The Syntactic Ngrams Corpus

This file describes the syntactic-ngrams dataset, the data organization and data format, as well as provide examples of the kinds of syntactic structures that are available.

OVERVIEW

The syntactic ngrams corpus contain dependency tree fragments from automatically parsed English text. The dependency trees follow the Stanford basic-dependencies scheme[1], and the underlying corpus is the Google English Books collection. Each syntactic-ngram is accompanied with a corpus-level occurrence count, as well as a time-series of counts over the years.

When creating the syntactic-grams corpus, we aimed to make it useful for many aspects of linguistic study and language modeling, including both syntactic and lexical-semantic aspects of language.

We provide several datasets, each with a different representation. The different representations provide different views of the data, and lend themselves to different kinds of analyses. Some of the representations are inspired by current work in corpus-based models, while others are more open ended and designed for facilitating higher-order explorations.

The kinds of data we provide should allow researchers to extract from them most, if not all, of the syntactic patterns that were used in previous works on syntax-based vector space models [2], and many others. In particular, the tree-fragments we provide include all of the syntactic paths used in Pado and Lapata 2007 as well as the syntactic patterns in Baroni and Lenci 2010.

Additional information can be found in the accompanying research paper:

"*A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books*", Yoav Goldberg and Jon Orwant, 2013.

The data is available for download at:
http://storage.googleapis.com/books/syntactic-ngrams/index.html

[1] With the exception that copular verbs are treated as heads of their expressions.
[2] Different work use different syntactic parsers and different dependency schemas to the ones used in this corpus, and so we could not guarantee a direct correspondence. However, patterns in the spirit of all the syntactic patterns in used in these previous efforts are all covered in this dataset.

DATA ORGANIZATION

This release is composed of several corpora (`eng`, `eng-1M`, `eng-fiction`, `eng-us`, `eng-gb`), and each corpus is divided into 14 collections, each collection contain different kinds of syntactic structures.

The available collections are: `nodes`, `arcs`, `biarcs`, `triarcs`, `quadarcs`, `extended-nodes`, `extended-arcs`, `extended-biarcs`, `extended-triarcs`, `extended-quadarcs`, `verbargs`, `unlex-verbargs`, `nounargs`, `unlex-nounargs`

Each collection is divided into 99 gzipped text files (numbered 00 to 98).

Each line in the files represent a unique syntactic ngram. The first element in each line is the head word of the ngram.

All the lines in a collection are sorted in lexicographic order.

DATA FORMAT

Each line represents one syntactic ngram. The format of a line is:

`head_word`<TAB>`syntactic-ngram`<TAB>`total_count`<TAB>`counts_by_year`

The `counts_by_year` format is a tab-separated list of `year<comma>count` items. Years are sorted in ascending order, and only years with non-zero counts are included.

The `syntactic-ngram` format is a space-separated list of tokens, each token format is "word/pos-tag/dep-label/head-index".

The *word* field can contain any non-whitespace character. The other fields can contain any non-whitespace character except for '/'.

pos-tag is a Penn-Treebank part-of-speech tag.

dep-label is a stanford-basic-dependencies label.

head-index is an integer, pointing to the head of the current token. "1" refers to the first token in the list, 2 the second, and 0 indicates that the head is the root of the fragment.

Example of a syntactic-ngram:

```
cease/VB/ccomp/0  for/IN/prep/1  some/DT/det/4  time/NN/pobj/2
```

Example of a complete line:

```
cease   cease/VB/ccomp/0 for/IN/prep/1 an/DT/det/4 instant/NN/pobj/2
56   1834,2    1835,1    1856,1    1863,1    1871,1    1872,1
1874,1    1875,3    1880,2    1883,2    1889,1    1904,7
1905,2    1915,5    1918,1    1961,1    1963,5    1973,2
1975,1    1977,1    1981,2    1987,2    1988,1    1989,1
1991,1    1996,5    2000,1    2008,2
```

TERMINOLOGY AND CONCEPTS

All data is lowercased.

Heads:
The root of a tree-fragments is called its *head.*

Functional markers vs. content words:
We distinguish relations between content-words from relations including a word is a functional marker. We identify the functional markers by the dependency-label assigned to them by the parser. A subset of the dependency-labels such as determiners, negators, auxiliary verbs and possessives are treated as indicating "functional markers" (for a complete list see appendix A), while the rest of the labels are treated as "content words". We only consider dependency arcs between content words. For some of the datasets (marked as "extended") we add all of the functional markers that modify any of the content words.

