Root and Rule-Based Natural Language Processing:

Subject-Predicate-Object Phrases

Jonah Tash and Peter Zhang

Motivation

- We have used our root and rule-based program to create terms and sentences from scientific articles
- We want to express the terms and sentences in subject-predicate-object groups

Example

The crystal packing is mainly stabilized by van der Waals interactions

crystal:0:packing

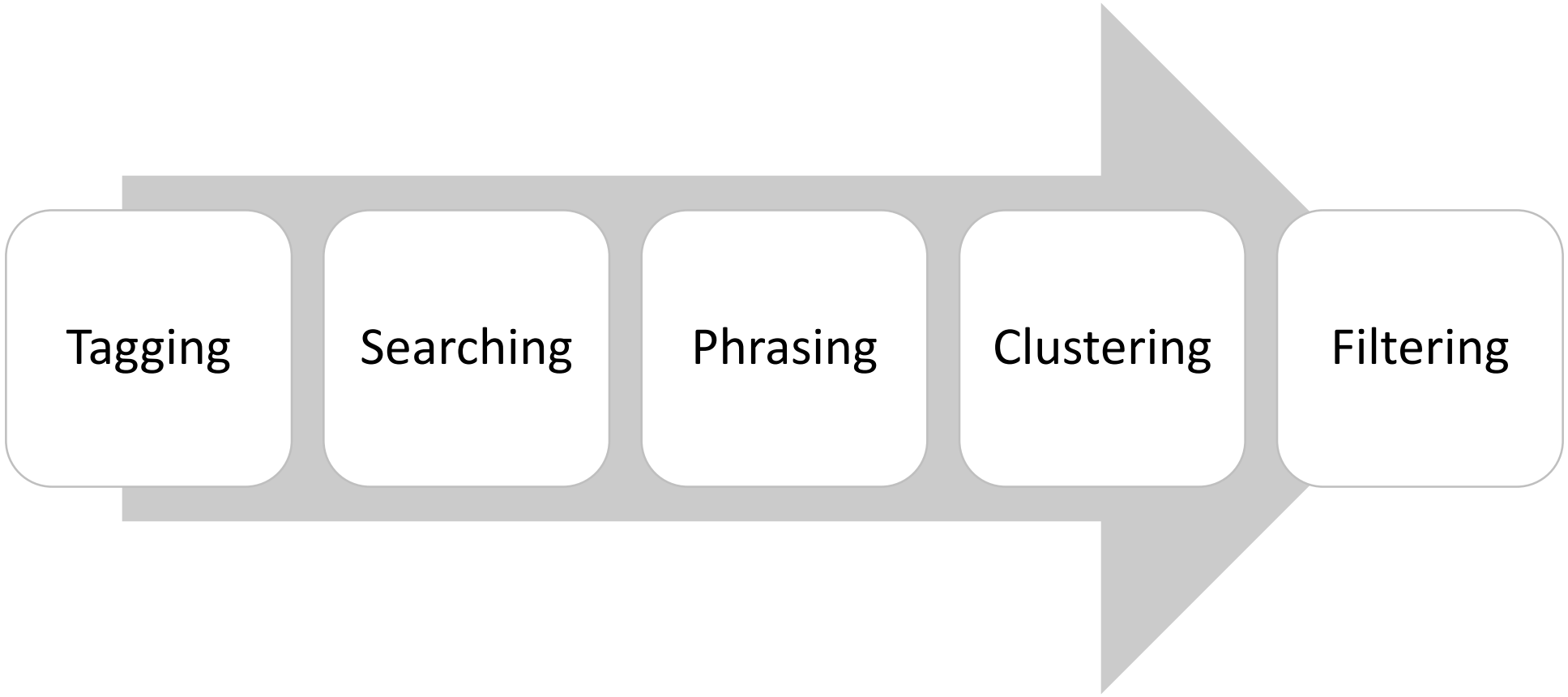
van:2:der:1:waals:0:interacitons

crystal packing — Subject

stabilized by — Predicate Phrase

Van der Waals interactions — Object

Tian, L. & Liu, L.-Z. (2004). Acta Cryst. E60, o1826-o1827.



