protect the reviewers from the potential resentments from the authors so that they can express their comments freely. Most of the journals also do not provide the reviewers' comments to the general public when a paper is published. Nevertheless, since reviewers' comments are important information for the general public to assess the published paper, more and more journals start to publish reviewer comments while holding the identity of reviewers confidential (Polka et al., 2018). Most importantly, to help the authors improve their work and to ensure the fairness of the peer review process, most reviewers' comments would (and should) be provided to the authors even when the reviewed manuscripts are rejected. However, the reviewers' comments unavoidably provide some clues about the identities of the reviewers, such as the research field that the reviewer might come from, their opinion toward certain statements, or the related publications that they are familiar with. These clues might be quite revealing about who the reviewers might be. With the advancement of natural language processing techniques (Argamon et al., 2009),more information, such as gender, age, and even native language can be obtained from the (unmodified) comments. Therefore, it is of interest to ask whether or not the anonymity really holds well during the reviewing process. In order to address this question, we manually collected reviewer comments of 25 prolific reviewers from the journal Biology Direct. We further manually collected their scientific publications from PubMed.We trained two simple models on their scientific publications, and we show that we can achieve relatively high accuracy ($40 \sim 50$ percent) in predicting the identities of the reviewers from their reviewer comments.

Our work will definitely be of interest for the academic community as a whole, given that the vast majority of the reviewers choose not to reveal their identities during the reviewing process(Bravo et al., 2019). Moreover, our work can provide information for designing algorithms to re-anonymize the reviewer comments, with a focus on re-anonymizing the comments of vulnerable reviewers whose comments disclose too much information about their identities.

(Disclaimer: we hold no position about whether or not the reviewing process should be fully transparent (in fact, at least one of the authors favor reviewers that signed their names during the reviewing process). However, we believe that any policy should be well informed by its potential outcomes.)

2 Problem Definition and data

Our problem is a closed-set authorship attribution problem (Stamatatos, 2009). Closed-set authorship attribution refers to tasks that require inferring the author of a corpus by selecting from a set of candidate authors. Specifically, our problem can be treated as cross-domain authorship attribution: we predict authors of reviewer comments by a classifier trained on their scientific publications, which are not reviewer comments. In our final dataset, we have 25 candidate authors. We collected their reviewer comments from Biology Direct (`https://biologydirect.biomedcentral.com/`), an open-access journal with unique publication philosophy.They published both the reviewers' comments and their names.We choose the 25 most prolific authors with at least ten reviewer comments as our candidates. For each candidate, we retrieved the ten longest reviewer comments. We then downloaded their publications from PubMed (`https://www.ncbi.nlm.nih.gov/pubmed/`) manually. For each candidate, we manually collected those papers that the candidate serves as the correspondent author or first author, prioritizing single-author papers. We divided the reviewer comments into a dev set and a test set (with fix random seed, to make the results across different algorithms comparable). Each set contains five reviewer comments. Our training set is the reviewers' published papers. Because we manually collect the data, the text is quite clean except for some copy-paste errors from pdf files, and misrepresentation of mathematical symbols.

We use table 1 to give one example of the 25 reviewers we collected. And table 2 describes some simple statistics of our dataset.

3 Related Work

There is very little previous research on identifying reviewers from their comments. The only publication (Nanavati, 2011) with regard to identifying reviewers from their comments formulate the question as an in-domain authorship attribution task: they try to predict reviewer identity from their other reviews with known authorship. This is not a very realistic setting because, in most cases, we do not have access to other reviewer comments of the candidates. Notice that although their results are not directly comparable to ours, they showed that simple algorithms can perform surprisingly well: a naive bayes classifier with a set of tf-idf words as features (which, by the way is not the most appropriate feature in authorship attribution) can reach 75 percent accuracy on conference reviews with ~ 15 candidates. This suggested that it might not be very hard to identify reviewers from their comments and henceforth put serious concern about the confidentiality of the peer-review process. Albeit there are no previous tasks that focus on identifying reviewers using classifiers trained on publicly available scientific publications, authorship attribution (Stamatatos, 2009) is an area with rich NLP research history. So there are a lot of relevant researches, such as predicting author from their citations (Hill and Provost, 2003), cross-domain fan-fiction authorship attribution (Kestemont, 2018), etc. The performance of authorship attribution depends a lot on the specific task, so it is hard to predict what should be the expected baseline performance, and it is generally hard to compare the performance for methodologies from different tasks. However, in previous studies, the vast majority of best author attribution algorithms use support vector machines, and henceforth support vector machine is widely used as a baseline. Therefore, in our task, we use support vector machine as one of the baselines.