Collapsed Relations:
Some dependency relations which involve more than one arc are considered as a single arc when constructing the tree fragments. That is, if one of the arcs in a complex relation is included in a tree-fragments, the other arcs in the complex relation are included as well, and are counted together as a single dependency arc. This is similar in spirit to the "collapsed representation" of the Stanford-dependencies, but the collapsing is reflected only in the set of nodes and arcs in the tree fragments, and not in the actual tree structure.

Specifically, collapsing applies to:

Prepositions: chains of (A prep B, B pobj C) or (A prep B, B pcomp C) are treated as a single arc. Whenever A and B are included, C will be included as well, and whenver B and C are included A is included as well.

Conjunctions: Words with the "cc" relation that have a sibling with a "conj" relation will be included whenever the word with the "conj" relation is included, and will never be included when the word with the conj relation is not included. Words with the "cc" relation and without a sibling with a "conj" relation are treated as regular content words.

Multiword Expressions: whenever a content word has modifiers with the "mwe" label, all of the modifiers are included in the structure together with the content word. This relies on the parser to assign the "mwe" label, which happens for some expressions seen in the training data.

Lexicalized vs. Unlexicalized:

A tree-fragment is composed of tokens. We refer to a token as “lexicalized” if it includes the word form, and “unlexicalized” otherwise. Information about unlexicalized tokens include their part-of-speech, the dependency relations with respect the their head token, and the identity of the head if it is part of the tree fragment, but not the word form which is replaced with a wildcard *W* symbol. Information about a lexicalized token includes all of the above mentioned information, as well as the word form. All the tokens are lexicalized unless we note otherwise.

Frequency Cutoff

In order to keep the size of the resource manageable, we employ a frequency cutoff of 10. This means that only syntactic ngrams appearing at least 10 times in the corpus are included.

THE COLLECTIONS

nodes / extended-nodes

Items in the “nodes” dataset represent single content words (as well as their part-of-speech tags and their dependency-relation to the word they modify).
These are useful for answering questions such as “in what syntactic context is this word being used” as well as “what are the most dominant subjects in a given year”.

In the “extended-nodes” dataset, the functional markers of the the main content word are included as well.

Examples

nodes:

```
matter  matter/NN/dobj/0        6202639
matter  matter/NN/nsubj/0       3938399
```

extended nodes:

```
matter  any/DT/det/2 matter/NN/nsubj/0  22318
matter  did/VBD/aux/3 not/RB/neg/3 matter/VB/ROOT/0     212222
```

arcs / extended-arcs

Items in the “arcs” dataset represent direct relations between two content words.

In the “extended-arcs” dataset, the functional markers of the the main content word are included as well.

The arcs datasets reflect direct head-modifier relations, which are the predominant source of information in current-day syntax-based lexical-semantic models.

Some example of queries that can be issued against this dataset include include adjecitivial modifiers of a given noun, verbs a given noun is object of, and conjoined nouns.

Examples

arcs:

```
efficiency      statistical/JJ/amod/2 efficiency/NN/pobj/0      1099
efficient       efficient/JJ/amod/0 or/CC/cc/1 effective/JJ/conj/1      1160
```

extended-arcs:

```
emerge  to/TO/aux/2 emerge/VB/xcomp/0 as/IN/prep/2 a/DT/det/5 force/NN/pobj/3   1434
```

biarcs / extended-biarcs

Items in the “biarcs” dataset reflect higher-order dependency relations that involve two arcs -- three connected content words.

In the “extended-biarcs” dataset, the functional markers of the the main content word are included as well.

The biarcs datasets allows modeling information involving interactions between two modifiers of the same head, e.g. subject and object of the same verb (boy, ate, cookies), as well as complex arguments of a head, e.g. adjectivial modifier of a verb’s argument ((small, boy), ate).

By abstracting over the middle element, we could also get second-order information, e.g. (boy, *, cookies) and ((small, *), ate).

By further abstracting over the words, one may uncover second-order syntactic phenomena: maybe some verbs are more likely to have adjectives to their subjects than others?