Process

Input

- Used the original phrases and the fragmented terms
- Considered only short terms
  - Terms of level 0, 1 and 2
  - Subjects and objects should be concise

Input

term	originals
crystal:0:pack	crystal packing
crystal	crystal
pack	packing
van:2:der:1:waal:0:interaction	van der Waals interactions
van	van
der:1:waal:0:interaction	der Waals interactions
der	der
waal:0:interaction	Waals interactions
waal	Waals
interaction	interactions
one:1:triazine:0:molecule	One triazine molecule
triazine:0:molecule	triazine molecule
triazine	triazine
crystal:0:conformation	two conformations within the crysta
conformation	conformations
crystal	crystal
occupancy	occupancies

The crystal packing is mainly stabilized by van der Waals interactions

One triazine molecule is disordered over two conformations within the crystal, the occupancies being 62 emsp14 (1) and 38 emsp14

Step 1: Tagging Verbs

- Use the part-of-speech tagger in the Natural Language Toolkit (NLTK) library to identify verbs
  - The model can isolate the sense of a word
  - E.g. “report” *noun* vs. “report” *verb*

Found at: https://www.nltk.org/

Step 1: Tagging Verbs

- NLTK is not perfect, ~95% accuracy
- First, remove “verbs” that appear in the terms
- Second, check results with Princeton’s WordNet lexicon
  - Removes words that are incorrectly classified as verbs
  - E.g. symbols, molecule names

Found at: https://wordnet.princeton.edu/

Step 2: Searching for Verbs

- Classify terms into Level 0, 1, and 2 by maximum colon number
  - Higher level means a longer phrase
  - E.g. “crystal:0:packing” ⇨ Level 0

    “van:2:der:1:waal:0:interaction” ⇨ Level 2
- Conduct searches first on *only* Level 0 terms, then on Level 0 *and* Level 1, and finally among *all* levels

Step 2: Searching for Verbs

- Search the span between subject and object for verbs
  - E.g. between “crystal packing” and “Van der Waal interactions”
- Count the number of characters and words in the span

The **crystal packing** is mainly stabilized by **van der Waals interactions**

Span

Characters: 23
Words: 4

Step 3: Phrasing

- A verb by itself is not enough information
- Use the Collaborative International Dictionary of English to capture conjugations of the verb
  - Hand-collected a list of English conjugations
  - E.g. *is* linking, *was* linked, *will* link

Found at: http://www.ibiblio.org/webster/

Step 3: Phrasing

- Reference the Pattern Dictionary of English Propositions to append neighboring prepositions
  - e.g. “stabilized” ⇨ “stabilized by”
    “connect” ⇨ “connect to”
    “extracting” ⇨ “extracting from”

Found at: https://www.clres.com/pdep.html

Step 4: Clustering

- Want to group similar verb phrases
- Reference the British National Corpus to find the “stem” of each verb phrase
  - E.g. “is” ⇨ “be”

    “stabilized by” ⇨ “stabilize”

Found at: https://github.com/skywind3000/lemma.en

Step 4: Clustering

- Reference the Princeton's WordNet lexicon to find synonyms for the stems
  - Use these to cluster similar phrases
  - E.g. "appear" ⇨ {"appear", "seem", "look"}
    "bond" ⇨ {"bond", "bind", "attach", "adhere", "stick"}

Found at: https://wordnet.princeton.edu/

Step 4: Clustering

{'organise', 'coordinate', 'align', 'organize', 'ordinate'}
{'dispatch', 'finish', 'complete', 'nail', 'discharge'}
{'concede', 'yield', 'grant', 'soften', 'generate', 'cede', 'succumb', 'render', 'afford', 'relent', 'give'}
{'dispatch', 'finish', 'complete', 'nail', 'discharge'}
{'concede', 'yield', 'grant', 'soften', 'generate', 'cede', 'succumb', 'render', 'afford', 'relent', 'give'}
{'acquire', 'borrow', 'embrace', 'assume', 'espouse', 'dramatise', 'adopt', 'dramatize', 'follow', 'take'}
{'connect', 'colligate', 'join', 'yoke', 'unite', 'associate', 'relate', 'link', 'tie'}
{'stabilise', 'steady', 'stabilize', 'brace'}
{'site', 'situate', 'settle', 'place', 'locate'}
{'betoken', 'argue', 'point', 'bespeak', 'show', 'signal', 'indicate', 'suggest'}
{'expose', 'exhibit', 'present', 'parade', 'demo', 'demonstrate', 'display', 'show', 'march'}
{'compile', 'write', 'indite', 'compose', 'frame', 'pen'}
{'bind', 'attach', 'adhere', 'stick', 'tie', 'bond'}
{'bind', 'attach', 'adhere', 'stick', 'tie', 'bond'}
{'organise', 'constitute', 'spring', 'form', 'organize', 'mould', 'imprint', 'forge', 'mold', 'shape'}

Step 5: Filtering

- Count the number of each subject, predicate, object, and verb stem
- Sort by frequency and identify phrases to filter
  - Some predicate phrases lack meaningful relation
  - E.g. “is”, “was”, “are”, “have”
- Compile a list of these phrases