4 Methodology

We implemented simple logistic regression in our task. We reasoned that since our data set is small, simple logistic regression might perform better than complicate learning techniques, which is also true for authorship attribution in general (Kestemont, 2018). Our features for classification are the top 500 most frequent words, the top 1000 most frequent 3-grams of English characters and punctuation, and 26 types of part of speech tags in Penn Treebank tagsets(Marcus et al., 1993). The top 500 words and top 1000 3-grams were extracted from all reviewer comments unsupervisedly. Notice we do not explicitly and systematically use dev set for tuning parameters, but we do use dev set to empirically help us narrow down efficient features (in that sense, the dev set is still dev set

Reviewer name	# of reviews	# of single author publications	# of correspondent author publications
Pierre Pontarotti	10	3	7

Table 1: One example of collected data

Simple statistics	
Total number of reviewers	25
Total number of reviews	25×10
Total number of publications	25×10
Average number of words per review	678.08
Average number of words per publication	4518.25

Table 2: Description of the dataset

for us). To prevent overfitting, we use AdamGrad as our optimizer, and we stop the training early– we only trained for ten epochs.

5 Evaluation and Results

We evaluated the overall performance of our algorithms by accuracy. Since the number of the corpus of each author equals to each other in our training set, dev set, and test set, the accuracy will be equal to micro-precision and micro-recall, and henceforth evaluating the accuracy is equivalent to evaluating the micro-f1 score. We have two baseline algorithms: 1) Randomly sample candidates based on their frequency; 2) Support vector machine with a linear kernel using the same set of features we used in our logistic regression. We use 5-fold cross-validation on training corpus (scientific publications) for selecting the best C parameters for support vector machine classification using micro f-score. Because of the evenness of our dataset, random sampling from the authors performed very badly. As expected, the random guess can only guess correctly around 4 percent of the time (table 3). Both the support vector machine baseline and our logistic regression algorithm perform much better (table 3). And the logistic regression can indeed outperform the support vector machine in this case, both for the dev set and the test set. In table 3, we showed the accuracy of different methodologies in both the dev set and test set. Besides the selected features of top 500 words-top 1000 3-grams-26 POS tags as we discussed, we also included the performance of our algorithms using a smaller dimension of selected features (top 50 words-top 50 3-grams-26 POS tags) as a comparison.

We also provide a plot of the confusion matrix of test set results of our logistic regression model (fig 1).

6 Discussion

Our simple logistic regression without complicated tuning outperforms both the random sample baseline and the support vector machine baseline. The main reason we think is because of the small scale nature of our dataset, for which simple algorithms generally perform better. Interestingly, we find a positive correlation between the total length of training corpus and the f-score on test-set for each class label (Spearman's rho=0.47, p=0.016), suggesting that one way to boost the performance is to collect more data. Moreover, we do not find a correlation between the fraction of words of the single-author corpus in the training set and the f1-scores for each class label. This can be due to: (1) the sample size is small (n=25),(2) the first author or senior author write a large portion of the published scientific corpus. If the second point is true, we can collect a lot of training corpus, especially for prolific scientists. And if the dataset is significantly enlarged, the support vector machine and even deep learning might be more suitable for those tasks. We have to emphasize that our results are preliminary, given the time limit of the course, although the close to 50 percent accuracy (we are surprised by that) is enough for sending a warning signal for the current scientific reviewing system. Nevertheless, to illustrate the issues, we would need to enlarge the dataset to include more candidates (with more training data, of course).

