Tweet NLP

We provide a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets, along with annotated corpora and web-based annotation tools.

Contributors: Archna Bhatia, Dipanjan Das, Chris Dyer, Jacob Eisenstein, Jeffrey Flanigan, Kevin Gimpel, Michael Heilman, Lingpeng Kong, Daniel Mills, Brendan O'Connor, Olutobi Owoputi, Nathan Schneider, Noah Smith, Swabha Swayamdipta and Dani Yogatama.

Quick Links

- **Part-of-speech Tagger and POS annotated data** - also **Twokenizer**: tokenizer software (part of tagger package) and **Tagging Models** -- **Download Link**
- **Tweeboparser and Tweebank**: Dependency parser software and dependency annotated data -- **Download Link**
- **Documentation, annotation guidelines, and papers** describing this work
- **Hierarchical Twitter Word Clusters**

Part-of-Speech Tagging

We provide a fast and robust Java-based **tokenizer** and **part-of-speech tagger** for tweets, its training data of manually labeled **POS annotated tweets**, a web-based annotation tool, and **hierarchical word clusters** from unlabeled tweets.

These were created by Olutobi Owoputi, Brendan O'Connor, Kevin Gimpel, Nathan Schneider, Chris Dyer, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah Smith.

What the tagger does

```
./runTagger.sh --output-format conll examples/casual.txt
```

These are real tweets.

```
ikr smh he asked fir yo last name so he can add u
on fb lololol
```

```
word    tag     confidence
ikr     !       0.8143
smh     G       0.9406
he      O       0.9963
asked   V       0.9979
fir     P       0.5545
yo      D       0.6272
last    A       0.9871
name    N       0.9998
```

```
:o :/ :'( >:o (: :) >.< XD -__-
o.O ;D :-) @_@ :P 8D :1 >:( :D =|
") :> ....
```

```
word    tag     confidence
:o      E       0.9387
:/      E       0.9983
:'(     E       0.9975
>:o     E       0.9964
(:      E       0.9994
:)      E       0.9997
>.<     E       0.9952
```

```
so      P       0.9838
he      O       0.9981
can     V       0.9997
add     V       0.9997
u       O       0.9978
on      P       0.9426
fb      ^       0.9453
lololol !       0.9664
```

- *"ikr"* means "I know, right?", tagged as an interjection.
- *"so"* is being used as a subordinating conjunction, which our coarse tagset denotes *P*.
- *"fb"* means "Facebook", a very common proper noun (*^*).
- *"yo"* is being used as equivalent to *"your"*; our coarse tagset has posessive pronouns as *D*.
- *"fir"* is a misspelling or spelling variant of the preposition *for*.
- Perhaps the only debatable errors in this example are for *ikr* and *smh* ("shake my head"): should they be *G* for miscellaneous acronym, or *!* for interjection?

```
XD      E       0.9938
-__-    E       0.9956
o.O     E       0.9899
;D      E       0.9995
:-)     E       0.9992
@_@     E       0.9964
:P      E       0.9996
8D      E       0.9961
:       E       0.6925
1       $       0.9194
>:(     E       0.9715
:D      E       0.9996
=|      E       0.9963
"       ,       0.6125
)       ,       0.9078
:       ,       0.7460
>       G       0.7490
...     ,       0.5223
.       ,       0.9946
```

Challenge case for emoticon segmentation/recognition: 20/26 precision, 18/21 recall.

TweeboParser and Tweebank

We provide a dependency parser for English tweets, **TweeboParser** . The parser is trained on a subset of a new labeled corpus for 929 tweets (12,318 tokens) drawn from the POS-tagged tweet corpus of Owoputi et al. (2013) , **Tweebank** .

These were created by Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith.

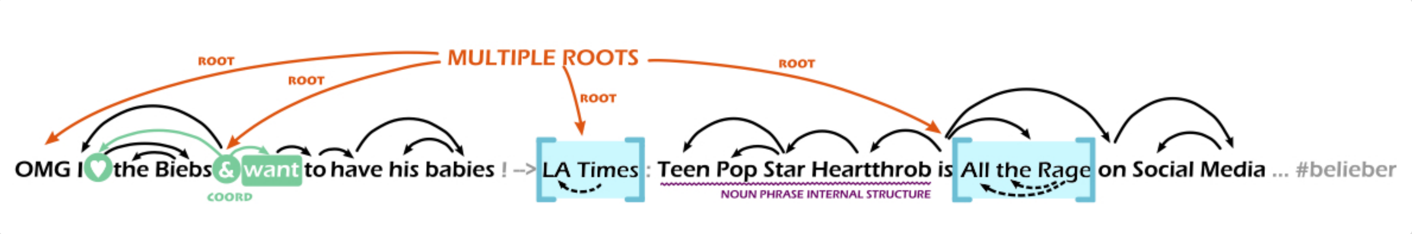
Thanks to Tweebank annotators: Waleed Ammar, Jason Baldridge, David Bamman, Dallas Card, Shay Cohen, Jesse Dodge, Jeffrey Flanigan, Dan Garrette, Lori Levin, Wang Ling, Bill McDowell, Michael Mordowanec, Brendan O'Connor, Rohan Ramanath, Yanchuan Sim, Liang Sun, Sam Thomson, and Dani Yogatama.

What TweeboParser does

Given a tweet, TweeboParser predicts its syntactic structure, represented by unlabeled dependencies. Since a tweet often contains more than one utterance, the output of TweeboParser will often be a multi-rooted graph over the tweet. Also, many elements in tweets have no syntactic function. These include, in many cases, hashtags, URLs, and emoticons. TweeboParser tries to exclude these tokens from the parse tree (grayed out in the example below).