Examples

biarcs:

```
experience      that/IN/compl/3 patients/NNS/nsubj/3 experience/VB/ccomp/0      3092
experience      very/RB/advmod/2 good/JJ/amod/3 experience/NN/attr/0    1400
experienced     experienced/VBD/ROOT/0 great/JJ/amod/3 difficulty/NN/dobj/1     4758
```

extended biarcs:

```
expect  therefore/RB/advmod/4 one/PRP/nsubj/4 would/MD/aux/4 expect/VB/ROOT/0   3831
expect  to/TO/aux/2 expect/VB/xcomp/0 that/IN/compl/5 will/MD/aux/5 be/VB/ccomp/2
9927
```

triarcs / extended-triarcs

Item in the “triarcs” dataset include all relations involving three arcs -- 4 content words.

In the “extended-triarcs” dataset, the functional markers of the the main content word are included as well.

The locality of the dependency representation causes this set of three-arcs structures to be large, sparse and noisy -- many of the relations may appear random because some arcs are in many cases almost independent given the others. However, some of the relations are known to be of interest, and we hope more of them will prove to be of interest in the future.

Some of the interesting relations include:
* modifiers of the head noun of a subject or object in a (subject, verb, object) construction: ((small,boy), ate, cookies), (boy, ate, (tasty, cookies)), and with abstraction: adjectives that a

boy likes to eat: (boy, ate, (tasty, *))

* arguments of an embeded verb (said, (boy, ate, cookie)), (said, ((small, boy), ate))

* modifiers of conjoined elements ((small, boy) (young, girl)) ((small, *) (young, *))

* relative clause constructions (boy, (girl, with-cookies, saw))

Examples

triarcs:

```
feel    feel/VBP/ROOT/0 that/IN/compl/3 is/VBZ/ccomp/1 wrong/JJ/acomp/3 1864
feel    how/WRB/advmod/3 you/PRP/nsubj/3 feel/VB/ROOT/0 now/RB/advmod/3 2451
feel    how/WRB/compl/3 i/PRP/nsubj/3 feel/VBP/ccomp/0 about/IN/prep/3 it/PRP/pobj/4
5976
feel    i/PRP/nsubj/2 feel/VB/ROOT/0 sorry/JJ/acomp/2 for/IN/prep/3 him/PRP/pobj/4
2566
```

extended triarcs

```
feel    how/WRB/advmod/4 do/VBP/aux/4 you/PRP/nsubj/4 feel/VB/ROOT/0 about/IN/prep/4
being/VBG/pcomp/5  1092

feel    i/PRP/nsubj/4 do/VBP/aux/4 not/RB/neg/4 feel/VB/ccomp/0 so/RB/advmod/6
good/JJ/acomp/4  1323

feel    feel/VBP/ROOT/0 something/NN/dobj/1 rising/VBG/partmod/2 in/IN/prep/3
my/PRP$/poss/6 breast/NN/pobj/4   228
```

quadarcs / extended-quadarcs

In contrast to the the previous datasets, the “quadarcs” dataset includes only a subset of the possible relations involving 4-arcs (5 content words).

We chose to focus on relations which are attested in the literature (Pado and Lapata 2007, appendix A), namely structures consisting of two chains of length 2 with a single head:
((small, boy), ate, (tasty, cookie)).

In the “extended-quadarcs” dataset, the functional markers of the the main content word are included as well.

Examples

quadarcs:

```
constitute   parts/NNS/nsubj/4 of/IN/prep/1 compilation/NN/pobj/2
constitute/VBP/ROOT/0 one/CD/num/6 work/NN/dobj/4  113
```

```
constituted  extraordinary/JJ/amod/2 vigour/NN/nsubj/3 constituted/VBD/rcmod/0
one/CD/dobj/3 of/IN/prep/4 features/NNS/pobj/5       116
```

extended-quadarcs:

```
constitute   a/DT/det/2 majority/NN/nsubj/7 of/IN/prep/2 the/DT/det/5 board/NN/pobj/3
shall/MD/aux/7 constitute/VB/ROOT/0 a/DT/det/9 quorum/NN/dobj/7 for/IN/prep/9
the/DT/det/12 transaction/NN/pobj/10    158
```

verbargs / unlex-verbargs

The “verbargs” datasets include verbs with all of their immediate modifiers, as well as functional-markers of the head word and all of its modifiers.