7 Conclusion

Our current results suggested that simple algorithms can be quite accurate in identifying anonymous reviewers, at least if we have a reasonable number of author candidates. In general, as the

Accuracy	dev-set	test-set
Random sample	4.8%	4.8%
SVM (top 50 words-top 50 3-grams-26 POS tags), C = 0.001	37.6%	44.8%
SVM (top 500 words-top 1000 3-grams-26 POS tags), C = 0.005	44.8%	53.6%
LR with AdamGrad (top 50 words-top 50 3-grams-26 POS tags)	29.6%	36%
LR with AdamGrad (top 500 words-top 1000 3-grams-26 POS tags)	**47.2%**	**55.2%**

Table 3: Results of different methodology and different feature selections

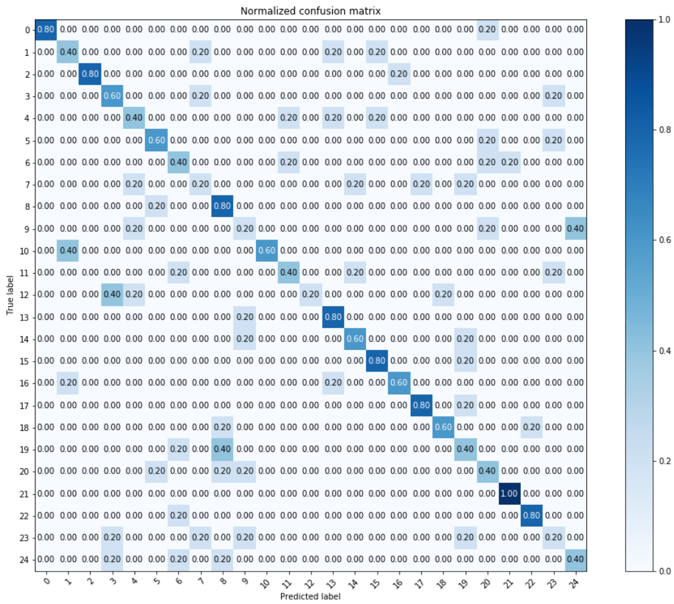

Figure 1: Confusion matrix of the 25 reviewers in test sets with the model LR with AdamGrad (top 500 words-top 1000 3-grams-26 POS tags)

candidate number increases, the attribution becomes less accurate, so whether or not the algorithm will perform well depends on whether or not in the real world we can narrow down the reviewers into a small set of candidates. The precise magnitude of a 'reasonable' number of candidates will be the interest of our future research.

8 Other Things We Tried

At the very beginning we try to write manuscript to automatically download the correspondent text and clean the data. However, given the heterogeneity of text data (which is generally true for NLP tasks), this simply didn't work well. Albeit so, during the process we figure out some efficient coding to at least get the names of the authors, which provide great convenience for our later manually collection, also trigger us to conceive about how to collect data efficiently in NLP task.

9 What You Would Have Done Differently or Next

In the beginning we plan to frame our problem as an authorship attribution problem. However, due to the time limit and the convenience, we cannot achieve that, so we change our project to a small scale authorship attribution task instead. That said, in the future we will definitely incorporate the authorship meta-data into prediction,which can be helpful and are not normally incorporate in pure authorship attribution tasks. Moreover, we are going to figure out some other ways for more efficient and systematic feature selection, which will be important for our future work.

10 Author contributions

Conceptualization: Mengyi Sun. Data curation: Mengyi Sun and Shang Zhang. Training methodology: Shang Zhang and Mengyi Sun. Training code implementation: Shang Zhang. Writing: Mengyi Sun and Shang Zhang.

Acknowledgments

We thank our academic mentors, Prof Jianzhi Zhang and Prof Xiaoming Mao for valuable comments. We also would like to express our gratitude to Prof David Jurgens for his suggestions about our projects. Finally, we would like to thank our colleague Liuxing Shen for providing us precious computational resource.

References

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52(2):119–123.

Giangiacomo Bravo, Francisco Grimaldo, Emilia López-Iñesta, Bahar Mehmani, and Flaminio Squazzoni. 2019. The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature communications* 10(1):322.

Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review?: author identification using only citations. *Acm Sigkdd Explorations Newsletter* 5(2):179–184.

Jacalyn Tara Sadeghieh Kelly and Khosrow Adeli. 2014. Peer review in scientific publications: benefits, critiques, and a survival guide.

Mike et al Kestemont. 2018. Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *CLEF 2012 Labs and Workshop, Notebook Papers*. Citeseer.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank .

Mihir et al Nanavati. 2011. Herbert west-deanonymizer. In *HotSec*. Citeseer.

Jessica K Polka, Robert Kiley, Boyana Konforti, Bodo Stern, and Ronald D Vale. 2018. Publish peer reviews.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3):538–556.

A Supplemental Material

A.1 Top 500 words selected from reviews data

```
the
,
.
of
to
and
in
a
is
that
)
(
be
not
this
are
for
I
The
it
as
with
by
:
have
on
authors
would
?
or
from
an
``
''
but
more
can
which
one
some
In
This
genes
these
evolution
paper
at
their
could
should
do
proteins
other
all
```

if
they
been
there
also
gene
It
manuscript
might
was
such
about
has
than
we
very
tree
between
analysis
only
different
may
two
what
data
were
model
interesting
think
does
important
any
many
will
present
evolutionary
[
]

origin

protein
most
no
how
1
see
’s
used
same
so

results
species
work
new
even
its
2
et
here
selection
life
case
much
hypothesis
DNA
;
eukaryotes
However
genomes
genome
sequence
you
because
study
first
into
sequences
my
-
possible
For
A
why
well
both
like
example
several
’
using
bacteria
number
al
discussion
RNA
complex
eukaryotic
e.g
cells
If
bacterial

based
evidence

system
point
out
whether
seems
domains
problem
then
use
domain
phylogenetic
where
rather
cell
way
3
author
when
make
clear
time
Koonin
methods
presented
likely
organisms
those
biological

function
me
fact
question
early
better
given
them
really
am
viruses
transfer
find
role
our
common
suggest
less
distribution

useful
What
ancestor
conserved
three
being
specific
since
scenario
level
particular
Figure
genetic
explain
still
general
provide
too
us
single
amino
functional
least
page
models
just
found
set
major
part
Archaea
family
good
There
code
including
within
cellular
related
information
main
membrane
view
need
trees
4
few
HGT
analyses
idea
similar

without
replication
lineages
before
mutations
binding
archaea
They
current
especially
comments
approach
...
state
form
seem
although
5
each
evolved
section
up
known
archaeal
show
either
quite
enzymes
instance
LUCA
comment
proposed
structure
cases
he
understand
process
involved
modern

article
know
another
sites
points
introns
already
literature
further
difficult
issue

multiple
One
history
novel
functions
families
while
method
made
context
shown
n’t
prokaryotes
discussed
viral
thus
high
host
expression
consider
considered
against
available
As
biology
sense
possibility
residues
large
systems
indeed
perhaps
group
i.e
argument
conclusions
Is
text
world
terms
transition
opinion
say
nucleus
your
true
human
To
structural
had
probably

did
second
described
theory
%
similarity
last
support
root
his
positive
ancestral
type
issues
therefore
mean
loss
features
due
order
over
small
ancient
prokaryotic
acids
mentioned
TOL
clearly
result
molecular
structures
s
interactions
paragraph
after
experimental
various
How
highly
yet
recent
search
now
reference
6
change
presence
help
however
review
But
itself
detailed
translation
identified
version
mitochondria
genomic
provides
Please
instead
interest
&
explanation
among
makes
claim
test
take
who
conclusion
originated
rates
vertical
We
Pol
community
7
natural
nature
published
certainly
types
On
studies
simple
concept
far
examples
relevant
mitochondrial
prediction
processes
through
principle
agree
mutation
under
best
long
hypotheses
done

additional
provided
though
correct
often
understanding
papers
appears
Comment
following
argue
course
groups
shows
short
observed
believe
alternative
distinct
actually
significant
title
propose
knowledge
needed
real
homology
simply
!
predicted
key
future
kind
acid
position
homologous
addition
reason