Please refer to the paper for more information.

An example of a dependency parse of a tweet is:


MULTIPLE ROOTS
ROOT
ROOT
ROOT
ROOT
OMG I ♥ the Biebs & want to have his babies ! —> LA Times : Teen Pop Star Heartthrob is All the Rage on Social Media … #belieber
COORD
NOUN PHRASE INTERNAL STRUCTURE

Corresponding CoNLL format representation of the dependency tree above:

```
1       OMG     _       !       !       _       0       _
2       I       _       O       O       _       6       _
3       ♥       _       V       V       _       6       CONJ
4       the     _       D       D       _       5       _
5       Biebs   _       N       N       _       3       _
6       &       _       &       &       _       0       _
7       want    _       V       V       _       6       CONJ
8       to      _       P       P       _       7       _
9       have    _       V       V       _       8       _
10      his     _       D       D       _       11      _
11      babies  _       N       N       _       9       _
12      !       _       ,       ,       _       -1      _
13      –>      _       G       G       _       -1      _
14      LA      _       ^       ^       _       15      MWE
15      Times   _       ^       ^       _       0       _
16      :       _       ,       ,       _       -1      _
17      Teen    _       ^       ^       _       19      _
18      Pop     _       ^       ^       _       19      _
19      Star    _       ^       ^       _       20      _
20      Heartthrob      _       ^       ^       _       21      _
21      is      _       V       V       _       0       _
22      All     _       X       X       _       24      MWE
23      the     _       D       D       _       24      MWE
24      Rage    _       N       N       _       21      _
25      on      _       P       P       _       21      _
26      Social  _       ^       ^       _       27      _
27      Media   _       ^       ^       _       25      _
28      …       _       ,       ,       _       -1      _
29      #belieber       _       #       #       _       -1      _
```

(HEAD = -1 means the word is not included in the tree)

Download

- For the part-of-speech tagger:
  - Releases of the tagger (and tokenizer), data, and annotation tool are available here on Google Code.
  - The tagger source code (plus annotated data and web tool) is on GitHub.
  - Tagger Models
    - To use an alternate model, download the one you want and specify the flag: `--model MODELFILENAME`
      - model.20120919 (2MB) -- the Twitter POS model with our coarse 25-tag tagset. This is included with the tagger release and used by default.
      - model.ritter_ptb_alldata_fixed.20130723 (1.5 MB) -- a model that gives a Penn Treebank-style tagset for Twitter. Trained from a fixed version of Ritter et al. EMNLP 2011's annotated data. **If you want Penn Treebank-style POS tags for Twitter, use this model.** We documented issues and changes here. Also, here is an accuracy evaluation to compare with other work.
      - model.irc.20121211 (3MB) -- a model trained on the NPSChat IRC corpus, with a PTB-style tagset.
      - A model trained on the English Web Treebank
    - The *ritter_ptb* and *irc* models are trained on datasets that were annotated separately from the work described here. Our tagging guidelines and various distinctions they describe (like constituent versus tag uses of hashtags) do not apply if you are using the tagger with these models.
  - The tokenizer code is self-contained in Twokenize.java; or use *twokenize.sh* in the tagger download. A Python port of the tokenizer is available from Myle Ott: ark-twokenize-py. (There is also an older Python version from 2010, also called "twokenize," here.)
- For the dependency parser:
  - Releases of the parser (including the POS tagger and the token selection tool), pre-trained models, and annotated data (Tweebank) are available here on Github.

To receive announcements about updates, join the ARK-tools mailing list.

Further Reading

Please cite the appropriate paper when using these resources in research.

- For the part-of-speech tagger:
  - The newest paper (version 0.3), is:
    Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters
    Olutobi Owoputi, Brendan O’Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith.
    In *Proceedings of NAACL 2013*.
    - Tech report version, with a few more and a few less details: **Owoputi et al. (2012)**. Technical Report, Machine Learning Department. CMU-ML-12-107. September 2012.
  - The original paper:
    Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments
    Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith.
    In *Proceedings of ACL 2011*.
  - See also:
    The Annotation Guidelines, extensively revamped for 0.3.
- For the dependency parser:
  - **A Dependency Parser for Tweets**
    Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. In *Proceedings of EMNLP 2014*.

Twitter Word Clusters

Here is an **HTML viewer of the word clusters**. Produced by an unsupervised HMM: Percy Liang's Brown clustering implementation on Lui and Baldwin's langid.py-identified English tweets; see Owoputi et al. (2012) for details.

We recommend the largest one:

filename	#wordtypes	#tweets	#tokens	#clusters	min count	tweet source
50mpaths2	216,856	56,345,753	847,372,038	1000	40	100k tweet/day sample, 9/10/08 to 8/14/12

Also, here are the smaller ones used in the experiments.

Links to Software Provided by Others

- Python wrappers: (1) by github.com/ianozsvald, (2) by github.com/kevinzzz007.
- PL/Java wrapper: gp-ark-tweet-nlp is: "a PL/Java Wrapper for Ark-Tweet-NLP, that enables you to perform part-of-speech tagging on Tweets, using SQL. If your environment is an MPP system like Pivotal's Greenplum Database you can piggyback on the MPP architecture and achieve implicit parallelism in your part-of-speech tagging tasks."

- Node.js wrapper by github.com/mbejda (npmjs.com)

Acknowledgments

This research was supported in part by: the NSF through CAREER grant IIS-1054319, grant IIS-1352440, the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533, Sandia National Laboratories (fellowship to K. Gimpel), and the U.S. Department of Education under IES grant R305B040063 (fellowship to M. Heilman), an REU supplement to NSF grant IIS-0915187 and Google's support of the Worldly Knowledge project at CMU.