In the “unlex-verbargs” dataset, the head token is always lexicalized, but the other tokens are lexicalized only if the word appears in the predefined list of words given in Appendix B, and unlexicalized otherwise. The word list is constructed by taking the top-1000 most frequent words in the eng-1M subset of the corpus.

The lexicalized version can be used for modeling interactions between the different modifiers of the verb, as well as to assist in dependency-based language modeling (as all of the modifiers are guaranteed to be present).

This unlexicalized version is meant for the study of subcategorization frames.

Examples

verbargs:

```
covering     hands/NNS/nsubj/2 covering/VBG/dep/0 her/PRP$/poss/4 face/NN/dobj/2
106
covers  as/IN/mark/3 water/NN/nsubj/3 covers/VBZ/advcl/0 the/DT/det/5 sea/NN/dobj/3
126
```

unlex-verbargs:

```
cut    he/PRP/nsubj/2 cut/VBD/conj/0 the/DT/det/4 *W*/NN/dobj/2 from/IN/prep/2
the/DT/det/7 *W*/NN/pobj/5  103

cut    the/DT/det/2 *W*/NN/nsubj/3 cut/VBD/ccomp/0 a/DT/det/5 *W*/NN/dobj/3     248
```

nounargs / unlex-nounargs

The “nounargs” datasets include nouns with all of their immediate modifiers, as well as functional-markers of the head word and all of its modifiers.

In the “unlex-nounargs” dataset, the head token is always lexicalized, but the other tokens are lexicalized only if the word appears in the predefined list of words given in Appendix B, and

unlexicalized otherwise. The word list is constructed by taking the top-1000 most frequent words in the eng-1M subset of the corpus.

The lexicalized version represent something very similar to NP-chunks, which are, broadly, “things” that people talk about. In addition, the lexicalized version can be used for modeling interactions between the different modifiers of a noun, as well as to assist in dependency-based language modeling (as all of the modifiers are guaranteed to be present).

This unlexicalized version can be used to study linguistic modification patterns for nominal words, answering questions like: are some words more likely to have more adjectives than others? Are nouns with a specific syntactic function (e.g. direct objects) are more likely to be modified than other nouns? and so on.

Examples

nounargs:

```
hinduism	the/DT/det/3 new/JJ/amod/3 hinduism/NN/pobj/0	119
hinduism	the/DT/det/2 hinduism/NN/pobj/0 of/IN/prep/2 bali/NNP/pobj/3	12
hinduism	rajput/NNP/nn/2 hinduism/NNP/conj/0 in/IN/prep/2 rajasthan/NNP/pobj/3
14
```

unlex-nounargs:

```
gutter	the/DT/det/4 *W*/JJ/amod/4 *W*/JJ/amod/4 gutter/NN/pobj/0	169
guy	a/DT/det/2 guy/NN/dobj/0 from/IN/prep/2 the/DT/det/5 *W*/NN/pobj/3 140
```

UNDERLYING CORPORA

This dataset is based on the English Google Books corpus. This is the same corpus used to derive the Google Books Ngrams, and is described in detail in Michel et.al 2011.
The corpus consists of the text of 3,473,595 English books which were published between 1520 and 2008, with the majority of the content published after 1800.
We provide counts based on the entire corpus, as well as on several subsets of it.

English	(eng)	All the available books.
English 1M	(eng-1M)	Uniformly sampled 1 million books.
Fiction	(eng-fiction)	Works of Fiction
American English	(eng-us)	Books published in the US.
British English	(eng-gb)	Books published in Britain.

References

Michel et.al. 2011. “Quantitative Analysis of Culture Using Millions of Digitized Books”, Michel, Jean-Baptiste and Shen, Yuan Kui and Aiden, Aviva Presser and Veres, Adrian and Gray, Matthew K. and Pickett, Joseph P. and Hoiberg, Dale and Clancy, Dan and Norvig, Peter and Orwant, Jon and Pinker, Steven and Nowak, Martin A. and Aiden, Erez Lieberman. *Science*.

Baroni and Lenci 2010. “Distributional Memory: A General Framework for Corpus-Based Semantics”. Marco Baroni and Alessandro Lenci. *Computational Linguistics*.

Pado and Lapata 2007. “Dependency-based construction of semantic space models”. Sebastian Pado and Mirella Lapata. *Computational Linguistics*.