Output

- Output to a CSV file with the following columns:
  - Subject, Object, Predicate
  - Sentence
  - Stem and Synonyms
  - Frequencies of Subject, Object, Predicates and Stems
  - R&R Terms for Subject, Object, and Predicate
  - Document ID

Output

- For an input of 67k terms, the program extracts 5700+ phrases in under 15 seconds

subject	object	predicate	sentence	pred freq	sub freq	obj freq	char dist	word dis	root	root freq	synonyms	pred term	sub term	obj term	level	doc ID
Further c	extensive	is accom	Further c	1	1	1	20	3	accompli	2	{'action', 'at	is:0:acco	crystal:0	extensiv	3	10.1107/S1600
hydroxy	mdash H	are involv	The hydr	47	7	128	19	4	involve	69	{'affect', 'as	are:0:inv	hydroxy:	mdash:0	1	10.1107/S1600
hydroxy	O mdash	are involv	The hydr	47	7	51	17	3	involve	69	{'affect', 'as	are:0:inv	hydroxy:	o:1:mdas	2	10.1107/S1600
Ni atom	dianionic	is chelate	The Ni at	5	2	1	18	4	chelate	7	{'chelate'}	is:0:chel	ni:0:atom	dianioni	1	10.1107/S1600
Ni atom	dianionic	is chelate	The Ni at	5	2	1	18	4	chelate	7	{'chelate'}	is:0:chel	ni:0:atom	dianioni	2	10.1107/S1600
ZnII atom	O atoms	is bonde	The ZnII	1	5	28	18	4	bond	15	{'bind', 'atta	is:0:bonc	znii:0:ato	o:0:aton	1	10.1107/S1600
NCCN g	dihedral	subtend	The two	2	1	44	11	2	subtend	2	{'delimit', 's	subtend	nccn:0:g	dihedral	1	10.1107/S1600
chelating	dihedral	subtend	The two	2	1	44	11	2	subtend	2	{'delimit', 's	subtend	chelate:1:	dihedral	2	10.1107/S1600
crystal st	mdash H	are conn	In the cry	17	155	128	70	10	connect	56	{'connect', '	are:0:cor	crystal:0	mdash:0	1	10.1107/S1600
crystal st	O mdash	are conn	In the cry	17	155	51	68	9	connect	56	{'connect', '	are:0:cor	crystal:0	o:1:mdas	2	10.1107/S1600
crystal st	intermol	are conn	In the cry	17	155	16	53	8	connect	56	{'connect', '	are:0:cor	crystal:0	intermol	3	10.1107/S1600
title com	recrystal	were obt	Crystals	27	302	2	36	4	obtain	84	{'get', 'recei	were:0:o	title:0:co	methano	1	10.1107/S1600
Crystals	recrystal	were obt	Crystals	27	4	2	36	4	obtain	84	{'get', 'recei	were:0:o	title:0:co	methano	2	10.1107/S1600
MoVI ato	O atoms	is coordi	The MoV	78	2	28	29	5	coordina	94	{'organise',	is:0:coor	movi:0:at	o:0:aton	1	10.1107/S1600
MoVI ato	oxide O	is coordi	The MoV	78	2	1	23	4	coordina	94	{'organise',	is:0:coor	movi:0:at	oxide:0:	2	10.1107/S1600
coordina	methanol	is compl	The coor	15	2	3	33	8	complete	18	{'dispatch',	is:0:com	sphere:0	methano	1	10.1107/S1600
methanol	distorted	yielding	The coor	3	1	3	12	2	yield	25	{'concede', '	yielding	methanol	distort:0	1	10.1107/S1600
coordina	O atom c	is compl	The coor	15	2	1	21	4	complete	18	{'dispatch',	is:0:com	sphere:0	methano	2	10.1107/S1600
O atom c	distorted	yielding	The coor	3	1	3	12	2	yield	25	{'concede', '	yielding	methanol	distort:0	2	10.1107/S1600
AuIII ato	coordina	adopts	In the titl	50	1	11	23	3	adopt	67	{'acquire', 'b	adopts	auiii:0:at	coordin	1	10.1107/S1600
hydroge	organic c	link	Cl hydro	61	47	2	10	2	link	178	{'connect', '	link	hydroge	organic:	1	10.1107/S1600
title com	mdash H	is stabili	The mole	156	302	128	52	7	stabilize	182	{'stabilise', '	is:0:stab	title:0:co	mdash:0	1	10.1107/S1600
methylsu	opposite	are locat	In the titl	18	1	2	16	3	locate	34	{'site', 'situa	are:0:loc	methylsu	opposit	1	10.1107/S1600