A.2 Top 1000 3-grams of English characters and punctuation selected from reviews data

the
ion
tio
ing
and
ent
hat
tha
ati
pro
ter
ere
res
nce
ons
her
gen
his
for
con
thi
ate
not
ver
tho
are
tic
ted
oul
uld
all
ica
ene
cti
ive
com
est
act
enc
ome
The
pre
ect
ote
int
ith
men
tin
ese
tra
sti
eri
per
nal
ain
ide
ase
wit
log
eve

ary
nte
hor
ist
str
cal
ers
ins
ore
uti
ess
vol
ble
lat
ant
oth
era
iti
rea
ors
sta
aut
tiv
rat
ort
ran
uth
ure
cte
sen
der
cha
hes
ave
ces
ple
nti
hav
lly
ral
pec
par
rot
ssi
one
evo
fer
ria
tes
rom
ear
tat
spe
use
mpl
igh
sis
ren
pos
lar
ses
dis
omp
ies
som
red
anc
ial
ine
tan
olu
nde
ein
lut
but
sed
nes
ell
ste
ity
olo
que
wou
man
sio
tur
can
ree
ind
rec
hou
imp
mor
ari
sho
inc
ona
fic
cri
ula
equ
exp

nts
fro
din
lea
abl
een
han
ifi
ans
ode
por
cat
out
eci
eas
ten
ght
pla
tei
ely
mat
ous
ana
kar
hin
nsi
lin
age
eno
ont
und
sit
scr
ryo
rch
met
ori
ult
wor
sin
tre
whi
art
bac
es.
cou
yot
hic
ite
den
ove

ich
ffe
app
end
mod
ele
nom
bio
cle
nta
rel
uen
es,
nat
dif
oti
min
uch
eth
orm
arc
dat
any
mes
des
tly
tor
lic
ara
iff
ape
ges
les
rti
esi
lit
rre
how
ali
clu
gin
ose
ntr
ost
oss
aly
del
cie
eti
onc
pli

ven
see
omm
ead
cul
als
ini
abo
rip
ani
vel
ser
mpo
rop
mai
rob
ina
nin
lec
eco
ndi
ues
Thi
pap
rig
cel
rib
ami
whe
tro
uct
seq
nct
usi
uka
ved
ame
ata
tai
ipt
ond
rep
vid
evi
ill
ili
sel
mil
ass
nar
hey

tte
nst
ery
tru
rac
tem
tri
unc
on.
ncl
sse
rta
ism
bet
hae
eme
ike
nse
rin
hei
eir
mic
lys
rou
imi
euk
sid
isc
rit
eta
ref
arg
lik
hen
ime
er,
rge
wer
has
oge
atu
sib
ong
net
our
rst
erm
ela
fun
eal
lso

erv	che
rio	cur
oma	ink
phy	usc
fin	tim
ext	pti
igi	hyl
ugh	xpl
cen	lis
on,	mol
ign	ert
sim	vir
ly,	tif
ctu	oin
ual	ogi
oug	ibl
nly	gan
tal	eed
suc	esp
ork	was
cor	mig
oun	ogy
ner	onl
ust	owe
pot	bas
ens	nit
ned	mos
iol	pen
ron	las
etw	mme
rov	iou
dom	cas
anu	eat
ysi	ila
its	arl
two	mer
ruc	mit
bou	rma
ien	esu
ern	may
tab	RNA
osi	cla
sto	ded
nis	lig
bee	rly
ile	mon
cus	ete
ire	now
nus	mal
ful	bil
ast	olv

ed.
oes
sul
ole
sig
ake
dic
scu
low
wha
hem
rev
iat
ept
ach
lie
rth
qui
cia
ici
nge
kin
pri
cod
lan
sub
wee
aea
twe
ntl
ath
tia
lve
duc
ssu
uss
har
lus
lem
lle
ns.
led
uni
yst
opo
exa
edi
hyp
iss
omo
aus