Appendix A -- functional marker relations

The following Stanford-dependencies relations are treated as functional-markers and not as content-bearing words:

det
poss
neg
aux
auxpass
ps
mark
complm
prt

Appendix B -- 1000 most frequent words in the eng-1M corpus.

, the . of and to in a is " that ; was The for be with as it - not by I his on are he) (which at from or this have had 's were but an ' they their : all one you -- been has we It In will her more him ? them can no would so He its who may there other than into A when my these only time This some do out if any two 1 she me But very about our up such what said should also first could made most must upon ! man then great now much same many did And over those after like They being before us well your 2 between We see life There men through good under work where own years people even little day without way found each If new part long make three every For never Mr. 3 used both place shall too God against New know down might however still how because As himself old another THE here use When does She His given What world called To case right water come came power take while last p de say few again present large back * number On year small 4 less form just general You whole At country go always left state different far / thought These during though hand often give OF think per That among point high end yet order off course am 5 things system others taken children find fact therefore nothing went means name side away ever seen above law important American within i known certain second once days took four get having mind house public set nature war York several One] best body done put 6 until since almost John head By themselves following thus better rather possible [itself s become true English words death young See 10 government No Lord light home subject All necessary matter question cases England love either common times half States thing human line became family whom land & brought later early five night effect period soon A. Then nor next already J. kind school least full $ • With quite heart together something enough London need father United whose et whether social action interest word value gave women change 8 example From eyes saw character history

position French city So II near face various age told view person held After feet 7 reason How going c child sense let air An seems political cause business money book free room ... result short further Some want received white began asked look group making usually M. State study account force mother Thus process says Of seemed St. along perhaps C. greater hands show knowledge AND sometimes pp. sent development got open General British able latter around hundred parts especially moment knew read heard party shown probably conditions ground woman son King felt service believe Sir alone tell natural particular generally 20 12 individual manner indeed real members blood cent act purpose passed control My century Now University church six help looked letter considered Church similar strong king France Fig town influence class self experience S. truth results Dr. 9 Mrs. information due poor return type method wife field clear rest keep idea third required care morning twenty carried surface single H. 15 lost forms language problem low amount mean followed turned terms according close ^ special written spirit really Christ material friends May America continued died attention condition why call persons length living appear rate placed Here longer miles earth friend anything area kept E. fire towards William art society hope although v. Such seem evidence level nearly ten past obtained No. object Government education difficult W. support works makes except lower turn leave hours formed higher produced live However B. sea black feel former described months increase office lines future ii policy appears hard army simple story doubt opinion till 14 late local property changes food myself answer complete company Do R. round % Christian fine court total Let 13 degree beyond 11 born paper respect trade below thousand added 'll D. situation German increased toward data established South appeared practice 30 personal gives treatment theory behind House authority pressure Is equal peace religious importance term voice met production bring national original door size Their acid modern 16 points cut chief sure I. hour play speak patient Why rule returned problems disease foreign river Not beginning reached becomes economic James b population dead la considerable solution Europe June books note table private Henry 25 direction religion particularly cost hold laws difference 18 observed movement pay CHAPTER main deep fall section immediately taking groups shows strength Charles soul faith understand hear states George proper red Paris neither published provided mentioned frequently distance stood across III growth gone plan places led wrote merely principle working entirely lay feeling front easily species current paid success comes March capital building meet coming 17 loss eye al July regard North effects determined doing extent based Many military circumstances n certainly health India Even Our structure countries direct moral remained produce author Press Indian presence lead employed wish sound built non fear run ready else developed road President beautiful brother existence Yet G. cells energy duty follow April ought pass using arms 19 die price stand series stage bad P. War space dark likely schools cold community Her looking Great operation applied enemy eight sufficient .... physical First L. activity Washington History normal e outside expected function available ancient addition weight attempt 24 appearance methods Two temperature Its nation t allowed seven greatest simply meaning wanted volume letters F. middle iron boy 100 Court herself remain principles none West everything questions fell lived analysis B impossible sort basis somewhat charge desire instance ways evening gold meeting test follows larger deal ideas forces thy relation Thomas expression During students ask wide average heat 21 thou

forward 'm .. writing member specific presented report sun measure laid north IN Miss Act tried entered takes major relations