benzofur	dihedral	indicatec	The 4chl	6	1	44	21	4	indicate	35	{'betoken', '	indicatec	benzofur	dihedral	1	10.1107/S1600
crystal st	C mdash	exhibits	The cryst	38	155	40	32	4	exhibit	46	{'expose', 'e	exhibits	crystal:0	c:0:mda	1	10.1107/S1600
title com	pyrazole	is compo	The title	10	302	2	44	6	composc	13	{'compile', '	is:0:com	title:0:co	pyrazolc	1	10.1107/S1600
S atom	tetrahedr	is bonde	In the titl	5	4	6	26	5	bond	15	{'bind', 'atta	is:0:bonc	s:0:atom	tetrahed	1	10.1107/S1600
S atom	distorted	is bonde	In the titl	5	4	2	16	4	bond	15	{'bind', 'atta	is:0:bonc	s:0:atom	distort:1	2	10.1107/S1600
pyrimidin	dihedral	forms	The esse	32	3	44	57	9	form	139	{'organise',	forms	pyrimidin	dihedral	1	10.1107/S1600
essential	dihedral	forms	The esse	32	1	44	57	9	form	139	{'organise',	forms	essential	dihedral	2	10.1107/S1600
pairs of r	mdash H	are linkec	In the cry	50	3	128	32	5	link	178	{'connect', '	are:0:link	molecule	mdash:0	1	10.1107/S1600
pairs of r	N mdash	are linkec	In the cry	50	3	50	30	4	link	178	{'connect', '	are:0:link	molecule	n:1:mdas	2	10.1107/S1600
pairs of r	intermol	are linkec	In the cry	50	3	11	15	3	link	178	{'connect', '	are:0:link	molecule	intermol	3	10.1107/S1600
pyrrole r	propano	are inclin	The pyrr	4	2	1	60	12	incline	10	{'lean', 'disp	are:0:incl	pyrrole:0	propanc	1	10.1107/S1600
propano	extended	is in	The pyrr	56	1	8	10	3	be	474	{'equal', 'be'	is in	propano	extend:0	1	10.1107/S1600
pyrrole r	the prop	are inclin	The pyrr	4	2	1	56	11	incline	10	{'lean', 'disp	are:0:incl	pyrrole:0	propanc	2	10.1107/S1600
NH group	mdash H	are involv	In the cry	47	1	128	34	5	involve	69	{'affect', 'as	are:0:inv	nh:0:gro	mdash:0	1	10.1107/S1600
two pyrr	N mdash	are involv	In the cry	47	2	50	32	4	involve	69	{'affect', 'as	are:0:inv	pyrrole:1	n:1:mdas	2	10.1107/S1600
two pyrr	intermol	are involv	In the cry	47	2	11	17	3	involve	69	{'affect', 'as	are:0:inv	pyrrole:1	intermol	3	10.1107/S1600
title mole	E confor	features	The title	32	21	3	24	3	feature	32	{'boast', 'sp	features	title:0:mc	e:0:conf	1	10.1107/S1600
title mole	E confor	features	The title	32	21	1	24	3	feature	32	{'boast', 'sp	features	title:0:mc	oxime:0:	2	10.1107/S1600
crystal st	mdash H	is consol	The cryst	11	155	128	22	4	consolid	18	{'consolidat	is:0:cons	crystal:0	mdash:0	1	10.1107/S1600
crystal st	C mdash	is consol	The cryst	11	155	61	20	3	consolid	18	{'consolidat	is:0:cons	crystal:0	c:1:mdas	2	10.1107/S1600
methylsu	aromatic	show	The cryst	48	2	3	37	5	show	150	{'reveal', 'pr	show	methylsu	aromatic	1	10.1107/S0108
methylsu	aromatic	stacking	The cryst	10	2	3	37	5	stack	15	{'heap', 'pile	stacking	methylsu	aromatic	1	10.1107/S0108
pyrrolidi	aromatic	show	The cryst	48	4	3	37	5	show	150	{'reveal', 'pr	show	methylsu	aromatic	2	10.1107/S0108
pyrrolidi	aromatic	stacking	The cryst	10	4	3	37	5	stack	15	{'heap', 'pile	stacking	methylsu	aromatic	2	10.1107/S0108
pyrrolidi	aromatic	show	The cryst	48	4	3	37	5	show	150	{'reveal', 'pr	show	methylsu	aromatic	3	10.1107/S0108

Implementation

- The program was written in Python with use of the NLTK, CSV and SQLite3 libraries
- Compilers for the verb, predicate, and lemma dictionaries are available as separate programs
  - Dictionaries can be easily expanded and re-integrated
- Filtered words are manually stored in a separated in a CSV file