org
amp
mpa
bra
thr
erg
ang
ane
ems
ypo
ail
att
stu
cer
rga
epe
pha
hos
lai
val
mbi
ory
ade
tud
ord
ivi
rve
ict
poi
ize
fac
gic
vin
eli
omi
emb
igu
ura
gre
div
oba
icu
oni
sam
vie
kno
ovi
way
uta
iew
own

lud
fam
hod
ima
eem
ise
reg
ard
ngl
ino
los
ife
bin
rie
ece
sys
dit
add
irs
ylo
mis
ed,
nci
ace
inf
bec
wil
esc
gge
doe
ily
lif
dep
ack
ope
asi
lex
son
cau
you
sup
mea
bse
gro
xam
How
ied
ena
rem
ras
roc

wev
det
mul
sur
pon
old
gra
ppe
sug
urr
ced
ugg
cif
non
lev
ber
oph
ume
ris
gur
orr
fir
pat
xpe
fie
het
yme
Fig
cep
al.
rgu
och
mbe
rim
lts
ucl
pol
ffi
eca
exi
don
hil
rol
ean
sts
inv
num
ish
typ
nuc
nda

acc	nto
lti	uit
set	tua
chi	rus
new	epl
lon	rod
bly	ipl
hig	hom
re,	lac
ano	els
ibu	ice
mem	bli
ram	ian
nco	nea
mak	ndo
var	lls
nfo	cid
odu	mbr
efe	spo
ret	opi
pea	oup
ype	muc
cho	gni
fol	sms
ric	ns,
ppr	ito
ema	umb
alo	hro
efi	win
tit	ts.
upp	ics
ppo	ves
nee	gue
oll	kel
oce	nsf
udi	ecu
dri	abi
hap	rai
wel	udy
mar	sum
sec	adi
efo	fec
DNA	cro
tel	vat
ccu	oso
ann	alt
llo	bst
sce	aci
obl	rsi
mut	tie
For	llu

too
sym
hol
hel
van
oli
se,
nds
rte
nor
med
nno
agr
phi
err
rok
ubs
isi
iru
ude
dea
eag
sfe
iso
bot
emp
zym
let
le,
ech
occ
oka
tar
giv
efu
vio
nic
gai
erf
syn
sea
oon
cit
xis
roa
200
ett
sev
sue
ify
er.

til
ega
req
cin
nth
top
rse
aps
ndr
hre
nec
def
rag
toc
qua
.g.
elo
ero
ibi
obs
aso
nme
ng.
alu
tex
ddi
e.g
why
ood
ubl
on-
col
jus
exc
eff
pic
hon
due
ier
itt
siv
ng,
ota
pin
ute
tis
nvi
ira
ita
onf
lab

dee
nd,
cre
uce
ods
odo
sol
ymb
ce,
yin
lul
apt
nvo
xte
is,
len
riv
idu
car
odi
tag
zed
rmi
fit
rva
ogo
eng
ibe
mpr
mmo
oac
egi
obi
roo
tle
tec
abs
ora
nt.
pub
fou
heo
cto
ios
shi
nzy
put
eac
fil
tid
eor

loo
urt
zat
gly
enz
onv
unt
she
liz
get
lth
re.
war
hed
cyt
gou
riz
uri
iza
ajo
jor
fte
eva
lue
mun
edu
ppl
uir
iga
ynt
es)
bel
rna
cts
avi
nif
mmu
ale
bab
tas
sco
rde
icr
epr
sci
nt,
arr
nve
oot
gle
lim

ros
ssa
rpr
Koo
nov
tak
emi
aga
cce
erl
ctl
bes
ibo
dev
adv
sma
mot
rap
oly
vis
dge
nam
RBS
VBD
IN
FW
RP
JJR
JJS
PDT
MD
VB
WRB
NNP
EX
NNS
SYM
CC
CD
POS

A.3 POS tag list from Penn Treebank

LS
TO
VBN
''
WP
UH
VBG
JJ
VBZ
--
VBP
NN
DT
PRP
:
WP$
NNPS
PRP$
WDT
(
)
.
,
``
$
RB
RBR